TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa UPC-UB

Título: Métodos de simulación de cohortes para la estimación de prevalencia a 5 años de cáncer de cuello de útero en Cataluña

Autor: Valerie Cayssials da Cunha

Director: Ramón Clèries Soler

Departamento: Estadística e Investigación Operativa

Universidad: Politécnica de Cataluña - UPC

Convocatoria: Junio 2015

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

UNIVERSIDAD POLITÉCNICA DE CATALUÑA
FACULTAD DE MATEMÁTICAS Y ESTADÍSTICA

Trabajo de fin de Máster Interuniversitario en Estadística e Investigación Operativa (UPC-UB)

Métodos de simulación de cohortes para la estimación de prevalencia a 5 años de cáncer de cuello de útero en Cataluña

Valerie Cayssials da Cunha

Director: Ramón Clèries Soler

Ponente: Klaus Langohr

Departamento de Estadística e Investigación Operativa
Junio 2015

Agradecimientos

A Ramón Clèries Soler por dirigirme a lo largo del proyecto, siempre con entusiasmo y buena disposición para atender mis inquietudes.

A Klaus Langohr por sus valiosos aportes, que contribuyeron tanto a la mejora del proyecto como del manuscrito final.

A Mireia Diaz por proporcionarme muchísimo material sobre cáncer de cuello de útero.

Este trabajo no hubiera sido posible sin la colaboración de los Registros de Cáncer de Girona y Tarragona, de los cuales proceden los datos utilizados. Mi agradecimiento especial a todos aquellos que gestionan los registros de cáncer para dichas áreas geográficas, ya que nos han permitido realizar estimaciones de cáncer en Cataluña.

Resumen

Palabras clave: prevalencia, cáncer de cuello de útero, simulación de cohortes, supervivencia.

MSC2010: 92D30, 62N02, 62P10.

La prevalencia de una enfermedad, así como la incidencia y la mortalidad, es una medida de gran utilidad para la planificación de los servicios de salud pública, particularmente para la prestación de servicios. La prevalencia mide el número absoluto, y la proporción relativa en la población, de individuos afectados por la enfermedad en un tiempo dado y que requieren alguna forma de atención médica. Por tanto expresa la demanda de cuidados. En la mayoría de los casos de cáncer la fase de tratamiento que involucra servicios de salud, se da dentro de los primeros 5 años desde el diagnóstico. Por esta razón, en general se consideran casos prevalentes de cáncer aquellos individuos que continúan vivos cuando fueron diagnosticados dentro de los últimos 5 años (prevalencia a 5 años). Dado que los recursos requeridos para el tratamiento de los pacientes diagnosticados varía de acuerdo a la fase de tratamiento en que se encuentran, a menudo se usan definiciones de prevalencia alternativas. Las estimaciones de prevalencia puntual a 1, 2-3, y 4-5 años describen el número de casos que continúan vivos luego de 1, 2-3, y 4-5 años de diagnosticados por cáncer. Estas estimaciones de prevalencia puntual son aplicables a tratamiento de evaluación inicial, seguimiento clínico y punto de cura respectivamente para la mayor parte de los tipos de cáncer. Las estimaciones de prevalencia pueden hacerse directamente desde los registros de cáncer de base poblacional (RCBP) por contar el número de casos aún vivos de la población en un punto de tiempo especificado. Sin embargo, esta aproximación requiere el registro y el seguimiento del estatus vital de los pacientes por varios años. Una alternativa, cuando no se tienen registros de seguimiento, es estimar prevalencia a partir de cohortes simuladas obtenidas desde proyecciones de casos incidentes de la población de interés y de los registros de cáncer de una población de referencia. En este trabajo en primer lugar reportamos medidas útiles a los servicios de planificación de salud referentes a la ocurrencia de cáncer de cuello de útero por intervalo de edad quinquenal en Girona y Tarragona entre 1999 y 2007 (incidencia anual, prevalencias puntuales a 1, 2-3, y 4-5 años, y prevalencia a 5 años), estimadas directamente desde sus RCBP. Posteriormente presentamos ocho métodos que permiten simular cohortes de mujeres diagnosticadas por cáncer de cuello de útero, y a partir de éstas estimar prevalencia a 5 años de la enfermedad. Estimamos prevalencia a 5 años en Cataluña por intervalo de edad quinquenal para cada año entre 2003 y 2007. Las cohortes simuladas fueron obtenidas desde estimaciones disponibles de casos incidentes para cáncer de cuello de útero en Cataluña durante el período 1999-2007 y de los dos registros de cáncer de base poblacional

catalanes, de Girona y Tarragona. El método basado en la supervivencia de las mujeres afectadas por esta enfermedad en Girona y Tarragona mostró ser el más apropiado en cuanto a las estimaciones de prevalencia a 5 años obtenidas y al coste en tiempo de cálculo.

Abstract

Keywords: prevalence, cancer of the cervix uteri, cohort simulation, survival.

MSC2014: 92D30, 62N02, 62P10.

The prevalence of a disease, as well as the incidence and the mortality, are useful measurements for health services planning, and particularly for the provision of services. The prevalence is the absolute number of individuals, or the proportion of the population, found to have a disease at a given time requiring some form of medical assistance. In this way, the prevalence some how express the demand of a given population for health care services. In most cases, most of the treatment phase, at least that involving the health services, takes place within the first 5 years from diagnosis. For this reason, prevalent cases of cancer are defined for those individuals that are alive and were diagnosed within the last 5 years (5-year prevalence). Since the resource requirements for specific phases of cancer care vary, alternative definitions of prevalence are commonly used. The estimates of 1, 2-3 and 4-5 years point prevalence describe the number of cases diagnosed alive at 1, 2-3 and 4-5 years after diagnosis. For the majority of cancers these estimates of point prevalence are applicable to the evaluation of initial treatment, clinical follow-up, and point of cure respectively. The cancer prevalence may be estimated directly from population-based cancer registries by counting the number of cases that are still alive at a specified point in time. However, this approach requires the registration and follow-up of vital status over many years. When follow-up registries are not available, prevalence can be alternatively estimated from simulated cohorts obtained from projections of incident cases of the target population, and from cancer registries of a reference population. We start this work reporting several measurements useful for health services planning related to the occurrence of cervix uteri cancer by 5-year age intervals from Girona and Tarragona between 1999 and 2007 (annual incidence, point prevalences, and 5-year prevalence). All these measures were estimated directly from their corresponding population-based cancer registries. A posteriori, we present 8 methods that simulate cohorts of women diagnosed with cervix uteri cancer, that allow to estimate the 5-year prevalence of the disease. Prevalence for Catalonia was estimated from 5-year age intervals for each year between 2003 and 2007. The simulated cohorts were obtained from available estimated data of cervical cancer incidence in Catalonia during the years 1999 to 2007, and from the two available Catalan population-based cancer registries from Girona and Tarragona. The method based on the survival of affected women from Girona and Tarragona, revealed to be the more appropriate due to 5-year prevalence estimates obtained and the computational cost of the simulation.

Abreviaturas

HIV	Virus de inmunodeficiencia humana
HPV	Virus del papiloma humano
ML	Máxima verosimilitud
RCBP	Registros de cáncer de base poblacional
RDm	Razón de discrepancia media
RDmT	Razón de discrepancia media total

Índice general

1

Introducción

La ocurrencia de una enfermedad describe la frecuencia de la enfermedad en una población. Entre las medidas de ocurrencia más usadas podemos distinguir: prevalencia, como la proporción de individuos en una población afectados por la enfermedad; incidencia acumulada, como la proporción de nuevos casos de la enfermedad (casos incidentes) en una población sana definida dentro de un período de tiempo especificado (es una medida de riesgo); y la tasa de incidencia, que es la tasa en la que ocurren nuevos casos de la enfermedad en la población (Porta 2014).

De acuerdo a las estimaciones de Globocan en 2012 se diagnosticaron en el mundo 14 millones de nuevos casos de algún tipo de cáncer y 8 millones de personas murieron por causa de esta enfermedad (Ferlay et al 2013).

En las mujeres, el cáncer de cuello de útero tiene un gran impacto. Se estima que cada año se diagnostican más de 527,000 nuevos casos, y más de 265,000 mueren. Es el cuarto tipo de cáncer más frecuente entre las mujeres del mundo. La incidencia de cáncer de cuello de útero varía mucho geográficamente (Figura 1.1), afectando más a los países en desarrollo; 84 % de los casos incidentes y 87 % de las muertes ocurren en éstos países (Globocan 2012: Ferlay et al 2013).

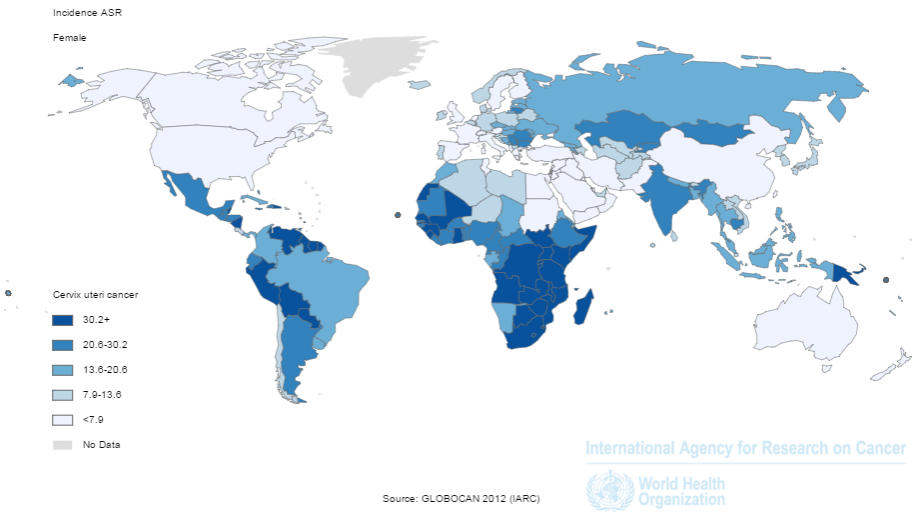

Figura 1.1: Tasa de incidencia estandarizada por edad según la población mundial (por cada 100.000 mujeres-año) de cáncer de cuello de útero a nivel de país.

La baja incidencia de cáncer de cuello de útero en España, así como en otros países desarrollados, se debe en parte a la efectividad de los programas de cribado (organizados y oportunista) basados en la citología cérvico-vaginal (prueba de Papanicolau). El objetivo del cribado es detectar lesiones precursoras en el epitelio cervical que serían el antecedente al cáncer invasor. La larga duración de las lesiones que lo preceden y su adecuado tratamiento, permiten la prevención del carcinoma invasor.

En España, la incidencia de cáncer de cuello de útero presenta unas tasas ajustadas por edad a la población estándar mundial de entre 3.8 y 8.5 casos por cada 100,000 mujeres-año de acuerdo a los datos publicados en *Cancer Incidence in Five Continents Vol. X* (Forman et al 2013). La incidencia de cáncer de cérvix es edad específica ya que son raros los casos en mujeres menores de 30 años pero rápidamente alcanzan una meseta después de los 40-45 años (Bosch 1999). En Cataluña, la incidencia de este tipo de cáncer se ha mantenido constante en los últimos 15 años (1993-2007; Figura 1.2a). Sin embargo, cuando se analiza la tendencia de la incidencia según la edad de las pacientes, se observa un incremento de la incidencia para las mujeres de 40-44 años en el periodo 1993-2002, mientras que en el período 2003-2007 dicho incremento se produce entre los 50 y 60 años (Figura 1.2b). Se ha indicado que el incremento de la incidencia de este tipo de cáncer ocurrió a partir de 1930, de la misma forma que se ha observado en otros países europeos excepto en Francia y en Suiza (Vaccarella et al 2013).

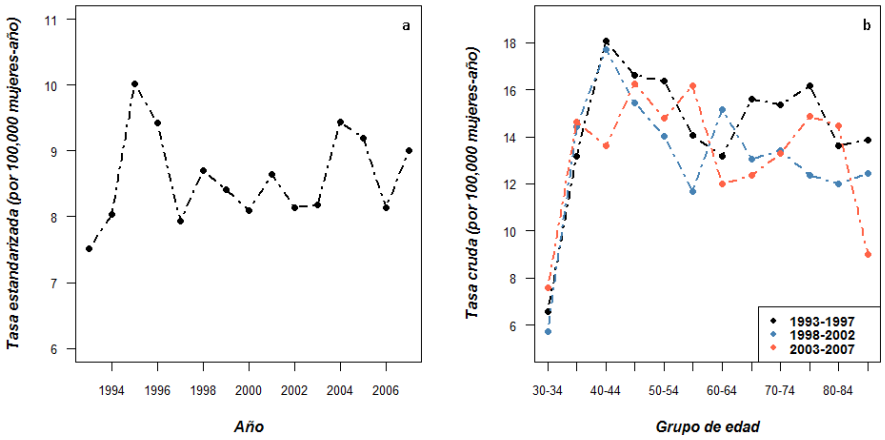

Figura 1.2: A la izquierda (a) tasa de incidencia anual estandarizada por población mundial para cáncer de cuello de útero en Cataluña (período 1993-2007) expresada en casos por cada 100,000 mujeres-año. A la derecha (b), tasa cruda (por cada 100,000 mujeres-año) de cáncer de cuello de útero en grupos de edad definidos por intervalos de cinco años entre 30-34 años hasta más de 85 años para diferentes períodos en España. (Fuente: CI5*plus*, Ferlay et al 2014).

En cuanto a la mortalidad por cáncer de cuello de útero, es más difícil de determinar ya que en España, en una parte considerable de los boletines estadísticos de defunción (BED) la causa de muerte se codifica como cáncer de útero no especificado (Loos et al 2004). Un estudio de revisión de los BEDs de pacientes diagnosticadas de cáncer de cuello uterino (1985-1989) muestra que el 24 % de los cánceres de útero no especificados (Clasificación Internacional de Enfermedades CIE, CIE-9: 179) corresponderían a cáncer de cuello uterino, el 29 % a cáncer del cuerpo de útero y el resto a otras localizaciones (Sánchez et al 1996a). Por este motivo en España frecuentemente no se analiza la tendencia de la mortalidad del cuello y cuerpo uterinos por separado, sino que se estudia como una única localización (útero).

Según los datos disponibles para Cataluña, la mortalidad por cáncer de cuello uterino se ha mantenido estable en esta comunidad en los últimos 20 años (Cléries et al 2014), aunque si se corrigiese por los tumores de útero no especificados (CIE-O: 179) posiblemente se observaría un descenso (Sánchez et al 1996a).

1.1. Cáncer de cuello de útero e infección por virus del papiloma humano (HPV)

Existe una relación causal entre el desarrollo de cáncer de cuello de útero y la infección por el virus del papiloma humano HPV (Bosch et al 2002, zur Hausen

2000), aunque la infección por HPV no es suficiente indicando que existen otros factores promotores (Trottier & Franco 2006).

En el desarrollo del cáncer de cuello de útero se pueden definir 4 fases: la infección por HPV en el epitelio, la persistencia viral del HPV, la progresión a lesiones precancerosas y la invasión a través de la membrana basal del epitelio.

1.1.1. Transmisión sexual del HPV

Se considera esencial para desarrollar la enfermedad. La infección genital por HPV es una de las enfermedades de transmisión sexual más comunes en el mundo. La trasmisión de la infección se produce a través de las relaciones sexuales cuando las lesiones causadas por el HPV padecen microtraumas durante el coito vaginal o anal, y el virus se desprende ingresando a través de la mucosa del compañero sexual (Winer et al 2003). Otra vía de trasmisión del HPV es el sexo oral (Hernandez et al 2008).

Los factores de riesgo que se asocian a la adquisición de la infección por HPV son el inicio de las relaciones sexuales a temprana edad, el elevado número de parejas sexuales nuevas y recientes, y el elevado número de parejas sexuales de la pareja masculina (Almonte et al 2008, Muñoz et al 2006).

Aunque un gran porcentaje de las mujeres y hombres se infectarán por HPV en algún momento de sus vidas (50 % o más) (Paavonen 2007), normalmente la infección es asintomática y desaparece espontáneamente. La mayoría de las infecciones por HPV son transitorias y en general remiten en unos meses; entorno al 90 % se eliminará en 2 años (Rodríguez et al 2010, Winer et al 2011). Solo una pequeña proporción de mujeres con infección persistente por HPV de alto riesgo desarrollará lesiones cervicales precancerosas y una proporción menor a cáncer de cuello de útero. Aunque la infección por HPV es necesaria para la carcinogénesis, ciertos cofactores ayudan en la progresión desde la infección hasta el cáncer (Bosch & de Sanjosé 2007).

1.1.2. Persistencia y progresión

No existe una única definición de persistencia, pero en general se determina como la detección del mismo tipo del HPV en dos o más ocasiones, con un intervalo de tiempo determinado entre exploraciones, aunque no está oficialmente reconocido ningún umbral entre la transitoriedad y la persistencia (Moscicki et al 2012).

La persistencia de tipos del HPV de alto riesgo es necesaria para el desarrollo, el mantenimiento y la progresión de las lesiones precancerosas. Solamente una pequeña proporción de las infecciones será persistente y el tiempo que transcurre entre la infección y las primeras evidencias microscópicas de la existencia de lesiones precancerosas puede ser sorprendentemente corto, a menudo de 5

años (Diaz Sanchis 2014).

La mayoría de las mujeres infectadas por HPV no desarrollan cáncer de cuello de útero lo que indica que existen ciertos factores que intervienen en el desarrollo del cáncer. Los factores de riesgo para la persistencia y progresión a cáncer de las mujeres infectadas con HPV no se han determinado con precisión pero si se han identificado algunos potenciales factores que pueden agruparse en tres categorías (Almonte et al 2008, Muñoz et al 2006):

Cofactores ambientales o exógenos (Castellsagué, Bosch & Muñoz 2002, IARC 2007) como consumo de tabaco (Zeng et al 2012, Louie et al 2011), uso prolongado de anticonceptivos orales, alta paridad, coinfección con otras enfermedades de transmisión sexual (infección por *Chlamydia trachomatis*, infección por virus del herpes simplex 2, infección por virus de inmunodeficiencia humana HIV), la edad del primer embarazo a término, la dieta y el sobrepeso.

Características virales (IARC 2007, Wang & Hildesheim 2003) como la infección por tipos específicos de HPV, coinfección con otros tipos de HPV, variantes del HPV, carga viral e integración viral. El tipo de HPV es el cofactor más importante que afecta el riesgo de persistencia viral y de progresión a lesiones precancerosas. Se han caracterizado más de 100 tipos de HPV a nivel molecular, 40 de éstos son capaces de infectar el tracto genital y 12 de ellos están clasificados como cancerígenos (Doorbar et al 2012). Los tipos genitales del HPV se han clasificado en diferentes grupos en función de su asociación con el desarrollo del cáncer de cuello de útero (Bouvard et al 2009, Muñoz et al 2004):

> De alto riesgo, considerados como carcinógenos humanos (HPVs 16,18, 31, 33, 35, 39, 45, 52, 56, 58 y 59). Los tipos 16 y 18 son la causa del 70 % de todos los casos de cáncer de cuello de útero, dada su mayor capacidad de transmisión y persistencia, y a una progresión más rápida a lesiones precanserosas.
>
> De probable alto riesgo, considerados como probablemente carcinógenos (HPV 68).
>
> De posible alto riesgo, considerados como posiblemente carcinógenos (HPVs 26, 30, 34, 53, 66, 67, 70, 73, 82, 85 y 97).
>
> De bajo riesgo, considerados como no carcinógenos y asociados principalmente a verrugas genitales y a epitelio normal (HPVs 6 y 11).

Cofactores del huesped (IARC 2007, Wang & Hildesheim 2003) que incluyen hormonas endógenas, factores genéticos y otros factores relacionados a la respuesta inmunológica (mayor riesgo en personas con inmunodepresión asociada a la infección por HIV, o al tratamiento de una patología autoinmune).

1.1.3. El cáncer de cuello de útero

Es la culminación de un largo proceso, a menudo entre 10-20 años, que comienza con la infección de HPV crónica de las células en la superficie del cuello uterino. Esto conduce inicialmente a anomalías celulares que pueden persistir durante años o simplemente desaparecer. Las infecciones persistentes pueden progresar a lesiones de bajo grado (donde la mayoría pueden regresar naturalmente) o pueden progresar a lesiones de alto grado (particularmente las asociadas a tipos de HPV de alto riesgo). Estas últimas lesiones tienen mayor probabilidad de adquirir un potencial invasivo extendiéndose más allá del epitelio cervical y causar la enfermedad maligna (Diaz Sanchis 2014).

Por lo tanto, el cáncer de cérvix se considera que es una complicación tardía y poco frecuente de una infección persistente por HPV y es el resultado final de una cadena de eventos que pueden tardar muchos años en desarrollarse. El riesgo de desarrollar este tipo de cáncer depende principalmente de la infección por HPV y actualmente, de la falta de programas efectivos de cribado que permitan la detección precoz de la enfermedad. El carcinoma de células escamosas, que se inicia en las células epiteliales, es el tipo histológico más frecuente (80-85 % de los casos), seguido del adenocarcinoma (15-20 % de los casos) que aparece en las células epiteliales glandulares (Diaz Sanchis 2014).

1.2. Estimaciones de prevalencia

La prevalencia de una enfermedad, así como la incidencia y la mortalidad, es una medida epidemiológica muy útil para la planificación de los servicios de salud pública. En particular en el desarrollo de estrategias para la prestación de servicios. La prevalencia mide el número absoluto, y la proporción relativa en la población, de individuos afectados por la enfermedad en un tiempo dado y que requieren alguna forma de atención médica. Por lo tanto expresa la demanda de cuidados, una información necesaria en el desarrollo de estrategias para la prestación de servicios de salud. La prevalencia de una enfermedad en la población usualmente se determina mediante la combinación de la tasa de incidencia y la supervivencia (Pisani et al 1997). Uno de los problemas de definir prevalencia de cáncer consiste en establecer cuando un paciente está curado, ya que puede haber una recaída aún luego de varios años libres de la enfermedad. En la mayoría de los casos de cáncer la fase de tratamiento que involucra los servicios de salud, se da dentro de los primeros 5 años desde el diagnóstico. Por esta razón, en general se consideran casos prevalentes de cáncer aquellos individuos que continúan vivos cuando fueron diagnosticados dentro de los últimos 5 años (prevalencia a 5 años). Los individuos que viven a los 5 años de diagnosticados son considerados "curados". A excepción de algunos tipos de cáncer (ej. cáncer de mama), en general los casos que sobreviven a los 5 años de diagnóstico muestran la misma sobrevivencia que la población general (Pisani et al 1997).

Los recursos requeridos para el tratamiento de pacientes diagnosticados más recientemente son muy diferentes a los requeridos por los pacientes que llevan más tiempo en tratamiento. Por ello, comúnmente se usan definiciones alternativas de prevalencia. Las estimaciones de prevalencia puntual a 1, 2-3, y 4-5 años (también llamada de duración limitada, Bray et al 2013) describen el número de casos que continúan vivos luego de 1, 2-3, y 4-5 años de diagnosticados por cáncer (Pisani et al 1997). Estas estimaciones de prevalencia puntual son aplicables a: (i) tratamiento de evaluación inicial (dentro del primer año), (ii) seguimiento clínico (a los 2-3 años) y (iii) punto de cura (a los 4-5 años), para la mayor parte de los tipos de cáncer. En el contexto de los servicios de planificación de salud, esta forma de definir prevalencia tiene un sentido muy útil dado que ayuda a identificar los recursos requeridos de acuerdo a la fase de tratamiento en que se encuentran los pacientes (Pisani et al 1997, Bray et al 2013). Por otra parte, además de la fase de tratamiento, otra variable de interés para identificar las necesidades de los servicios de salud, es la edad de los pacientes ya que ésta puede determinar el tipo de tratamiento requerido (por ejemplo, un tratamiento paliativo en pacientes diagnosticados con edad avanzada).

Las estimaciones de prevalencia pueden hacerse directamente desde los registros de cáncer de base poblacional (RCBP) por contar el número de casos aún vivos de la población en un punto de tiempo especificado (Feldman et al 1986, Gail et al 1999, Krogh & Micheli 1996). Sin embargo, esta aproximación requiere el registro y el seguimiento del estatus vital de los pacientes por varios años, y únicamente proporciona resultados de prevalencia para las áreas cubiertas por el registro de cáncer de base poblacional. Por esta razón, las estimaciones de prevalencia a menudo se obtienen mediante métodos indirectos basados en su relación matemática con la incidencia y la supervivencia (Bray et al 2013, Capocaccia & De Angelis 1997, Colonna et al 2008, Pisani et al 2002, Verdecchia et al 2002).

Una aproximación alternativa y muy poco explorada es generar mediante métodos de simulación, cohortes de casos incidentes de la enfermedad de interés durante un período de tiempo determinado, que permitan estimar prevalencia por edad, aun cuando no se tienen registros de seguimiento. Asimismo, dado que la prevalencia combina tasa de incidencia y supervivencia, una adecuada estimación de prevalencia puede indicar la calidad de las simulaciones obtenidas y así, la eficacia del método desarrollado. Si las cohortes simuladas reproducen bien el patrón de la enfermedad en la población, esta metodología permitiría además realizar otros tipos de análisis (ej: análisis de supervivencia, mortalidad).

En este trabajo presentamos ocho métodos diferentes para simular cohortes de mujeres incidentes de cáncer de cuello de útero que permiten realizar estimaciones de prevalencia a 5 años por intervalo de edad, usando como base los registros de cáncer de una población de referencia. Como casos a simular usamos estimaciones de incidencia de cáncer de cérvix para Cataluña entre 1999 y 2007, y como población de referencia, los datos de los dos registros de cáncer

catalanes de base poblacional: de las provincias de Girona y Tarragona. Si bien los métodos planteados se aplican al cáncer de cuello de útero, la metodología desarrollada podría ser utilizada como base para realizar estimaciones de prevalencia de otros tipos de cánceres. Por otra parte, las funciones programadas en R para la implementación de los métodos presentados podrán ser usadas en un futuro para crear una librería o paquete estadístico.

1.3. Objetivos

1. Estimar incidencia anual, prevalencia (acumulada a 5 años) y prevalencia puntual a 1, 2-3 y 4-5 años de cáncer de cuello de útero en Girona y Tarragona a partir de sus registros de cáncer entre el 1 de enero de 1999 y el 31 de diciembre de 2007.

2. Describir la función de supervivencia por grupos de edad de las mujeres diagnosticadas por cáncer de cuello de útero en Girona y Tarragona entre 1999 y 2007 (y seguidas hasta el 30 de junio de 2010) con el fin de establecer una definición de grupos de edad que mejor resuma la supervivencia de las pacientes.

3. Desarrollar diferentes métodos de simulación sobre el tiempo de seguimiento que permitan generar una cohorte de mujeres diagnosticadas por cáncer de cérvix a partir de estimaciones de incidencia para un período dado y como base referencia, las mujeres de los registros de cáncer de Girona y Tarragona que fueron diagnosticadas en el período 1999-2007 y seguidas hasta el 30 de junio de 2010.

4. Identificar el mejor método de simulación a través de la comparación entre las estimaciones de prevalencia a 5 años realizadas sobre las simulaciones y las obtenidas directamente de los registros de Girona y Tarragona para cada año entre 2003 y 2007.

5. Estimar la prevalencia a 5 años de cáncer de cérvix en Cataluña para 2003, 2004, 2005, 2006 y 2007 utilizando simulaciones de cohortes de mujeres incidentes de cáncer de cérvix en Cataluña entre 1999-2007 y los registros de cáncer de Girona y Tarragona durante el mismo período con seguimiento hasta el final de junio de 2010.

6. Implementar en R todos los métodos desarrollados mediante la programación de funciones propias.

1.4. Estructura del trabajo

En la Introducción se presentaron los antecedentes, las motivaciones y los objetivos de este trabajo. En la siguiente sección se encuentran las definiciones y el marco teórico, así como la descripción de los datos utilizados, necesarias para realizar el análisis descriptivo de los datos de cáncer de cuello de útero en lo que es nuestra población de referencia, las provincias de Girona y Tarragona (1999-2007). En primer lugar lo que se refiere a medidas de utilidad para los servicios de planificación de salud como la tasa de incidencia anual, la prevalencia (acumulada a 5 años) y las prevalencias puntuales a 1, 2-3 y 4-5 años. Y en segundo lugar, lo referente al análisis de la supervivencia observada por grupos de edad, así como la evaluación del grado de ajuste de los datos a diferentes modelos paramétricos de distribución para el tiempo de muerte, y la estimación de sus parámetros.

La sección 3 está dedicada a presentar los diferentes métodos desarrollados para simular cohortes de mujeres incidentes de cáncer de cuello de útero; cuatro basados en remuestreo y cuatro basados en modelos de supervivencia. También se muestra cómo se realizan las estimaciones de prevalencia a 5 años de cáncer de cuello de útero en Cataluña mediante la simulación de cohortes basada en la población de Girona y Tarragona, y por último, la validación de los métodos realizada sobre la base de referencia.

En la sección Resultados se presentan tanto las medidas de ocurrencia de la enfermedad y el análisis de la supervivencia para la población de Girona y Tarragona, como las estimaciones de prevalencia a 5 años obtenidas por los métodos de simulación de cohortes en la misma población de referencia (Validación), y en Cataluña para cada año entre 2003 y 2007.

Finalmente se presenta la discusión y las conclusiones.

2

Métodos para el análisis descriptivo de los datos de cáncer de cuello uterino en Girona y Tarragona (1999-2007)

2.1. Medidas de ocurrencia de la enfermedad - Definiciones

Incidencia acumulada o proporción de incidencia es la proporción de nuevos casos dentro de un período de tiempo t especificado entre la población inicialmente sana:

$$I_{ac}(t) = \frac{I}{N_0}$$

donde N_0 es el tamaño de la población inicialmente sana e I el número de nuevos casos (**casos incidentes**) durante el período de tiempo. La incidencia acumulada puede interpretarse como la probabilidad de que un individuo sano de la población se enferme durante el período de tiempo especificado.

Tasa de incidencia o tasa cruda (Porta 2014) es la tasa en la que ocurren nuevos casos de la enfermedad, y se expresa como número de nuevos casos por unidad de persona-tiempo a riesgo durante un período definido:

$$I_r = \frac{I}{\Delta T}$$

donde I son los caso incidentes y ΔT es el tiempo total bajo riesgo de la población entera (suma de unidades persona-tiempo a riesgo de todas las personas

durante el período).

Si la tasa de incidencia es baja y constante, la incidencia acumulada en un período de tiempo de duración Δ puede ser estimada como:

$$I_{ac}(\Delta) \approx Ir \cdot \Delta$$

Por tanto, para un período de un año, se puede estimar la **tasa de incidencia anual** como:

$$I_r = \frac{I}{N^\star}$$

donde I son los casos incidentes dentro del año y $N^\star$ la población a mitad de año que es usada como una aproximación para ΔT.

Prevalencia o proporción de prevalencia (Porta 2014) es la proporción de individuos en una población afectados por la enfermedad en un tiempo especificado:

$$P = \frac{X}{N}$$

donde X representa la población enferma en un tiempo t particular (o período) y N es el tamaño de la población en el mismo tiempo t o a mitad del período. La prevalencia puede interpretarse como la probabilidad de que un individuo de la población tenga la enfermedad en el tiempo t.

En cáncer, usualmente se usan definiciones de prevalencia de duración limitada: número de pacientes diagnosticados con cáncer dentro de un período de tiempo fijo en el pasado (Bray et al 2013). Se considera población enferma de cáncer (X) en un tiempo t a los individuos diagnosticados dentro de los últimos 5 años (prevalencia a 5 años, Pisani et al 1997).

Prevalencia puntual a 1, 2-3, y 4-5 años corresponde a la proporción de casos prevalentes (afectados por cáncer) que llevan hasta un año (X_1), entre 2 y 3 años (X_{2-3}), y entre 4 y 5 años (X_{4-5}) de diagnosticados respectivamente entre la población (N) en el tiempo t (Pisani et al 1997).

Habitualmente las estimaciones de prevalencia de cáncer de duración limitada son obtenidas mediante el producto entre la incidencia y la supervivencia (Pisani et al 1997, Bray et al 2013). Es decir, los casos prevalentes a n años para la edad k se estiman en base a tasas de incidencia y probabilidades de supervivencia año-específicas de acuerdo a la siguiente fórmula:

$$\sum_{i=1}^{n} I_{k-i} \cdot S_{k-i}(i - 0{,}5)$$

donde I_k es el número anual de nuevos casos de edad k y S_k es la proporción de casos diagnosticados a edad k y vivos en el tiempo t después del diagnóstico,

con n el número de años como caso prevalente siguientes al diagnóstico (Pisani et al 1997, Bray et al 2013). Esta forma de estimar prevalencia de cáncer es la más comúnmente usada en áreas que no disponen de registros de cáncer de base poblacional, para la cual en general se utilizan estimaciones indirectas de incidencia y supervivencia (Ferlay et al 2015).

2.2. Análisis del tiempo de supervivencia

Sea T una variable aleatoria, continua y no negativa, que representa el tiempo hasta la muerte. **Supervivencia** es la probabilidad de que un paciente continúe vivo luego de transcurrido t unidades de tiempo desde un tiempo cero definido (Klein & Moeschberger 2003), en este caso desde que fue diagnosticado por cáncer de cuello de útero. Dicho de otro modo, es la probabilidad de que la muerte (evento) ocurra después de t unidades de tiempo desde el diagnóstico para cáncer:

$$S(t) = Prob(T > t)$$

La función supervivencia $S(t)$ está definida para $t \geq 0$, parte de 1, decrece monotónicamente y converge a cero cuando el tiempo t tiende a infinito.

La distribución del tiempo de supervivencia T puede ser caracterizada por otras funciones de utilidad como lo son la función de densidad de probabilidad $f(t)$, la función de riesgo $h(t)$ y la función de riesgo acumulada $H(t)$ (Klein & Moeschberger 2003).

La **función de densidad de probabilidad** $f(t)$ representa la probabilidad de que ocurra la muerte en un pequeño intervalo de tiempo y se define como el siguiente límite:

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

La función de densidad $f(t)$ es no negativa y se cumple:

$$S(t) = Prob(T > t) = \int_t^{\infty} f(u)\, du; \qquad f(t) = \frac{-\, dS(t)}{dt}$$

La **función de riesgo o tasa de fallo** $h(t)$ describe la probabilidad condicionada a morir en un pequeño intervalo dado que la persona estaba viva al inicio del mismo:

$$h(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} Prob[t \leq T < t + \Delta t \mid T \geq t]$$

La función de riesgo $h(t)$ es no negativa y se relaciona a la función de supervivencia $S(t)$ de la siguiente manera:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\ln S(t))$$

La **función de riesgo acumulada** $H(t)$ aunque no tiene una interpretación intuitiva es muy útil gráficamente. Se define como:

$$H(t) = \int_0^t h(u)\,du = -\ln S(t) \tag{2.1}$$

La función de riesgo acumulada $H(t)$ es no negativa y monótona creciente.

2.2.1. Estimación de la función supervivencia $S(t)$

Hay diferentes maneras de estimar la función supervivencia $S(t)$, una de ellas es la estimación de Kaplan-Meier (1958). Este estimador es un estimador no paramétrico que tiene en cuenta la censura por la derecha. Para K tiempos ordenados, $t_1 < t_2 < \cdots < t_K$

$$\hat{S}_{KM}(t) = \begin{cases} 1, & \text{si } t < t_1 \\ \prod_{t_i \leq t}(\frac{n_i - d_i}{n_i}), & \text{si } t_1 \leq t \end{cases}$$

donde d_i es el número de muertes en el momento t_i, y n_i es el número de sujetos a riesgo justo antes de t_i. Cuando el último tiempo no corresponde a una muerte sino a una censura, el estimador de la supervivencia no está bien definido y debe redefinirse $S(t)$ para tiempos mayores al último tiempo observado (por ejemplo, $S(t) = 0$ para todo t mayor al último tiempo observado).

Otra forma de estimar la función de supervivencia $S(t)$ es mediante el estimador de riego acumulado $H(t)$ de Nelson-Aalen, introducido por Nelson en 1972 y más tarde derivado por Aalen en 1978 utilizando técnicas de procesos contadores. Entre las propiedades de este estimador se encuentra su conveniencia en situaciones donde el tamaño de muestra es pequeño (Klein & Moeschberger 2003). La estimación de la supervivencia $S(t)$ se obtiene de acuerdo a la relación 2.1 o equivalentemente $S(t) = \exp[-H(t)]$, a partir del estimador de riesgo acumulado $H(t)$ de Nelson-Aalen,

$$\hat{H}_{NA}(t) = \begin{cases} 0, & \text{si } t \leq t_1 \\ \sum_{t_i \leq t}(\frac{d_i}{n_i}), & \text{si } t_1 \leq t \end{cases}$$

donde d_i/n_i es un estimador de la función de riesgo $h(t)$ en el momento t_i.

La estimación de la función de supervivencia por Kaplan-Meier y por Nelson-Aalen puede obtenerse mediante la función `survfit` del paquete `survival` (Therneau 2015) implementado en R.

La estimación empírica de la función de riesgo acumulada $H_{NA}(t)$ puede ser usada en un análisis exploratorio para seleccionar el modelo paramétrico que ajuste los datos de tiempo de muerte (Klein & Moeschberger 2003). En

estudios de cáncer de base poblacional los modelos paramétricos más usados son Exponencial, Weibull y log-Logístico (De Angelis et al 1999).

2.2.2. Modelos paramétricos

Distribución Exponencial

Bajo un modelo donde el tiempo de muerte se distribuye de forma exponencial la función de riesgo es contante $h(t) = \lambda$, por lo que las funciones de supervivencia $S(t)$ y de densidad $f(t)$ quedan definidas respectivamente como:

$$S(t) = e^{-\lambda t}; \qquad f(t) = \lambda e^{-\lambda t}$$

Y la función de riesgo acumulado $H(t)$:

$$H(t) = \lambda\, t \tag{2.2}$$

De acuerdo a la ecuación 2.2 la representación gráfica de la función de riesgo acumulado $H(t)$ vs tiempo t describe una recta de pendiente λ que pasa por el origen. De esta manera se puede evaluar el grado de ajuste de los datos al modelo exponencial y estimar su parámetro λ. Dado que la función de riesgo $h(t)$ es constante a menudo se denomina al parámetro λ tasa de mortalidad o tasa de fallo.

Distribución de Weibull

En un modelo donde el tiempo de fallo (muerte) sigue una distribución Weibull, la función de riesgo depende de una potencia del tiempo $h(t) = \lambda\rho t^{\rho-1}$ para todo $t > 0$, con forma $\rho > 0$ y escala $\lambda > 0$. Las funciones de supervivencia $S(t)$ y de densidad $f(t)$ quedan definidas respectivamente como:

$$S(t) = \exp[-\lambda t^{\rho}]; \qquad f(t) = \lambda\rho t^{\rho-1}\exp[-\lambda t^{\rho}]$$

Y la función de riesgo acumulado $H(t)$:

$$H(t) = \lambda\, t^{\rho}$$

Al aplicar la transformación a logaritmos sobre la función de riesgo acumulado $H(t)$ se obtiene:

$$\ln H(t) = \ln\lambda + \rho\, \ln t \tag{2.3}$$

por lo que la representación gráfica de $\ln H(t)$ vs $\ln t$ muestra una recta de ordenada en el origen $\ln\lambda$ y pendiente ρ (ecuación 2.3). De esta forma se puede evaluar el grado de ajuste de los datos al modelo Weibull y realizar la estimación de sus parámentros de forma ρ y escala λ.

Cuando $\rho = 1$ la distribución corresponde a una Exponencial ya que la distribución Exponencial constituye un caso particular de la distribución de Weibull.

Distribución log-Logística

Si $Y = \ln T$ presenta una distribución logística, T sigue una distribución log-Logística con la siguiente función de densidad:

$$f(t) = \frac{\rho t^{\rho-1}\lambda}{[1+\lambda t^{\rho}]^2}$$

Las funciones de supervivencia $S(t)$ y de riesgo $h(t)$ quedan definidas respectivamente como:

$$S(t) = \frac{1}{1+\lambda t^{\rho}}; \qquad h(t) = \frac{\rho t^{\rho-1}\lambda}{1+\lambda t^{\rho}}$$

Y la función de riesgo acumulado $H(t)$:

$$H(t) = \ln(1+\lambda t^{\rho})$$

Al aplicar la transformación exponencial sobre la función de riesgo acumulado $H(t)$ y reordenado términos se tiene,

$$e^{H(t)} = 1+\lambda t^{\rho} \Longrightarrow e^{H(t)} - 1 = \lambda t^{\rho}$$

Y al transformar a logaritmos,

$$\ln\left(e^{H(t)} - 1\right) = \ln\lambda + \rho \ln t \tag{2.4}$$

La ecuación 2.4 corresponde a una recta con pendiente ρ y ordenada en el origen $\ln\lambda$. Por lo que la representación gráfica de $\ln(e^{H(t)} - 1)$ vs el $\ln t$ puede ser utilizada para evaluar el grado de ajuste de los datos al modelo de distribución log-Logístico y estimar sus parámetros ρ y λ.

Estimación de parámetros por máxima verosimilitud

Si bien la estimación de los parámetros de los modelos paramétricos puede realizarse mediante métodos gráficos, usualmente se obtienen mediante el método de máxima verosimilitud (ML) ya que son más precisos, son consistentes y asintóticamente eficientes.

Los tres modelos de distribución, Exponencial, Weibull y log-Logístico, admiten una representación log-lineal:

$$Y = \ln T = \mu + \sigma W$$

Si la variable aleatoria T sigue una distribución de Weibull de parámetros ρ y λ, W sigue una distribución estándar del valor extremo (Gumbel), y parámetros $\rho = \sigma^{-1}$ y $\lambda = e^{-\mu/\sigma}$. Igualmente si T sigue una distribución Exponencial, ya que ésta corresponde a un caso particular de una distribución de Weibull con $\rho = 1$.

Las funciones de densidad y de supervivencia de la variable aleatoria W con distribución del valor extremo estándar son, respectivamente:

$$f_W(w) = exp[w - e^w]; \qquad S_W(w) = exp[-e^w]$$

Y la supervivencia de $Y = \ln T$ es:

$$S_Y(y) = exp[-\lambda e^{\rho y}]$$

Por otra parte, si la variable aleatoria T sigue una distribución log-Logística de parámetros ρ y λ, W sigue una distribución logística, y parámetros $\rho = \sigma^{-1}$ y $\lambda = e^{-\mu/\sigma}$.

Las funciones de densidad y de supervivencia de la variable aleatoria W con distribución Logística son, respectivamente:

$$f_W(w) = \frac{e^w}{(1+e^w)^2}; \qquad S_W(w) = \frac{1}{1+e^w}$$

Y la supervivencia de $Y = \ln T$ es:

$$S_Y(y) = \frac{1}{1+\lambda e^{y\rho}}$$

Entonces, los estimadores de máxima verosimilitud de los parámetros μ y σ se obtienen por maximizar (por métodos numéricos) la función de verosimilitud para datos censurados por la derecha:

$$L(\mu, \sigma) = \prod_{j \in Muertes} f_Y(y_j) \prod_{j \in Censuras} S_Y(y_j)$$

En el modelo log-lineal $Y = \ln T = \mu + \sigma W$ la función de densidad $f_Y(y)$ y de supervivencia $S_Y(y)$ son, respectivamente:

$$f_Y(y) = \frac{1}{\sigma} f_W(\frac{y-\mu}{\sigma}); \qquad S_Y(y) = S_W(\frac{y-\mu}{\sigma})$$

Por lo que la función de verosimilitud puede expresarse de la siguiente manera:

$$L(\mu, \sigma) = \prod_{i=1}^{n} [\frac{1}{\sigma} f_W(\frac{y_i - \mu}{\sigma}]^{\delta_i} [S_W(\frac{y_i - \mu}{\sigma})]^{1-\delta_i}$$

donde $\delta_i = 1$ si la observación es una muerte o $\delta_i = 0$ si corresponde a una censura, y los términos f_W y S_W representan las expresiones de densidad y supervivencia correspondientes a la distribución de W; del valor extremo estándar si T sigue una Weibull o Exponencial, o Logística si T sigue una log-Logística.

Dada la propiedad de invarianza de los estimadores de máxima verosimilitud, a partir de las estimaciones de μ y σ podemos obtener las estimaciones máximo verosímiles de los parámetros ρ y λ. La función `survreg` del paquete `survival` (Therneau 2015) implementada en R permite ajustar el modelo log-lineal y obtener los estimadores de máxima verosimilitud de μ y σ.

2.3. Fuentes de información

En Cataluña hay dos registros de cáncer de base poblacional, provincias de Girona y Tarragona, que abarcan el 20 % de la población catalana (`http://www20.gencat.cat/portal/site/cancer/`). Los registros de cáncer de base poblacional recopilan información sobre los tumores malignos en una población determinada con la finalidad de evaluar el impacto del cáncer en esta población. Se utilizaron los registros de mujeres diagnosticadas por cáncer de cuello de útero entre el primero de enero de 1999 y el 31 de diciembre de 2007 en Girona y Tarragona. El seguimiento de dichas mujeres finalizó el 30 de junio de 2010 (fecha de cierre). La información registrada es la edad, el mes y el año de diagnóstico de la enfermedad, el mes y el año del último seguimiento, y el estatus de la paciente en el último seguimiento (0 si la paciente estaba viva y 1 en caso de muerte). A esta base le agregamos la variable seguimiento con el tiempo en años desde el diagnóstico hasta la fecha de muerte o, si la paciente permaneció viva, hasta la fecha de cierre de la base (30 de junio de 2010). La base de datos contiene el registro de 596 pacientes de entre 22 y 93 años de edad, con tiempos de seguimiento que van desde 1 mes a 11 años y 5 meses. En la Tabla 2.1 se presenta el número de mujeres diagnosticadas por cáncer de cuello uterino por intervalo de edad quinquenal incluidas en la base de datos.

Intervalo de edad	Edad	Número de pacientes
1	0-4	0
2	5-9	0
3	10-14	0
4	15-19	0
5	20-24	5
6	25-29	15
7	30-34	50
8	35-39	90
9	40-44	64
10	45-49	70
11	50-54	48
12	55-59	52
13	60-64	37
14	65-69	39
15	70-74	43
16	75-79	41
17	80-84	25
18	85≤	17
Total		596

Tabla 2.1: Número de mujeres por intervalo de edad incluidas en los registros de cáncer de base poblacional de las provincias de Girona y Tarragona entre 1 de enero de 1999 y 31 de diciembre de 2007.

La población femenina de las provincias de Girona y Tarragona, y su distribución por edad para el período de estudio (1999 y 2007) fue obtenida del Instituto de Estadística de Cataluña (IDESCAT, `http://www.idescat.cat/cat/mapa.html#poblacio`). Se consideraron 18 grupos de edad, cada uno a intervalos de 5 años, desde 0 a 4 años hasta más de 85 años.

Las estimaciones de casos incidentes de cáncer de cuello de útero por intervalo de edad quinquenal en la comunidad de Cataluña para cada año entre 1999 y 2007 fueron proporcionadas por Clèries et al (2014) y se presentan en la Tabla 2.2. Dichas estimaciones fueron realizadas en base a las tasas de incidencia en las provincias de Tarragona y Girona, y la tasa de mortalidad para todo Cataluña con métodos de modelado desarrollado por Clèries et al (2012).

	Año								
Edad	1999	2000	2001	2002	2003	2004	2005	2006	2007
0-4	0	0	0	0	0	0	0	0	0
5-9	0	0	0	0	0	0	0	0	0
10-14	0	0	0	0	0	0	0	0	0
15-19	0	0	0	0	0	0	0	0	0
20-24	0	0	0	0	5	5	5	5	5
25-29	12	6	11	17	6	5	11	0	15
30-34	28	22	28	17	33	38	49	27	31
35-39	32	48	43	64	75	53	68	47	46
40-44	44	54	27	43	37	21	41	36	35
45-49	23	28	49	54	37	37	52	51	40
50-54	36	29	23	34	33	22	11	26	51
55-59	30	13	35	35	6	28	99	21	26
60-64	35	46	12	0	28	56	0	22	11
65-69	39	11	44	6	33	33	33	11	5
70-74	28	16	43	16	27	38	21	21	21
75-79	11	22	22	22	5	32	37	26	42
80-84	0	0	11	16	42	11	21	21	11
85≤	23	0	11	16	0	16	5	11	11
Total	341	294	359	340	367	395	453	325	350

Tabla 2.2: Casos incidentes de cáncer de cuello de útero estimados por año y grupo de edad quinquenal para Cataluña entre 1999 y 2007 por Clèries et al 2014.

2.4. Análisis descriptivo para el cáncer de cuello de útero en Girona y Tarragona (1999-2007)

En cuanto a las medidas de interés para los servicios de planificación de salud, se realizaron estimaciones directas de la tasa de incidencia anual de cáncer de cuello de útero en Girona y Tarragona entre 1999 y 2007 a partir de los datos de sus registros de cáncer de base poblacional. También se realizaron estimaciones de prevalencia de duración limitada; prevalencia a 5 años (entre 2003 y 2007), y prevalencias puntuales a 1, 2-3, y 4-5 años.

Implementación en R: Para realizar dichas estimaciones de incidencia anual, prevalencia a 5 años y prevalencia puntual (para casos que llevan diferente tiempo en tratamiento) se implementaron dos funciones en R (versión 3.1.2, R Development Core Team 2014), ***carga*** y ***prevalencia*** (ver detalles en Apéndice A.1.1). Ambas funciones utilizan como información de entrada la base de datos de los registros de cáncer de base poblacional (aquí de Girona y Tarragona), y los datos con la distribución de la población femenina por año y edad.

Los paquetes requeridos son `Hmisc` (Harrell 2014) y `Epi` (Carstensen et al 2014). La función ***carga*** contabiliza desde la base de datos de los registros los casos incidentes en número absoluto y expresado en casos por cada 100,000 mujeres-año (tasa cruda de incidencia anual) para cada intervalo de edad quinquenal (1 a 18) que se acumulan a lo largo del año especificado (por ej. casos incidentes en 1999). La función ***prevalencia*** permite calcular los dos tipos de prevalencia definidas; acumulada y puntual. Aquí se contabilizan los casos incidentes que continúan vivos al final del año especificado que llevan en tratamiento un determinado número de años; si se quiere la prevalencia a 5 años se contabilizan los casos que están vivos al final del año de interés que llevan entre 1 y 5 años de tratamiento, si se quiere prevalencia puntual 2-3 años se contabilizan los casos que continúan vivos a final del año de interés que incidieron en los 2 y 3 años previos. La prevalencia se expresa en número de casos prevalentes y en casos por cada 100,000 mujeres.

Con el fin de describir la supervivencia de las pacientes incluidas en la base de datos, estimamos la función de supervivencia $S(t)$ de Kaplan-Meier por grupos de edad. Para ello usamos la función `survfit` del paquete `survival` (Therneau 2015). En este punto, se exploraron diferentes definiciones de grupos de edad para encontrar aquellos que mejor resuman las diferencias en la supervivencia de la población de referencia. Finalmente, se optó por trabajar con 6 grupos de edad: de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años.

Posteriormente, con el objetivo de explorar cuál sería el modelo paramétrico (Exponencial, Weibull o log-Logístico) más apropiado para la distribución del tiempo de supervivencia, evaluamos mediante métodos gráficos la relación entre la estimación de la función de riesgo acumulado $H_{NA}(t)$ y el tiempo de muerte t (o transformaciones de estas variables) por grupo de edad definido previamente. Este tipo de análisis es exploratorio y si bien no permite encontrar la distribución correcta, resulta útil para descartar modelos claramente no apropiados. Los parámetros de las distribuciones ajustadas fueron estimados mediante métodos gráficos y por máxima verosimilitud (con la función `survreg` del paquete de R `survival`, Therneau 2015).

3

Métodos para la simulación de cohortes de mujeres incidentes de cáncer de cérvix y estimaciones de prevalencia a 5 años

El objetivo de todos los métodos planteados fue generar bases de datos simuladas de iguales características a las llevadas en los registros de cáncer de base poblacional (es decir con las mismas variables), para las estimaciones de casos incidentes de cáncer de cuello de útero durante cada año entre 1999 y 2007 en Cataluña (realizadas por Cleries et al 2014), y a partir de los datos de los registros de cáncer de Girona y Tarragona del mismo período como población de referencia. Por tanto, implícitamente asumimos que la población catalana se comporta igual a la población de estas dos provincias. En la Figura 3.1 se presenta un esquema de los ocho métodos desarrollados; cuatro que involucran remuestreo y cuatro basados en el modelado de la función de supervivencia. En la descripción de los métodos que sigue se distingue *tiempo de muerte* de *tiempo de seguimiento*, donde éste último debe interpretarse como el tiempo en el cual una paciente esta bajo observación; desde el diagnóstico hasta la muerte o hasta la fecha de cierre del estudio (observación censurada por la derecha).

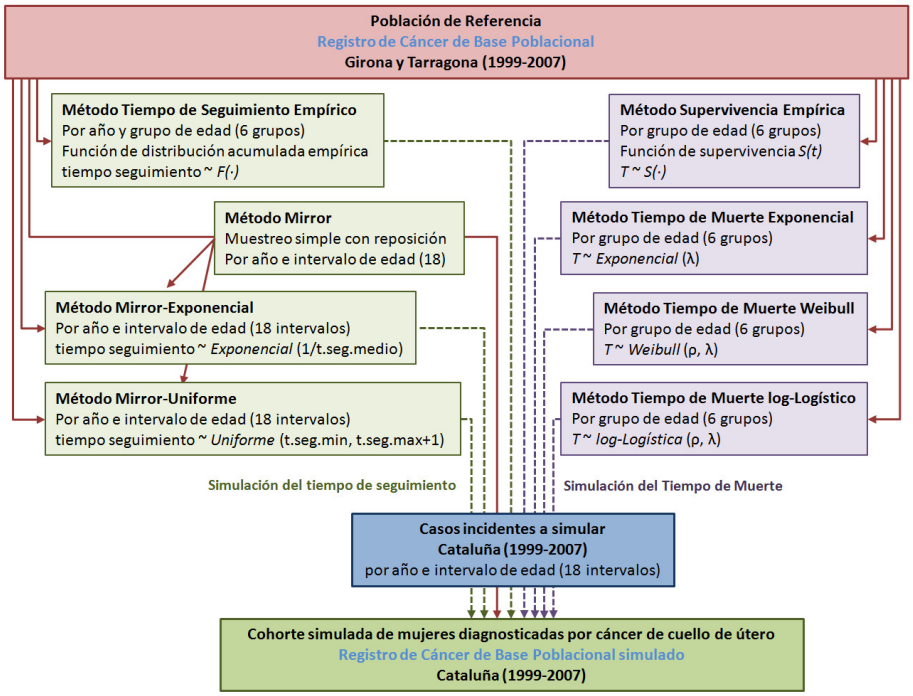

Figura 3.1: Métodos para la simulación de cohortes de mujeres diagnosticadas por cáncer de cuello de útero en Cataluña (1999-2007) a partir de los registros de cáncer de base poblacional (RCBP) de Girona y Tarragona, y de los casos incidentes estimados para Cataluña en igual período por Cleries et al (2014). El tiempo de muerte es diferente al tiempo de seguimiento ya que éste último corresponde al tiempo en que una paciente esta bajo observación; hasta la muerte o hasta la fecha de cierre del estudio (observación censurada por la derecha).

Cada uno de los métodos presentados fue implementado con el software R versión 3.1.2 (R Development Core Team 2014) a través de la programación de funciones propias.

3.1. Método de simulación del tiempo de seguimiento: "Método Tiempo de Seguimiento Empírico"

Para este método se utilizaron las distribuciones de tiempos de seguimiento observadas en 6 grupos de edad previamente definidos (de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años) en las mujeres incidentes de cáncer de cuello de útero en Girona y Tarragona para cada año entre 1999 y 2007, y con seguimiento hasta el 30 de junio de 2010. A partir de estas distribuciones empíricas se extraen de forma aleatoria los tiempos de seguimiento de los casos incidentes a simular de acuerdo a su grupo de edad y al año de incidencia. Por ejemplo, para simular los tiempos de seguimiento de los casos incidentes estimados para Cataluña en

1999 se utilizan las distribuciones de tiempo observadas en Girona y Tarragona en 1999 en el grupo de edad que corresponda, los tiempos de seguimiento de los casos catalanes del 2000 se simulan con las distribuciones observadas en la base de referencia en el 2000 y así sucesivamente. El mes de diagnóstico se fijó en julio para todos los casos, y junto con el tiempo de seguimiento simulado se definió el mes y el año del último seguimiento. El estatus de la paciente fue otorgado a partir de la fecha simulada para el último seguimiento; estatus 1 (muerte) si la misma fue anterior a la fecha de cierre en la base de datos de referencia (Girona y Tarragona, 30 de junio de 2010), o estatus 0 (vivo) en caso contrario.

Implementación en R

Para implementar este método se construyeron varias funciones (ver detalles en Apéndice A.2.1). Una primera función ***t.seg.int*** extrae desde una base de datos los tiempos de seguimiento por grupos de edad definidos mediante un vector de corte con los límites de los intervalos. La función ***simulate.fu*** genera un vector de tiempos de seguimientos simulados de longitud especificada (número de simulaciones) tras computar la función de distribución acumulada empírica (con la función `ecdf` del paquete `stats`) sobre un vector de tiempos de seguimiento dado (por ejemplo la salida de ***t.seg.int***). La función ***tiempo.seg.sim*** incorpora las funciones anteriores y permite generar tiempos de seguimiento simulados a cada caso incidente de un *data frame* dado (correspondiente a un año particular), según su grupo de edad, a partir de las distribuciones de tiempo de seguimiento observadas en el mismo grupo de edad en una base de datos de registros especificada (de referencia). La función ***base.sim.tse*** genera la base de datos simulada a partir de la base de referencia con los registros de cáncer (aquí de pacientes diagnosticadas por cáncer de cuello de útero en Girona y Tarragona entre 1999 y 2007, seguidas hasta junio de 2010), un vector de corte que define los grupos de edad para los cuales se obtienen las distribuciones de tiempo de seguimiento empíricas, y un *data frame* con los casos a simular de cada intervalo de edad por año de incidencia. Esta función incorpora las funciones anteriores y genera un "registro"para cada caso incidente a simular con edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en julio del año de incidencia, el tiempo de seguimiento generado desde las distribuciones observadas en el grupo de edad y año correspondiente, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de referencia (al 30 de junio de 2010).

3.2. Método por remuestreo: "Método Mirror"

En este caso la cohorte simulada para las estimaciones de casos incidentes de mujeres con cáncer de cuello de útero fue generada mediante un remuestreo simple en la base de datos de los registros de Girona y Tarragona. Cada caso incidente a simular fue obtenido mediante un simple remuestreo con reposición en la base de referencia de acuerdo al año de incidencia y al intervalo de edad

quinquenal correspondiente. Es decir, si en la estimación de casos incidentes para cáncer de cuello de útero se tienen 10 casos en 2002 pertenecientes al intervalo de edad 20-24 años, se recogen aleatoriamente 10 casos por remuestreo con reposición desde el conjunto de pacientes de Girona y Tarragona con edades del mismo intervalo quinquenal que incidieron en 2002. Los datos de cada caso remuestreado (ej. edad, fecha de diagnóstico, última fecha de seguimiento, estatus) se mantuvieron en la base simulada sin modificar.

Implementación en R

Para este método se implementó la función ***dat.sim.mirr*** que requiere como argumentos la base de datos de referencia, y un *data frame* con los casos a simular por intervalo de edad quinquenal y año de incidencia (ver en Apéndice A.2.2). El proceso es iterativo y va generando una base simulada por año incluido en el *data frame* de casos a simular, que posteriormente se unifican en una única base. Por cada año en los casos a simular, se realiza un remuestreo con reposición de tamaño igual al número de casos de cada intervalo de edad quinquenal sobre los registros de la base de referencia de la misma edad.

3.3. Método por remuestreo y tiempo de seguimiento exponencial: "Método Mirror Exponencial"

Aquí el procedimiento fue básicamente el mismo que en el método "Mirror"con la variante de que no se conservaron todos los datos de los casos originales remuestrados sino que la última fecha de seguimiento y el estatus fueron redefinidos. Los datos para éstas variables fueron obtenidos a partir de un tiempo de seguimiento generado aleatoriamente mediante una distribución exponencial de tasa igual al inverso del tiempo de seguimiento medio observado en la base de Girona y Tarragona (base de referencia) para el intervalo de edad (quinquenal) y el año correspondiente. Es decir, los datos de cada caso remuestreado de edad y fecha de diagnóstico se mantuvieron incambiados pero los datos para la última fecha de seguimiento y de ahí el estatus a la fecha de cierre de la base de referencia, fueron recalculados a partir de los tiempos de seguimiento generados.

Implementación en R

La función creada para obtener una base simulada por este método es ***dat.sim.mirrexp*** que lleva por argumentos la base de referencia (desde donde se hará el remuestreo según año e intervalo de edad, y se estimará la tasa de las distribuciones exponenciales utilizadas para generar los tiempos de seguimiento) y un *data frame* con los casos incidentes a simular por año y edad. El proceso es iterativo y se genera una base simulada por año incluido en el *data frame* de casos a simular por vez. Luego de obtener por remuestreo los casos del año según el intervalo de edad se sustituye el tiempo de seguimiento de cada "registro"de

acuerdo a un modelo exponencial con tasa igual al inverso del tiempo medio observado en la población de referencia en el año y la edad correspondiente, y de ahí la fecha del último seguimiento y el estatus a la fecha de cierre de la base de referencia (Apéndice A.2.3).

3.4. Método por remuestreo y tiempo de seguimiento uniforme: "Método Mirror Uniforme"

El procedimiento es similar al realizado en el método anterior ("Mirror-Exponencial") pero en este caso los tiempos de seguimiento son generados aleatoriamente mediante una distribución uniforme de mínimo y máximo tomados del tiempo de seguimiento mínimo y máximo más uno observado en el año e intervalo de edad correspondiente de la base de referencia (Girona y Tarragona). Es decir, los datos de cada caso remuestreado correspondientes a la edad y fecha de diagnóstico se mantienen incambiados pero los datos para la última fecha de seguimiento y de ahí el estatus a la fecha de cierre (30 de junio de 2010), son recalculados con los tiempos de seguimiento generados mediante las distribuciones uniformes.

Implementación en R

Para la implementar este método se creó la función ***dat.sim.mirrunif***, con la base de datos de referencia y un *data frame* con los casos a simular como argumentos. Operativamente es igual a la función creada para el método "Mirror-Exponencial"(***dat.sim.mirrexp***) con la variante que el tiempo de seguimiento de los casos a simular se genera a partir de distribuciones uniformes, pero al igual que antes, con parámetros específicos del intervalo de edad y el año de incidencia tomados desde la base de referencia (Apéndice A.2.4).

3.5. Método de simulación por supervivencia empírica: "Método Supervivencia KM (o NA)"

Aquí, cada base de datos (cohorte) simulada se construyó a partir de la función de supervivencia estimada por Kaplan-Meier (KM) o Nelson-Aalen (NA) para grupos de edad definidos previamente (de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años) sobre las pacientes de la base de referencia, Girona y Tarragona. A cada caso a simular se le asignó una probabilidad de supervivencia U mediante una distribución uniforme de parámetros 0 y 1. Posteriormente, el tiempo hasta la muerte se obtuvo por el método de la transformación inversa o transformación integral de probabilidad (Mood et al 1974, Bender et al 2005). Es decir, se buscó el tiempo T que correspondería a dicha supervivencia U en la estimación

empírica realizada por grupo de edad sobre las pacientes de Girona y Tarragona.

El **Teorema de Transformación Integral de Probabilidad** dice que si X es una variable aleatoria con función de distribución acumulada $F_X(x)$, entonces la variable aleatoria $U = F_x(X)$ tiene distribución uniforme en el intervalo (0,1). A la inversa, si U es distribuida de manera uniforme en (0,1), entonces $X = F_X^{-1}(U)$ tiene función de distribución acumulada $F_X(\cdot)$ (Mood et al 1974). Por otra parte, si U sigue una distribución uniforme (0,1), entonces $(1 - U)$ también se distribuye de forma uniforme en (0,1) (Mood et al 1974).

Por lo tanto, sea T el tiempo de supervivencia en el modelo, entonces dado que $F(t) = 1 - S(t)$ se deduce que $U = S(T)$ sigue una Uniforme(0,1). Entonces $S(T)$ puede ser invertida y obtenerse el tiempo de muerte T para la supervivencia U bajo el modelo (Bender et al 2005). Esto es la función inversa o recíproca, $S(T) = U \Longleftrightarrow T = S^{-1}(U)$.

Para los casos en que la probabilidad de supervivencia U asignada desde la distribución uniforme fue menor a las supervivencias definidas empíricamente, establecimos un tiempo de muerte igual al mayor tiempo de seguimiento observado en la base de referencia más un mes. Para la construcción de la base simulada, la edad de cada caso simulado fue igual al punto medio de su intervalo de edad quinquenal, se fijó la incidencia de las pacientes en julio del año de incidencia, y junto con el tiempo de muerte simulado se definió el mes y el año del último seguimiento. El estatus de la paciente fue redefinido a la fecha de cierre en la base de referencia, Girona y Tarragona (30 de junio de 2010); estatus 1 si el tiempo simulado determinó una fecha de muerte anterior a la fecha de cierre, o estatus 0 para tiempos que determinaron fechas de muerte posteriores.

Implementación en R

Para generar la base simulada por este método se creó la función ***base.sim.tsup*** con los siguientes argumentos: la base de datos de referencia, un vector de corte que define los grupos de edad, un *data frame* con los casos a simular por año e intervalo de edad quinquenal, y el estimador que se utilizará para la función de supervivencia empírica (por defecto Kaplan-Meier pero también puede utilizarse el estimador de Nelson-Aalen). La función lo que hace es ajustar la función de supervivencia empírica (KM o NA) por grupo de edad definido en el vector de corte, sobre todos los datos de la base de referencia, mediante la función `survfit` del paquete `survival` (Therneau 2015). El proceso es iterativo y va generando una base simulada por año incluido en el *data frame* de casos a simular, que posteriormente se unifican en una única base. Las bases anuales se generan también por un proceso iterativo, donde de acuerdo al intervalo de edad del caso a simular, la función de supervivencia ajustada que se utilizará para asignar un tiempo de muerte a una probabilidad de supervivencia obtenida previamente desde una Uniforme(0,1). Posteriormente se completa el resto de variables de la base simulada (edad, fecha de diagnóstico, fecha del últi-

mo seguimiento, y estatus) de acuerdo a los criterios explicados anteriormente. Para ver más detalles Apéndice A.2.5.

3.6. Método de simulación bajo un modelo de supervivencia Exponencial: "Método Tiempo de Muerte Exponencial"

En este método, las bases simuladas fueron construidas asumiendo que el tiempo de muerte de las pacientes sigue una distribución exponencial con parámetro λ (tasa de muerte) propio de su grupo de edad. La estimación de la tasa de muerte λ para cada grupo de edad (definido previamente, de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años) se realizó sobre la base de referencia Girona y Tarragona mediante dos métodos; a partir del método gráfico, y a través del método de máxima verosimilitud. A cada caso incidente a simular se le asignó una probabilidad de supervivencia U desde una distribución uniforme de parámetros 0 y 1. Por el método de la transformación inversa se generó un tiempo de muerte T para la supervivencia U con distribución exponencial de parámetro λ estimado para su grupo de edad (Bender et al 2005).

En un modelo exponencial la función supervivencia:

$$S(t) = e^{-\lambda t} \Longrightarrow U = S(T) = e^{-\lambda T} \sim \text{U(0,1)}$$

Entonces, a partir de la función inversa de $S(T) = U$ se obtiene el tiempo de supervivencia $T = S^{-1}(U)$ en el modelo exponencial con tasa de muerte λ edad-específica:

$$U = e^{-\lambda T} \Longrightarrow \ln U = -\lambda T$$

$$T = \frac{-\ln U}{\lambda}$$

La base de pacientes simulada (cohorte) se construyó con fecha de inicio (diagnóstico) en julio del año de incidencia del caso a simular, mes y año del último seguimiento a partir del tiempo de muerte generado desde el modelo exponencial, y estatus redefinido a la fecha de cierre en la base de referencia, Girona y Tarragona (30 de junio de 2010); estatus 1 si el tiempo simulado determinó una fecha de muerte anterior a la fecha de cierre, o estatus 0 para tiempos que determinaron fechas de muerte posteriores.

Implementación en R

La función creada que permite obtener una base de registros simulada por este método es ***base.sim.texp*** y su descripción se presenta más adelante (en sección 3.8; y detalles en Apéndice A.2.6).

3.7. Método de simulación bajo un modelo de supervivencia Weibull: "Método Tiempo de Muerte Weibull"

Los tiempos de muerte de los casos incidentes a simular fueron generados asumiendo una distribución Weibull de parámetros de forma ρ y escala λ específicos para su grupo de edad (definidos previamente, de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años). La estimación de los parámetros en cada grupo de edad se realizó sobre la base de datos de referencia (Girona y Tarragona) a través del método gráfico y por máxima verosimilitud. A cada caso incidente a simular se le asignó una probabilidad de supervivencia U desde una distribución uniforme con parámetros 0 y 1. Posteriormente, siguiendo el método de transformación inversa, se asignó un tiempo de muerte T en función de los parámetros de forma ρ y escala λ estimados para el grupo de edad correspondiente, y la supervivencia U (Bender et al 2005).

En el modelo Weibull la función de supervivencia es:

$$S(t) = \exp[-\lambda t^{\rho}] \Longrightarrow U = S(T) = \exp[-\lambda T^{\rho}] \sim \mathrm{U}(0,1)$$

De donde se deduce el tiempo de muerte $T = S^{-1}(U)$ para una valor de de supervivencia $S(T) = U$ y los parámetros ρ y λ estimados en el modelo Weibull ajustado por grupo de edad:

$$U = \exp[-\lambda T^{\rho}] \Longrightarrow \ln U = -\lambda T^{\rho} \Longrightarrow T^{\rho} = \frac{-\ln U}{\lambda}$$

$$T = (\frac{-\ln U}{\lambda})^{1/\rho}$$

En la construcción de la base simulada (cohorte) a cada caso incidente se le asignó una fecha de diagnóstico situada en julio del año de incidencia, mes y año de último seguimiento según el tiempo de muerte generado mediante la distribución Weibull del grupo de edad correspondiente, y estatus al 30 de junio de 2010 (fecha de cierre de seguimientos en la base de referencia).

Implementación en R

La generación de una base de registros simulada por este método se implementó mediante la creación de la función ***base.sim.twei***, cuya descripción se presenta más adelante (sección 3.8; detalles en Apéndice A.2.7).

3.8. Método de simulación bajo un modelo de supervivencia log-Logístico: "Método Tiempo de Muerte log-L"

Los tiempos de muerte de los casos incidentes a simular fueron generados asumiendo una distribución log-Logística de parámetros ρ y λ en cada grupo de edad (definidos previamente, de 0-34, 35-44, 45-54, 55-64, 65-74, y 75 o más años). La estimación de los parámetros se realizó mediante métodos gráficos (tras ajustar el modelo lineal $\ln(e^{H_{NA}(t)} - 1) = \ln\lambda + \rho \ln t$ por grupo de edad) y por el método de máxima verosimilitud (tras ajustar el modelo log-lineal $Y = \ln T = \mu + \sigma W$, donde W sigue una ley logística), sobre los datos de la base de referencia (Girona y Tarragona). A cada caso incidente a simular se le asignó una probabilidad de supervivencia U desde una distribución uniforme con parámetros 0 y 1. Posteriormente, siguiendo el método de transformación inversa, se asignó un tiempo de muerte T en función de la supervivencia U, y los parámetros ρ y λ estimados para el grupo de edad correspondiente (Bender et al 2005).

Esto es, dada la función de supervivencia $S(t)$ en un modelo log-Logístico:

$$S(t) = \frac{1}{(1+\lambda t^{\rho})} \Longrightarrow U = S(T) = \frac{1}{1+\lambda T^{\rho}} \sim \mathrm{U}(0,1)$$

A partir de la función inversa de $S(T) = U$ se deduce el tiempo de supervivencia $T = S^{-1}(U)$ bajo el modelo log-Logístico con los parámetros ρ y λ estimados por grupo de edad:

$$U = \frac{1}{1+\lambda T^{\rho}} \Longrightarrow 1 + \lambda T^{\rho} = \frac{1}{U} \Longrightarrow \lambda T^{\rho} = \frac{1}{e^{\ln U}} - 1$$

$$\lambda T^{\rho} = e^{-\ln U} - 1 \Longrightarrow T^{\rho} = \frac{e^{-\ln U} - 1}{\lambda}$$

$$T = (\frac{e^{-\ln U} - 1}{\lambda})^{1/\rho}$$

En la base simulada (cohorte) a cada caso incidente se le asignó edad de diagnóstico igual al punto medio de su intervalo de edad quinquenal, fecha de diagnóstico situada en julio del año de incidencia, mes y año de último seguimiento según el tiempo de muerte generado mediante la distribución log-Logística del grupo de edad correspondiente, y estatus al 30 de junio de 2010 (fecha de cierre de seguimientos en la base de referencia).

Implementación en R

Para implementar los últimos tres métodos se crearon las funciones siguientes: ***base.sim.texp***, ***base.sim.twei***, ***base.sim.tllog***, cada una asociada a un

modelo de distribución para el tiempo de muerte, Exponencial, Weibull o log-Logístico respectivamente (Ver Apéndice A.2.6-8 por detalles de las funciones). Las tres funciones tienen los mismos argumentos: una base de datos de referencia, un vector de corte que define grupos de edad, un *data frame* con los casos incidentes a simular por año e intervalo de edad, y el método para la estimación de los parámetros del modelo por grupo de edad; por métodos gráficos ("OLS") o por máxima verosimilitud ("ML"). El procedimiento es igual en las tres funciones, solo varían en la distribución usada para modelizar el tiempo de muerte y por tanto en la formulación utilizada para la estimación de los parámetros. El proceso es iterativo y se genera una base simulada por año incluido en el *data frame* de casos a simular por vez, que posteriormente se unifican en una sola base. En primer lugar se asigna una supervivencia U a cada caso a simular desde una uniforme de parámetros 0 y 1. Luego, se hace la estimación de los parámetros por grupo de edad sobre la base de registros de referencia, de acuerdo al modelo de distribución del tiempo de muerte utilizado. Esta estimación puede realizarse mediante métodos gráfico ("OLS", por *ordinary least squares*) o mediante el método de máxima verosimilitud ("ML"). Si la estimación de los parámetros es por métodos gráficos, se ajusta un modelo lineal mediante mínimos cuadrados ordinario entre el riesgo acumulado estimado $H_{NA}(t)$ y el tiempo t de muerte, o sobre las transformaciones de estas variables que correspondan según el modelo de distribución considerado. La estimación de riesgo acumulado $H(t)$ de Nelson-Aalen por grupo de edad se hace mediante la función `survfit` (con el argumento `type= 'fleming-harrington'`) del paquete `survival` (Therneau 2015). Para el método con modelo de distribución Exponencial, se ajusta un modelo lineal para cada grupo de edad, forzado a pasar por el origen, entre $H_{NA}(t)$ y el t, de forma que el parámetro λ se obtiene desde la estimación de la pendiente; para el método con modelo de distribución Weibull, se ajusta un modelo lineal por grupo de edad entre $\ln H_{NA}(t)$ y el $\ln t$, de manera que los parámetros se obtienen desde las estimaciones para la pendiente (ρ) y la ordenada en el origen ($\lambda = e^{[constante]}$); para el método con modelo de distribución log-Logístico, se ajusta un modelo lineal por grupo de edad entre $\ln(e^{H_{NA}(t)} - 1)$ y el $\ln t$, y los parámetros se obtienen desde las estimaciones para la pendiente (ρ) y la ordenada en el origen ($\lambda = e^{[constante]}$). Por otra parte, si el método de estimación es por máxima verosimilitud se ajusta un modelo log-lineal $Y = \ln T = \mu + \sigma W$ con los datos de las pacientes de cada grupo de edad por separado mediante la función `survreg` (con el argumento `distribution = 'exponential', 'weibull'` o `'loglogistic'` según corresponda) del paquete `survival` (Therneau 2015), la cual proporciona los estimadores de máxima verosimilitud de los parámetros μ y σ. Si el ajuste es a un modelo exponencial el parámetro σ estará fijado en 1. Una vez estimados los parámetros μ y σ de máxima verosimilitud se estima el parámetro λ de acuerdo a la relación $\lambda = e^{-\mu/\sigma}$ y el parámetro ρ según la relación $\rho = \sigma^{-1}$. Posteriormente el proceso es similar a las funciones de los métodos anteriores. Para cada caso a simular se obtiene un tiempo de muerte según: el modelo de distribución que corresponda, los parámetros del mismo estimados para su grupo de edad, y la supervivencia U previamente asignada desde una uniforma de parámetros 0 y 1. Se completa el resto de variables; edad

como punto medio del intervalo quinquenal que le corresponde al caso simulado, mes de diagnóstico fijado en julio para todos los casos, año de incidencia, mes y año del último seguimiento determinados de acuerdo al tiempo de muerte asignado, y estatus a la fecha de cierre de la base de referencia.

3.9. Estimaciones de prevalencia a 5 años en Cataluña

Las estimaciones de prevalencia en Cataluña de un año dado (entre 2003 y 2007) se realizaron a partir de 1000 bases simuladas con los métodos presentados anteriormente para las estimaciones de casos de mujeres con diagnóstico de cáncer de cervix entre 1999 y 2007. En cada base simulada se calcularon los casos prevalentes y la prevalencia (casos prevalentes por cada 100,000 mujeres) para cada intervalo de edad (quinquenal). Las estimaciones de prevalencia con sus intervalos de confianza 95 % en un año concreto fueron obtenidas a partir del cálculo de la mediana y los percentiles 2.5 % y 97.5 % de casos prevalentes y de prevalencia para cada intervalo de edad en las 1000 simulaciones.

Implementación en R

Cada método de simulación de cohortes de mujeres diagnosticadas por cáncer tiene su propia función asociada para la estimación de prevalencia a 5 años (***prev.AC.V.sim...***, ver Apéndice A.2.1-8) y llevan más o menos los mismos argumentos: base de datos de referencia, un vector de corte para los grupos de edad (si es necesario), un *data frame* con los casos a simular, el año para el cálculo de prevalencia a 5 años, la distribución de mujeres por año e intervalo de edad en la población de referencia, estimador para la función de supervivencia empírica o método de estimación de los parámetros (si es necesario), y número de bases simuladas. Entonces, para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

3.10. Validación

Antes de estimar la prevalencia en Cataluña se realizó una prueba de validación de los métodos desarrollados para la simulación de cohortes tomando como casos incidentes a simular aquellos registrados en la misma base de Girona y Tarragona entre 1999 y 2007. Es decir, se estimó la prevalencia a 5 años en 2003, 2004, 2005, 2006 y 2007 a través de simulaciones de los casos contenidos en la misma población de referencia.

Para evaluar comparativamente los diferentes métodos se calculó la *Razón de Discrepancia media* (*RDm*) (Moller et al 2005) adaptada a la estimación de

prevalencia a 5 años por intervalo de edad, para Girona y Tarragona en un año especificado. La *Razón de Discrepancia media* (RDm) se calculó como,

$$RDm = \frac{1}{18} \sum_{i=1}^{18} \frac{|O_i - E_i|}{1{,}96\,(sd_i + 0{,}001)}$$

donde O_i representa la prevalencia observada, E_i la prevalencia estimada (correspondiente a la prevalencia mediana tomada de las 1000 simulaciones) y sd_i el desvío estándar de la estimación (calculado sobre las simulaciones) para cada intervalo de edad i. La *Razón de Discrepancia* (RD) es definida como la distancia absoluta entre el número de casos observado y predicho, respecto a la distancia entre el valor predicho y el límite del intervalo de predicción (Moller et al 2005). Cuando el valor de RD es más grande que 1, ésta medida representa cuanto más amplio debe ser el intervalo de predicción para cubrir los casos observados. Aquí nosotros calculamos una *Razón de Discrepancia media* por año, y posteriormente una *Razón de Discrepancia media Total* ($RDmT$) para cada método como la suma de *Razones de Discrepancia media* correspondientes a cada año entre 2003 y 2007.

Por otra parte, con el fin de comparar el coste computacional de los métodos presentados, medimos el tiempo necesario para estimar la prevalencia a 5 años en Girona y Tarragona en 2003 con 1000 bases de registros simuladas. Para ello usamos la función `system.time` implementada en R.

4

Resultados

4.1. Ocurrencia de cáncer de cuello de útero en Girona y Tarragona (1999-2007)

La tasa de incidencia anual de cáncer de cuello de útero en Girona y Tarragona durante el período 1999-2007 se ha mantenido relativamente constante; ésta osciló entre 8.9 y 12.6 casos por cada 100,000 mujeres-año (Tabla 4.1). En general, la enfermedad afecta a mujeres de 20 años de edad o más, haciéndose más frecuente a partir de los 35 años.

En cuanto a la prevalencia de cáncer de cuello de útero a un año de diagnóstico (mujeres en tratamiento de evaluación inicial), ésta también se mantuvo relativamente constante a lo largo de todo el período de estudio y osciló entre 8.2 a 11.1 casos por cada 100,000 mujeres (Tabla 4.2), lo que muestra una baja letalidad al año.

La proporción de pacientes con cáncer de cuello de útero en fase de seguimiento clínico, prevalencia puntual a 2-3 años de diagnóstico, se ubicó entre 14.2 y 17.5 casos por cada 100,000 mujeres (Tabla 4.3).

Para la proporción de mujeres diagnosticadas por cáncer de cuello de útero que llegan a la fase de tratamiento designada como punto de cura (prevalencia puntual a 4-5 años) ésta se encontró entre 12.5 a 13.8 por cada 100,000 mujeres (Tabla 4.4). Si se ajusta por intervalo de edad, es posible observar que la prevalencia puntual a 4-5 años de estas pacientes es menor en mujeres jóvenes, de menos de 35 años de edad.

	Año																	
	1999		2000		2001		2002		2003		2004		2005		2006		2007	
Edad	C	T.C.	C	T.C.	C	T.C.	C	T.C.	C	T.C.	C	T.C.	C	T.C.	C	T.C.	C	T.C.
0-4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5-9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
15-19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20-24	0	0.00	0	0.00	0	0.00	0	0.00	1	2.24	1	2.27	1	2.27	1	2.29	1	2.27
25-29	2	4.62	1	2.22	2	4.25	3	6.02	1	1.90	1	1.82	2	3.48	0	0.00	3	5.03
30-34	5	11.45	4	8.86	5	10.84	3	6.28	6	11.96	7	13.31	9	16.07	5	8.45	6	9.59
35-39	6	13.73	9	19.84	8	17.24	12	24.90	14	27.92	10	19.42	13	24.34	9	16.25	9	15.67
40-44	8	19.79	10	23.79	5	11.44	8	17.58	7	14.77	4	8.15	8	15.55	7	13.22	7	12.74
45-49	4	11.12	5	13.43	9	23.24	10	24.81	7	16.54	7	15.85	10	21.50	10	20.53	8	15.80
50-54	6	17.59	5	14.16	4	11.12	6	16.11	6	15.60	4	10.25	2	4.91	5	11.75	10	22.56
55-59	5	17.75	2	6.62	6	19.51	6	18.47	1	2.88	5	13.69	18	47.13	4	10.22	5	12.34
60-64	6	21.78	8	30.03	2	7.32	0	0.00	5	18.68	10	34.81	0	0.00	4	12.05	2	5.70
65-69	7	22.26	2	6.28	8	25.30	1	3.14	6	18.92	6	19.94	6	21.50	2	6.95	1	3.56
70-74	5	17.65	3	10.38	8	26.92	3	9.93	5	16.21	7	22.47	4	12.72	4	12.78	4	12.63
75-79	2	8.46	4	16.29	4	15.93	4	15.54	1	3.83	6	22.53	7	25.79	5	17.84	8	28.19
80-84	0	0.00	0	0.00	2	11.69	3	16.61	8	41.82	2	10.07	4	19.11	4	18.66	2	9.03
85≤	4	29.96	0	0.00	2	13.01	3	19.03	0	0.00	3	17.86	1	5.75	2	10.93	2	10.32
Total	60	10.53	53	9.06	65	10.89	62	10.11	68	10.74	73	11.22	85	12.57	62	8.88	68	9.43

Tabla 4.1: Incidencia anual de cáncer de cuello de útero en las provincias de Girona y Tarragona (Cataluña) para el período 1999-2007. Se presentan los casos incidentes (C) y la tasa cruda (T.C., número de nuevos casos por cada 100,000 mujeres-año).

	Año																	
	1999		2000		2001		2002		2003		2004		2005		2006		2007	
Edad	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P
0-4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5-9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
15-19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20-24	0	0.00	0	0.00	0	0.00	0	0.00	1	2.24	1	2.27	1	2.27	1	2.29	1	2.27
25-29	2	4.62	1	2.22	2	4.25	3	6.02	1	1.90	1	1.82	2	3.48	0	0.00	3	5.03
30-34	5	11.45	4	8.86	5	10.84	3	6.28	6	11.96	7	13.31	9	16.07	5	8.45	6	9.59
35-39	5	11.45	8	17.63	8	17.24	12	24.90	13	25.92	9	17.48	12	22.47	9	16.25	8	13.93
40-44	8	19.79	11	26.17	5	11.44	7	15.39	7	14.77	4	8.15	8	15.55	6	11.33	5	9.10
45-49	4	11.12	5	13.43	9	23.24	10	24.81	7	16.54	7	15.85	10	21.50	10	20.53	9	17.78
50-54	5	14.66	3	8.50	4	11.12	6	16.11	4	10.40	4	10.25	2	4.91	5	11.75	9	20.30
55-59	3	10.65	2	6.62	6	19.51	5	15.39	2	5.77	4	10.95	16	41.90	4	10.22	4	9.87
60-64	4	14.52	8	30.03	2	7.32	0	0.00	5	18.68	10	34.81	2	6.19	4	12.05	2	5.70
65-69	6	19.08	2	6.28	8	25.30	1	3.13	6	18.92	5	16.61	4	14.34	2	6.95	1	3.56
70-74	3	10.59	2	6.92	7	23.56	2	6.62	3	9.73	6	19.26	3	9.54	4	12.78	3	9.47
75-79	2	8.46	2	8.15	5	19.91	3	11.65	1	3.83	5	18.78	4	14.74	5	17.84	6	21.14
80-84	0	0.00	0	0.00	0	0.00	2	11.08	4	20.91	0	0.00	2	9.55	4	18.66	1	4.51
85≤	2	14.98	0	0.00	1	6.51	0	0.00	0	0.00	0	0.00	0	0.00	1	5.46	1	5.16
Total	49	8.60	48	8.21	62	10.39	54	8.81	60	9.47	63	9.68	75	11.09	60	8.59	59	8.19

Tabla 4.2: Prevalencia a un año de cáncer de cuello de útero en las provincias de Girona y Tarragona (Cataluña) durante el período 1999-2007. Se presenta el número de casos prevalentes (C) y la prevalencia puntual a 1 año (P, número de casos prevalentes a un año por cada 100,000 mujeres).

	Año															
	2000⋆		2001		2002		2003		2004		2005		2006		2007	
Edad	C	P	C	P	C	P	C	P	C	P	C	P	C	P	C	P
0-4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5-9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
15-19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20-24	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	1	2.29	2	4.54
25-29	1	2.22	2	4.25	2	4.01	1	1,90	2	3.64	4	6.96	3	5.06	1	1.68
30-34	4	8.86	6	13.01	5	10.46	6	11,96	8	15.21	8	14.28	10	16.89	9	14.38
35-39	5	11.02	13	28.01	14	29.05	17	33,90	20	38.85	10	18.72	15	27.08	19	33.08
40-44	6	14.28	16	36.60	19	41.76	17	35,87	15	30.57	17	33.04	19	35.88	15	27.29
45-49	8	21.49	12	30.99	14	34.73	10	23,63	13	29.43	13	27.95	16	32.85	17	33.58
50-54	4	11.33	9	25.03	7	18.79	14	36,41	9	23.07	6	14.72	5	11.75	10	22.56
55-59	4	13.24	6	19.51	7	21.55	9	25,95	9	24.64	6	15.71	16	40.87	11	27.15
60-64	4	15.02	8	29.29	8	29.98	2	7,47	4	13.92	10	30.93	11	33.15	10	28.52
65-69	2	6.28	4	12.65	8	25.10	7	22,08	5	16.61	13	46.59	10	34.75	4	14.23
70-74	4	13.83	5	16.83	8	26.49	8	25,94	2	6.42	7	22.27	5	15.98	4	12.63
75-79	2	8.15	4	15.93	7	27.19	5	19,17	5	18.78	7	25.79	6	21.41	7	24.66
80-84	0	0.00	0	0.00	1	5.54	2	10,46	4	20.14	2	9.55	5	23.32	6	27.08
85≤	0	0.00	0	0.00	0	0.00	0	0.00	1	5.95	1	5.75	0	0.00	1	5.16
Total	44	7.52	85	14.24	100	16.31	98	15,47	97	14.91	104	15.38	122	17.47	116	16.09

Tabla 4.3: Prevalencia puntual a 2-3 años de cáncer de cuello de útero para casos diagnosticados durante el período 1999-2007 en las provincias de Girona y Tarragona (Cataluña). Se presenta el número de casos prevalentes (C) y la prevalencia puntual a 2-3 años (P, número de casos prevalentes con 2 a 3 años de diagnóstico por cada 100,000 mujeres). ⋆ es la prevalencia puntual a solo 2 años, corresponde a casos diagnosticados durante 1999 que prevalecen al 31 de diciembre de 2000.

	Año											
	2002⋆		2003		2004		2005		2006		2007	
Edad	C	P	C	P	C	P	C	P	C	P	C	P
0-4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5-9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
15-19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20-24	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
25-29	0	0.00	0	0.00	0	0.00	0	0.00	1	1.69	2	3.35
30-34	3	6.28	4	7.97	4	7.61	4	7.14	7	11.82	5	7.99
35-39	5	10.38	11	21.94	10	19.42	11	20.59	8	14.44	10	17.41
40-44	6	13.19	17	35.87	15	30.57	15	29.15	20	37.77	15	27.29
45-49	5	12.40	11	25.99	12	27.16	12	25.80	13	26.69	10	19.75
50-54	6	16.11	10	26.01	11	28.20	11	26.98	11	25.84	12	27.07
55-59	3	9.24	6	17.30	7	19.16	11	28.80	7	17.88	6	14.81
60-64	2	7.50	7	26.15	7	24.37	4	12.37	3	9.04	8	22.82
65-69	3	9.41	5	15.77	7	23.26	4	14.34	5	17.38	9	32.02
70-74	4	13.24	6	19.46	8	25.68	6	19.09	2	6.39	7	22.11
75-79	1	3.88	1	3.83	6	22.53	7	25.79	5	17.84	6	21.14
80-84	1	5.54	3	15.68	3	15.10	4	19.11	4	18.66	5	22.57
85≤	0	0.00	0	0.00	0	0.00	0	0.00	1	5.46	1	5.16
Total	39	6.36	81	12.79	90	13.83	89	13.16	87	12.46	96	13.32

Tabla 4.4: Prevalencia puntual a 4-5 años de cáncer de cuello de útero para casos diagnosticados durante el período 1999-2007 en las provincias de Girona y Tarragona (Cataluña). Se presenta el número de casos prevalentes (C) y la prevalencia puntual a 4-5 años (P, número de casos prevalentes con 4 a 5 años de diagnóstico por cada 100,000 mujeres). ⋆ es la prevalencia solo a 4 años, corresponde a casos diagnosticados durante 1999 que prevalecen al 31 de diciembre de 2002.

La prevalencia de la enfermedad (acumulada a 5 años) en la población femenina de Girona y Tarragona entre 2003 y 2007 se mantuvo constante, entre 37.6 y 39.6 casos por cada 100,000 mujeres (Tabla 4.5). Ajustando por edad, la prevalencia de cáncer de cuello de útero es más alta en mujeres de más de 35 años comenzando a bajar aproximadamente a partir de los 55 años de edad.

	Año									
	2003		2004		2005		2006		2007	
Edad	C	P	C	P	C	P	C	P	C	P
0-4	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
5-9	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
10-14	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
15-19	0	0.00	0	0.00	0	0.00	0	0.00	0	0.00
20-24	1	2.24	1	2.27	1	2.27	2	4,57	3	6.81
25-29	2	3.80	3	5.47	6	10.43	4	6,75	6	10.05
30-34	16	31.89	19	36.12	21	37.50	22	37,16	20	31.97
35-39	41	81.76	39	75.75	33	61.78	32	57,77	37	64.43
40-44	41	86.52	34	69.29	40	77.73	45	84,98	35	63.68
45-49	28	66.15	32	72.44	35	75.24	39	80,08	36	71.11
50-54	28	72.82	24	61.52	19	46.60	21	49,33	31	69.93
55-59	17	49.01	20	54.75	33	86.41	27	68,96	21	51.84
60-64	14	52.29	21	73.10	16	49.49	18	54,24	20	57.04
65-69	18	56.77	17	56.48	21	75.26	17	59,08	14	49.82
70-74	17	55.13	16	51.36	16	50.89	11	35,15	14	44.21
75-79	7	26.84	16	60.09	18	66.31	16	57,08	19	66.95
80-84	9	47.05	7	35.24	8	38.21	13	60,64	12	54.17
85≤	0	0.00	1	5.95	1	5.75	2	10,93	3	15.48
Total	239	37.74	250	38.42	268	39.63	269	38,53	271	37.60

Tabla 4.5: Prevalencia (acumulada a 5 años) de cáncer de cuello de útero para casos diagnosticados durante el período 1999-2007 en las provincias de Girona y Tarragona (Cataluña). Se presenta el número de casos prevalentes (C) y la prevalencia (P, número de casos prevalentes, diagnosticados en los últimos 5 años, por cada 100,000 mujeres).

4.2. Análisis de la supervivencia en mujeres diagnosticadas por cáncer de cuello uterino en Girona y Tarragona

La función de supervivencia empírica de Kaplan-Meier por grupo de edad muestra diferencias entre los grupos considerados (menor de 35, 35-44, 45-54, 55-64, 65-74 y 75 o más años). En la Figura 4.1 se puede ver que conforme aumenta la edad de las mujeres que son diagnosticadas para cáncer de cuello de útero en Girona y Tarragona, disminuye la supervivencia. Una prueba de log-rango evidencia dichas diferencias ($\chi^2 = 115$; $gl = 5$; $p < 0{,}05$). A los 5 años desde el diagnóstico para cáncer de cuello de útero, la probabilidad de que una paciente continúe viva es de 88.3 % (IC95 %: 80.9-96.3) si fue diagnosticada con menos de 35 años. En cambio, en las mujeres diagnosticadas con 75 años o más la supervivencia a los 5 años es de tan solo 37.3 % (IC95 %: 28.3-49.4). Por otra parte, puede notarse que pasados los 5 años de tratamiento, no se observan

grandes saltos en la supervivencia de las pacientes de los diferentes grupos de edad, excepto para el grupo de 65-69 años donde la supervivencia no parece estabilizarse hasta luego de 7 años de tratamiento (Figura 4.1).

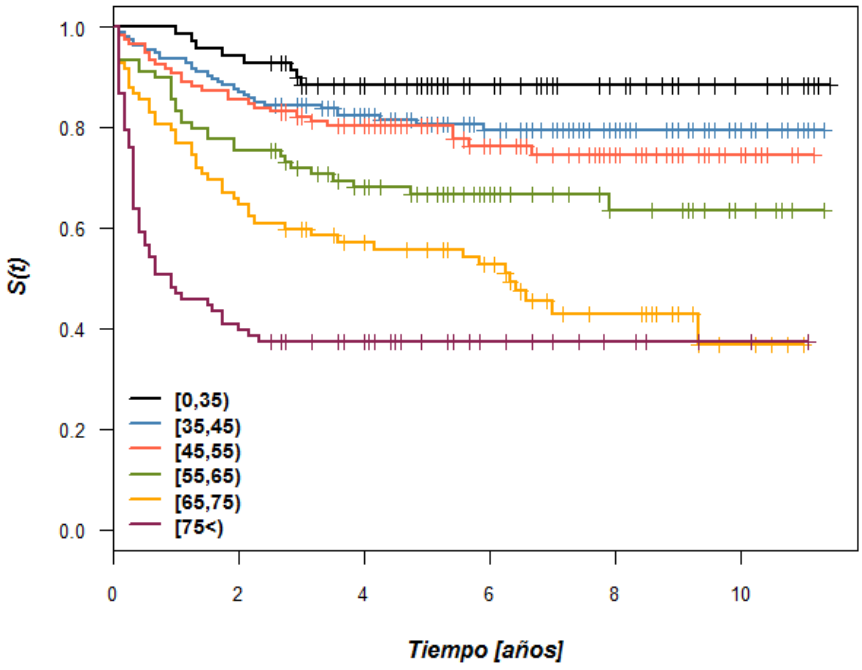

Figura 4.1: Función de supervivencia Kaplan-Meier por grupo de edad en pacientes diagnosticadas por cáncer de cuello de útero en Girona y Tarragona durante el período 1 de enero de 1999 y 31 de diciembre de 2007, y seguidas hasta el 30 de junio de 2010.

Cuando exploramos mediante métodos gráficos, el grado de ajuste de los datos observados en Girona y Tarragona a los modelos de distribución para el tiempo de muerte Exponencial, Weibull y log-Logístico, encontramos que las distribuciones Weibull y log-Logística serían la más apropiadas (Figura 4.2). El modelo de distribución exponencial no ajusta bien los datos en ningún grupo de edad considerado ya que, a pesar de los altos valores para el coeficiente de determinación (r^2) obtenidos, hay una clara falta de linealidad entre el riesgo acumulado $H(t)$ y el tiempo de supervivencia t (Figura 4.2).

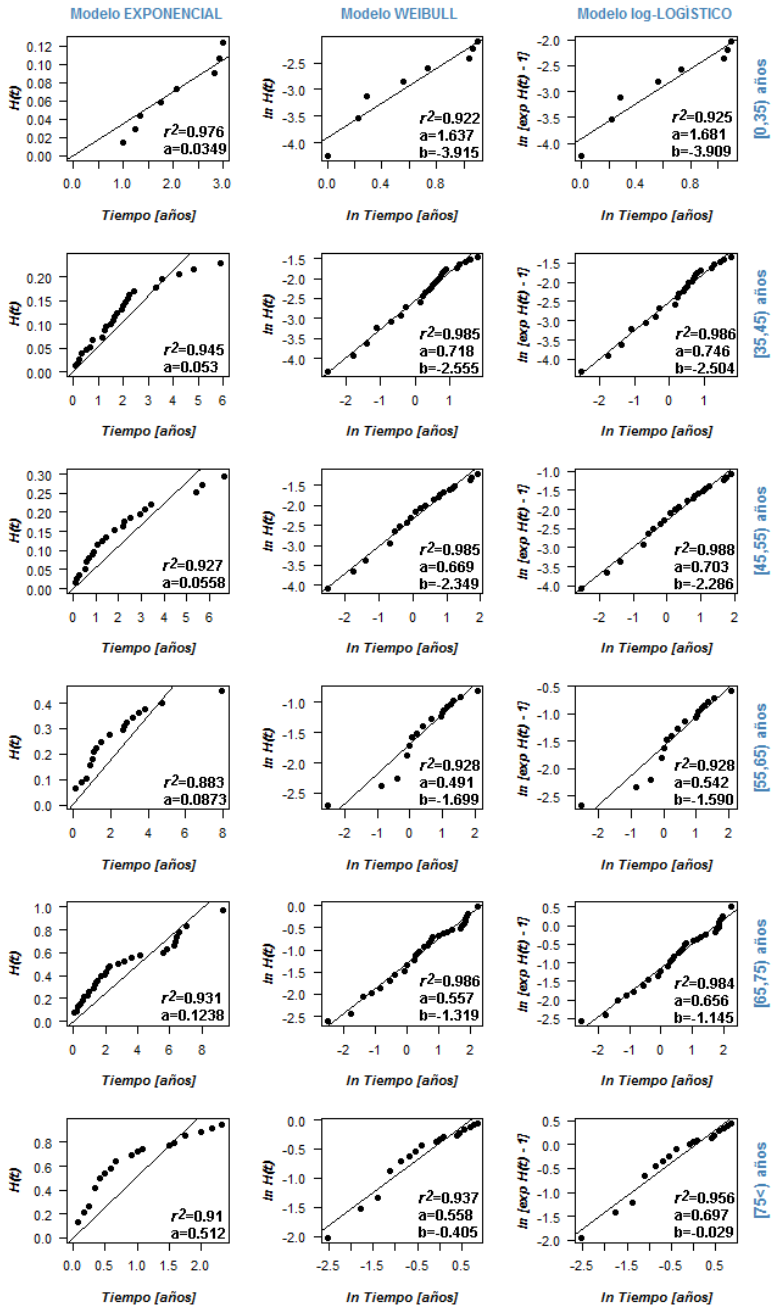

Figura 4.2: Relación entre la función de riesgo acumulado $H(t)$ estimada por Nelson-Alen y el tiempo de muerte t de acuerdo a la transformación lineal correspondiente bajo un modelo de distribución Exponencial (izquierda), Weibull (centro) y log-Logístico (derecha) para cada grupo de edad de pacientes diagnosticadas por cáncer de cérvix en Girona y Tarragona entre 1999 y 2007, seguidas hasta junio de 2010. Se presenta el coeficiente de determinación r^2 y los parámetros estimados para la recta ajustada $y = ax + b$ (nótese que en el modelo exponencial la recta pasa por el origen $b = 0$).

La estimación de los parámetros por métodos gráficos y por máxima verosimilitud en cada modelo de distribución explorado se presenta en la Tabla 4.6. Nótese que en el grupo de edad [0,35), con menor número de casos y mayor supervivencia, la diferencia en la estimación de los parámetros mediante métodos gráficos y por máxima verosimilitud es mayor si se compara con las diferencias observadas en otros grupos de edad.

	Exponencial $H(t) = \lambda t$		Weibull $H(t) = \lambda t^\rho$				log-Logística $H(t) = \ln(1 + \lambda t^\rho)$			
	λ		ρ		λ		ρ		λ	
Edad	gráf.	ML	gráf.	ML	gráf.	ML	gráf.	ML	gráf.	ML
[0-35)	0.035	0.019	1.637	0.802	0.020	0.028	1.681	0.845	0.020	0.028
[35-45)	0.053	0.033	0.718	0.589	0.078	0.072	0.747	0.633	0.082	0.075
[45-55)	0.056	0.042	0.669	0.611	0.095	0.085	0.703	0.663	0.102	0.090
[55-65)	0.087	0.068	0.491	0.548	0.183	0.152	0.542	0.620	0.204	0.170
[65-75)	0.124	0.122	0.557	0.626	0.267	0.233	0.656	0.752	0.318	0.280
[75<)	0.512	0.253	0.558	0.485	0.667	0.494	0.697	0.691	0.971	0.783

Tabla 4.6: Parámetros estimados para modelos de supervivencia con distribución Exponencial, Weibull y log-Logística por grupo de edad a partir de métodos gráficos y por máxima verosimilitud (ML) sobre los tiempos de muerte observados en pacientes diagnosticadas por cáncer de cérvix en Girona y Tarragona entre 1999 y 2007, seguidas hasta junio de 2010.

Por otra parte, si se utiliza el criterio AIC para evaluar cuál es el modelo paramétrico de mayor ajuste a los datos observados, en sus representaciones log-lineales $Y = \ln T = \mu + \sigma W$ (utilizada para la estimación de parámetros máximo verosímiles), la distribución log-Logística muestra el mejor ajuste para todos los grupos de edad a excepción del primero (menores de 35 años) donde el modelo Exponencial sería el mejor.

En la Figura 4.3 se presenta la función de supervivencia $S(t)$ ajustada bajo los diferentes modelos paramétricos para cada grupo de edad de las pacientes de Girona y Tarragona, utilizando los estimadores máximo verosímiles (ML). El modelo de distribución Exponencial en general es el que muestra peor ajuste respecto a la función de supervivencia empírica de Kaplan-Meier. Por otro lado, los modelos de distribución Weibull y log-Logístico se comportan relativamente bien en los grupos de edad intermedios (de más de 35 y menos de 75 años).

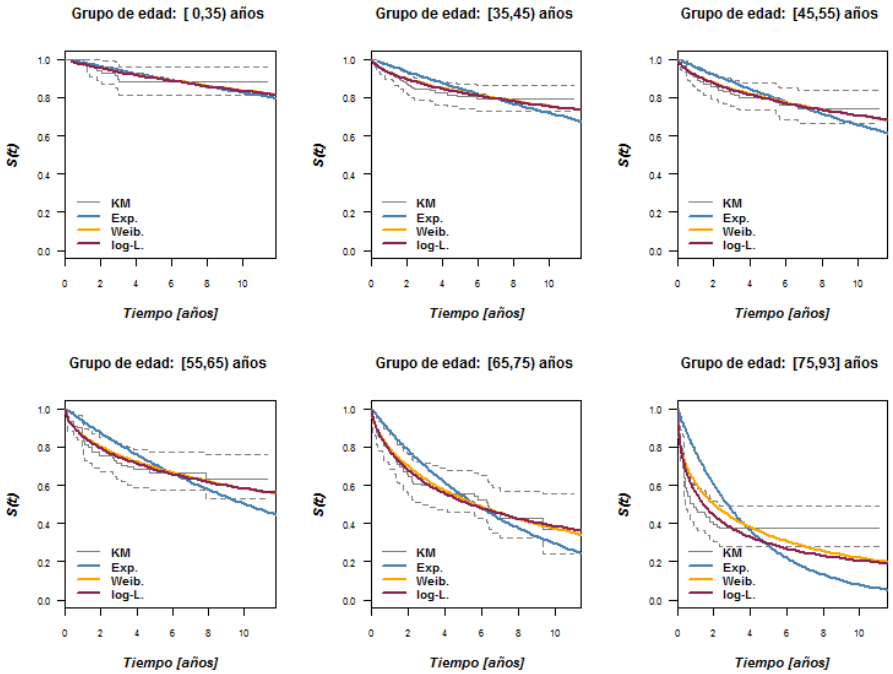

Figura 4.3: Función de supervivencia $S(t)$ bajo diferentes modelos paramétricos de distribución, Exponencial, Weibull y log-Logística ajustados (por máxima verosimilitud) por grupo de edad en pacientes diagnosticadas por cáncer de cérvix en Girona y Tarragona entre 1999 y 2007, seguidas hasta junio de 2010. A modo de referencia se presenta la función de supervivencia de Kaplan-Meier y sus intervalos de confianza 95 % en gris.

4.3. Estimación de prevalencia a 5 años en Girona y Tarragona por métodos de simulación de cohortes - Validación

Las estimaciones de prevalencia a 5 años obtenidas para Girona y Tarragona a partir de cada uno de los diferentes métodos de simulación de cohortes de mujeres diagnosticadas por cáncer de cuello de útero entre 1999 y 2007 se presentan en las siguientes ocho Figuras (4.4-4.11); las primeras cuatro corresponden a los métodos que simulan tiempo de seguimiento (involucran remuestreo) y las cuatro siguientes corresponden a los métodos que simulan tiempos de muerte desde los diferentes modelos de supervivencia ajustados (empírica y modelos paramétricos).

4.3.1. Resultados para métodos de simulación del tiempo de seguimiento

En el método de simulación para el tiempo de seguimiento desde las distribuciones empíricas del mismo en seis grupos de edad (método "Tiempo de Seguimiento Empírico"), si bien la prevalencia a 5 años estimada dibuja una trayectoria similar a la prevalencia observada, los intervalos de confianza 95 % generados no siempre cubren los valores observados (Figura 4.4).

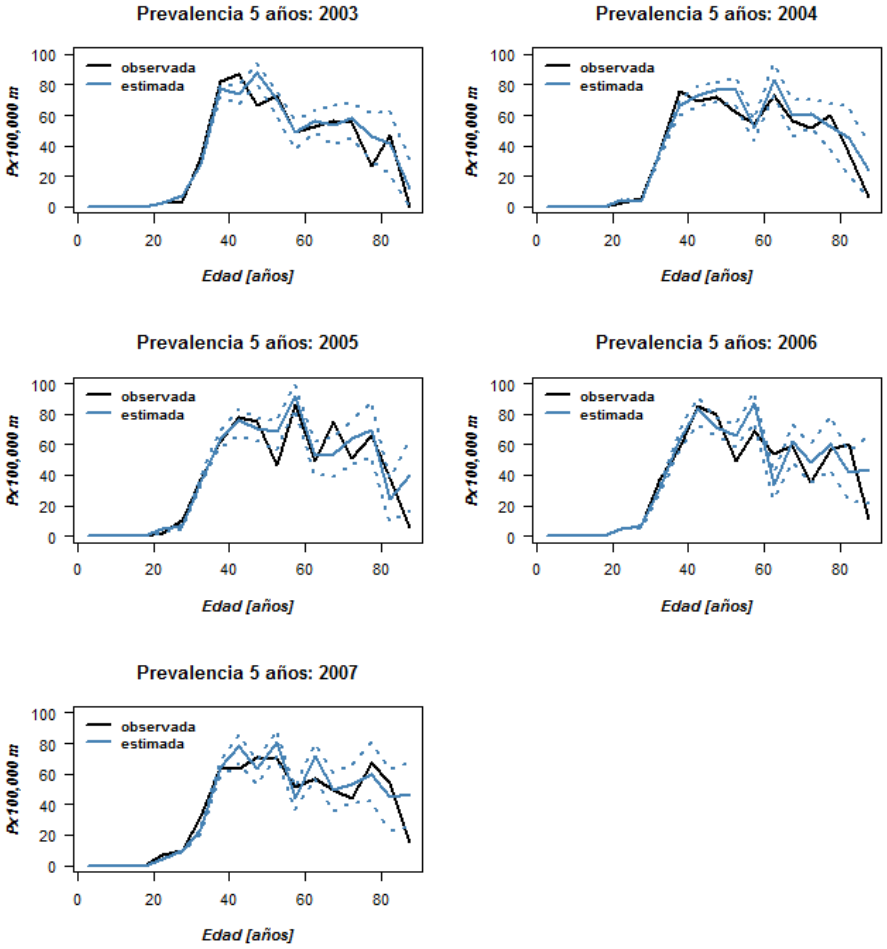

Figura 4.4: Prevalencia a 5 años de cáncer de cuello de útero en Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada (línea azul) a partir del método de simulación del tiempo de seguimiento desde las distribuciones empíricas del mismo en 6 grupos de edad, método "Tiempo de Seguimiento Empírico". Las líneas punteadas en color azul representan el límite inferior y superior de intervalos de confianza 95 %.

Las estimaciones de prevalencia 5 años obtenidas desde el método de re-

muetreo por año de incidencia e intervalo de edad quinquenal, Método "Mirror"(Figura 4.5) coinciden perfectamente a las calculadas directamente de los registros de cáncer. No obstante, cabe señalar que los intervalos de confianza 95 % obtenidos son relativamente amplios.

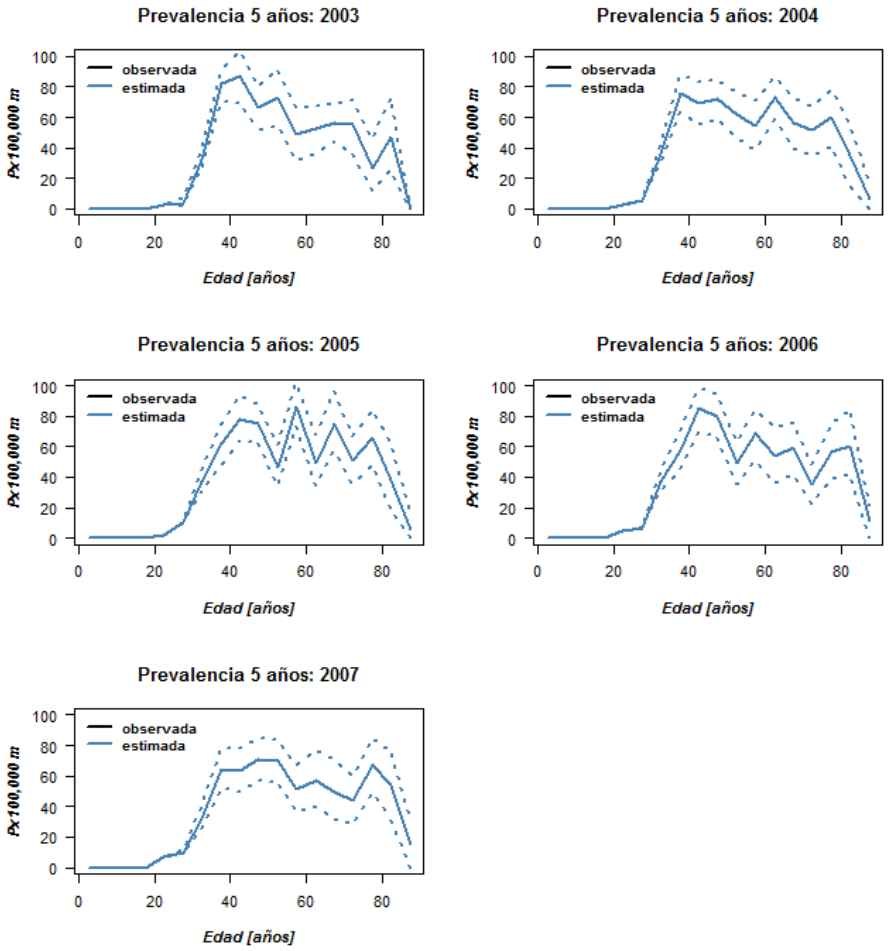

Figura 4.5: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 estimada a partir del método de remuestreo "Mirror". La prevalencia a 5 años observada queda oculta bajo la línea azul (correspondiente a la mediana de prevalencia en las simulaciones). Las líneas punteadas representan los percentiles 2,5 y 97,5 % de la prevalencia estimada.

La estrategia de remuestreo por año de incidencia e intervalo de edad quinquenal de las pacientes, y posterior estimación del tiempo de seguimiento desde una distribución exponencial propia del intervalo de edad quinquenal correspondiente, método "Mirror-Exponencial", no logra emular muy bien los valores de

prevalencia observados en Girona y Tarragona (Figura 4.6). La prevalencia a 5 años de cáncer de cuello de útero es sistemáticamente subestimada con este método, incluso se acentúa en los últimos años del período de estudio. Por otra parte, los intervalos de confianza 95 % parecen ser levemente más estrechos que los generados por el método "Mirror"simple.

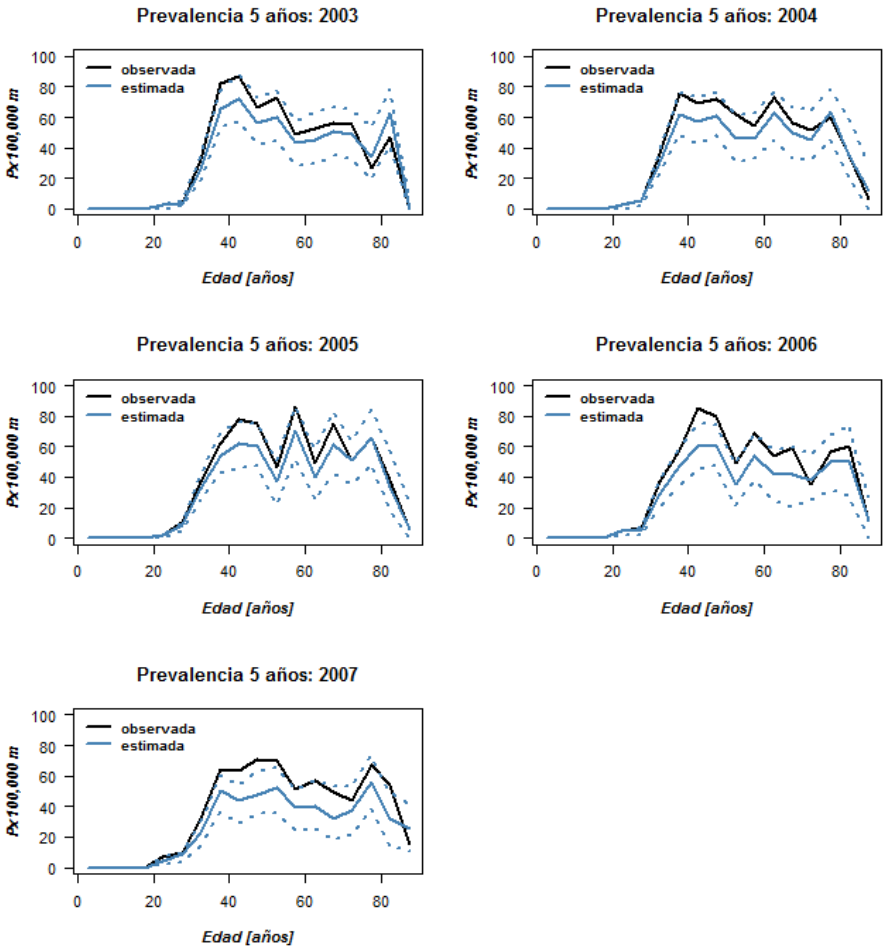

Figura 4.6: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada a partir del método "Mirror-Exponencial"(línea azul). Las líneas punteadas azul indican los percentiles 2.5 y 97.5 % para la prevalencia simulada.

Para el último de los métodos que involucran remuestreo, la simulación del tiempo de seguimiento a partir de una distribución uniforme para cada intervalo de edad quinquenal, método "Mirror-Uniforme", produce estimaciones de prevalencia a 5 años bastante cercanas a las observadas, con intervalos de con-

fianza 95 % que en general contienen a los valores observados y que además son relativamente estrechos (Figura 4.7). Cabe notar que en los intervalos de mayor edad (a partir de los 70 años) puede haber una tendencia a sobrestimar la prevalencia.

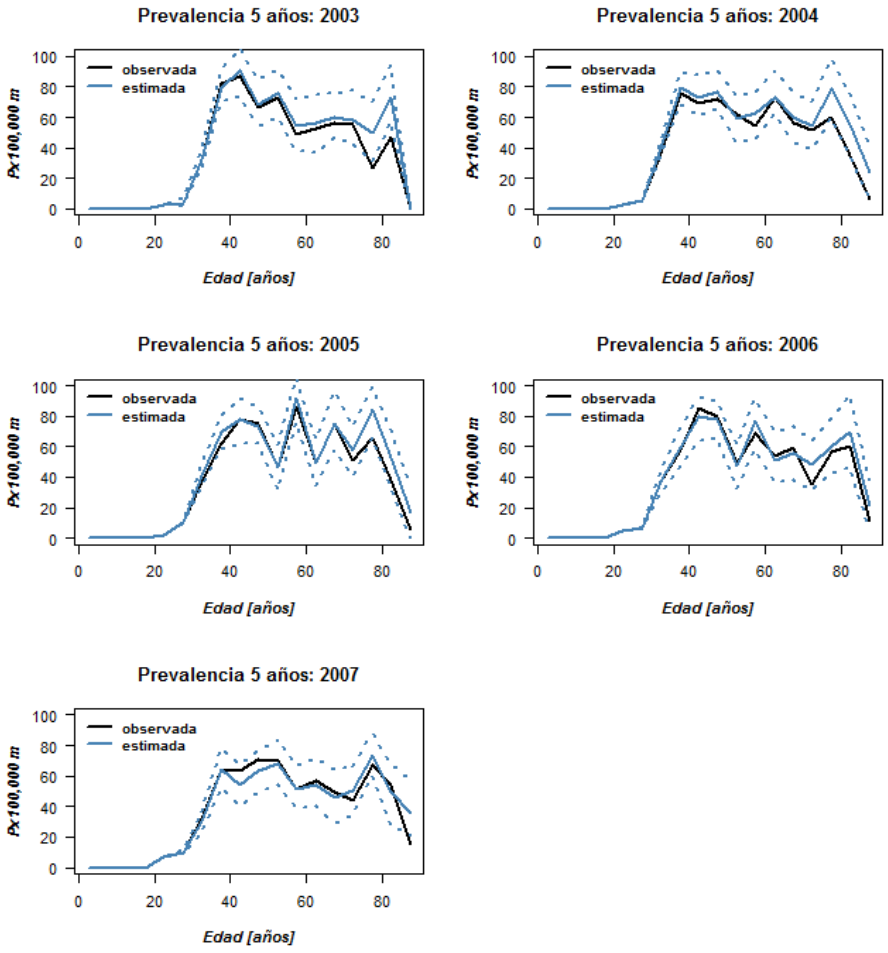

Figura 4.7: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada mediante el método "Mirror-Uniforme" (línea azul). Los percentiles 2.5 y 97.5 % de prevalencia simulada están representados por las líneas punteadas de color azul.

4.3.2. Resultados para métodos de simulación basados en supervivencia

Las estimaciones de prevalencia a 5 años obtenidas mediante la generación de cohortes de mujeres incidentes de cáncer de cuello de útero en Girona y

Tarragona entre 1999 y 2007 con tiempos de supervivencia generados a partir de la función de supervivencia empírica de Kaplan-Meier estimada en seis grupos de edad, método de "Supervivencia KM", fueron relativamente cercanas a las prevalencias observadas (Figura 4.8). En cuanto a los intervalos de confianza 95 % de las estimaciones, éstos son relativamente estrechos y en general logran cubrir los valores de prevalencia a 5 años observados.

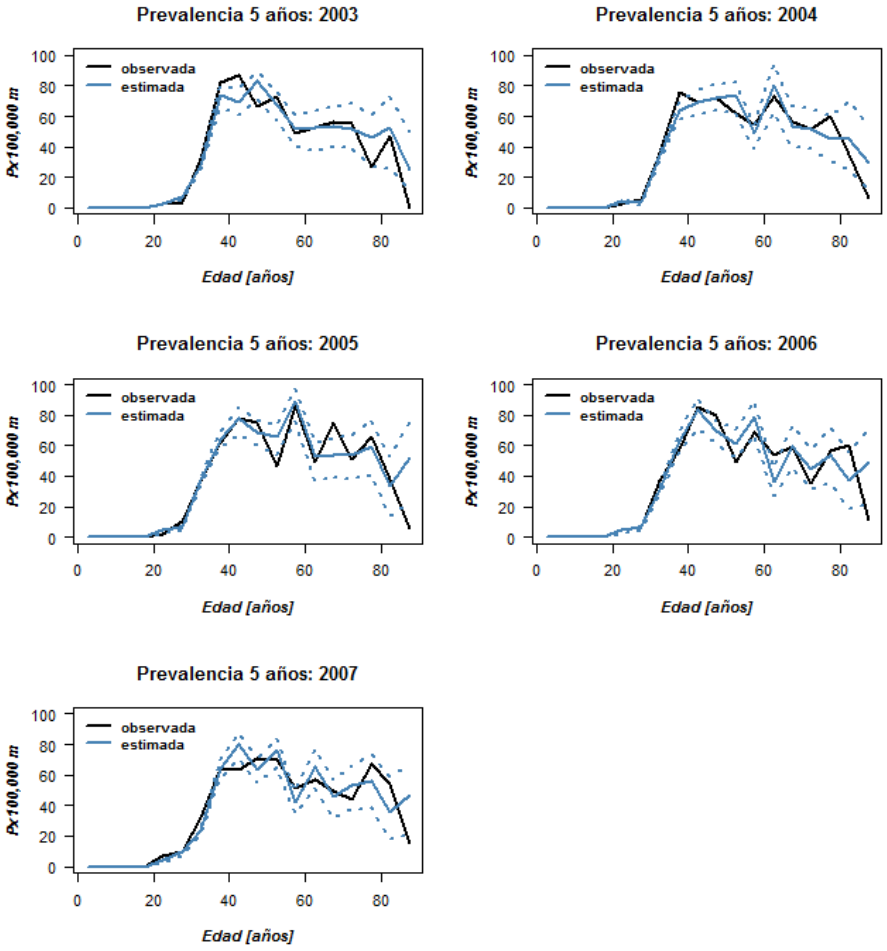

Figura 4.8: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007, observada (línea negra) y estimada (línea azul) a partir del método "Supervivencia empírica KM". Los percentiles 2.5 y 97.5 % de prevalencia simulada están representados por las líneas punteadas de color azul.

En cuanto a los métodos de simulación que utilizan modelos de supervivencia paramétricos, Exponencial, Weibull y log-Logística, con parámetros estimados en cada uno de los 6 grupos de edad definidos, los resultados no son notoria-

mente contrastantes. Bajo los tres modelos las estimaciones de prevalencia a 5 años en Girona y Tarragona para los años 2003 a 2007 son relativamente cercanas a las observadas. Sin embargo, los método con tiempo de muerte Weibull y log-Logístico parecen generar intervalos de confianza 95 % de mejor de cobertura (Figuras 4.10 y 4.11) que los generados mediante el método con tiempos de muerte exponencial (Figura 4.9). Por otra parte, las estimaciones de prevalencia bajo el modelo exponencial parecen tener cierta tendencia a quedar por encima de las observadas (sobreestimación). Estos resultados son coherentes con lo encontrado anteriormente en el análisis gráfico exploratorio, donde la distribución exponencial mostró ser inapropiada para ajustar los datos observados (Figura 4.2).

Por otro lado, el método utilizado para la estimación de los parámetros de los modelos paraméticos ajustadas (gráfico vs máxima verosimilitud) no afecta claramente a las estimaciones de prevalencia (ver comparación de gráficos a la izquierda y derecha en Figuras 4.9, 4.10 y 4.11). Dependiendo del año y el intervalo de edad las estimaciones de prevalencia obtenidas con una u otra estimación de parámetros se parecen más o menos a las observadas. Por ejemplo, las estimaciones de prevalencia para el 2003 en los intervalos de edad de 70 a más años con el método "Tiempo de Muerte Exponencial"(Figura 4.9) se parecen más a las observadas cuando las estimaciones de la tasa de fallo por grupo de edad son obtenidas por métodos gráficos respecto a las estimaciones obtenidas por máxima verosimilitud. En cambio, en las estimaciones de prevalencia para el 2007 en los intervalos de edad de 75 a menos de 85 obtenidas con el método "Tiempo de Muerte log-Logístico"son más cercanas a las observadas cuando se utilizan parámetros estimados por máxima verosimilitud (Figura 4.11).

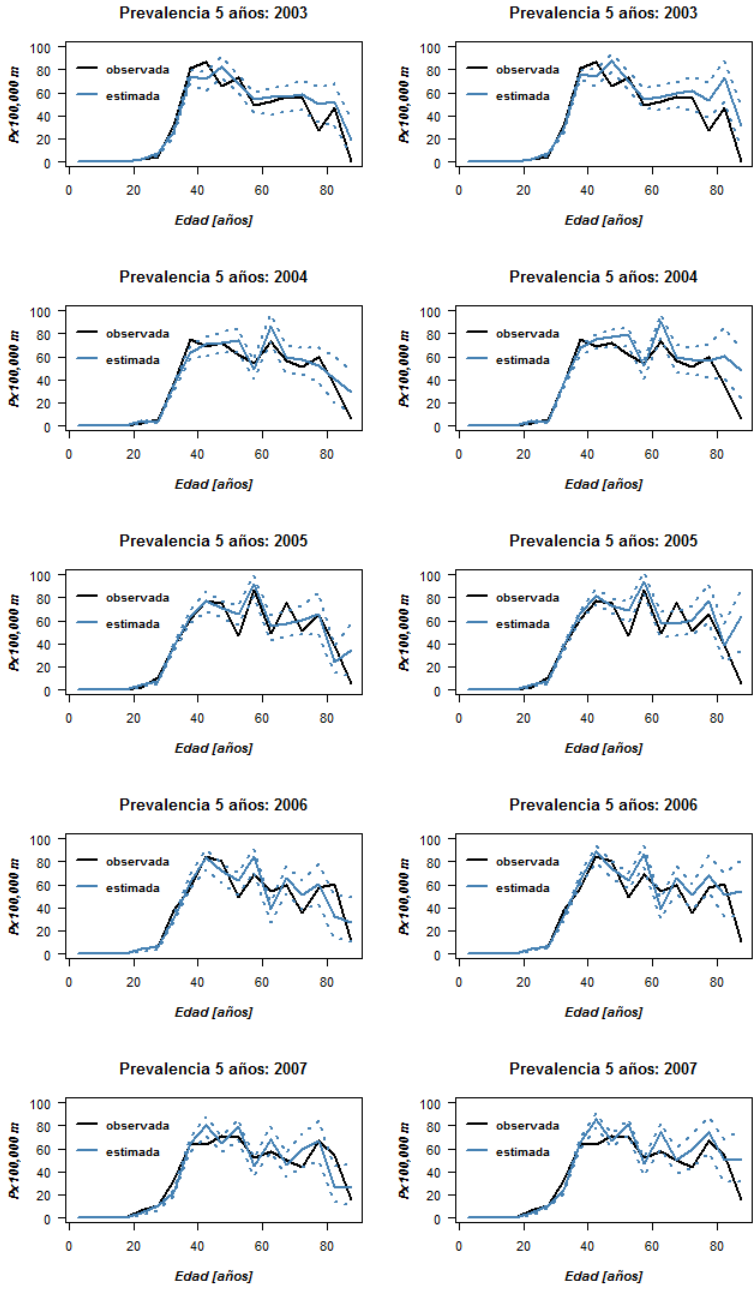

Figura 4.9: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada (línea azul, continua para la mediana y punteada para los percentiles 2,5 y 97,5 %) a partir del método "Tiempo de Muerte Exponencial". A la izquierda corresponde a simulaciones con parámetros de las distribuciones exponenciales por grupos de edad estimados por métodos gráficos y a la derecha con estimadores máximo verosímiles.

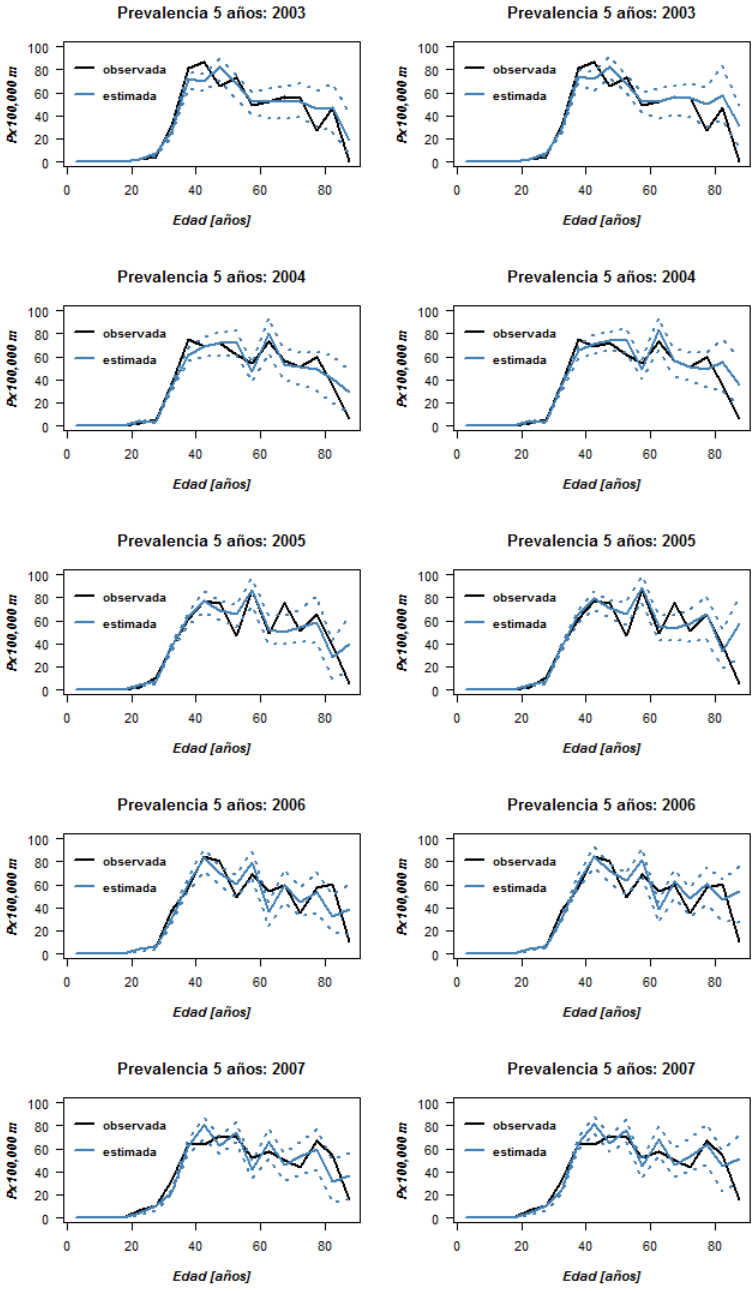

Figura 4.10: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada (línea azul, continua para la mediana y punteada para los percentiles 2,5 y 97,5 %) a partir del método "Tiempo de Muerte Weibull". A la izquierda corresponde a simulaciones con parámetros de las distribuciones Weibull por grupos de edad estimados por métodos gráficos y a la derecha con estimadores máximo verosímiles.

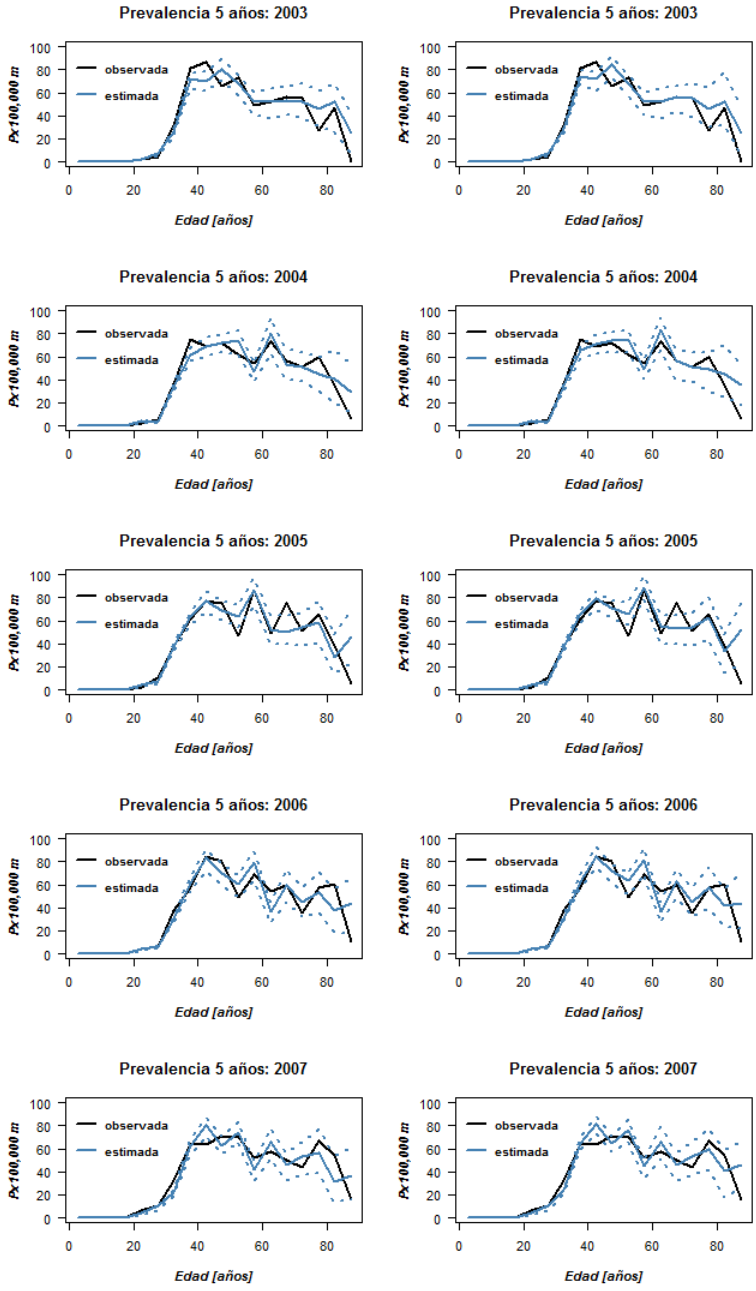

Figura 4.11: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007 observada (línea negra) y estimada (línea azul, continua para la mediana y punteada para los percentiles 2,5 y 97,5 %) a partir del método "Tiempo de Muerte log-L". A la izquierda corresponde a simulaciones con parámetros de las distribuciones log-Logísticas por grupos de edad estimados por métodos gráficos y a la derecha por método de máxima verosimilitud.

Continuando con el análisis de validación de los métodos desarrollados para estimar prevalencia a 5 años, en la Tabla 4.7 se presentan los valores de *Razón de Discrepancia media* (*RDm*) por año, para cada uno de los métodos desarrollados. De acuerdo a esta medida (y a lo percibido en las figuras precedentes), el método que produce mejores estimaciones de prevalencia a 5 años corresponde al método "Mirror"(ya que muestra la menor discrepancia entre los valores observados y estimados), seguido por el método "Mirror-Uniforme". A su vez, los métodos que utilizan modelización empírica o paramétrica de la función de supervivencia $S(t)$ para estimar el tiempo de muerte, muestran valores de *Razón de Discrepancia media* y *total* levemente mayores a los obtenidos con los métodos de remuestreo "Mirror", siendo los métodos de tiempo de muerte Weibull y log-Logístico los de menor discrepancia entre la prevalencia observada y la estimada.

	Año					*RDm*	Coste
Método de simulación	2003	2004	2005	2006	2007	*Total*	relativo
T. de Seg. Emp.	221.309	116.442	0.787	0.927	64.966	404.431	12.71
Mirror	0.057	0.204	0.252	0.166	0.134	0.812	1
Mirror-Exponencial	0.396	0.387	0.386	0.559	0.765	2.493	1.96
Mirror-Uniforme	0.289	0.528	0.500	0.376	0.359	2.052	2.09
Supervivencia Emp. Kaplan-Meier	0.790	0.711	0.728	0.687	0.837	3.753	10.82
Supervivencia Emp. Nelson-Aalen	0.813	0.739	0.730	0.684	0.825	3.791	10.84
T. de Muerte Exp. (ML)	0.963	0.942	0.983	0.856	1.017	4.761	54.4
T. de Muerte Exp. (gráfico)	0.761	0.699	0.731	0.749	0.832	3.772	34.60
T. de Muerte Weibull (ML)	0.754	0.699	0.767	0.699	0.795	3.714	55.33
T. de Muerte Weibull (gráfico)	0.731	0.655	0.739	0.633	0.846	3.604	34.97
T. de Muerte log-L. (ML)	0.717	0.702	0.755	0.679	0.805	3.659	54.67
T. de Muerte log-L. (gráfico)	0.756	0.680	0.712	0.641	0.846	3.635	35.67

Tabla 4.7: *Razón de discrepancia media* (*RDm*) para cada uno de los ocho métodos utilizados para estimar la prevalencia a 5 años en Girona y Tarragona entre 2003 y 2007. Para los métodos basados en el modelado de la supervivencia, se incluyen los valores de *RDm* obtenidos para sus variaciones en los métodos de estimación usados. También se presenta el coste en recursos computacionales relativo al método "Mirror"por ser el más eficiente en términos de tiempo de cálculo.

Las mayores discrepancias fueron obtenidas con el método "Tiempo de Seguimiento Empírico", que produce valores de *RDm* extremadamente grandes (Tabla 4.7). Esto se debe, al menos en parte, a que en algunos intervalos de edad los intervalos de confianza 95 % de las estimaciones son muy estrechos (baja variabilidad en las bases simuladas), provocando que la diferencia absoluta entre el valor observado y el estimado sea varios ordenes de magnitud más grande que la amplitud del intervalo de confianza 95 % de la estimación. Por consiguiente, la *Razón de discrepancia* (*RD*) correspondiente al intervalo de edad es muy grande.

En cuanto al coste computacional en la implementación de los diferentes métodos presentados para la simulación de bases y la posterior estimación de prevalencia a 5 años, el método "Mirror"es el más rápido y por tanto, el más eficiente. En la Tabla 4.7 se puede ver el coste relativo de cada método, en términos de tiempo de cálculo requerido respecto al método más rápido ("Mirror"), para estimar prevalencia a 5 años en Girona y Tarragona en 2003 con mil bases simuladas. Los métodos de mayor coste son los basados en la modelización paramétrica de la supervivencia con estimación de parámetros por máxima verosimilitud, que requieren entorno a 55 veces más tiempo de cálculo (en Apéndice B.2 se encuentran los tiempos requeridos por cada método para estimar prevalencia a 5 años en Girona y Tarragona en 2003 con 1000 simulaciones en un ordenador de uso personal).

4.4. Estimaciones de prevalencia a 5 años para Cataluña

Como se explicó en la sección 3.9, las estimaciones de prevalencia a 5 años en Cataluña se realizaron a partir de 1000 bases (cohortes de mujeres) simuladas sobre las estimaciones de incidencia en Cataluña entre 1999 y 2007, usando como referencia los registros de cáncer de base poblacional de Girona y Tarragona en igual período. Un resultado que se repite en todas las estimaciones de prevalencia obtenidas para Cataluña (mediante los diferentes métodos de simulación desarrollados) es que sus intervalos de confianza 95 % son más estrechos que en las estimaciones para Girona y Tarragona con igual método. Esto se debe a que en Cataluña hay más casos incidentes a simular, ya que la población de Girona y Tarragona representa aproximadamente el 20 % de la población total de la comunidad.

Para el primero de los métodos presentados, "Tiempo de Seguimiento Empírico"(basado en las distribuciones empíricas del tiempo de seguimiento observado en seis grupos de edad en Girona y Tarragona), las estimaciones de prevalencia a 5 años en Cataluña oscilan entorno a las observadas en la población de referencia Girona y Tarragona, a lo largo del mismo período (Apéndice B.3: Figura B.2). Sin embargo, en ciertas ocasiones parece haber una tendencia a la sobreestimación de la prevalencia; por ejemplo para el 2004 la prevalencia estimada en Cataluña tiende a estar por encima de la observada en Girona y Tarragona en los intervalos de edad intermedios (entre los 40 y los 70 años), y de forma similar en el 2007.

Claramente el método "Mirror"basado en el remuestreo por año e intervalo de edad quinquenal sobre la base de referencia es el que produce estimaciones de prevalencia a 5 años en Cataluña que mejor emulan a las observadas en Girona y Tarragona (Figura 4.12). Al igual que con el los otros métodos, los intervalos

de confianza para las estimaciones de Cataluña son bastante más estrechos que los obtenidos para Girona y Tarragona (Figura 4.5) debido a que en Cataluña hay más casos a simular.

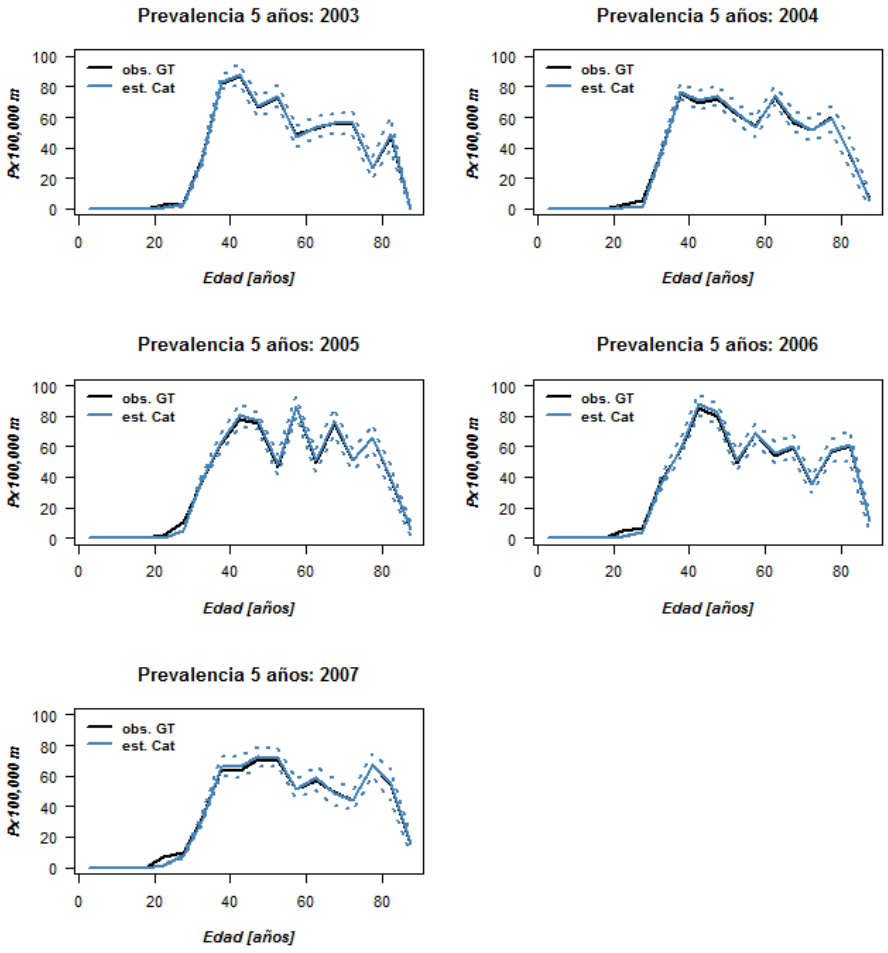

Figura 4.12: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 estimada mediante el método de remuestreo desde la base de Girona y Tarragona según año de incidencia e intervalo de edad, "Método Mirror". En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

De manera similar a lo observado para Girona y Tarragona, el método "Mirror Exponencial"subestima la prevalencia a 5 años en Cataluña en todos los años, siendo ésta subestimación más acentuada en 2006 y 2007 (Apéndice B.3: Figura B.3).

Por otra parte, los resultados obtenidos por el método "Mirror Uniforme" indican que la estrategia de remuestreo y posterior simulación del tiempo de seguimiento desde distribuciones uniformes para cada intervalo de edad quinquenal es una estrategia relativamente buena para la generación de cohortes de mujeres diagnosticadas por cáncer de cuello de útero. Las estimaciones de prevalencia obtenidas logran emular bastante bien el patrón de prevalencia observado en Girona y Tarragona, aunque hay cierta tendencia a sobrestimar la prevalencia en los intervalos de mayor edad (Apéndice B.3: Figura B.4).

En cuanto a los métodos que utilizan modelos de supervivencia ajustados en la población de referencia para simular las cohortes de mujeres, a excepción del método con tiempos de muerte exponenciales, todos producen resultados similares y pueden ser igualmente válidos. En el método donde se modela en base a la supervivencia empírica de Kaplan-Meier, las estimaciones de prevalencia a 5 años en Cataluña reproducen la tendencia en el patrón de ocurrencia de la enfermedad en Girona y Tarragona aunque los intervalos de confianza 95 % no cubren completamente lo observado en dichas provincias (Figura 4.13).

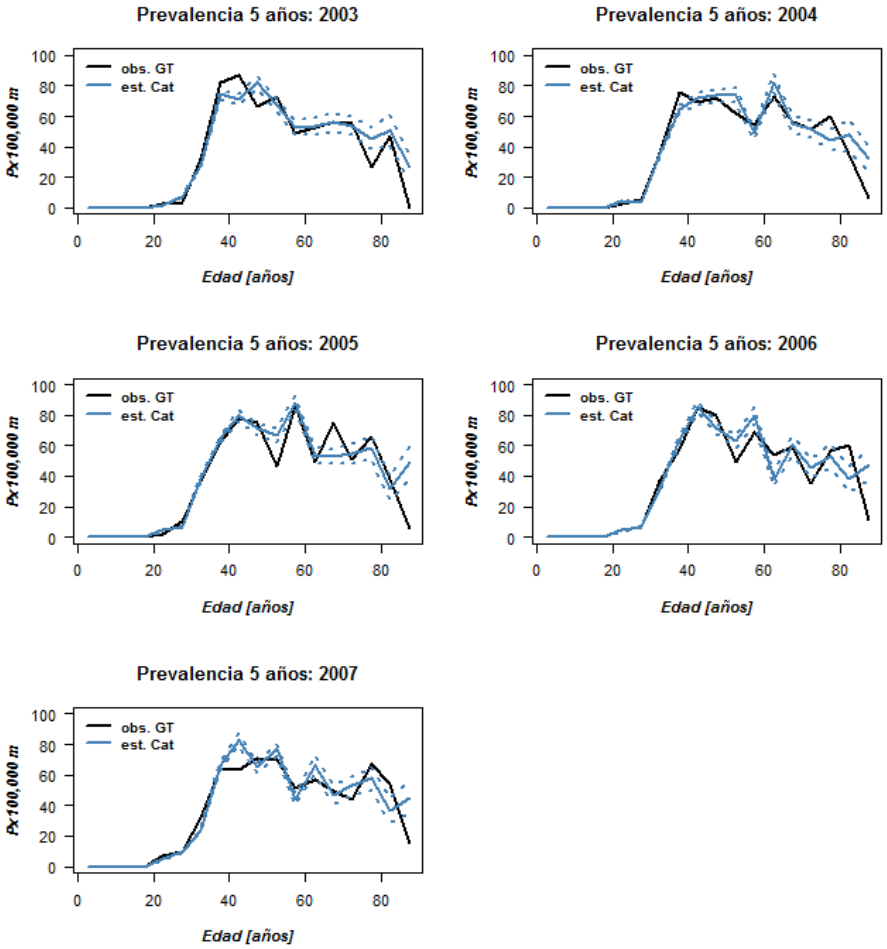

Figura 4.13: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 realizadas en base al método de simulación de cohortes por Supervivencia empírica Kaplan-Meier. En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

Las estimaciones de prevalencia a 5 años para Cataluña basadas en modelos exponenciales ajustados por grupo de edad sobre la población de Girona y Tarragona, son frecuentemente sobrestimadas. Esta tendencia ya notada en las estimaciones realizadas para la población de referencia Girona y Tarragona, se hace más evidente en Cataluña, en particular en 2003 y 2007 (Apéndice B.3: Figura B.5).

Por último, no se observan prácticamente diferencias en los resultados obtenidos mediante los métodos que utilizan la función de supervivencia ajustada

utilizando un modelo de distribución log-Logístico (Figura 4.14) o un modelo Weibull (Apéndice B.3: Figura B.6). Este resultado no llama la atención ya que en la exploración gráfica realizada para evaluar el grado de ajuste de los datos a los diferentes modelos paramétricos, los modelos Weibull y log-Logístico mostraron ser posiblemente adecuados (Figura 4.2).

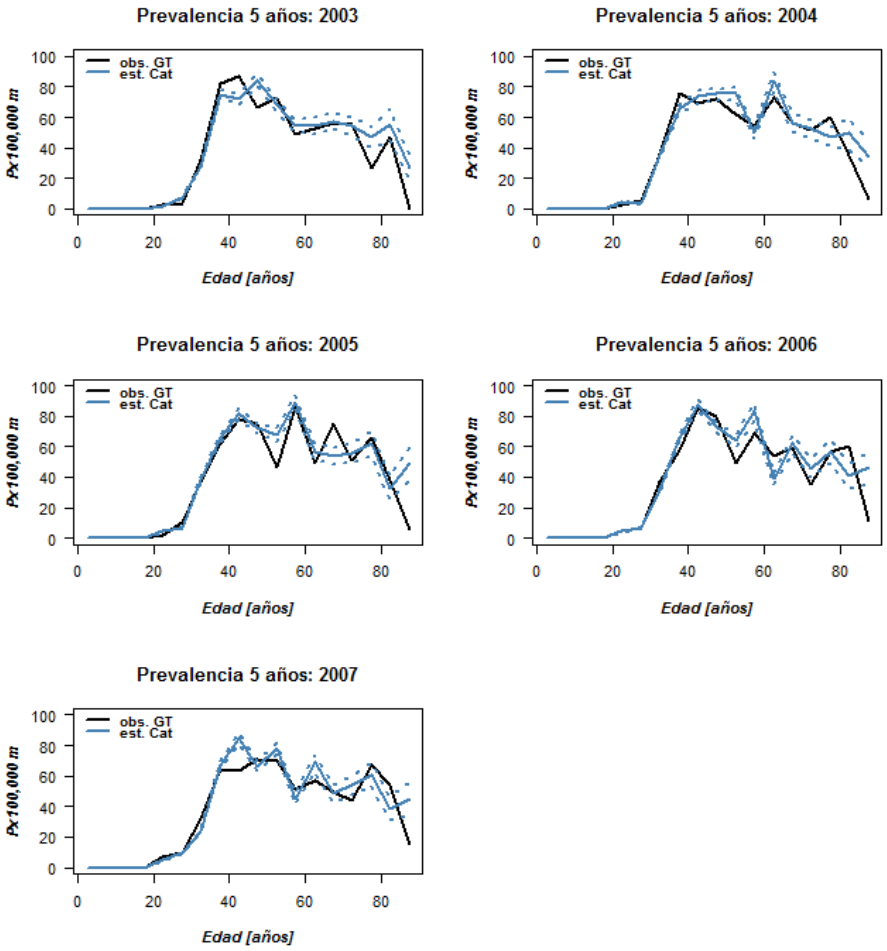

Figura 4.14: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 realizadas en base a modelos de supervivencia con distribución log-Logística ajustados (mediante ML) en seis grupos de edad de la población de Girona y Tarragona, método "Tiempo de Muerte log-Logístico". En negro se muestra la prevalencia observada en la población de referencia, Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

5

Discusión y Conclusiones

En este trabajo proporcionamos medidas de ocurrencia de cáncer de cuello de útero en Girona y Tarragona, útiles para la identificación de los recursos requeridos en la prestación de servicios de salud. La tasa de incidencia anual, y las prevalencias de duración limitada (acumulada a 5 años, y puntuales a 1, 2-3 y 4-5 años) en cada intervalo de edad quinquenal fueron calculadas directamente sobre los registros de cáncer de base poblacional. La utilidad de estas medidas radica en que informan sobre las necesidades sanitarias requeridas dada la edad y la fase de tratamiento en que se encuentran las pacientes. Por otra parte, presentamos ocho métodos para la generación de cohortes simuladas de mujeres incidentes de esta enfermedad en Cataluña que permiten realizar estimaciones de prevalencia a 5 años por intervalo de edad.

En Girona y Tarragona, tanto la incidencia anual como la prevalencia de la enfermedad se ha mantenido relativamente constante durante el período de estudio, 1999-2007. Estos resultados coinciden con la tendencia reportada para la incidencia en Cataluña y España (Bosch 1999, Clèries et al 2014).

Los casos de cáncer de cuello de útero comienzan a aparecer en las mujeres de Girona y Tarragona con 20 o más años de edad, y es a partir de los 35 años donde se observa un claro aumento de la incidencia de esta enfermedad, hasta alcanzar una meseta después de los 40-45 años. La edad también muestra una relación con la supervivencia de las mujeres que son diagnosticadas por este tipo de cáncer, ya que los grupos de mayor edad muestran menor supervivencia que las mujeres más jóvenes (Figura 4.1). Luego de transcurridos 5 años desde la fecha de diagnóstico para cáncer de cuello uterino, la probabilidad de que una paciente diagnosticada con menos de 35 años continúe viva es el doble que en las mujeres diagnosticadas con 75 años o más. Un estudio previo realizado por Sánchez et al (1996b) reporta resultados similares para Girona, donde las mujeres de mayor edad tendieron a mostrar peor supervivencia que las más jóvenes aunque la diferencia no fue estadísticamente significativa.

En cuanto a los métodos planteados, las diferentes estrategias utilizadas para simular cohortes de mujeres incidentes para cáncer de cérvix, permiten alcanzar estimaciones de prevalencia a 5 años que varían en el grado de acuerdo con las prevalencias observadas y en el coste computacional (Tabla 4.7). A excepción del método "Tiempo de Seguimiento Empírico"(que utiliza las distribuciones de tiempo de seguimiento observadas por grupos de edad en la base de referencia para generar las cohortes simuladas), la *Razón de Discrepancia media* entre la prevalencia observada y la estimada de los intervalos de edad quinquenales es menor o igual a 1 para todos los años y métodos utilizados. La interpretación de este resultado sería que la amplitud de los intervalos de confianza 95 % obtenidos para las estimaciones de prevalencia de cada intervalo de edad cubren en promedio los valores observados. Esta interpretación debe tomarse con cierta precaución ya que la medida *Razón de Discrepancia* asume que la prevalencia simulada tiene distribución normal y que los intervalos de predicción son simétricos (Moller et al 2005), lo cual aquí no se cumple necesariamente. Por otro lado, lo que sucede en el método "Tiempo de Seguimiento Empírico"es que en ciertos intervalos de edad las predicciones de prevalencia a 5 años tienen intervalos de predicción demasiado estrechos, por lo tanto el cociente (razón) entre la distancia de la estimación a la observación, y la distancia de la estimación al límite de su intervalo es extremadamente grande.

Los métodos basados en remuestreo, específicamente "Mirror"y "Mirror Uniforme", producen estimaciones de prevalencia cercanas a las observadas (excepto para los tres intervalos de mayor edad entre 2003-2005 con el método "Mirror Uniforme", Figura 4.7), aunque algo imprecisas dado que los intervalos de confianza son relativamente amplios. Desde el punto de vista del coste computacional ambos métodos son los más eficientes, permitiendo obtener estimaciones de prevalencia en muy poco tiempo (Tabla 4.7). La gran limitante de éstos métodos radica en que todos los casos incidentes a simular deben tener un "análogo"en cuanto a la edad y al año de incidencia en la base de referencia. Es decir, si en los casos incidentes para las que queremos generar una base simulada existe algún caso perteneciente a un intervalo de edad que no aparece en la base de referencia, éste no podrá ser simulado a partir del remuestreo. Aquí esto no fue un problema porque los casos incidentes a simular de Cataluña son proyecciones basadas en la información contenida en los registros de Girona y Tarragona (Cléries et al 2014) y no hay casos a simular en Cataluña con edad y año de incidencia que no pueda ser "remuestreado"en la población de referencia. De la limitación mencionada se desprende que estos métodos no permiten hacer proyecciones a futuro, ya que el año o los años de incidencia de los casos a simular deben estar siempre presentes en los registros de cáncer de la población de referencia. En este sentido, los métodos basados en la modelización de la supervivencia, no tienen restricción, ya que la supervivencia es ajustada por grupo de edad sobre el conjunto de datos de los registros de cáncer independientemente de que abarque el año de incidencia de cada caso a simular. De hecho, estos métodos podrían ser usados para simular cohortes de mujeres con proyecciones a futuro de los casos incidentes a simular.

Los métodos basados en la supervivencia, ya sea empírica o bajo un modelo de distribución paramétrico apropiado (para el caso de estudio, Weibull o log-Logístico), muestran resultados aceptables en cuanto a que reproducen el patrón de ocurrencia de la enfermedad, para algunos intervalos de edad de forma no muy exacta pero bastante precisos en general (intervalos de confianza relativamente estrechos). La gran diferencia entre éstos últimos métodos de estimación de prevalencia radica en el coste computacional, donde los métodos basados en modelos paramétricos de distribución del tiempo de muerte son extremadamente lentos (requieren entre 35 y 55 veces más tiempo que el método más rápido, Tabla 4.7). Otra desventaja es la posibilidad de elegir un modelo de distribución inapropiado, lo que produciría estimaciones de prevalencia equivocadas (por ejemplo, la sobreestimación de la prevalencia a 5 años producida con el método "Tiempo de Muerte Exponencial", Apéndice B.3: Figura B.5). Además, si bien estos métodos son flexibles en el sentido de que permiten ajustar el modelo elegido por grupo de edad y así obtener estimaciones de los parámetros específicas de la edad, se restringen a un único modelo de distribución para el tiempo de muerte. Esto puede ser una desventaja ya que no necesariamente el tiempo de muerte se distribuye de la misma manera en cada grupo de edad. En este sentido, el método de estimación de prevalencia basado en la supervivencia empírica de la población de referencia es ventajoso ya que en cada grupo de edad se estima una función de supervivencia sin ceñirse a un modelo de distribución particular.

Una consideración a tener en cuenta, es que la supervivencia es ajustada de manera tradicional, utilizando todas las pacientes diagnosticadas dentro del período de estudio, y donde los grupos de edad son definidos unicamente por la edad en el momento del diagnóstico (tiempo cero). Una consecuencia es que para largos períodos de tiempo, ésta supervivencia refleja escencialmente la esperanza de supervivencia de pacientes diagnosticadas varios años antes (Brenner et al 2004). Si a lo largo del período de estudio la eficacia de las técnicas de detección y/o de los tratamientos mejora, entonces la supervivencia estará subestimada. En este sentido, el ajuste de la supervivencia con el método de períodos (Brenner & Gefeller 1997, Brenner et al 2004) podría ser más apropiado (ya que proporciona supervivencias más actualizadas) y debería ser explorado.

Si bien en este trabajo la estrategia de simulación de cohortes de mujeres incidentes de cáncer de cuello uterino fue utilizada para estimar prevalencia a 5 años, la metodología es fácilmente adaptable para estimar prevalencia puntual a 1, 2-3 o 4-5 años por intervalo de edad.

En conclusión, parece claro que las estimaciones de prevalencia a partir de cohortes simuladas es una opción factible además de novedosa. De hecho, en este proyecto generamos estimaciones de prevalencia a 5 años para Cataluña que resultan razonables, en particular con los métodos "Mirror", "Supervivencia Empírica"y "Tiempo de Muerte log-Logístico". A juzgar entre las venta-

jas y desventajas de cada uno de ellos, el método basado en la supervivencia empírica produce los mejores resultados, ya que puede utilizarse para estimar prevalencia en cualquier conjunto de casos a simular (incluso proyecciones de casos incidentes), es relativamente eficiente en cuanto al tiempo de cálculo y no hay posibilidad de ajustar un modelo de distribución erróneo.

Referencias

Aalen OO. 1978. Nonparametric Inference for a Family of Counting Processes. *The Annals of Statistics* 6: 701-726.

Almonte M, Albero G, Molano M, Carcamo C, García PJ & Pérez G. 2008. Risk factors for human papillomavirus exposure and co-factors for cervical cancer in Latin America and the Caribbean. *Vaccine* 26(11): L16-36.

Bender R, Augustin T & Blettner M. 2005. Generating survival times to simulate Cox proportional hazards models. 2005. *Statistics in Medicine* 24(11): 1713-1723.

Bray F, Ren J-S, Masuyer E & Ferlay J. 2013. Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *International Journal of Cancer* 132(5): 1133-1145.

Bosch FX. 1999. Trends in cervical cancer mortality. *Journal of Epidemiology and Community Health* 53(7):392.

Bosch FX & de Sanjosé S. 2007. The epidemiology of human papillomavirus infection and cervical cancer. *Disease Markers* 23(4): 213-227.

Bosch FX, Lorincz AN, Muñoz N, Meijer CJLM & Shah KV. 2002. The causal relation between human papillomavirus and cervical cancer. *Journal of Clinical Pathology* 55(4): 244-265.

Bouvard V, Baan R, Straif K, Grosse Y, Secretan B, El Ghissassi F, Benbrahim-Tallaa L, Guha N, Freeman C, Galichet L & Cogliano V; WHO International Agency for Research on Cancer Monograph Working Group. 2009. A review of human carcinogens-Part B: biological agents. *The Lancet Oncology* 10(4): 321-322.

Brenner H & Gefeller O. 1997. Deriving more up-to-date estimates of long term patient survival. *Journal of Clinical Epidemiology* 50(2): 211-216.

Brenner H, Gefeller O & Hakulinen T. 2004. Period analysis for 'up-to-date'

cancer survival data: theory, empirical evaluation, computational realisation and applications. *European Journal of Cancer* 40(3): 326-335.

Capocaccia R & De Angelis R. 1997. Estimating the completeness of prevalence based on cancer registry data. *Statistics in Medicine* 16(4): 425-440.

Castellsagué X, Bosch FX & Muñoz N. 2002. Environmental co-factors in HPV carcinogenesis. *Virus Research* 89(2): 191-199.

Carstensen B, Plummer M, Laara E & Hills M. 2014. Epi: A Package for Statistical Analysis in Epidemiology. R package version 1.1.67. URL `http://CRAN.R-project.org/package=Epi`

Clèries R, Esteban L, Borràs J, Marcos-Gragera R, Freitas A, Carulla M, Buxó M, Puigdefàbregas A, Izquierdo Á, Gispert R, Galceran J & Ribes J. 2014. Time trends of cancer incidence and mortality in Catalonia during 1993-2007. *Clinical and Translational Oncology* 16(1): 18-28.

Clèries R, Ribes J, Buxo M, Ameijide A, Marcos-Gragera R, Galceran J, Miguel Martínez J & Yasui Y. 2012. Bayesian approach to predicting cancer incidence for an area without cancer registration by using cancer incidence data from nearby areas. *Statistics in Medicine* 31(10): 978-987.

Colonna M, Danzon A, Delafosse P, Mitton N, Bara S, Bouvier AM, Ganry O, Guizard AV, Launoy G, Molinie F, Sauleau EA, Schvartz C, Velten M, Grosclaude P & Tretarre B. 2008. Cancer prevalence in France: time trend, situation in 2002 and extrapolation to 2012. *European Journal of Cancer* 44(1): 115-122.

De Angelis R, Capocaccia R, Hakulinen T, Soderman B & Verdecchia A. 1999. Mixture models for cancer survival analysis: application to population-based data with covariates. *Statistics in Medicine* 18(4): 441-454.

Diaz Sanchis M. 2014. *Modelos de coste-efectividad en la prevención del cáncer de cuello de útero en países en desarrollo.* Tesis doctoral. Universidad Autónoma de Barcelona, Facultad de Medicina.

Doorbar J, Quint W, Banks L, Bravo IG, Stoler M, Broker TR & Stanley MA. 2012. The biology and life-cycle of human papillomaviruses. *Vaccine* 30(5): F55-70.

Feldman AR, Kessler L, Myers MH & Naughton MD. 1986. The prevalence of cancer. Estimates based on the Connecticut tumor registry. *The New England Journal of Medicine* 315(22): 1394-1397.

Ferlay J, Bray F, Steliarova-Foucher E and Forman D. 2014. *Cancer Incidence in Five Continents*, CI5*plus*. IARC CancerBase Nº 9. International Agency

for Research on Cancer, Lyon. URL http://ci5.iarc.fr, acceso [26/04/2015].

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D & Bray F. 2013. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide*. IARC CancerBase Nº 11 [Internet]. International Agency for Research on Cancer, Lyon. URL http://globocan.iarc.fr, acceso [26/03/2015].

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D & Bray F. 2015. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. International Journal of Cancer 136(5): E359-E386.

Forman D, Bray F, Brewster DH, Gombe Mbalawa C, Kohler B, Piñeros M, Steliarova-Foucher E, Swaminathan R & Ferlay J, editores. 2013. *Cancer Incidence in Five Continents, Vol. X* [Internet]. International Agency for Research on Cancer, Lyon. URL http://ci5.iarc.fr, acceso [25/03/2015].

Gail MH, Kessler L, Midthune D & Steven Scoppa S. 1999. Two approaches for estimating disease prevalence from populationbased registries of incidence and total mortality. *Biometrics* 55(4): 1137-1144.

Harrell FE Jr con contribuciones de Dupont C y otros. 2014. Hmisc: Harrell Miscellaneous. R package version 3.16-0. URL http://CRAN.R-project.org/package=Hmisc

Hernandez BY, Wilkens LR, Zhu X,Thompson P,McDuffie K,Shvetsov YB, Kamemoto LE, Killeen J, Ning L & Goodman MT. 2008. Transmission of human papillomavirus in heterosexual couples. *Emerging Infectious Diseases* 14(6): 888-894.

IARC. 2007. Human papillomaviruses. *IARC Monographs on Evaluation of Carcinogenic Risks to Humans* 90: 1-636.

Kaplan EL & Meier P. 1958. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* 53(282): 457-481.

Klein JP & Moeschberger ML. 2003. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York.

Krogh V & Micheli A. 1996. Measure of cancer prevalence with a computerized program: an example on larynx cancer. *Tumori* 82(3): 287-290.

Loos AH, Bray F, McCarron P, Weiderpass E, Hakama M & Parkin DM. 2004. Sheep and goats: separating cervix and corpus uteri from imprecisely coded uterine cancer deaths, for studies of geographical and temporal variations

in mortality. *European Journal of Cancer* 40(18): 2794-2803.

Louie KS, Castellsague X, de Sanjose S, Herrero R, Meijer CJ, Shah K, Muñoz N & Bosch X; International Agency for Research on Cancer Multicenter CERVICAL CANCER Study Group. 2011. Smoking and passive smoking in cervical cancer risk: pooled analysis of couples from the IARC multicentric case-control studies. *Cancer Epidemiology, Biomarkers & Prevention* 20: 1379?1390.

Moller B, Weedon-Fekjaer H & Haldorsen T. 2005. Empirical evaluation of prediction intervals for cancer incidence. *BMC Medical Research Methodology* 5:21.

Mood AM, Graybill FA & Boes DC. 1974. *Introduction to the Theory of Statistics* (3rd ed). McGraw-Hill, New York.

Moscicki AB, Schiffman M, Burchell A, Albero G, Giuliano A, Goodman MT, Kjaer SK & Palefsky J. 2012. Updating the natural history of human papillomavirus and anogenital cancers. *Vaccine* 30(5): F24-33.

Muñoz N, Bosch FX, Castellsagué X, Díaz M, de Sanjose S, Hammouda D, Shah KV & Meijer CJ. 2004. Against which human papillomavirus types shall we vaccinate and screen? The international perspective. *International Journal of Cancer* 111(2): 278-285.

Muñoz N, Castellsagué X, de González AB & Gissmann L. 2006. Chapter 1: HPV in the etiology of human cancer. *Vaccine* 24(3): 1-10.

Nelson, W. 1972. Theory and applications of hazard plotting for censored failure data. *Technometrics* 14(4): 945-965.

Paavonen J. 2007. Human papillomavirus infection and the development of cervical cancer and related genital neoplasias. *International Journal of Infectious Diseases* 11(2): S3?9.

Pisani P, Bray F & Parkin DM. 2002. Estimates of the world-wide prevalence of cancer for 25 sites in the adult population. *International Journal of Cancer* 97(1): 72-81.

Porta M. 2014. *A Dictionary of Epidemiology* (6th ed). Oxford University Press, New York.

R Development Core Team. 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL `http://www.R-project.org`

Rodríguez AC, Schiffman M, Herrero R, Hildesheim A, Bratti C, Sherman ME, Solomon D, Guillén D, Alfaro M, Morales J, Hutchinson M, Katki H, Cheung L, Wacholder S & Burk RD. 2010. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. *Journal of the Natlional Cancer Institute* 102(5): 315-324.

Sánchez MV, Izquierdo A, Beltran M, Bosch FX & Viladiu P. 1996a. Tendencias temporales de la mortalidad por cáncer de cérvix en Cataluña 1975-1992: análisis del Boletín Estadístico de Defunción y del Registro de Cáncer de Girona. *Gaceta Sanitaria* 10(53):67-72.

Sánchez MV, Izquierdo A, Beltran M, Bosch FX & Viladiu P. 1996b. Epidemiología del cáncer invasor de cérvix en el área sanitaria de Girona durante el período 1980-1989. Registro poblacional de cáncer de Gerona. *Revista Española de Salud Pública* 71(1):19-26.

Therneau T. 2015. A Package for Survival Analysis in S. version 2.38, URL `http://CRAN.R-project.org/package=survival`

Trottier H & Franco EL. 2006. The epidemiology of genital human papillomavirus infection. *Vaccine* 24(1): 1-15.

Vaccarella S, Lortet-Tieulent J, Plummer M, Franceschi S & Bray F. 2013. Worldwide trends in cervical cancer incidence: Impact of screening against changes in disease risk factors. *European Journal of Cancer* 49(15): 3262-3273.

Verdecchia A, De Angelis G & Capocaccia R. 2002. Estimation and projections of cancer prevalence from cancer registry data. *Statistics in Medicine* 21(22):3511-3526.

Wang SS & Hildesheim A. 2003. Chapter 5: Viral and host factors in human papillomavirus persistence and progression. *Journal of the National Cancer Institute. Monographs* 31: 35-40.

Winer RL, Lee SK, Hughes JP, Adam DE, Kiviat NB & Koutsky LA. 2003. Genital human papillomavirus infection: incidence and risk factors in a cohort of female university students. *American Journal of Epidemiology* 157(3): 218-226.

Winer RL, Hughes JP, Feng Q, Xi LF, Cherne S, O'Reilly S, Kiviat NB & Koutsky LA. 2011. Early natural history of incident, type-specific human papillomavirus infections in newly sexually active young women. *Cancer Epidemiology, Biomarkers & Prevention* 20(4): 699-707.

Zeng XT, Xiong PA, Wang F, Li CY, Yao J & Guo Y. 2012. Passive smoking and cervical cancer risk: a meta-analysis based on 3,230 cases and 2,982

controls. *Asian Pacific Journal of Cancer Prevention* 13(6): 2687-2693.

zur Hausen H. 2000. Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis. *Journal of the National Cancer Institute* 92(9): 690-698.

Apéndice A

Funciones propias

A.1.1. Funciones *carga* y *prevalencia*

carga contabiliza nuevos casos para la enfermedad por intervalo de edad quinquenal acumulados en el período en años especificado. Si el período es de 1 año 'Lx100000' corresponde a la Tasa de incidencia anual (casos incidentes por cada 100,000 mujeres-año).

Argumentos:

`datos`:	base de los registros de cáncer en formato de *data frame* con la edad ('edat.inc'), el mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us') y año ('any.us) del último seguimiento, y estatus ('status') a la fecha del último seguimiento.
`año`:	año para el cual se calcula la incidencia.
`período`:	se refiere a la cantidad de años para los que se quiere acumular. si se quiere la tasa de incidencia anual período debe ser 1.
`pob`:	*data frame* con la población total ('population') por año ('year') e intervalo de edad quinquenal ('age.group', 18 intervalos).

Salida: Un *data frame* con intervalo de edad quinquenal, número de casos diagnosticados por cáncer (C), población a riesgo (Y) y tasa cruda de incidencia anual (L, casos por cada 100,000 mujeres a riesgo-año).

```
> carga <- function(datos, año, periodo, pob){
+   a <- año
+   require("Hmisc")
+   require("Epi")
+   set <- (subset(datos, any.inc>=a-(periodo-1) & any.inc<=a))
+
+   corr.age<-function(x.age)
```

```
+   {
+     if (x.age>85) x.age<-85
+     x.age
+   }
+   edat.inc.tmp <- as.numeric(lapply(set$edat.inc, corr.age))
+
+   intervalo.edat <- cut2(edat.inc.tmp, seq(0, 90, 5))
+   set.c <- cbind(set, intervalo.edat)
+
+   C <- stat.table(list(Edad=intervalo.edat),
+                   list(C=count()), margins=TRUE, data=set.c)
+
+   Ytemp<-pob[pob$year>=a-(periodo-1) & pob$year<=a,]
+   Y <- stat.table(list(GrupoEdad=age.group),list(Y=sum(population)),
+                   margins=TRUE, data=Ytemp)
+   result <- as.data.frame(t(rbind(C, Y)),row.names=c(1:19))
+   result$Lx100000 <- result$C/result$Y*100000
+   result$Age <- c(levels(intervalo.edat),'Total')
+   result[,c(4,1:3)]
+ }
> # Ejemplo: casos incidentes en 2007
> carga07.1a <- carga(fit.surv,2007,1,fit.pop.F)
> head(carga07.1a,6)

      Age C     Y Lx100000
1       0 0 38869 0.000000
2       5 0 34918 0.000000
3      10 0 33232 0.000000
4      15 0 34825 0.000000
5 [20,25) 1 44087 2.268242
6 [25,30) 3 59678 5.026978

>
> #-----------------------------------------------------------------
```

prevalencia permite calcular la prevalencia acumulada o puntual (por intervalos), pacientes vivos que han sido diagnosticados dentro de un período de tiempo especificado.

Argumentos:

`datos`:	base de los registros de cáncer en formato de *data frame* con la edad ('edat.inc'), el mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us') y año ('any.us) del último seguimiento, y estatus ('status') a la fecha del último seguimiento.
`año.t`:	año para el que se quiere calcular prevalencia.
`int1` e `int2`:	indican el número de años que llevan en tratamiento los pacientes, primero el tiempo mínimo en tratamiento y segundo el máximo (ej. si se quiere calcular prevalencia en el `año.t` para pacientes que incidieron en los últimos 5 años se debe definir `int.1=1` e `int.2=5`, si se quieren los casos prevalentes que llevan entre 2 y 3 años en tratamiento se debe definir `int.1=2` e `int.2=3`).
`pob`:	*data frame* con la población total ('population') por año ('year') e intervalo de edad quinquenal ('age.group', 18 intervalos).

Salida: Un *data frame* con intervalo de edad quinquenal, 'C' son los casos incidentes que llevan entre `int.1` e `int.2` años de diagnóstico y que prevalecen al final del `año.t`, 'Y' son las personas del `año.t`, y 'Px100000' es la prevalencia (casos prevalentes por cada 100,000 mujeres).

```
> prevalencia <- function(datos, año.t, int1, int2, pob){
+   a <- año.t
+
+   require("Hmisc")
+   require("Epi")
+
+   set <- (subset(datos, any.inc>=a-(int2-1) & any.inc<=a-(int1-1) &
+                    (!(any.us<=a & status==1))))
+
+   edat.us <- with(set,(12-mes.inc)*(1/12)+(a-any.inc)+edat.inc)+0.111
+
+     corr.age<-function(x.age){
+     if (x.age>85) x.age<-85
+     x.age
+     }
+
+   edat.us <- as.numeric(lapply(edat.us, corr.age))
+
+   intervalo.edat.us <- cut2(edat.us, seq(0, 90, 5))
+   set.c <- cbind(set, edat.us, intervalo.edat.us)
+
+   C <- stat.table(list(Edad=intervalo.edat.us),list(C=count()),
+                   margins=TRUE, data=set.c)
+   Y<-pob[pob$year==a,]
+   Y<-Y$population
```

```
+   Y[19] <- sum(Y)
+   result <- as.data.frame(t(rbind(C, Y)))
+   result$Px100000 <- round(result$C/result$Y*100000,3)
+   result
+ }
> # Ej. 1: Prevalencia puntual a 1 año, son incidentes de
> # 2007 que prevalencen al final del año
>
> AC07.1a <- prevalencia(datos=fit.surv, año=2007, int1=1,
+                        int2=1, pob=fit.pop.F)
> head(AC07.1a)

        C     Y Px100000
 0      0 38869    0.000
 5      0 34918    0.000
10      0 33232    0.000
15      0 34825    0.000
[20,25) 1 44087    2.268
[25,30) 3 59678    5.027

> # Ej. 2: Prevalencia acumulada en 5 años, pacientes de
> # 2007 que llevan hasta 5 años de diagnóstico
>
> AC07.5a <- prevalencia(datos=fit.surv, año=2007, int1=1,
+                        int2=5, pob=fit.pop.F)
> head(AC07.5a)

        C     Y Px100000
 0      0 38869    0.000
 5      0 34918    0.000
10      0 33232    0.000
15      0 34825    0.000
[20,25) 3 44087    6.805
[25,30) 6 59678   10.054

> # Ej. 3: Prevalencia puntual 2-3 años, casos prevalentes
> # en 2006 que llevan entre 2 y 3 años de diagnosticados
>
> Int06.23a <- prevalencia(datos=fit.surv, año=2006, int1=2,
+                          int2=3, pob=fit.pop.F)
> head(Int06.23a)

        C     Y Px100000
 0      0 36608    0.000
 5      0 33232    0.000
10      0 32429    0.000
15      0 33976    0.000
```

```
[20,25) 1 43731    2.287
[25,30) 3 59238    5.064

>
> #-----------------------------------------------------------
```

A.2.1. Funciones para el "Método Tiempo de Seguimiento Empírico": *t.seg.int*, *simulate.fu*, *tiempo.seg.sim*, *base.sim.tse* y *prev.AC.V.sim.tse*

t.seg.int extrae desde una base de datos los tiempos de seguimiento observados en cada grupo de edad definido mediante un vector de corte.

Argumentos:

- `data`: base de registros de cáncer para un año dado de donde se extraen los tiempos de seguimiento observados en cada uno de los grupos de edad definidos por `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `cut.age`: un vector de corte para definir los límites de los grupos de edad.

Salida: lista por grupo de edad con tiempos de seguimiento observado.

```
> t.seg.int <- function(data, cut.age){
+   a <- cut.age
+   b <- length(a)+1
+   int.edad<- cut2(data$edat.inc, a)
+
+   t.seg <- list()
+   for(i in 1:b){
+     d <- which(as.numeric(int.edad)==i)
+     t.seg[[i=levels(int.edad)[i]]] <- data$follow.up[d]
+   }
+   t.seg
+ }

> # Ej: tiempos de seg. observados en 1999 en 6 grupos de edad
>
> t.seg.obs.99 <- t.seg.int(datos.II[datos.II$any.inc==1999,],
+                           c(35,45,55,65,75))
> t.seg.obs.99
```

```
$`[27,35)`
[1] 11.00 11.42 11.17 11.08 10.75 10.67 11.25

$`[35,45)`
 [1] 11.08 11.08  5.92 11.00 10.75 11.33 10.92 11.08  0.08 10.67 11.17 11.08
[13]  2.17  3.58

$`[45,55)`
 [1] 10.92 10.92 10.92 10.83  3.17  5.42  0.50  5.42 11.17 10.83

$`[55,65)`
 [1] 10.67  0.08 11.33  7.92  0.08 10.58 10.67  0.08  1.92  0.42 10.83

$`[65,75)`
 [1]  0.58  1.33  0.25  3.58 11.00 10.75 10.50  6.58  2.75  0.08  0.42  0.67

$`[75,87]`
[1] 11.08  0.33  1.58  0.67  0.08  0.08

>
> #------------------------------------------------------------
```

simulate.fu genera un vector de tiempos de seguimientos simulados de longitud especificada (número de simulaciones) tras computar la función de distribución acumulada empírica (con la función `ecdf` del paquete `stats`) sobre un vector de tiempos de seguimiento dado.

Argumentos:

`T.tmp`: un vector de tiempos observados.
`n.sim`: número de tiempos simulados.

Salida: un vector con `n.sim` tiempos simulados.

```
> simulate.fu<-function(T.tmp,n.sim=100)
+ {
+   ecdf.T <- ecdf(T.tmp)
+   # now simulate 100 numbers from this ECDF...
+   as.numeric(quantile(ecdf.T, runif(n.sim)))
+ }
> # Ej.: 7 tiempos de seguimiento extraídos de distribución
> # acumulada empírica para mujeres menores de 35 años
> # diagnosticadas en 1999.
>
> sim.T.0.35<-simulate.fu(t.seg.obs.99[[1]],n.sim=7)
> sim.T.0.35

[1] 10.82616 10.68638 11.20034 11.20820 11.16941 10.71411 11.41152
```

```
>
> #-----------------------------------------------------------
```

tiempo.seg.sim permite generar tiempos de seguimiento simulados a cada caso incidente de un *data frame* según su grupo de edad (definido en un vector de corte) a partir de las distribuciones de tiempo de seguimiento observadas del mismo grupo de edad en una base de datos de registros especificada (de referencia) de un año particular.

Argumentos:

`data`:	es la base de registros de un año particular usada para extraer las distribuciones acumuladas empíricas del tiempo de seguimiento en cada intervalo de edad definido por `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	es el vector de corte que define los grupos de edad.
`casos`:	es un *data frame* con el número de casos ('C') de un año particular por intervalo de edad quinquenal ('age.group', de 1 a 18) para los cuales se desea simular tiempos de seguimiento.

Salida: tiempos de seguimiento simulados para cada caso según su grupo de edad.

```
> tiempo.seg.sim <- function(data, cut.age, casos){
+
+   a <- cut.age
+   casos$Age2 <- seq(0, 85, 5)
+   casos$Age2<- cut2(casos$Age2, a)
+   casos.proy <- stat.table(list(Age2),list(C=sum(C)),margins=TRUE,
+                            data=casos)
+
+   Int.Edad <- rep(levels(casos$Age2),
+                   casos.proy[,1:length(levels(casos$Age2))])
+   t.sim <- rep(NA, casos.proy[,'Total'])
+   salida <- data.frame(Int.Edad, t.sim)
+
+   t.emp <- t.seg.int(data, a)
+
+   for (i in 1:casos.proy[,'Total']) {
+     j<- as.numeric(salida[i,1])
+     salida[i,2] <- simulate.fu(t.emp[[j]], n.sim=1 )
+   }
+   salida
+ }
```

```
> # Ej.: se simulan los tiempos de seguimiento de cada caso
> # incidente en 'cas (incidentes de 1999 en Girona y Tarragona);
> # 'dat' son los registros de Girona y Tarragona de 1999 de donde
> # se extraen las distrib. emp. por grupo de edad definido por
> # el vector de corte c(35,45,55,65,75)
>
> t.seg.GT.1999 <- tiempo.seg.sim(dat, c(35,45,55,65,75), cas)
> head(t.seg.GT.1999,10)

   Int.Edad    t.sim
1   [ 0,35) 11.33271
2   [ 0,35) 11.36194
3   [ 0,35) 11.01060
4   [ 0,35) 10.74924
5   [ 0,35) 11.23508
6   [ 0,35) 11.13831
7   [ 0,35) 10.98742
8   [35,45)  3.37706
9   [35,45) 10.74262
10  [35,45)  2.39292

>
> #-------------------------------------------------------------
```

base.sim.tse genera una base de datos simulada a partir de la base de referencia con los registros de cáncer (ej. mujeres diagnosticadas por cáncer en Girona y Tarragona entre 1999 y 2007, seguidas hasta junio de 2010), un vector de corte que define los grupos de edad para los cuales se obtienen las distribuciones acumuladas empíricas de tiempo de seguimiento, y un *data frame* con los casos a simular por año e intervalo de edad. Esta función genera un "registro" para cada caso incidente a simular con edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en 1 de julio del año de incidencia, el tiempo de seguimiento generado desde las distribuciones observadas en el grupo de edad y año correspondiente, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de referencia (al 30 de junio de 2010).

Argumentos:

`data.orig`: es la base de registros desde donde se extraen las distrib. empíricas (ej: Girona y Tarragona) por año y grupo de edad definido en `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').

`cut.age`: vector de corte para la generación de grupos de edad.

`casos.sim`: un *data frame* con el número de casos incidentes a simular ('C') por año de incidencia ('year') e intervalo de edad ('age.groupo', de 1 a 18).

Salida: una base de registros de cáncer simulada.

```
> base.sim.tse <- function(data.orig, cut.age, casos.sim){
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base <- NULL
+   for(i in t0:tf) {
+     casos <- subset(casos.sim, year==i)
+     dist <- subset(data.orig, any.inc==i)
+
+     follow.up <- tiempo.seg.sim(dist, cut.age, casos)[,'t.sim']
+     edat.inc <- rep(seq(2.5, 87.5, 5), casos$C)
+     mes.inc <- rep_len(7,sum(casos$C))
+     any.us0 <- i+follow.up
+     any.us <- floor(any.us0)
+     mes.us0 <- ((any.us0-any.us)*12)+6
+
+     cambia.año <- which(mes.us0>12)
+     any.us[cambia.año] <- any.us[cambia.año]+1
+     mes.us0[cambia.año] <- mes.us0[cambia.año]-12
+
+     mes.us <- ceiling(mes.us0)
+
+     status <- ifelse((any.us>max(data.orig$any.us) |
+                         (any.us==max(data.orig$any.us)&mes.us>6)),
+                      0,1)
+     id.pacient <- c(1:sum(casos$C))+i*0.0001
+     sexe <- rep_len('X',sum(casos$C))
+     loc <-  rep_len('X',sum(casos$C))
+
+     any.inc <- rep_len(i,sum(casos$C))
+
+     base[[i]] <- data.frame(id.pacient,sexe,loc,edat.inc,mes.inc,
+                             any.inc,mes.us,any.us,status,follow.up)
```

```
+   }
+   base <- do.call(rbind, base)
+   base
+ }
>
> # Ej.: Una base simulada para los casos que incidieron en Girona
> # y Tarragona entre 1999 y 2007 (in.GT), usando distribuciones
> # de tiempo de seguimiento observadas en 6 grupos de edad (con
> # corte en 35,45,55,65,75) en la base de registros de Girona
> # y Tarragona (datos.II)

> base.sim.GT <- base.sim.tse(datos.II, c(35,45,55,65,75), in.GT)
> head(base.sim.GT)

  id.pacient sexe loc edat.inc mes.inc any.inc mes.us any.us status follow.up
1     1.1999    X   X     27.5       7    1999      8   2010      0  11.08681
2     2.1999    X   X     27.5       7    1999     10   2010      0  11.32898
3     3.1999    X   X     32.5       7    1999      7   2010      0  11.03170
4     4.1999    X   X     32.5       7    1999      9   2010      0  11.19091
5     5.1999    X   X     32.5       7    1999     10   2010      0  11.31576
6     6.1999    X   X     32.5       7    1999      4   2010      1  10.77582

>
> #---------------------------------------------------------------
```

prev.AC.V.sim.tse permite estimar prevalencia a 5 años desde *nsim* cohortes simuladas. Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`:	base de referencia con los registros de cáncer que se usa para extraer las distribuciones de seguimiento empírico por año y grupo de edad de las pacientes.El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us) y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	vector de corte que define los grupos de edad.
`casos.sim`:	un *data frame* con los 'C' casos incidentes a simular (ej: incidentes de Cataluña), por intervalo de edad quinquenal ('age.group', de 1 a 18), año de incidencia ('year'), y población total ('population').
`t.AC`:	año para la estimación de prevalencia a 5 años.
`pob.orig`:	distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia.
`nsim`:	número de bases (cohortes) que se desean simular para estimar prevalencia a 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población total ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.tse <- function(data.orig, cut.age, casos.sim,
+                               t.AC, pob.orig, nsim=100){
+     conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- base.sim.tse(data.orig, cut.age, casos.sim)
+     c <- prevalencia(datos=base, año=t.AC, int1=1, int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
```

```
+
+   salida <- cbind(emp[,1:3], cuantiles, c$Y,
+                   round(cuantiles/c$Y*100000,2), round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf',
+                         'Csup','Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: se simulan casos incidentes de Girona y Tarragona (in.GT)
> # 100 veces y se obtiene prevalencia a 5 años 2003 con intervalos
> # de confianza 95%. Los datos de referencia provienen de los
> # registros de cáncer de las mismas provincias (datos.II), se
> # definen 6 grupos de edad (35,45,55,65,75), y distrib de mujeres
> # en población de referencia (fit.pop.F).
>
> AC.5a.tse.GT.2003 <-prev.AC.V.sim.tse(datos.II,c(35,45,55,65,75),
+                                       in.GT,2003,fit.pop.F,nsim=100)
> AC.5a.tse.GT.2003

        C.emp  Y.emp  P.emp  Cm Cinf Csup      Y     P  Pinf  Psup   P_sd
 0          0  30018  0.000   0    0    0  30018  0.00  0.00  0.00  0.000
 5          0  29056  0.000   0    0    0  29056  0.00  0.00  0.00  0.000
10          0  30245  0.000   0    0    0  30245  0.00  0.00  0.00  0.000
15          0  32911  0.000   0    0    0  32911  0.00  0.00  0.00  0.000
[20,25)     1  44603  2.242   1    1    1  44603  2.24  2.24  2.24  0.000
[25,30)     2  52610  3.802   4    4    4  52610  7.60  7.60  7.60  0.000
[30,35)    16  50174 31.889  14   14   14  50174 27.90 27.90 27.90  0.000
[35,40)    41  50147 81.760  39   36   40  50147 77.77 71.79 79.77  2.361
[40,45)    41  47388 86.520  35   32   38  47388 73.86 67.53 80.19  3.900
[45,50)    28  42328 66.150  37   34   39  42328 87.41 80.33 92.14  3.285
[50,55)    28  38451 72.820  27   22   29  38451 70.22 57.22 75.42  4.836
[55,60)    17  34684 49.014  17   14   20  34684 49.01 40.36 57.66  5.322
[60,65)    14  26773 52.291  15   13   17  26773 56.03 48.56 63.50  4.721
[65,70)    18  31706 56.772  18   13   20  31706 56.77 41.00 63.08  6.209
[70,75)    17  30839 55.125  18   13   21  30839 58.37 42.15 68.10  7.177
[75,80)     7  26080 26.840  12    8   16  26080 46.01 30.67 61.35  7.682
[80,85)     9  19130 47.047   8    4   12  19130 41.82 20.91 62.73 10.332
[85,90]     0  16193  0.000   2    0    5  16193 12.35  0.00 30.88  7.706
Total     239 633336 37.737 246  236  257 633336 38.84 37.26 40.58  0.802
```

A.2.2. Funciones para el "Método Mirror": *dat.sim.mirr* y *prev.AC.V.sim.mirr*

dat.sim.mirr permite simular una base de registros de cáncer (cohorte de mujeres diagnosticadas por cáncer) por remuestreo con reposición (por año e intervalo de edad quinquenal) desde los registros de una base de referencia.

Argumentos:

- `data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007). El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').

Salida: una base de registros simulados por remustreo en base de referencia.

```
> dat.sim.mirr <- function(data.orig, casos.sim){
+
+   data.orig$age.group <- as.numeric(cut2(data.orig$edat.inc,
+                                          seq(0,85,5)))
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base.sim <- NULL
+   for(i in t0:tf) {
+     data.o <- subset(data.orig, any.inc==i)
+     casos.s <- subset(casos.sim, year==i)
+
+     m <- NULL
+     for (j in 1:18){s <- which(data.o$age.group==j)
+                      if(length(s)==1){m[[j]] <- s}
+                      else {m[[j]] <- sample(s, casos.s[j,'C'],
+                                             replace=TRUE)}
+                      m[[j]]
+                     }
+     vector <- unlist(m)
+     base.sim[[i]] <-data.o[vector,]
+   }
+   base.sim <- do.call(rbind, base.sim)
+   base.sim[,1:10]
+ }

> # Ej.: base simulada de casos incidentes (in.GT) en Girona y Tarragona
> # 1999-2007, por remuestreo en la misma base de referencia (datos.II)
> # por año e interv. de edad.
>
> base.mirr <- dat.sim.mirr(datos.II, in.GT)
> head(base.mirr)
```

```
     id.pacient sexe       loc edat.inc mes.inc any.inc mes.us any.us status
29        67400    2 Coll uter       29       4    1999      6   2010      0
29.1      67400    2 Coll uter       29       4    1999      6   2010      0
377       67934    2 Coll uter       33      10    1999      6   2010      0
21        67392    2 Coll uter       31       6    1999      6   2010      0
21.1      67392    2 Coll uter       31       6    1999      6   2010      0
24        67395    2 Coll uter       31       1    1999      6   2010      0
     follow.up
29       11.17
29.1     11.17
377      10.67
21       11.00
21.1     11.00
24       11.42

>
> #------------------------------------------------------------------------------
```

prev.AC.V.sim.mirr permite estimar prevalencia a 5 años desde `nsim` bases simuladas por metodo "Mirror" (remuestreo por año e intervalo de edad quinquenal sobre una base de referencia). Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007).El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').
`t.AC`:	año para la estimación de prevalencia a 5 años.
`pob.orig`:	distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia.
`nsim`:	número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y

97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.mirr <- function(data.orig, casos.sim, t.AC,
+                                pob.orig, nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- dat.sim.mirr(data.orig,casos.sim)
+     c <- prevalencia(datos=base, año=t.AC, int1=1, int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf',
+                         'Csup','Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: se simulan 100 veces los casos incidentes de Girona y
> # Tarragona (in.GT) por remuestreo según año e intervalo de edad
> # en la base de referencia y se estima prevalencia a 5 años
> # 2003 con intervalos de confianza 95%. Los datos de referencia
> # provienen de los registros de cáncer de las mismas provincias
> # (datos.II). fit.pop.F contiene el total riesgo por año
> # e intervalo de edad de la población de referencia.
>
> AC.5a.mirr.GT.2003 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2003,
+                                          fit.pop.F, nsim=100)
> AC.5a.mirr.GT.2003
```

	C.emp	Y.emp	P.emp	Cm	Cinf	Csup	Y	P	Pinf	Psup	P_sd
0	0	30018	0.000	0	0	0	30018	0.00	0.00	0.00	0.000
5	0	29056	0.000	0	0	0	29056	0.00	0.00	0.00	0.000
10	0	30245	0.000	0	0	0	30245	0.00	0.00	0.00	0.000

```
15          0  32911  0.000   0    0    0  32911  0.00  0.00   0.00  0.000
[20,25)     1  44603  2.242   1    1    1  44603  2.24  2.24   2.24  0.000
[25,30)     2  52610  3.802   2    1    4  52610  3.80  1.90   7.60  1.641
[30,35)    16  50174 31.889  16   13   19  50174 31.89 25.91  37.87  3.241
[35,40)    41  50147 81.760  41   36   47  50147 81.76 71.79  93.72  5.571
[40,45)    41  47388 86.520  41   34   49  47388 86.52 71.75 103.40  8.633
[45,50)    28  42328 66.150  28   20   35  42328 66.15 47.25  82.69  8.580
[50,55)    28  38451 72.820  28   21   34  38451 72.82 54.61  88.42  9.543
[55,60)    17  34684 49.014  18   12   23  34684 51.90 34.60  66.31  8.944
[60,65)    14  26773 52.291  14   10   18  26773 52.29 37.35  67.23  7.849
[65,70)    18  31706 56.772  18   14   23  31706 56.77 44.16  72.54  7.823
[70,75)    17  30839 55.125  17   12   22  30839 55.13 38.91  71.34  8.266
[75,80)     7  26080 26.840   7    2   11  26080 26.84  7.67  42.18  8.278
[80,85)     9  19130 47.047   9    4   13  19130 47.05 20.91  67.96 12.020
[85,90]     0  16193  0.000   0    0    0  16193  0.00  0.00   0.00  0.000
Total     239 633336 37.737 239  226  248 633336 37.74 35.68  39.16  0.895
```

A.2.3. Funciones para el "Método Mirror Exponencial": *dat.sim.mirrexp* y *prev.AC.V.sim.mirrexp*

dat.sim.mirrexp permite simular a un conjunto de casos incidentes, una base de registros de cáncer (cohorte de mujeres diagnosticadas por cáncer) por remuestreo con reposición (por año e intervalo de edad quinquenal) desde los registros de la base de referencia, y tiempo de seguimiento reasignado desde una exponencial con tasa igual al inverso del tiempo de seguimiento medio observado en el año y el intervalo de edad quinquenal correspondiente.

Argumentos:

- `data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007).El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').

Salida: una base de registros simulados por remustreo en base de referencia y reasignación del tiempo de seguimiento mediante exponenciales con parámetros estimados para el año y la edad correspondiente sobre los tiempos de seguimiento observados en la base de referencia.

```
> dat.sim.mirrexp <- function(data.orig, casos.sim){
+
+   data.orig$age.group <- as.numeric(cut2(data.orig$edat.inc,
+                                          seq(0, 85, 5)))
```

```
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base.sim <- NULL
+   for(i in t0:tf) {
+     data.o <- subset(data.orig, any.inc==i)
+     casos.s <- subset(casos.sim, year==i)
+
+     tsm.o <- stat.table(list(age.group),
+                         list(m=mean(follow.up)), margins=FALSE,
+                         data=data.o)
+     m <- NULL
+     for (j in 1:18){
+       s <- which(data.o$age.group==j)
+       if(length(s)==1){m[[j]] <- s}
+       else {m[[j]] <- sample(s, casos.s[j,'C'], replace = TRUE)}
+       m[[j]]
+     }
+     vector <- unlist(m)
+     base.sim[[i]] <-data.o[vector,]
+
+     c.sim.ag <- stat.table(list(age.group), list(C=count()),
+                            margins=FALSE, data=base.sim[[i]])
+     vec_tsm <- rep(tsm.o['m',],c.sim.ag['C',])
+
+     base.sim[[i]]$follow.up <- round(rexp(sum(c.sim.ag['C',]),
+                                           1/vec_tsm),3)
+     base.sim[[i]]$any.us <- floor(base.sim[[i]]$any.inc+base.sim[[i]]$follow.up)
+     base.sim[[i]]$mes.us <- (base.sim[[i]]$any.inc+base.sim[[i]]$follow.up-
+                                base.sim[[i]]$any.us)*12+base.sim[[i]]$mes.inc
+
+     cambia.año <- which(base.sim[[i]]$mes.us>12)
+     base.sim[[i]]$any.us[cambia.año] <- base.sim[[i]]$any.us[cambia.año]+1
+     base.sim[[i]]$mes.us[cambia.año] <- base.sim[[i]]$mes.us[cambia.año]-12
+     base.sim[[i]]$mes.us <- ceiling(base.sim[[i]]$mes.us)
+     base.sim[[i]]$status <- ifelse(base.sim[[i]]$any.us>max(data.orig$any.us)|
+                                      (base.sim[[i]]$any.us==max(data.orig$any.us)&
+                                         base.sim[[i]]$mes.us>6),0,1)
+   }
+   base.sim <- do.call(rbind, base.sim)
+   base.sim[,1:10]
+ }
> # Ej.: base simulada de casos incidentes en Girona y Tarragona 1999-2007,
> # por remuestreo en la misma base de referencia por año e intervalo de edad, y
> # tiempo de seguimiento reasignado desde exponencial según año y edad.
>
> base.mirrexp <- dat.sim.mirrexp(datos.II, in.GT)
> head(base.mirrexp)

      id.pacient sexe       loc edat.inc mes.inc any.inc mes.us any.us status
```

```
29         67400    2 Coll uter       29       4    1999       8    2001       1
29.1       67400    2 Coll uter       29       4    1999       2    2010       1
372        67929    2 Coll uter       34       9    1999       8    2005       1
372.1      67929    2 Coll uter       34       9    1999       3    2011       0
24         67395    2 Coll uter       31       1    1999      10    1999       1
366        67923    2 Coll uter       34       5    1999       9    2004       1
      follow.up
29        2.311
29.1     10.812
372       5.907
372.1    11.469
24        0.728
366       5.332

>
> #---------------------------------------------------------------------------
```

prev.AC.V.sim.mirrexp para un conjunto de casos incidentes a simular por año e intervalo de edad, se estima prevalencia a 5 años desde `nsim` bases simuladas por metodo "Mirror Exponencial" (remuestreo en base de referencia por año e intervalo de edad quinquenal, y posterior reasignación del tiempo de seguimiento por exponencial correspondiente al año y el intervalo de edad de cada caso). Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007).El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').
`t.AC`:	año para la estimación de prevalencia a 5 años.
`pob.orig`:	distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia.
`nsim`:	número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp')

y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.mirrexp <- function(data.orig, casos.sim, t.AC,
+                                   pob.orig, nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- dat.sim.mirrexp(data.orig,casos.sim)
+     c <- prevalencia(datos=base, año=t.AC, int1=1, int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf',
+                         'Csup','Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: se simulan casos incidentes de Girona y Tarragona (in.GT)
> # por método mirror-exponencial 100 veces y se estima prevalencia
> # a 5 años 2003 con intervalos de confianza 95%. Los datos de
> # referencia provienen de los registros de cáncer de las mismas
> # provincias (datos.II). fit.pop.F es el total de mujeres por
> # año e intervalo de edad de la población de referencia.
>
> AC.5a.mirrexp.GT.2003 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2003,
+                                                fit.pop.F, nsim=100)
> AC.5a.mirrexp.GT.2003

       C.emp  Y.emp  P.emp  Cm Cinf Csup      Y     P  Pinf  Psup  P_sd
 0         0  30018  0.000   0    0    0  30018  0.00  0.00  0.00 0.000
 5         0  29056  0.000   0    0    0  29056  0.00  0.00  0.00 0.000
```

```
10          0  30245  0.000   0    0    0  30245  0.00  0.00  0.00 0.000
15          0  32911  0.000   0    0    0  32911  0.00  0.00  0.00 0.000
[20,25)     1  44603  2.242   1    0    1  44603  2.24  0.00  2.24 0.491
[25,30)     2  52610  3.802   2    1    3  52610  3.80  1.90  5.70 1.369
[30,35)    16  50174 31.889  12    9   17  50174 23.92 17.94 33.88 3.971
[35,40)    41  50147 81.760  33   27   39  50147 65.81 53.84 77.77 6.049
[40,45)    41  47388 86.520  34   27   40  47388 71.75 56.98 84.41 7.542
[45,50)    28  42328 66.150  24   17   30  42328 56.70 40.16 70.88 7.718
[50,55)    28  38451 72.820  23   17   29  38451 59.82 44.21 75.42 8.133
[55,60)    17  34684 49.014  15    9   22  34684 43.25 25.95 63.43 8.501
[60,65)    14  26773 52.291  12    7   17  26773 44.82 26.15 63.50 9.382
[65,70)    18  31706 56.772  16   12   21  31706 50.46 37.85 66.23 7.680
[70,75)    17  30839 55.125  15   10   20  30839 48.64 32.43 64.85 8.168
[75,80)     7  26080 26.840  10    4   15  26080 38.34 15.34 57.52 9.381
[80,85)     9  19130 47.047  11    8   15  19130 57.50 41.82 78.41 9.986
[85,90]     0  16193  0.000   0    0    0  16193  0.00  0.00  0.00 0.618
Total     239 633336 37.737 208  195  228 633336 32.84 30.79 36.00 1.270
```

A.2.4. Funciones para el "Método Mirror Uniforme": *dat.sim.mirrunif* y *prev.AC.V.sim.mirrunif*

dat.sim.mirrunif permite simular una base (cohorte) para un conjnto de casos incidentes, a partir del remuestreo en la base de registros de referencia por año e intervalo de edad y reasignación del tiempo de seguimiento mediante distrib. uniforme con mínimo y máximo+1 tomados del tiempo observado en la base de referencia en el año y el intervalo de edad quinquenal correspondiente.

Argumentos:

- `data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007). El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us) y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').

Salida: una base de registros simulados por remustreo en base de referencia y resaignación del tiempo de seguimiento desde uniformes con parámetros tomados desde los seguimientos observados por año e intervalo de edad quinquenal en la población de referencia.

```
> dat.sim.mirrunif <- function(data.orig, casos.sim){
+
+   data.orig$age.group <- as.numeric(cut2(data.orig$edat.inc,
+                                          seq(0, 85, 5)))
```

```
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base.sim <- NULL
+   for(i in t0:tf) {
+     data.o <- subset(data.orig, any.inc==i)
+     casos.s <- subset(casos.sim, year==i)
+
+     t.seg.o <- stat.table(list(age.group),
+                           list(Max=max(follow.up),min=min(follow.up)),
+                           margins=FALSE, data=data.o)
+
+     m <- NULL
+     for (j in 1:18){
+       s <- which(data.o$age.group==j)
+       if(length(s)==1){m[[j]] <- s}
+       else {m[[j]] <- sample(s, casos.s[j,'C'], replace = TRUE)}
+       m[[j]]
+     }
+     vector <- unlist(m)
+     base.sim[[i]] <-data.o[vector,]
+
+     c.sim.ag <- stat.table(list(age.group), list(C=count()),
+                            margins=FALSE, data=base.sim[[i]])
+     vec_tsMax <- rep(t.seg.o['Max',],c.sim.ag['C',])
+     vec_tsmin <- rep(t.seg.o['min',],c.sim.ag['C',])
+
+     base.sim[[i]]$follow.up <- round(runif(sum(c.sim.ag['C',]),
+                                            vec_tsmin,vec_tsMax+1),3)
+     base.sim[[i]]$any.us<-floor(base.sim[[i]]$any.inc+base.sim[[i]]$follow.up)
+     base.sim[[i]]$mes.us<-(base.sim[[i]]$any.inc+base.sim[[i]]$follow.up-
+                             base.sim[[i]]$any.us)*12+base.sim[[i]]$mes.inc
+
+     cambia.año <- which(base.sim[[i]]$mes.us>12)
+     base.sim[[i]]$any.us[cambia.año]<-base.sim[[i]]$any.us[cambia.año]+1
+     base.sim[[i]]$mes.us[cambia.año]<-base.sim[[i]]$mes.us[cambia.año]-12
+     base.sim[[i]]$mes.us <- ceiling(base.sim[[i]]$mes.us)
+     base.sim[[i]]$status <- ifelse(base.sim[[i]]$any.us>max(data.orig$any.us)|
+                                      (base.sim[[i]]$any.us==max(data.orig$any.us)&
+                                         base.sim[[i]]$mes.us>6),0,1)
+   }
+   base.sim <- do.call(rbind, base.sim)
+   base.sim[,1:10]
+ }
> # Ej.: se simula una base (cohorte) para todos los casos incidentes de G+T
> # entre 1999 y 2007 (in.GT) a partir de remuestreo y reasignación del
> # tiempo de seguimiento (Unif) desde la base de referencia (datos.II)
>
> base.mirrunif <- dat.sim.mirrunif(datos.II, in.GT)
> head(base.mirrunif)
```

```
      id.pacient sexe       loc edat.inc mes.inc any.inc mes.us any.us status
401        67979    2 Coll uter       27       3    1999      6   2011      0
401.1      67979    2 Coll uter       27       3    1999      5   2011      0
372        67929    2 Coll uter       34       9    1999      2   2011      0
24         67395    2 Coll uter       31       1    1999     11   2009      1
366        67923    2 Coll uter       34       5    1999      5   2010      1
21         67392    2 Coll uter       31       6    1999     11   2011      0
      follow.up
401      12.176
401.1    12.117
372      11.361
24       10.775
366      10.925
21       12.336

>
> #------------------------------------------------------------------------
```

prev.AC.V.sim.mirrunif permite estimar prevalencia a 5 años para un año dado, a partir de `nsim` simulaciones de bases (cohortes) generadas para un conjunto de casos incidentes por método "Mirror-Uniforme". Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Arguentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007). El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').
`t.AC`:	año para la estimación de prevalencia a 5 años.
`pob.orig`:	distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia.
`nsim`:	número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y

97.5 %), luego mujeres en población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.mirrunif <- function(data.orig, casos.sim, t.AC,
+                                    pob.orig, nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- dat.sim.mirrunif(data.orig,casos.sim)
+     c <- prevalencia(datos=base, año=t.AC, int1=1,
+                      int2=5, pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2), round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf',
+                         'Csup','Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: Prevalencia a 5 años Girona y Tarragona en 2003 estimada
> # con 100 bases simuladas por método Mirror-uniforme para los
> # casos incidentes entre 1999-2007 (in.GT), tomando como base
> # de referencia los registros de las mismas provincias (datos.II).
> # fit.pop.F contiene el total de mujeres por año e intervalo de
> # edad de la pob. de referencia.
>
> AC.5a.mirrunif.GT.2003 <- prev.AC.V.sim.mirrunif(datos.II,in.GT,
+                                                  2003,fit.pop.F,nsim=100)
> AC.5a.mirrunif.GT.2003
```

	C.emp	Y.emp	P.emp	Cm	Cinf	Csup	Y	P	Pinf	Psup	P_sd
0	0	30018	0.000	0	0	0	30018	0.00	0.00	0.00	0.000
5	0	29056	0.000	0	0	0	29056	0.00	0.00	0.00	0.000
10	0	30245	0.000	0	0	0	30245	0.00	0.00	0.00	0.000
15	0	32911	0.000	0	0	0	32911	0.00	0.00	0.00	0.000

[20,25)	1	44603	2.242	1	1	1	44603	2.24	2.24	2.24	0.000
[25,30)	2	52610	3.802	2	1	3	52610	3.80	1.90	5.70	1.395
[30,35)	16	50174	31.889	16	13	20	50174	31.89	25.91	39.86	3.560
[35,40)	41	50147	81.760	41	34	46	50147	81.76	67.80	91.73	5.586
[40,45)	41	47388	86.520	42	36	50	47388	88.63	75.97	105.51	7.722
[45,50)	28	42328	66.150	28	24	35	42328	66.15	56.70	82.69	6.847
[50,55)	28	38451	72.820	29	24	35	38451	75.42	62.42	91.02	8.082
[55,60)	17	34684	49.014	18	13	24	34684	51.90	37.48	69.20	7.583
[60,65)	14	26773	52.291	15	10	20	26773	56.03	37.35	74.70	8.899
[65,70)	18	31706	56.772	19	15	24	31706	59.93	47.31	75.70	7.004
[70,75)	17	30839	55.125	18	13	22	30839	58.37	42.15	71.34	7.763
[75,80)	7	26080	26.840	13	8	19	26080	49.85	30.67	72.85	10.220
[80,85)	9	19130	47.047	15	11	19	19130	78.41	57.50	99.32	10.614
[85,90]	0	16193	0.000	0	0	0	16193	0.00	0.00	0.00	0.000
Total	239	633336	37.737	258	249	267	633336	40.74	39.32	42.16	0.771

A.2.5. Funciones para el "Método Supervivencia KM (o NA)": *base.sim.tsup* y *prev.AC.V.sim.tsup*

base.sim.tsup permite generar una base de datos simulada (una cohorte de mujeres) para un conjunto de casos incidentes por año e intervalo de edad quinquenal en un período (ej: 1999-2007) tomando las supervivencias empíricas Kaplan-Meier (o Nelson-Aalen) ajustadas por grupos de edad (especificados) sobre una base de referencia (ej. Girona y Tarragona 1999-2007). A cada caso a simular se le asigna una supervivencia desde una uniforme (0,1) y luego se busca el tiempo que tocaría a esa supervivencia en las empíricas de la base de referencia. Para aquellas supervivencias no definidas se define un t.max (muerte) igual al seguimiento máximo observado en la base de referencia más un mes. Cada caso incidente simulado tendrá edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en 1 de julio del año de incidencia, el tiempo de muerte generado, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de registros de referencia.

Argumentos:

`data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se toman las funciones de supervivencias empíricas por grupo de edad definido en *cut.age*. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').

`cut.age`: un vector de corte que define los grupos de edad para los que se ajusta supervivencia empírica

`casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').

`estim.sup`: estimador que se usará para ajustar la supervivencia empírica, Kaplan Meier ('KM' por defecto) o Nelson-Aalen ('NA').

Salida: una base de registros simulados para un conjunto de casos incidentes a partir de la supervivencia empírica de una base de referencia.

```
> base.sim.tsup <- function(data.orig, cut.age, casos.sim, estim.sup='KM'){
+
+   require("Hmisc")
+   require("survival")
+   require("Epi")
+
+   # obtengo supervivencias para cut.age
+   data.orig$int.edad.inc <- cut2(data.orig$edat.inc, cut.age)
+   sup <- with(data.orig, Surv(follow.up, status))  # elemento con la sobrev.
+
+   if (estim.sup=='KM'){
+     #Supervivencia Kaplan-Meier (distinguiendo grupos de edad)
+     svf.K_M <- summary(survfit(sup~int.edad.inc, data.orig))
+     svf <- svf.K_M # objeto con superv. emp. K-M por grupo de edad de cut.age
+   }
+
+   if (estim.sup=='NA'){
+     #Supervivencia Nelson-Aalen (distinguiendo grupos de edad)
+     svf.N_A <- summary(survfit(sup~int.edad.inc, data.orig,
+                                type="fleming-harrington"))
+     svf <- svf.N_A # objeto con superv. emp. N-A por grupo de edad de cut.age
+   }
+
+   # un data.frame con edad, t y S tomado de svf (superv. emp.)
+   svf.tmp <- data.frame(grupo.edad=as.numeric(svf$strata),
+                         tiempo=svf$time, S=svf$surv)
+
+   # tiempo de muerte que se asignará para supervi. no definidas
+   t.max <- max(data.orig$follow.up)+0.0833
+
```

```
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base <- NULL
+   for(i in t0:tf) {
+     casos <- subset(casos.sim, year==i)
+
+     a <- cut.age
+     casos$Age2 <- seq(0, 85, 5)
+     casos$Age2<- cut2(casos$Age2, a)
+     casos.proy <- stat.table(list(Age2),list(C=sum(C)),
+                              margins=TRUE,data=casos)
+
+     Int.Edad <- rep(levels(casos$Age2),
+                     casos.proy[,1:length(levels(casos$Age2))])
+     t.sim    <- rep(NA, casos.proy[,'Total'])
+
+     S.unif <- runif(casos.proy[,'Total'], min = 0, max = 1)
+
+     salida <- data.frame(Int.Edad, t.sim, S.unif)
+
+     for (l in 1:casos.proy[,'Total']) {
+
+       j<- as.numeric(salida[l,'Int.Edad'])
+
+       S.edad <- subset(svf.tmp[dim(svf.tmp)[1]:1,],grupo.edad==j) # subset edad
+       fila <- findInterval(salida[l,'S.unif'], S.edad$S)
+
+       if (fila==0) salida[l,'t.sim'] <- t.max else
+         salida[l,'t.sim'] <- S.edad[fila,'tiempo']
+     }
+     salida
+
+     follow.up <- salida[,'t.sim']  # es tiempo hasta la muerte
+
+     edat.inc <- rep(seq(2.5, 87.5, 5), casos$C)
+
+     mes.inc <- rep_len(7,sum(casos$C))
+
+     any.us0 <- i+follow.up
+     any.us <- floor(any.us0)
+     mes.us0 <- ((any.us0-any.us)*12)+6
+
+     cambia.año <- which(mes.us0>12)
+     any.us[cambia.año] <- any.us[cambia.año]+1
+     mes.us0[cambia.año] <- mes.us0[cambia.año]-12
+
+     mes.us <- ceiling(mes.us0)
+
+     status <- ifelse( (any.us>max(data.orig$any.us)|
+                          (any.us==max(data.orig$any.us) & mes.us>6)),0,1)
```

```
+
+     id.pacient <- c(1:sum(casos$C))+i*0.0001
+     sexe <- rep_len('X',sum(casos$C))
+     loc <-  rep_len('X',sum(casos$C))
+     any.inc <- rep_len(i,sum(casos$C))
+
+     base[[i]] <- data.frame(id.pacient,sexe,loc,edat.inc,mes.inc,
+                             any.inc,mes.us,any.us,status,follow.up)
+   }
+   base <- do.call(rbind, base)
+   base
+ }
> # Ejemplo: Se simula una base para todos los casos incidentes de Girona
> # y Tarragona entre 1999 y 2007 (in.GT) según supervivencia observada
> # (estimador Nelson-Aalen) en 6 grupos de edad (con corte en 35,45,55,65,75)
> # en los registros de referencia (datos.II)
>
> base.sup.GT <- base.sim.tsup(datos.II, c(35,45,55,65,75), in.GT, 'NA')
> head(base.sup.GT)

  id.pacient sexe loc edat.inc mes.inc any.inc mes.us any.us status follow.up
1     1.1999    X   X     27.5       7    1999      1   2011      0   11.5033
2     2.1999    X   X     27.5       7    1999      1   2011      0   11.5033
3     3.1999    X   X     32.5       7    1999      1   2011      0   11.5033
4     4.1999    X   X     32.5       7    1999      1   2011      0   11.5033
5     5.1999    X   X     32.5       7    1999      1   2011      0   11.5033
6     6.1999    X   X     32.5       7    1999      1   2011      0   11.5033

>
> #-------------------------------------------------------------------------
```

prev.AC.V.sim.tsup permite estimar prevalencia a 5 años en un año dado (`t.AC`) a partir de `nsim` bases (cohortes) simuladas para un conjunto de casos incidentes, por método "Supervivencia empírica". Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se toman las funciones de supervivencias empíricas por grupo de edad definido en `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us) y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	un vector de corte que define los grupos de edad para los que se ajusta supervivencia empírica
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').
`t.AC`:	año para el que se estima prevalencia a 5 años.
`pob.orig`:	distribución de población femenina ('population') por año (year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en población de referencia (ej. Girona y Tarragona).
`estim.sup`:	estimador que se usará para ajustar la supervivencia empírica, Kaplan Meier ('KM' por defecto) o Nelson-Aalen ('NA').
`nsim`:	número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.tsup <- function(data.orig, cut.age, casos.sim, t.AC,
+                                pob.orig, estim.sup='KM', nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- base.sim.tsup(data.orig, cut.age, casos.sim, estim.sup)
+     c <- prevalencia(datos=base,año=t.AC,int1=1,int2=5,pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
```

```
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
+
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf','Csup',
+                         'Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: Prevalencia a 5 años en Girona y Tarragona en 2003 estimada con
> # 100 bases simuladas por método "Supervivencia empírica NA".
>
> AC.5a.sup.GT.2003.na <-prev.AC.V.sim.tsup(datos.II,c(35,45,55,65,75),
+                                           in.GT, 2003,fit.pop.F,'NA',
+                                           nsim=100)
> AC.5a.sup.GT.2003.na
```

	C.emp	Y.emp	P.emp	Cm	Cinf	Csup	Y	P	Pinf	Psup	P_sd
0	0	30018	0.000	0	0	0	30018	0.00	0.00	0.00	0.000
5	0	29056	0.000	0	0	0	29056	0.00	0.00	0.00	0.000
10	0	30245	0.000	0	0	0	30245	0.00	0.00	0.00	0.000
15	0	32911	0.000	0	0	0	32911	0.00	0.00	0.00	0.000
[20,25)	1	44603	2.242	1	1	1	44603	2.24	2.24	2.24	0.000
[25,30)	2	52610	3.802	4	3	4	52610	7.60	5.70	7.60	0.573
[30,35)	16	50174	31.889	14	12	14	50174	27.90	23.92	27.90	1.447
[35,40)	41	50147	81.760	37	33	40	50147	73.78	65.81	79.77	3.194
[40,45)	41	47388	86.520	33	29	37	47388	69.64	61.20	78.08	4.302
[45,50)	28	42328	66.150	35	30	38	42328	82.69	70.88	89.78	4.283
[50,55)	28	38451	72.820	26	22	29	38451	67.62	57.22	75.42	4.395
[55,60)	17	34684	49.014	18	14	21	34684	51.90	40.36	60.55	5.419
[60,65)	14	26773	52.291	14	10	17	26773	52.29	37.35	63.50	6.344
[65,70)	18	31706	56.772	17	13	21	31706	53.62	41.00	66.23	6.507
[70,75)	17	30839	55.125	16	12	22	30839	51.88	38.91	71.34	8.353
[75,80)	7	26080	26.840	12	8	17	26080	46.01	30.67	65.18	9.661
[80,85)	9	19130	47.047	10	7	15	19130	52.27	36.59	78.41	11.771
[85,90]	0	16193	0.000	4	1	8	16193	24.70	6.18	49.40	10.904
Total	239	633336	37.737	240	227	252	633336	37.89	35.84	39.79	1.044

A.2.6. Funciones para el "Método Tiempo de Muerte Exponencial": *base.sim.texp* y *prev.AC.V.sim.texp*

base.sim.texp permite generar una base de datos simulada (una cohorte de mujeres) para un número de casos incidentes por año e intervalo de edad en un período (1999-2007) donde los tiempos de muerte son simulados a partir de un modelo exponencial ajustado por grupo de edad (definido con un vector de

corte) sobre los registros de una población de referencia (ej. Girona y Tarragona). Cada caso incidente simulado tendrá edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en 1 de julio del año de incidencia, el tiempo de muerte generado, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de registros de referencia.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros de las distribuciones exponenciales ajustadas para el tiempo de muerte en los grupos de edad definidos en `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	un vector de corte que define los grupos de edad para los que se ajusta cada modelo exponencial y se estima su parámetro λ
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').
`met.pars`:	es el método usado para la estimación del parámetro λ, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').

Salida: una base de registros simulados para un conjunto de casos incidentes a partir de modelos de supervivencia exponenciales ajustados por edad sobre una base de registros de cáncer de referencia.

```
> base.sim.texp <- function(data.orig, cut.age, casos.sim, met.pars='OLS'){
+
+   # obtengo supervivencias para cut.age
+   require("Hmisc")
+   require("survival")
+   require("Epi")
+
+   data.orig$int.edad.inc <- cut2(data.orig$edat.inc, cut.age)
+
+   #Supervivencia Nelson-Aalen (distinguiendo intervalos de edad)
+   svf.N_A <- summary(survfit(Surv(follow.up,status)~int.edad.inc,
+                              data.orig, type="fleming-harrington"))
+   svf.N_A # objeto con supervivencia N-A por grupo de edad de cut.age
+
+   # un data.frame con variables edad, tiempo y H(t) tomadas de svf.N_A
+   dat.tmp <- data.frame(grupo.edad=as.numeric(svf.N_A$strata),
+                         t=svf.N_A$time, RA=-log(svf.N_A$surv))
+
```

```
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base <- NULL
+   for(i in t0:tf) {
+     casos <- subset(casos.sim, year==i)
+
+     a <- cut.age
+     casos$Age2 <- seq(0, 85, 5)
+     casos$Age2<- cut2(casos$Age2, a)
+     casos.proy <- stat.table(list(Age2), list(C=sum(C)),
+                              margins=TRUE, data=casos)
+
+     Int.Edad <- rep(levels(casos$Age2),
+                     casos.proy[,1:length(levels(casos$Age2))])
+     t.exp <- rep(NA, casos.proy[,'Total'])
+     unif <- runif(casos.proy[,'Total'], min = 0, max = 1)
+     salida <- data.frame(Int.Edad, t.exp, unif)
+
+     if (met.pars=='OLS'){
+
+       for (l in 1:casos.proy[,'Total']){
+
+         j<- as.numeric(salida[l,'Int.Edad'])
+
+         dat.tmp.j <- subset(dat.tmp,grupo.edad==j) # subset edad
+
+         # parámetro: tasa de fallo lambda
+         fit.j <- summary(lm(RA~t-1, data=dat.tmp.j))
+         tasa.j <- fit.j$coeff[1] # solo pendiente
+
+         salida[l,'t.exp'] <- (-log(salida[l,'unif'])/tasa.j)
+       }
+       salida
+     }
+
+     if (met.pars=='ML') {
+
+       for (l in 1:casos.proy[,'Total']){
+         j<- as.numeric(salida[l,'Int.Edad'])
+         # subset de edad
+         dat.tmp.j <- subset(data.orig,as.numeric(int.edad.inc)==j)
+
+         # parámetro lambda o tasa = 1/b
+         fit.e.j <- survreg(Surv(follow.up, status)~1,
+                            dist="exponential",data=dat.tmp.j)
+         # lambda=1/b= 1/exp(survreg'intercept)
+         tasa.j <- 1/exp(fit.e.j$coeff)
+
+         salida[l,'t.exp'] <- (-log(salida[l,'unif'])/tasa.j)
+       }
```

```
+       salida
+     }
+
+     follow.up <- salida[,'t.exp']  # es tiempo hasta muerte
+
+     edat.inc <- rep(seq(2.5, 87.5, 5), casos$C)
+     mes.inc <- rep_len(7,sum(casos$C))
+
+     any.us0 <- i+follow.up
+     any.us <- floor(any.us0)
+     mes.us0 <- ((any.us0-any.us)*12)+6
+
+     cambia.año <- which(mes.us0>12)
+     any.us[cambia.año] <- any.us[cambia.año]+1
+     mes.us0[cambia.año] <- mes.us0[cambia.año]-12
+
+     mes.us <- ceiling(mes.us0)
+
+     status <- ifelse((any.us>max(data.orig$any.us)|
+                         (any.us==max(data.orig$any.us)& mes.us>6)),0,1)
+
+     id.pacient <- c(1:sum(casos$C))+i*0.0001
+     sexe <- rep_len('X',sum(casos$C))
+     loc <-  rep_len('X',sum(casos$C))
+     any.inc <- rep_len(i,sum(casos$C))
+
+     base[[i]] <- data.frame(id.pacient,sexe,loc,edat.inc,mes.inc,
+                             any.inc,mes.us,any.us,status,follow.up)
+   }
+   base <- do.call(rbind, base)
+   base
+ }
> # Ej.: Se simula una base para todos los casos incidentes de Girona y
> # Tarragona entre 1999 y 2007 (in.GT) según un modelo exponencial para
> # el tiempo de muerte ajustado (por máxima verosimilitud, 'ML') en
> # cada uno de los 6 grupos de edad (definidos por corte en 35,45,55,65,75)
> # sobre los registros de referencia (datos.II)
>
> base.texp.GT <- base.sim.texp(datos.II, c(35,45,55,65,75), in.GT, 'ML')
> head(base.texp.GT)

  id.pacient sexe loc edat.inc mes.inc any.inc mes.us any.us status follow.up
1     1.1999    X   X     27.5       7    1999      9   2035      0  36.24200
2     2.1999    X   X     27.5       7    1999      6   2017      0  17.98708
3     3.1999    X   X     32.5       7    1999      9   2009      1  10.20958
4     4.1999    X   X     32.5       7    1999      6   2060      0  60.92886
5     5.1999    X   X     32.5       7    1999      6   2047      0  47.99238
6     6.1999    X   X     32.5       7    1999      7   2112      0 113.05887

>
> #---------------------------------------------------------------------------
```

prev.AC.V.sim.texp permite estimar prevalencia a 5 años en un año dado (`t.AC`) a partir de `nsim` bases simuladas para un conjunto de casos incidentes, por método "Tiempo de Muerte Exponencial". Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros de las distribuciones exponenciales ajustadas para el tiempo de muerte en los grupos de edad definidos en `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	un vector de corte que define los grupos de edad para los que se ajusta cada modelo exponencial y se estima su parámetro λ
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población ('population').
`t.AC`:	año para el que se estima prevalencia a 5 años.
`pob.orig`:	distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en población de referencia (ej. Girona y Tarragona).
`met.pars`:	es el método usado para la estimación del parámetro λ, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').
`nsim`:	número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.texp <- function(data.orig, cut.age, casos.sim, t.AC,
+                                pob.orig, met.pars='OLS', nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
```

```
+     print("Sim")
+     print(sim)
+
+     base <- base.sim.texp(data.orig, cut.age, casos.sim, met.pars)
+     c <- prevalencia(datos=base, año=t.AC, int1=1, int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1,
+                      int2=5, pob=pob.orig)
+
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf','Csup',
+                         'Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: Prevalencia a 5 años Girona y Tarragona en 2003 estimada con
> # 100 bases simuladas por método "Tiempo de Muerte Exponencial" con
> # estimación de parámetros por máxima verosimilitud en 6 grupos de
> # edad (definidos con corte, 35,45,55,65,75) de base de referencia
> # (Girona y Tarragona).
>
> AC.5a.texp.GT.2003.ml<-prev.AC.V.sim.texp(datos.II,c(35,45,55,65,75),
+                                           in.GT, 2003,fit.pop.F,'ML',
+                                           nsim=100)
> AC.5a.texp.GT.2003.ml
```

	C.emp	Y.emp	P.emp	Cm	Cinf	Csup	Y	P	Pinf	Psup	P_sd
0	0	30018	0.000	0	0	0	30018	0.00	0.00	0.00	0.000
5	0	29056	0.000	0	0	0	29056	0.00	0.00	0.00	0.000
10	0	30245	0.000	0	0	0	30245	0.00	0.00	0.00	0.000
15	0	32911	0.000	0	0	0	32911	0.00	0.00	0.00	0.000
[20,25)	1	44603	2.242	1	1	1	44603	2.24	2.24	2.24	0.315
[25,30)	2	52610	3.802	4	3	4	52610	7.60	5.70	7.60	0.416
[30,35)	16	50174	31.889	14	12	14	50174	27.90	23.92	27.90	1.279
[35,40)	41	50147	81.760	38	36	40	50147	75.78	71.79	79.77	2.311
[40,45)	41	47388	86.520	35	32	38	47388	73.86	67.53	80.19	3.027
[45,50)	28	42328	66.150	37	34	39	42328	87.41	80.33	92.14	3.803
[50,55)	28	38451	72.820	27	24	30	38451	70.22	62.42	78.02	4.114
[55,60)	17	34684	49.014	20	16	22	34684	57.66	46.13	63.43	4.229

```
[60,65)    14  26773 52.291  15   12   18  26773 56.03 44.82 67.23 5.465
[65,70)    18  31706 56.772  19   15   22  31706 59.93 47.31 69.39 6.278
[70,75)    17  30839 55.125  18   15   23  30839 58.37 48.64 74.58 6.997
[75,80)     7  26080 26.840  14   10   17  26080 53.68 38.34 65.18 6.982
[80,85)     9  19130 47.047  14   10   17  19130 73.18 52.27 88.87 9.671
[85,90]     0  16193  0.000   6    3    9  16193 37.05 18.53 55.58 9.511
Total     239 633336 37.737 261  251  272 633336 41.21 39.63 42.95 0.900
```

A.2.7. Funciones para el "Método Tiempo de Muerte Weibull": *base.sim.twei* y *prev.AC.V.sim.twei*

base.sim.twei permite generar una base de datos simulada (una cohorte de mujeres) para un número de casos incidentes por año e intervalo de edad en un período (1999-2007) donde los tiempos de muerte son simulados a partir de un modelo Weibull ajustado por grupo de edad (definido con un vector de corte) sobre los registros de la población de referencia (ej. Girona y Tarragona). Cada caso incidente simulado tendrá edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en 1 de julio del año de incidencia, el tiempo de muerte generado, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de registros de referencia.

Argumentos:

- `data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros de forma y escala de las distribuciones Weibull ajustadas en cada grupo de edad definido en el vector de corte `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `cut.age`: un vector de corte que define los grupos de edad para los que se ajusta cada modelo Weibull y se estima sus parámetros ρ y λ.
- `casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').
- `met.pars`: es el método usado para la estimación de los parámetros de forma y escala, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').

Salida: una base de registros simulados para un conjunto de casos incidentes, a partir de modelos de supervivencia Weibull ajustados por edad sobre una base de registros de cáncer de referencia.

```
> base.sim.twei <- function(data.orig, cut.age, casos.sim, met.pars='OLS'){
+
```

```
+   # obtengo supervivencias para cut.age
+   require("Hmisc")
+   require("survival")
+   require("Epi")
+
+   data.orig$int.edad.inc <- cut2(data.orig$edat.inc, cut.age)
+
+   #Supervivencia Nelson-Aalen (distinguiendo intervalos de edad)
+   svf.N_A <- summary(survfit(Surv(follow.up,status)~int.edad.inc,
+                              data.orig, type="fleming-harrington"))
+   svf.N_A # objeto con supervivencia N-A por grupo de edad de cut.age
+
+   # un data.frame con variables edad, ln(t) y ln(RA)=ln(-lnS(t))
+   # tomadas de svf.N_A
+   dat.tmp <- data.frame(grupo.edad=as.numeric(svf.N_A$strata),
+                         ln_t=log(svf.N_A$time),ln_RA=log(-log(svf.N_A$surv)))
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base <- NULL
+   for(i in t0:tf) {
+     casos <- subset(casos.sim, year==i)
+
+     a <- cut.age
+     casos$Age2 <- seq(0, 85, 5)
+     casos$Age2<- cut2(casos$Age2, a)
+     casos.proy <- stat.table(list(Age2), list(C=sum(C)), margins=TRUE,
+                              data=casos)
+
+     Int.Edad <- rep(levels(casos$Age2),
+                     casos.proy[,1:length(levels(casos$Age2))])
+     t.wei    <- rep(NA, casos.proy[,'Total'])
+     unif <- runif(casos.proy[,'Total'], min = 0, max = 1)
+     salida <- data.frame(Int.Edad, t.wei, unif)
+
+     if (met.pars=='OLS'){
+       for (l in 1:casos.proy[,'Total']){
+
+         j<- as.numeric(salida[l,'Int.Edad'])
+         dat.tmp.j <- subset(dat.tmp,grupo.edad==j) # subset edad
+
+         # parámetros: forma rho y escala lambda
+         fit.j <- summary(lm(ln_RA~ln_t, data=dat.tmp.j))
+         rho.j <- fit.j$coeff[2] # pendiente
+         lambda.j <- exp(fit.j$coeff[1]) # exp(intercept)
+
+         salida[l,'t.wei'] <- (-log(salida[l,'unif'])/lambda.j)^(1/rho.j)
+       }
+       salida
+     }
```

```
+
+     if (met.pars=='ML'){
+       for (l in 1:casos.proy[,'Total']){
+         j<- as.numeric(salida[l,'Int.Edad'])
+         dat.tmp.j <- subset(data.orig,as.numeric(int.edad.inc)==j) # subset edad
+
+         # parámetros: forma rho y escala lambda
+         fit.w.j <- survreg(Surv(follow.up, status) ~ 1,
+                            dist="weibull", data=dat.tmp.j)
+         rho.j <- fit.w.j$scale^-1
+         lambda.j <- (exp(fit.w.j$coefficients))^(-rho.j)
+
+         salida[l,'t.wei'] <- (-log(salida[l,'unif'])/lambda.j)^(1/rho.j)
+       }
+       salida
+     }
+
+     follow.up <- salida[,'t.wei']  # es tiempo de superv. (hasta muerte)
+
+     edat.inc <- rep(seq(2.5, 87.5, 5), casos$C)
+     mes.inc <- rep_len(7,sum(casos$C))
+
+     any.us0 <- i+follow.up
+     any.us <- floor(any.us0)
+     mes.us0 <- ((any.us0-any.us)*12)+6
+
+     cambia.año <- which(mes.us0>12)
+     any.us[cambia.año] <- any.us[cambia.año]+1
+     mes.us0[cambia.año] <- mes.us0[cambia.año]-12
+
+     mes.us <- ceiling(mes.us0)
+
+     status <- ifelse( (any.us>max(data.orig$any.us)|
+                          (any.us==max(data.orig$any.us) & mes.us>6)),0,1)
+
+     id.pacient <- c(1:sum(casos$C))+i*0.0001
+     sexe <- rep_len('X',sum(casos$C))
+     loc <-  rep_len('X',sum(casos$C))
+     any.inc <- rep_len(i,sum(casos$C))
+
+     base[[i]] <- data.frame(id.pacient,sexe,loc,edat.inc,mes.inc,
+                             any.inc,mes.us,any.us,status,follow.up)
+   }
+   base <- do.call(rbind, base)
+   base
+ }
> # Ej.: Se simula una base para todos los casos incidentes de Girona y
> # Tarragona entre 1999 y 2007 (in.GT) según un modelo Weibull para el
> # tiempo de muerte ajustado (por máxima verosimilitud, 'ML') en cada
> # uno de los 6 grupos de edad (definidos por corte en 35,45,55,65,75)
```

```
> # sobre los registros de referencia (datos.II)
>
> base.wei.GT <- base.sim.twei(datos.II, c(35,45,55,65,75), in.GT, 'ML')
> head(base.wei.GT)

  id.pacient sexe loc edat.inc mes.inc any.inc mes.us any.us status  follow.up
1     1.1999    X   X     27.5       7    1999      2   2020      0  20.603194
2     2.1999    X   X     27.5       7    1999      7   2141      0 142.054634
3     3.1999    X   X     32.5       7    1999      5   2036      0  36.888806
4     4.1999    X   X     32.5       7    1999      6   2373      0 373.924735
5     5.1999    X   X     32.5       7    1999      2   2009      1   9.615949
6     6.1999    X   X     32.5       7    1999     11   2088      0  89.372702

>
> #---------------------------------------------------------------------
```

prev.AC.V.sim.twei permite estimar prevalencia a 5 años en un año dado (`t.AC`) a partir de `nsim` bases simuladas para un conjunto de casos incidentes, por método "Tiempo de Muerte Weibull". Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

`data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros de forma y escala de las distribuciones Weibull ajustadas en cada grupo de edad definido en el vector de corte `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').

`cut.age`: un vector de corte que define los grupos de edad para los que se ajusta cada modelo Weibull y se estima sus parámetros ρ y λ

`casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').

`t.AC`: año para el que se estima prevalencia a 5 años.

`pob.orig`: distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia (ej. Girona y Tarragona).

`met.pars`: es el método usado para la estimación de los parámetros de forma y escala, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').

`nsim`: número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.twei <- function(data.orig,cut.age,casos.sim,t.AC,
+                                pob.orig,met.pars='OLS',nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- base.sim.twei(data.orig,cut.age,casos.sim,met.pars)
+     c <- prevalencia(datos=base, año=t.AC,int1=1,int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
```

```
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig,año=t.AC,int1=1,int2=5,
+                      pob=pob.orig)
+
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf',
+                         'Csup','Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: Prevalencia a 5 años de cáncer de cérvix en Girona y
> # Tarragona en 2003 estimada con 100 bases simuladas por método
> # "Tiempo de Muerte Weibull" con estimadores máximo verosímiles
> # de los parámetros en 6 grupos de edad (definidos con corte,
> # 35,45,55,65,75) sobre la base de referencia (Girona y Tarragona).
>
> AC.5a.wei.GT.2003.ml <- prev.AC.V.sim.twei(datos.II,c(35,45,55,65,75),
+                                            in.GT,2003,fit.pop.F,'ML',
+                                            nsim=100)
> AC.5a.wei.GT.2003.ml

       C.emp  Y.emp  P.emp  Cm Cinf Csup      Y     P  Pinf  Psup   P_sd
 0         0  30018  0.000   0    0    0  30018  0.00  0.00  0.00  0.000
 5         0  29056  0.000   0    0    0  29056  0.00  0.00  0.00  0.000
10         0  30245  0.000   0    0    0  30245  0.00  0.00  0.00  0.000
15         0  32911  0.000   0    0    0  32911  0.00  0.00  0.00  0.000
[20,25)    1  44603  2.242   1    1    1  44603  2.24  2.24  2.24  0.224
[25,30)    2  52610  3.802   4    3    4  52610  7.60  5.70  7.60  0.634
[30,35)   16  50174 31.889  14   12   14  50174 27.90 23.92 27.90  1.520
[35,40)   41  50147 81.760  38   34   40  50147 75.78 67.80 79.77  3.056
[40,45)   41  47388 86.520  34   29   37  47388 71.75 61.20 78.08  4.003
[45,50)   28  42328 66.150  36   32   39  42328 85.05 75.60 92.14  4.733
[50,55)   28  38451 72.820  26   22   29  38451 67.62 57.22 75.42  4.347
[55,60)   17  34684 49.014  19   15   22  34684 54.78 43.25 63.43  5.042
[60,65)   14  26773 52.291  15   11   17  26773 56.03 41.09 63.50  6.907
[65,70)   18  31706 56.772  18   13   21  31706 56.77 41.00 66.23  6.816
[70,75)   17  30839 55.125  17   13   21  30839 55.13 42.15 68.10  7.335
[75,80)    7  26080 26.840  12    7   17  26080 46.01 26.84 65.18  9.126
[80,85)    9  19130 47.047  12    6   16  19130 62.73 31.36 83.64 12.466
[85,90]    0  16193  0.000   5    2    8  16193 30.88 12.35 49.40  9.784
Total    239 633336 37.737 248  236  259 633336 39.16 37.26 40.89  0.951
```

A.2.8. Funciones para el "Método Tiempo de Muerte log-Logístico": *base.sim.tllog* y *prev.AC.V.sim.tllog*

base.sim.tllog permite generar una base de datos simulada (una cohorte de mujeres) para un número de casos incidentes por año e intervalo de edad en un período (1999-2007) donde los tiempos de muerte son simulados a partir de un modelo de distribución log-Logístico ajustado por grupo de edad (definido con un vector de corte) sobre los registros de la población de referencia (ej. Girona y Tarragona). Cada caso incidente simulado tendrá edad igual al punto medio de su intervalo de edad quinquenal, la fecha de diagnóstico fijada en 1 de julio del año de incidencia, el tiempo de muerte generado, y de ahí el mes y el año del último seguimiento, y el estatus a la fecha de cierre de la base de registros de referencia.

Argumentos:

`data.orig`:	registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros de ρ y λ de las distribuciones log-L ajustadas en cada grupo de edad definido en el vector de corte `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
`cut.age`:	un vector de corte que define los grupos de edad para los que se ajusta cada modelo log-L y se estima sus parámetros rho y lambda
`casos.sim`:	*data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').
`met.pars`:	es el método usado para la estimación de los parámetros ρ y λ, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').

Salida: una base de registros simulados para un conjunto de casos incidentes, a partir de modelos de supervivencia log-L ajustados por edad sobre una base de registros de cáncer de referencia.

```
> base.sim.tllog <- function(data.orig, cut.age, casos.sim, met.pars='OLS'){
+
+   # obtengo supervivencias para cut.age
+   require("Hmisc")
+   require("survival")
+   require("Epi")
+
+   data.orig$int.edad.inc <- cut2(data.orig$edat.inc, cut.age)
+
```

```
+   #Supervivencia Nelson-Aalen (distinguiendo intervalos de edad)
+   svf.N_A <- summary(survfit(Surv(follow.up,status)~int.edad.inc,
+                              data.orig, type="fleming-harrington"))
+   svf.N_A # objeto con supervivencia N-A por grupo de edad de cut.age
+
+   # data frame con variables edad, t, ln(t) y otras transf. de svf.N_A
+   dat.tmp <- data.frame(grupo.edad=as.numeric(svf.N_A$strata),
+                         t=svf.N_A$time, ln_t=log(svf.N_A$time),
+                         RA=-log(svf.N_A$surv),
+                         y=log(exp(-log(svf.N_A$surv))-1))
+
+   t0 <- min(casos.sim$year)
+   tf <- max(casos.sim$year)
+   base <- NULL
+   for(i in t0:tf) {
+     casos <- subset(casos.sim, year==i)
+
+     a <- cut.age
+     casos$Age2 <- seq(0, 85, 5)
+     casos$Age2<- cut2(casos$Age2, a)
+     casos.proy <- stat.table(list(Age2), list(C=sum(C)),margins=TRUE,
+                              data=casos)
+
+     Int.Edad <- rep(levels(casos$Age2),
+                     casos.proy[,1:length(levels(casos$Age2))])
+     t.llog    <- rep(NA, casos.proy[,'Total'])
+     unif <- runif(casos.proy[,'Total'], min = 0, max = 1)
+     salida <- data.frame(Int.Edad, t.llog, unif)
+
+     if (met.pars=='OLS'){
+       for (l in 1:casos.proy[,'Total']){
+
+         j<- as.numeric(salida[l,'Int.Edad'])
+         dat.tmp.j <- subset(dat.tmp,grupo.edad==j) # subset edad
+
+         # parámetros: rho y lambda
+         fit.j <- summary(lm(y~ln_t, data=dat.tmp.j))
+         rho.j <- fit.j$coefficients[2] # pendiente=rho
+         lambda.j <- exp(fit.j$coefficients[1]) # exp(intercepto)=lambda
+
+         salida[l,'t.llog'] <-((exp(-log(salida[l,'unif']))-1)/lambda.j)^(1/rho.j)
+       }
+       salida
+     }
+
+     if (met.pars=='ML'){
+       for (l in 1:casos.proy[,'Total']){
+
+         j<- as.numeric(salida[l,'Int.Edad'])
+         dat.tmp.j <- subset(data.orig,as.numeric(int.edad.inc)==j) #subset edad
```

```
+
+         # parámetros lambda y rho estimados con función survreg
+
+         fit.llog.j <- survreg(Surv(follow.up, status) ~ 1,
+                               dist="loglogistic",data=dat.tmp.j)
+         # rho.j = (survreg' scale)^-1 = sigma^-1 = a
+         rho.j <- fit.llog.j$scale^-1
+         # lambda=1/b^a=(exp(survreg'intercept))^-a=exp(-survreg'intercept/sigma)
+         lambda.j <- exp(-fit.llog.j$coefficients/fit.llog.j$scale)
+
+         salida[l,'t.llog'] <-((exp(-log(salida[l,'unif']))-1)/lambda.j)^(1/rho.j)
+       }
+       salida
+     }
+
+     follow.up <- salida[,'t.llog']  # es tiempo hasta la muerte
+
+     edat.inc <- rep(seq(2.5, 87.5, 5), casos$C)
+     mes.inc <- rep_len(7,sum(casos$C))
+
+     any.us0 <- i+follow.up
+     any.us <- floor(any.us0)
+     mes.us0 <- ((any.us0-any.us)*12)+6
+
+     cambia.año <- which(mes.us0>12)
+     any.us[cambia.año] <- any.us[cambia.año]+1
+     mes.us0[cambia.año] <- mes.us0[cambia.año]-12
+
+     mes.us <- ceiling(mes.us0)
+
+     status <- ifelse((any.us>max(data.orig$any.us)|
+                         (any.us==max(data.orig$any.us) & mes.us>6)),0,1)
+
+     id.pacient <- c(1:sum(casos$C))+i*0.0001
+     sexe <- rep_len('X',sum(casos$C))
+     loc <-  rep_len('X',sum(casos$C))
+     any.inc <- rep_len(i,sum(casos$C))
+
+     base[[i]] <- data.frame(id.pacient,sexe,loc,edat.inc,mes.inc,
+                             any.inc,mes.us,any.us,status,follow.up)
+   }
+   base <- do.call(rbind, base)
+   base
+ }
> # Ejemplo: Se simula una base para todos los casos incidentes de Girona
> # y Tarragona entre 1999 y 2007 (in.GT) según un modelo log-L para el
> # tiempo de muerte ajustado (por máxima verosimilitud, 'ML') en cada
> # uno de los 6 grupos de edad (definidos por corte en 35,45,55,65,75)
> # sobre los registros de referencia (datos.II)
>
```

```
> base.llog.GT <- base.sim.tllog(datos.II, c(35,45,55,65,75),in.GT, 'ML')
> head(base.llog.GT)

  id.pacient sexe loc edat.inc mes.inc any.inc mes.us any.us status   follow.up
1     1.1999    X   X     27.5       7    1999      5   2015      0   15.840817
2     2.1999    X   X     27.5       7    1999      7   2054      0   55.067078
3     3.1999    X   X     32.5       7    1999      2   2009      1    9.611195
4     4.1999    X   X     32.5       7    1999      7   2834      0  835.063356
5     5.1999    X   X     32.5       7    1999      1   2001      1    1.504033
6     6.1999    X   X     32.5       7    1999     10   3376      0 1377.310755

>
> #---------------------------------------------------------------------
```

prev.AC.V.sim.tllog permite estimar prevalencia a 5 años en un año dado (`t.AC`) a partir de `nsim` bases simuladas para un conjunto de casos incidentes, por método "Tiempo de Muerte log-Logístico". Para cada base simulada se calcula prevalencia a 5 años por intervalo de edad, y luego sobre el total de simulaciones se extrae la mediana junto a los percentiles 2,5 % y 97,5 % de prevalencia, y se calcula el desvío estándar por intervalo de edad.

Argumentos:

- `data.orig`: registros de cáncer de la población de referencia (ej. Girona y Tarragona 1999-2007) desde donde se estiman los parámetros ρ y λ de las distribuciones log-L ajustadas en cada grupo de edad definido en el vector de corte `cut.age`. El *data frame* debe contener la siguientes variables: edad ('edat.inc'), mes ('mes.inc') y año ('any.inc') de diagnóstico, mes ('mes.us'), año ('any.us') y estatus ('status') en el último seguimiento, y tiempo de seguimiento ('follow.up').
- `cut.age`: un vector de corte que define los grupos de edad para los que se ajusta cada modelo log-L y se estima sus parámetros rho y lambda
- `casos.sim`: *data frame* con los casos incidentes a simular ('C'), por año de incidencia ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18), y correspondiente población total ('population').
- `t.AC`: año para el que se estima prevalencia a 5 años.
- `pob.orig`: distribución de mujeres ('population') por año ('year') e intervalo de edad quinquenal ('age.group', de 1 a 18) en la población de referencia (ej. Girona y Tarragona).
- `met.pars`: es el método usado para la estimación de los parámetros ρ y λ, método gráfico ('OLS', por defecto) o método de máxima verosimilitud ('ML').
- `nsim`: número de bases (cohortes) que se desean simular para la estimación de prevalencia 5 años.

Salida: primera columna corresponde al intervalo de edad, siguientes 3 columnas corresponden a casos prevalentes ('C.emp'), población femenina ('Y.emp') y prevalencia observada ('P.emp', casos por cada 100,000 mujeres) en población de referencia; siguientes 3 columnas para casos simulados ('C'; mediana, 2.5 y 97.5 %), luego mujeres en la población para la que se simula ('Y'), 3 columnas para percentiles de prevalencia simulada ('P'; mediana, 2.5 y 97.5 %), y por último el desvío de prevalencia simulada ('P sd').

```
> prev.AC.V.sim.tllog <- function(data.orig, cut.age, casos.sim, t.AC,
+                                 pob.orig, met.pars='OLS', nsim=100){
+   conteos <- matrix(NA, 19, nsim)
+
+   for(sim in 1:nsim){
+     print("Sim")
+     print(sim)
+
+     base <- base.sim.tllog(data.orig, cut.age, casos.sim, met.pars)
+     c <- prevalencia(datos=base, año=t.AC, int1=1, int2=5,
+                      pob=casos.sim)
+     conteos[,sim] <- c$C
+   }
+
+   cuantiles <- round(t(apply(conteos,1,quantile,
+                              probs=c(0.5, 0.025, 0.975))),0)
+   desvios <- apply(conteos/c$Y*100000, 1, sd)
+
+   emp <- prevalencia(datos=data.orig, año=t.AC, int1=1, int2=5,
+                      pob=pob.orig)
+
+   salida <- cbind(emp[,1:3],cuantiles,c$Y,
+                   round(cuantiles/c$Y*100000,2),round(desvios,3))
+   rownames(salida) <- rownames(c)
+   colnames(salida) <- c('C.emp','Y.emp','P.emp','Cm','Cinf','Csup',
+                         'Y','P','Pinf','Psup','P_sd')
+   salida
+ }
> # Ej.: Prevalencia a 5 años de cáncer de cérvix en Girona y Tarragona
> # en 2003 estimada con 100 bases simuladas por método "Tiempo de Muerte
> # log-L" con estimadores máximo verosímiles de los parámetros en 6
> # grupos de edad (definidos con corte, 35,45,55,65,75) sobre la base
> # de referencia (Girona y Tarragona)
>
> AC.5a.tllog.GT.2003.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2003,fit.pop.F,'ML',
+                                               nsim=100)
> AC.5a.tllog.GT.2003.ml
```

	C.emp	Y.emp	P.emp	Cm	Cinf	Csup	Y	P	Pinf	Psup	P_sd
0	0	30018	0.000	0	0	0	30018	0.00	0.00	0.00	0.000
5	0	29056	0.000	0	0	0	29056	0.00	0.00	0.00	0.000
10	0	30245	0.000	0	0	0	30245	0.00	0.00	0.00	0.000
15	0	32911	0.000	0	0	0	32911	0.00	0.00	0.00	0.000
[20,25)	1	44603	2.242	1	0	1	44603	2.24	0.00	2.24	0.384
[25,30)	2	52610	3.802	4	3	4	52610	7.60	5.70	7.60	0.767
[30,35)	16	50174	31.889	14	12	14	50174	27.90	23.92	27.90	1.637
[35,40)	41	50147	81.760	37	35	40	50147	73.78	69.79	79.77	2.801
[40,45)	41	47388	86.520	34	30	37	47388	71.75	63.31	78.08	3.613
[45,50)	28	42328	66.150	36	30	39	42328	85.05	70.88	92.14	4.917
[50,55)	28	38451	72.820	27	22	29	38451	70.22	57.22	75.42	4.721
[55,60)	17	34684	49.014	18	15	21	34684	51.90	43.25	60.55	5.096
[60,65)	14	26773	52.291	14	10	17	26773	52.29	37.35	63.50	6.632
[65,70)	18	31706	56.772	17	14	21	31706	53.62	44.16	66.23	6.009
[70,75)	17	30839	55.125	17	13	21	30839	55.13	42.15	68.10	7.765
[75,80)	7	26080	26.840	13	7	16	26080	49.85	26.84	61.35	8.081
[80,85)	9	19130	47.047	10	6	15	19130	52.27	31.36	78.41	12.612
[85,90]	0	16193	0.000	4	1	7	16193	24.70	6.18	43.23	9.596
Total	239	633336	37.737	245	235	256	633336	38.68	37.11	40.42	0.919

Apéndice B

Script de R

B.1. Análisis descriptivo del cáncer de cérvix en Girona y Tarragona (1999-2007)

```
> # Directorio y archivos
>
> dir.in<-"C:/Users/user/Documents/Vale/MESIO/TFM mesio/Paso 2/"
> file.surv<-"Coll Uter 1999-2007.txt"
> file.pop<-"GiT_Population_Distribution.txt"
> # Datos
> fit.surv<-as.data.frame(read.table(paste(dir.in,file.surv,sep=""),
+                                    header=T))
> fit.pop<-as.data.frame(read.table(paste(dir.in,file.pop,sep=""),
+                                   header=T))
> # Distribución de mujeres por año e intervalo de edad Girona y Tarragona
> fit.pop.F<-fit.pop[fit.pop$sex==2,]
> # Incidencia estimada en Catalunya (1999-2007)
> file.incid.Cat<-"Incidencia Cervix Catalunya 1999-2007.txt"
> fit.incid.Cat<-as.data.frame(read.table(paste(dir.in,
+                                  file.incid.Cat,sep=""),header=T))
> # cambio nombres de columnas por lo que esperan las funciones
> colnames(fit.incid.Cat)<- c('age.group','year','C','population')
> ############
> # Librerías
>
> library("Hmisc")
> library("Epi")
> library("survival")
> ############
> # Función para llevar todos los tiempo de seg. con status 0 hasta 6/2010
> corr.año.cierre.II <- function(dat, año.cierre, mes.cierre){
+   a <- año.cierre
+
```

```
+   b <- which(dat$any.us>a)
+   dat$mes.us[b] <- mes.cierre
+   dat$status[b] <- 0
+
+   dat$any.us[b] <- a
+
+   c <- which(dat$status==0)
+   dat$mes.us[c] <- mes.cierre
+   dat$any.us[c] <- a
+
+   dat$follow.up <- round(with(dat, (mes.us-mes.inc)*(1/12)+(any.us-any.inc)),2)
+   dat$follow.up[which(dat$follow.up==0)] <- round(1/12,2)
+   dat
+ }
> datos.II <- corr.año.cierre.II(fit.surv,2010,6)
```

Tablas de ocurrencia de cáncer de cuello de útero en Girona y Tarragona

```
> #Tabla 1. Incidencia anual de cáncer de cuello de útero en Girona y Tarragona
>
> carga.1999 <- carga(datos.II,1999,1, fit.pop.F)
> carga.2000 <- carga(datos.II,2000,1, fit.pop.F)
> carga.2001 <- carga(datos.II,2001,1, fit.pop.F)
> carga.2002 <- carga(datos.II,2002,1, fit.pop.F)
> carga.2003 <- carga(datos.II,2003,1, fit.pop.F)
> carga.2004 <- carga(datos.II,2004,1, fit.pop.F)
> carga.2005 <- carga(datos.II,2005,1, fit.pop.F)
> carga.2006 <- carga(datos.II,2006,1, fit.pop.F)
> carga.2007 <- carga(datos.II,2007,1, fit.pop.F)
> carga.1999
> carga.2000
> carga.2001
> carga.2002
> carga.2003
> carga.2004
> carga.2005
> carga.2006
> carga.2007
> #
> # Tabla 2. Prevalencia a un año de cáncer de cuello de útero en Girona y Tarragona
>
> AC99.1a <- prevalencia(datos=datos.II, año=1999, int1=1, int2=1, pob=fit.pop.F)
> AC00.1a <- prevalencia(datos=datos.II, año=2000, int1=1, int2=1, pob=fit.pop.F)
> AC01.1a <- prevalencia(datos=datos.II, año=2001, int1=1, int2=1, pob=fit.pop.F)
> AC02.1a <- prevalencia(datos=datos.II, año=2002, int1=1, int2=1, pob=fit.pop.F)
> AC03.1a <- prevalencia(datos=datos.II, año=2003, int1=1, int2=1, pob=fit.pop.F)
> AC04.1a <- prevalencia(datos=datos.II, año=2004, int1=1, int2=1, pob=fit.pop.F)
> AC05.1a <- prevalencia(datos=datos.II, año=2005, int1=1, int2=1, pob=fit.pop.F)
```

```
> AC06.1a <- prevalencia(datos=datos.II, año=2006, int1=1, int2=1, pob=fit.pop.F)
> AC07.1a <- prevalencia(datos=datos.II, año=2007, int1=1, int2=1, pob=fit.pop.F)
> AC99.1a
> AC00.1a
> AC01.1a
> AC02.1a
> AC03.1a
> AC04.1a
> AC05.1a
> AC06.1a
> AC07.1a
> #
> # Tabla 3. Prevalencia de cáncer para casos con 2-3 años de diagnosticados en
> # Girona y Tarragona
>
> Int00.2a <- prevalencia(datos=datos.II, año=2000, int1=2, int2=3, pob=fit.pop.F)
> Int01.23a <- prevalencia(datos=datos.II, año=2001, int1=2, int2=3, pob=fit.pop.F)
> Int02.23a <- prevalencia(datos=datos.II, año=2002, int1=2, int2=3, pob=fit.pop.F)
> Int03.23a <- prevalencia(datos=datos.II, año=2003, int1=2, int2=3, pob=fit.pop.F)
> Int04.23a <- prevalencia(datos=datos.II, año=2004, int1=2, int2=3, pob=fit.pop.F)
> Int05.23a <- prevalencia(datos=datos.II, año=2005, int1=2, int2=3, pob=fit.pop.F)
> Int06.23a <- prevalencia(datos=datos.II, año=2006, int1=2, int2=3, pob=fit.pop.F)
> Int07.23a <- prevalencia(datos=datos.II, año=2007, int1=2, int2=3, pob=fit.pop.F)
> Int00.2a
> Int01.23a
> Int02.23a
> Int03.23a
> Int04.23a
> Int05.23a
> Int06.23a
> Int07.23a
> #
> # Tabla 4. Prevalencia de cáncer para casos con 4-5 años de diagnosticados en
> # Girona y Tarragona.
>
> Int02.4a <- prevalencia(datos=datos.II, año=2002, int1=4, int2=5, pob=fit.pop.F)
> Int03.45a <- prevalencia(datos=datos.II, año=2003, int1=4, int2=5, pob=fit.pop.F)
> Int04.45a <- prevalencia(datos=datos.II, año=2004, int1=4, int2=5, pob=fit.pop.F)
> Int05.45a <- prevalencia(datos=datos.II, año=2005, int1=4, int2=5, pob=fit.pop.F)
> Int06.45a <- prevalencia(datos=datos.II, año=2006, int1=4, int2=5, pob=fit.pop.F)
> Int07.45a <- prevalencia(datos=datos.II, año=2007, int1=4, int2=5, pob=fit.pop.F)
> Int02.4a
> Int03.45a
> Int04.45a
> Int05.45a
> Int06.45a
> Int07.45a
> #
> # Tabla 5. Prevalencia acumulada a 5 años de casos de cáncer incidentes en Girona
> # y Tarragona durante el período 1999-2007.
```

```
>
> AC03.5a <- prevalencia(datos=datos.II, año=2003, int1=1, int2=5, pob=fit.pop.F)
> AC04.5a <- prevalencia(datos=datos.II, año=2004, int1=1, int2=5, pob=fit.pop.F)
> AC05.5a <- prevalencia(datos=datos.II, año=2005, int1=1, int2=5, pob=fit.pop.F)
> AC06.5a <- prevalencia(datos=datos.II, año=2006, int1=1, int2=5, pob=fit.pop.F)
> AC07.5a <- prevalencia(datos=datos.II, año=2007, int1=1, int2=5, pob=fit.pop.F)
> AC03.5a
> AC04.5a
> AC05.5a
> AC06.5a
> AC07.5a
>
```

Supervivencia en Girona y Tarragona

```
> # Supervivencia Empírica K-M en 6 grupos de edad:
> # menor de 35, cada 10 años hasta mayor de 75 años
> datos <- datos.II
> datos$int.edad.inc <- cut2(datos$edat.inc, c(0, 35, 45, 55, 65, 75))
> #creamos el elemento con la sobrevivencia de las pacientes
> sup99.07 <- with(datos, Surv(follow.up, status))
> sup99.07 # son los tiempos de seguimiento (varios con censura por derecha, ok!)
> # Estimamos supervivencia con datos censurados usando K-M
> (svf.edad <- survfit(sup99.07~int.edad.inc, datos))
> summary(svf.edad)   # funcion de supervivencia por grupo de edad
> #
> # Gráfico con función de supervivencia por grupo de edad definido
>
> par(mfrow=c(1,1), font=2, font.lab=4, las=1)
> plot(svf.edad, col=c('black','steelblue','tomato','olivedrab','orange','violetred4'),
+      conf.int=F, mark.time=T, lwd=2, ylab = 'S(t)', xlab='Tiempo [años]',
+      cex.axis=0.8)
> #title('Función de supervivencia por grupo de edad (99/2007)')
> legend('bottomleft', c("[0,35)","[35,45)","[45,55)","[55,65)","[65,75)","[75<)"),
+        col=c('black','steelblue','tomato','olivedrab','orange','violetred4'),
+        bty="n", lwd=2, cex=0.9)
> #
> # Test pare evaluar si hay diferencias en la supervivencia por grupos de edad
> # Test logrank (rho=0, por defecto)
> TestlogR <- survdiff(sup99.07~int.edad.inc, datos)
> TestlogR
> # Test Peto-Peto / menos peso a las diferencias tardías
> TestPP <- survdiff(sup99.07~int.edad.inc, datos)
> TestPP
> #######
> # Exploracion de ajuste a modelos paramétricos
> # y estimación de parámetros
> #
> # Método GRÁFICO
> # con distribución Exponencial: H(t) vs t
```

```
> # con distribución de Weibull: ln[H(t)] vs ln(t)
> # con distribución log-logística: ln(exp[H(t)]-1) vs ln(t)
>
> #Supervivencia Nelson-Aalen distinguiendo 6 grupos de edad
> # (pacientes del periodo 1999-2007 con seguimiento a 6/2010)
> N_A <- summary(survfit(Surv(follow.up,status)~int.edad.inc, datos,
+                        type="fleming-harrington"))
> N_A
> # un data.frame con las variables de interés por grupo de edad (numérico)
> dat <- data.frame(grupo.edad=as.numeric(N_A$strata),t=(N_A$time),
+                   ln_t=log(N_A$time),S=N_A$surv,RA=-log(N_A$surv),
+                   ln_RA=log(-log(N_A$surv)),y=log(exp(-log(N_A$surv))-1))
> # Modelo Exponencial
> # h(t)= lambda , tasa de fallo
> # Riesgo acumulado H(t)= lambda t  => -LnS(t)= lambda t
> # Evaluamos RiesgoAcum vs tiempo (ajuste forzado a pasar por origen)
>
> par(mfrow=c(2,3), font=2, font.lab=4, las=1, pch = 16, cex.main=1.1)
> summary(lm(RA~t-1, subset(dat,grupo.edad==1)))
> plot(RA~t, subset(dat,grupo.edad==1), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,3), ylim=c(0,0.125))
> title('Grupo de edad: [0,35) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==1)))
> summary(lm(RA~t-1, subset(dat,grupo.edad==2)))
> plot(RA~t, subset(dat,grupo.edad==2), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,6), ylim=c(0,0.24))
> title('Grupo de edad: [35,45) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==2)))
> summary(lm(RA~t-1, subset(dat,grupo.edad==3)))
> plot(RA~t, subset(dat,grupo.edad==3), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,6.6), ylim=c(0,0.30))
> title('Grupo de edad: [45,55) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==3)))
> summary(lm(RA~t-1, subset(dat,grupo.edad==4)))
> plot(RA~t, subset(dat,grupo.edad==4), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,8), ylim=c(0,0.45))
> title('Grupo de edad: [55,65) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==4)))
> summary(lm(RA~t-1, subset(dat,grupo.edad==5)))
> plot(RA~t, subset(dat,grupo.edad==5), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,9.3), ylim=c(0,1))
> title('Grupo de edad: [65,75) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==5)))
> summary(lm(RA~t-1, subset(dat,grupo.edad==6)))
> plot(RA~t, subset(dat,grupo.edad==6), xlab = "Tiempo [años]",
+      ylab = 'H(t)', xlim=c(0,2.35), ylim=c(0,0.96))
> title('Grupo de edad: [75<) años')
> abline(lm(RA~t-1, subset(dat,grupo.edad==6)))
> #
> # con distribución de Wiebull: ln(RiesgoAcum) vs ln(t)
```

```
>
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==1)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==1), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [0,35) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==1)))
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==2)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==2), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [35,45) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==2)))
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==3)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==3), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [45,55) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==3)))
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==4)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==4), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [55,65) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==4)))
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==5)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==5), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [65,75) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==5)))
> summary(lm(ln_RA~ln_t, subset(dat,grupo.edad==6)))
> plot(ln_RA~ln_t, subset(dat,grupo.edad==6), xlab = "ln Tiempo [años]",
+      ylab = 'ln H(t)')
> title('Grupo de edad: [75<) años')
> abline(lm(ln_RA~ln_t, subset(dat,grupo.edad==6)))
> #
> # con distribución log-logística: ln(exp[RiesgoAcum]-1) vs ln(t)
>
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==1)))
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==1), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [0,35) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==1)))
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==2)))
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==2), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [35,45) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==2)))
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==3)))
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==3), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [45,55) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==3)))
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==4)))
```

```
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==4), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [55,65) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==4)))
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==5)))
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==5), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [65,75) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==5)))
> summary(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==6)))
> plot(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==6), xlab = "ln Tiempo [años]",
+      ylab = 'ln [exp H(t) - 1]')
> title('Grupo de edad: [75<) años')
> abline(lm(log(exp(RA)-1)~ln_t, subset(dat,grupo.edad==6)))
> ###
> # Método ML
> # Ajustamos un modelo log-lineal Y = Ln (T) = intercepto + scala W = mu + sigma W
> # (en exponencial se fija escala, sigma=1)
> # (rho=1/sigma  lambda=exp(-mu/sigma)
> # trabajo con todos los datos (status 1 y 0)
>
> # "Exponencial" para cada grupo de edad por separado
> # Grupo de edad 1
> datos.1 <- subset(datos,as.numeric(int.edad.inc)==1)  #datos de la base
> fit.1.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.1)
> fit.1.e
> (rho.1 <- 1/fit.1.e$scale)  # tiene que ser 1 porque exponencial
> (lambda.1 <- (exp(fit.1.e$coeff))^-rho.1)
> (exp(-fit.1.e$coeff/fit.1.e$scale))
> # Grupo de edad 2
> datos.2 <- subset(datos,as.numeric(int.edad.inc)==2)  #datos de la base
> fit.2.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.2)
> fit.2.e
> (rho.2 <- 1/fit.2.e$scale)  # tiene que ser 1 porque es exponencial
> (lambda.2 <- (exp(fit.2.e$coeff))^-rho.2)
> (exp(-fit.2.e$coeff/fit.2.e$scale))
> # Grupo de edad 3
> datos.3 <- subset(datos,as.numeric(int.edad.inc)==3)  #datos de la base
> fit.3.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.3)
> fit.3.e
> (rho.3 <- 1/fit.3.e$scale)  # tiene que ser 1 porque es exponencial
> (lambda.3 <- (exp(fit.3.e$coeff))^-rho.3)
> (exp(-fit.3.e$coeff/fit.3.e$scale))
> # Grupo de edad 4
> datos.4 <- subset(datos,as.numeric(int.edad.inc)==4)  #datos de la base
> fit.4.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.4)
> fit.4.e
> (rho.4 <- 1/fit.3.e$scale)  # tiene que ser 1 porque exponencial
> (lambda.4 <- (exp(fit.4.e$coeff))^-rho.4)
> (exp(-fit.4.e$coeff/fit.4.e$scale))
```

```
> # Grupo de edad 5
> datos.5 <- subset(datos,as.numeric(int.edad.inc)==5)
> fit.5.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.5)
> fit.5.e
> (rho.5 <- 1/fit.5.e$scale ) # tiene que ser 1 porque exponencial
> (lambda.5 <- (exp(fit.5.e$coeff))^-rho.5)
> (exp(-fit.5.e$coeff/fit.5.e$scale))
> # Grupo de edad 6
> datos.6 <- subset(datos,as.numeric(int.edad.inc)==6)
> fit.6.e <- survreg(Surv(follow.up, status) ~ 1,dist="exponential",data=datos.6)
> fit.6.e
> (rho.6 <- 1/fit.6.e$scale ) # tiene que ser 1 porque es exponencial
> (lambda.6 <- (exp(fit.6.e$coeff))^-rho.6)
> (exp(-fit.6.e$coeff/fit.6.e$scale))
> #
> # "weibull" para cada grupo de edad por separado
> # Grupo de edad 1
> datos.1 <- subset(datos,as.numeric(int.edad.inc)==1)
> fit.1.w <- survreg(Surv(follow.up, status) ~ 1,dist="weibull",data=datos.1)
> fit.1.w
> (rho.1 <- 1/fit.1.w$scale)
> (lambda.1 <- (exp(fit.1.w$coeff))^-rho.1)
> exp(-fit.1.w$coeff/fit.1.w$scale)
> # Grupo de edad 2
> datos.2 <- subset(datos,as.numeric(int.edad.inc)==2)
> fit.2.w <- survreg(Surv(follow.up, status) ~ 1,dist="weibull",data=datos.2)
> fit.2.w
> (rho.2 <- 1/fit.2.w$scale)
> (lambda.2 <- (exp(fit.2.w$coeff))^-rho.2)
> exp(-fit.2.w$coeff/fit.2.w$scale)
> # Grupo de edad 3
> datos.3 <- subset(datos,as.numeric(int.edad.inc)==3)
> fit.3.w <- survreg(Surv(follow.up, status) ~ 1,dist="weibull",data=datos.3)
> fit.3.w
> (rho.3 <- 1/fit.3.w$scale)
> (lambda.3 <- (exp(fit.3.w$coeff))^-rho.3)
> exp(-fit.3.w$coeff/fit.3.w$scale)
> # Grupo de edad 4
> datos.4 <- subset(datos,as.numeric(int.edad.inc)==4)
> fit.4.w <- survreg(Surv(follow.up, status) ~ 1,dist="weibull",data=datos.4)
> fit.4.w
> (rho.4 <- 1/fit.4.w$scale)
> (lambda.4 <- (exp(fit.4.w$coeff))^-rho.4)
> exp(-fit.4.w$coeff/fit.4.w$scale)
> # Grupo de edad 5
> datos.5 <- subset(datos,as.numeric(int.edad.inc)==5)
> fit.5.w <- survreg(Surv(follow.up, status) ~ 1,dist="weibull",data=datos.5)
> fit.5.w
> (rho.5 <- 1/fit.5.w$scale)
> (lambda.5 <- (exp(fit.5.w$coeff))^-rho.5)
```

```
> exp(-fit.5.w$coeff/fit.5.w$scale)
> # Grupo de edad 6
> datos.6 <- subset(datos,as.numeric(int.edad.inc)==6)
> fit.6.w <- survreg(Surv(follow.up, status) ~ 1, dist="weibull",data=datos.6)
> fit.6.w
> (rho.6 <- 1/fit.6.w$scale)
> (lambda.6 <- (exp(fit.6.w$coeff))^-rho.6)
> exp(-fit.6.w$coeff/fit.6.w$scale)
> #
> # "log-logistico" para cada grupo de edad por separado
> # Grupo de edad 1
> datos.1 <- subset(datos,as.numeric(int.edad.inc)==1)
> fit.1.llog <- survreg(Surv(follow.up, status) ~ 1,dist="loglogistic",data=datos.1)
> fit.1.llog
> (rho.1 <- 1/fit.1.llog$scale)
> (lambda.1 <- (exp(fit.1.llog$coefficients))^-rho.1)
> exp(-fit.1.llog$coefficients/fit.1.llog$scale)   # todo Ok!!
> # Grupo de edad 2
> datos.2 <- subset(datos,as.numeric(int.edad.inc)==2)
> fit.2.llog <- survreg(Surv(follow.up, status) ~ 1, dist="loglogistic",data=datos.2)
> fit.2.llog
> (rho.2 <- 1/fit.2.llog$scale) #
> (lambda.2 <- (exp(fit.2.llog$coefficients))^-rho.2)
> exp(-fit.2.llog$coefficients/fit.2.llog$scale)   # todo Ok!!
> # Grupo de edad 3
> datos.3 <- subset(datos,as.numeric(int.edad.inc)==3)
> fit.3.llog <- survreg(Surv(follow.up, status) ~ 1, dist="loglogistic",data=datos.3)
> fit.3.llog
> (rho.3 <- 1/fit.3.llog$scale) #
> (lambda.3 <- (exp(fit.3.llog$coefficients))^-rho.3)
> exp(-fit.3.llog$coefficients/fit.3.llog$scale)   # todo Ok!!
> # Grupo de edad 4
> datos.4 <- subset(datos,as.numeric(int.edad.inc)==4)
> fit.4.llog <- survreg(Surv(follow.up, status) ~ 1, dist="loglogistic",data=datos.4)
> fit.4.llog
> (rho.4 <- 1/fit.4.llog$scale) #
> (lambda.4 <- (exp(fit.4.llog$coefficients))^-rho.4)
> exp(-fit.4.llog$coefficients/fit.4.llog$scale)   # todo Ok!!
> # Grupo de edad 5
> datos.5 <- subset(datos,as.numeric(int.edad.inc)==5)
> fit.5.llog <- survreg(Surv(follow.up, status) ~ 1, dist="loglogistic",data=datos.5)
> fit.5.llog
> (rho.5 <- 1/fit.5.llog$scale) #
> (lambda.5 <- (exp(fit.5.llog$coefficients))^-rho.5)
> exp(-fit.5.llog$coefficients/fit.5.llog$scale)   # todo Ok!!
> # Grupo de edad 6
> datos.6 <- subset(datos,as.numeric(int.edad.inc)==6)
> fit.6.llog <- survreg(Surv(follow.up, status) ~ 1, dist="loglogistic",data=datos.6)
> fit.6.llog
> (rho.6 <- 1/fit.6.llog$scale) #
```

```
> (lambda.6 <- (exp(fit.6.llog$coefficients))^-rho.6)
> exp(-fit.6.llog$coefficients/fit.6.llog$scale)   # todo Ok!!
> # Los diferentes modelos de supervivencia paramétricos ajustados
> # y AIC
>
> # lista con datos por grupo de edad
> datos.0.35<-datos[datos$int.edad.inc=="[ 0,35)",]
> datos.35.45<-datos[datos$int.edad.inc=="[35,45)",]
> datos.45.55<-datos[datos$int.edad.inc=="[45,55)",]
> datos.55.65<-datos[datos$int.edad.inc=="[55,65)",]
> datos.65.75<-datos[datos$int.edad.inc=="[65,75)",]
> datos.75.T<-datos[datos$int.edad.inc=="[75,93]",]
> lista.datos<-list(datos.0.35,datos.35.45,datos.45.55,datos.55.65,datos.65.75,datos.75.T)
> #
> # Figura Supervivencia en modelos paramétricos
>
> plot.sel<-function(data.sel.t)
+ {
+   data.sel.t$dummy<-1
+   fKM <- survfit(Surv(follow.up,status)~dummy,data=data.sel.t)
+   plot(fKM, conf.int=T,  mark.time=F, ylab = 'S(t)',
+        xlab='Tiempo [años]', cex.axis=0.8, col='gray50')
+   fexp <-survreg(Surv(follow.up,status) ~ dummy,dist='weibull',
+                  data=data.sel.t,scale=1)
+   fwei <-survreg(Surv(follow.up,status) ~ dummy,dist='weibull',
+                  data=data.sel.t)
+   flogl <-survreg(Surv(follow.up,status) ~ dummy,dist='logl',
+                   data=data.sel.t)
+   lines(predict(fexp, newdata=list(dummy=1), type="quantile",
+                 p=seq(.01,.99,by=.01)),seq(.99,.01,by=-.01),
+         col="steelblue",lwd=2)
+   lines(predict(fwei, newdata=list(dummy=1), type="quantile",
+                 p=seq(.01,.99,by=.01)),seq(.99,.01,by=-.01),
+         col="orange",lwd=2)
+   lines(predict(flogl, newdata=list(dummy=1), type="quantile",
+                 p=seq(.01,.99,by=.01)),seq(.99,.01,by=-.01),
+         col="violetred4",lwd=2)
+   legend('bottomleft', c("KM","Exp.","Weib.","log-L."),
+          col=c('gray50','steelblue','orange','violetred4'),
+          bty="n", lwd=c(1,2,2,2), cex=0.9)
+ }
> # Comparación modelos por AIC
> aic.model<-function(tiempo.t,status.t)
+ {
+   exp.t<-survreg(Surv(tiempo.t, status.t)~1,dist="w",scale=1)
+   wei.t<-survreg(Surv(tiempo.t, status.t)~1,dist="w")
+   lgl.t<-survreg(Surv(tiempo.t, status.t)~1,dist="logl")
+
+   df.AIC<-(matrix(0,3,2))
```

```
+   df.AIC<-as.data.frame(rbind(extractAIC(exp.t),
+                               extractAIC(wei.t),extractAIC(lgl.t)))
+   names(df.AIC)<-c("df","AIC")
+   row.names(df.AIC)<-c("Exp.","Weib.","log-L")
+
+   df.AIC
+ }
> # Obtengo gráfico y AIC sobre modelos log-lineales Y=lnT=mu+sigma W
> par(mfrow=c(2,3), font=2, font.lab=4, las=1)
> for (i.data in 1:length(lista.datos))
+ {
+   data.sel<-lista.datos[[i.data]]
+   print(aic.model(data.sel$follow.up,data.sel$status))
+   plot.sel(data.sel)
+   print(i.data)
+   title(paste("Grupo de edad: ",
+               levels(datos$int.edad.inc)[i.data],"años","\n"),cex.main=1.1)
+ }
      df      AIC
Exp.   1 81.40498
Weib.  2 82.91817
log-L  2 82.62847
[1] 1
      df      AIC
Exp.   1 266.5281
Weib.  2 256.9468
log-L  2 255.9307
[1] 2
      df      AIC
Exp.   1 227.3560
Weib.  2 220.3536
log-L  2 219.5142
[1] 3
      df      AIC
Exp.   1 222.9981
Weib.  2 208.9263
log-L  2 207.4922
[1] 4
      df      AIC
Exp.   1 275.0296
Weib.  2 262.4571
log-L  2 260.9833
[1] 5
      df      AIC
Exp.   1 249.0334
Weib.  2 198.2856
log-L  2 185.6787
[1] 6

>
```

B.2. Estimaciones de prevalencia a 5 años en Girona y Tarragona - Validación

Método Tiempo de Seguimiento Empírico - GT

```
> set.seed(1234)
> # Prevalencia a 5 años 2003 GT estimada usando distirb de G+T
> AC.5a.tse.GT.2003 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),in.GT,
+                                        2003,fit.pop.F, nsim=1000)
> # Prevalencia a 5 años 2004 GT estimada usando distirb de G+T
> AC.5a.tse.GT.2004 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),in.GT,
+                                        2004,fit.pop.F,nsim=1000)
> # Prevalencia a 5 años 2005 GT estimada usando distirb de G+T
> AC.5a.tse.GT.2005 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),in.GT,
+                                        2005,fit.pop.F,nsim=1000)
> # Prevalencia a 5 años 2006 GT estimada usando distirb de G+T
> AC.5a.tse.GT.2006 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),in.GT,
+                                        2006,fit.pop.F,nsim=1000)
> # Prevalencia a 5 años 2007 GT estimada usando distirb de G+T
> AC.5a.tse.GT.2007 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),in.GT,
+                                        2007,fit.pop.F,nsim=1000)

> ###
> # Función para obtener Figs. de simulaciones
>
> plot.prev.sim<-function(data.tmp, ylim.t=c(0,40),
+                         title.t="Prevalencia 5 años: 2003",
+                         ylab.t="Px100,000 m", xlab.t="Edad [años]",leg=1)
+ {
+   data.tmp <- as.data.frame(data.tmp)[1:18,]
+   data.tmp$Age<-seq(2.5,87.5,5)
+
+   plot(data.tmp$Age,data.tmp$Px100000,type="n",ylim=ylim.t,xlab=xlab.t,
+        ylab=ylab.t)
+   lines(data.tmp$Age,data.tmp$Px100000.emp, col=1,lty=1,lwd=2)
+   lines(data.tmp$Age,data.tmp$Px100000, col='steelblue',lty=1,lwd=2)
+   lines(data.tmp$Age,data.tmp$Pinf,col='steelblue',lty=3,lwd=2)
+   lines(data.tmp$Age,data.tmp$Psup,col='steelblue',lty=3,lwd=2)
+   title(title.t)
+   if (leg==1) {legend('topleft', c("observada","estimada"),
+                       col=c('black','steelblue'),  bty="n",
+                       lwd=2, cex=0.9)}
+   if (leg==2) {legend('topleft', c("obs. GT","est. Cat"),
+                       col=c('black','steelblue'),  bty="n",
+                       lwd=2, cex=0.9)}
+ }
> ####
>
> # Gráficos prev. 5 años G+T estimada en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
```

```
> plot.prev.sim(data.tmp=AC.5a.tse.GT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.tse.GT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.tse.GT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.tse.GT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.tse.GT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Tiempo de Seguimiento Empírico - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
```

Método Mirror - GT

```
> set.seed(1234)
> # Prevalencia acumulada a 5 años de 2003 GT estimada remuestreo en G+T
> AC.5a.mirr.GT.2003 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2003, fit.pop.F,
+                                          nsim=1000)
> # Prevalencia acumulada a 5 años de 2004 GT estimada remuestreo en G+T
> AC.5a.mirr.GT.2004 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2004, fit.pop.F,
+                                          nsim=1000)
> # Prevalencia acumulada a 5 años de 2005 GT estimada remuestreo en G+T
> AC.5a.mirr.GT.2005 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2005, fit.pop.F,
+                                          nsim=1000)
> # Prevalencia acumulada a 5 años de 2006 GT estimada remuestreo en G+T
> AC.5a.mirr.GT.2006 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2006, fit.pop.F,
+                                          nsim=1000)
> # Prevalencia acumulada a 5 años de 2007 GT estimada remuestreo en G+T
> AC.5a.mirr.GT.2007 <- prev.AC.V.sim.mirr(datos.II, in.GT, 2007, fit.pop.F,
+                                          nsim=1000)
> #
> # Gráficos prevalencia a 5 años GT estimada en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirr.GT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.mirr.GT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.mirr.GT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.mirr.GT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.mirr.GT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Mirror - Girona y Tarragona', outer=TRUE,
+       col.main='steelblue')
```

Método Mirror-Exponencial - GT

```
> set.seed(1234)
> # Prevalencia a 5 años de GT estimada por remuestreo y t.s.exp en G+T
> AC.5a.mirrexp.GT.2003 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2003,
+                                                fit.pop.F, nsim=1000)
> AC.5a.mirrexp.GT.2004 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2004,
+                                                fit.pop.F, nsim=1000)
> AC.5a.mirrexp.GT.2005 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2005,
+                                                fit.pop.F, nsim=1000)
> AC.5a.mirrexp.GT.2006 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2006,
+                                                fit.pop.F, nsim=1000)
> AC.5a.mirrexp.GT.2007 <- prev.AC.V.sim.mirrexp(datos.II, in.GT, 2007,
+                                                fit.pop.F, nsim=1000)
> #
> # Gráficos GT estim. mirr-exp en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.GT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.mirrexp.GT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.mirrexp.GT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.mirrexp.GT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.mirrexp.GT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Mirror Exponencial - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
```

Método Mirror-Uniforme - GT

```
> set.seed(1234)
> #Prevalencia 5 años GT estim. por metodo mirror-unif sobre datos de G+T
> AC.5a.mirrunif.GT.2003 <- prev.AC.V.sim.mirrunif(datos.II, in.GT, 2003,
+                                                  fit.pop.F, nsim=1000)
> AC.5a.mirrunif.GT.2004 <- prev.AC.V.sim.mirrunif(datos.II, in.GT, 2004,
+                                                  fit.pop.F, nsim=1000)
> AC.5a.mirrunif.GT.2005 <- prev.AC.V.sim.mirrunif(datos.II, in.GT, 2005,
+                                                  fit.pop.F, nsim=1000)
> AC.5a.mirrunif.GT.2006 <- prev.AC.V.sim.mirrunif(datos.II, in.GT, 2006,
+                                                  fit.pop.F, nsim=1000)
> AC.5a.mirrunif.GT.2007 <- prev.AC.V.sim.mirrunif(datos.II, in.GT, 2007,
+                                                  fit.pop.F, nsim=1000)
> #
> # Gráficos GT estim. mirr-unif en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.GT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.mirrunif.GT.2004, ylim.t=c(0,100),
```

```
+              title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.mirrunif.GT.2005, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.mirrunif.GT.2006, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.mirrunif.GT.2007, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2007")
> title('Método Mirror Uniforme - Girona y Tarragona', outer=TRUE,
+       col.main='steelblue')
```

Método Supervivencia Empírica KM (o NA) - GT

```
> # Prevalencia a 5 años estim. en GT en base a superv.emp. en G+T
> ## con KM
> set.seed(1234)
> AC.5a.sup.GT.2003 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                         in.GT, 2003,fit.pop.F,nsim=1000)
> AC.5a.sup.GT.2004 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                         in.GT, 2004,fit.pop.F,nsim=1000)
> AC.5a.sup.GT.2005 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                         in.GT, 2005,fit.pop.F,nsim=1000)
> AC.5a.sup.GT.2006 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                         in.GT, 2006,fit.pop.F,nsim=1000)
> AC.5a.sup.GT.2007 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                         in.GT, 2007,fit.pop.F,nsim=1000)
> #
> # Gráficos GT estimada por superv.KM en G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2003, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2004, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2005, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2006, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2007, ylim.t=c(0,100),
+              title="Prevalencia 5 años: 2007")
> title('Método Supervivencia empírica (Kaplan-Meier) - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
> ## con NA ##
> set.seed(1234)
> AC.5a.sup.GT.2003.na <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2003,fit.pop.F,'NA',nsim=1000)
> AC.5a.sup.GT.2004.na <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2004,fit.pop.F,'NA',nsim=1000)
> AC.5a.sup.GT.2005.na <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2005,fit.pop.F,'NA',nsim=1000)
> AC.5a.sup.GT.2006.na <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2006,fit.pop.F,'NA',nsim=1000)
```

```
> AC.5a.sup.GT.2007.na <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2007,fit.pop.F,'NA',nsim=1000)
```

```
> # Gráficos GT estimada por superv.NA en G+T
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2003.na, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2004.na, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2005.na, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2006.na, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.sup.GT.2007.na, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Supervivencia empírica (NA) - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
```

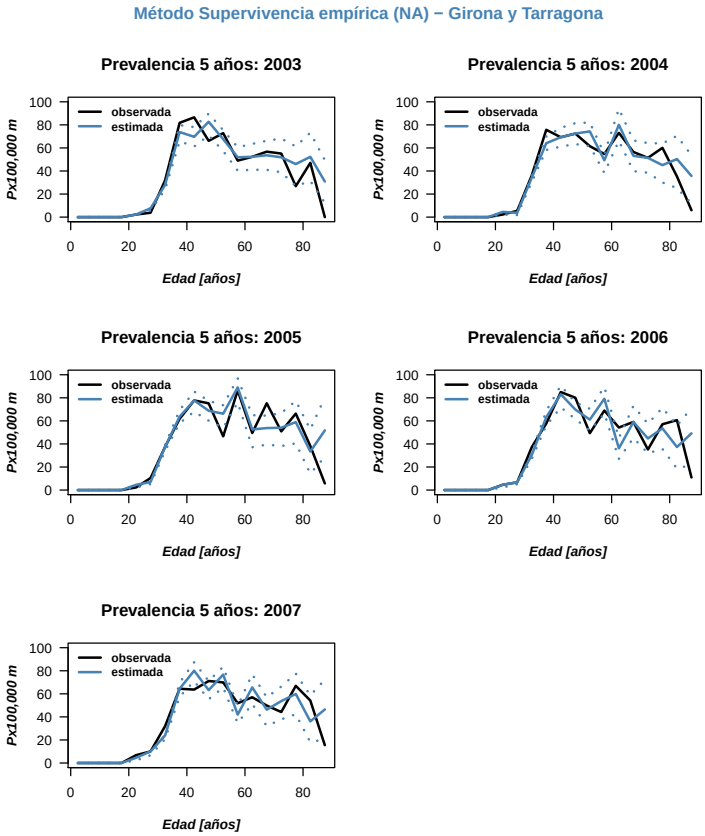

Figura B.1: Prevalencia a 5 años de cáncer de cuello de útero en la población femenina de Girona y Tarragona para cada año entre 2003 y 2007, observada (línea negra) y estimada (línea azul) a partir del método "Supervivencia NA". Los percentiles 2.5 y 97.5 % de prevalencia simulada están representados por las líneas punteadas de color azul.

Método Tiempo de Muerte Exponencial - GT

```
> # Por simulación de casos de GT obtengo prevalencia a 5 años estim. en GT
> # de acuerdo a modelos de supervivencia exponenciales ajustados en G+T
> # estimacion de parámetros de modelo exponencial por métodos gráficos
> set.seed(1234)
> AC.5a.texp.GT.2003.ols <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                              in.GT,2003,fit.pop.F,nsim=1000)
> AC.5a.texp.GT.2004.ols <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                              in.GT,2004,fit.pop.F,nsim=1000)
> AC.5a.texp.GT.2005.ols <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                              in.GT,2005,fit.pop.F,nsim=1000)
> AC.5a.texp.GT.2006.ols <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                              in.GT,2006,fit.pop.F,nsim=1000)
> AC.5a.texp.GT.2007.ols <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                              in.GT,2007,fit.pop.F,nsim=1000)
> #
> # estimacion de parámetros de modelo exponencial por ML
> set.seed(1234)
> AC.5a.texp.GT.2003.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                             2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.GT.2004.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                             2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.GT.2005.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                             2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.GT.2006.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                             2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.GT.2007.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                             2007,fit.pop.F,'ML',nsim=1000)
> #
> # Gráficos prev 5 años estim. GT en base a mod exp ajustaso en G+T
> par(mfrow=c(5,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2003.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2004.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2005.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2006.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2007.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
```

```
> plot.prev.sim(data.tmp=AC.5a.texp.GT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Tiempo de Muerte Exponencial (gráfico/ML) - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
```

Método Tiempo de Muerte Weibull - GT

```
> # Prevalencia a 5 años estimada para GT en base a modelos de superv. Weibull
> # ajustados sobre G+T
> # parámetros de weibull ajustados por métodos gráficos
> set.seed(1234)
> AC.5a.wei.GT.2003.ols <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                             in.GT, 2003,fit.pop.F,nsim=1000)
> AC.5a.wei.GT.2004.ols <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                             in.GT, 2004,fit.pop.F,nsim=1000)
> AC.5a.wei.GT.2005.ols <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                             in.GT, 2005,fit.pop.F,nsim=1000)
> AC.5a.wei.GT.2006.ols <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                             in.GT, 2006,fit.pop.F,nsim=1000)
> AC.5a.wei.GT.2007.ols <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                             in.GT, 2007,fit.pop.F,nsim=1000)
> # parámetros de weibull ajustados por ML
> set.seed(1234)
> AC.5a.wei.GT.2003.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.GT.2004.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.GT.2005.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.GT.2006.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.GT.2007.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                            in.GT, 2007,fit.pop.F,'ML',nsim=1000)
> #
> # Gráficos prev 5 años GT estimada en base a mod.Weibull ajust en G+T
> par(mfrow=c(5,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2003.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2004.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2005.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2006.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
```

```
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2007.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> plot.prev.sim(data.tmp=AC.5a.wei.GT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Tiempo de Muerte Weibull (gráfico/ML) - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
```

Método Tiempo de Muerte log-Logístico - GT

```
> # Prevalencia a 5 años estimada para GT desde modelos log-L ajustados en G+T
> # estimacion de parámetros log-L métodos gráficos
> set.seed(1234)
> AC.5a.tllog.GT.2003.ols <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT, 2003,fit.pop.F, nsim=1000)
> AC.5a.tllog.GT.2004.ols <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT, 2004,fit.pop.F,nsim=1000)
> AC.5a.tllog.GT.2005.ols <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT, 2005,fit.pop.F,nsim=1000)
> AC.5a.tllog.GT.2006.ols <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT, 2006,fit.pop.F,nsim=1000)
> AC.5a.tllog.GT.2007.ols <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT, 2007,fit.pop.F,nsim=1000)
> # estimacion de parámetros log-L por ML
> AC.5a.tllog.GT.2003.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.GT.2004.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.GT.2005.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.GT.2006.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.GT.2007.ml <- prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT, 2007,fit.pop.F,'ML',nsim=1000)
> #
> # Gráficos prev.5 años GT estimada tras ajuste de superv. log-L en G+T
> par(mfrow=c(5,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2003.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2004.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2005.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005")
```

```
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2006.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2007.ols, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> plot.prev.sim(data.tmp=AC.5a.tllog.GT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007")
> title('Método Tiempo de Muerte log-logístico (gráfico/ML) - Girona y Tarragona',
+       outer=TRUE, col.main='steelblue')
>
```

Razón de Discrepancia y Coste computacional

```
> #### Función para calcular Razón de Discrepancia por media por año.
>
> DR <- function(result){
+
+   nums <- abs(with(result[1:18,], Px100000.emp-Px100000))
+   dens <- 1.96*(with(result[1:18,], P_sd+0.001))
+   matriz <- data.frame(nums, dens, r_int=nums/dens)
+   print(matriz)
+   dr <- sum(matriz$r_int)/18
+   print(cat('DR=',dr,fill = TRUE))
+   dr
+ }
> DR.03.s <- DR(AC.5a.tse.GT.2003)
> DR.04.s <- DR(AC.5a.tse.GT.2004)
> DR.05.s <- DR(AC.5a.tse.GT.2005)
> DR.06.s <- DR(AC.5a.tse.GT.2006)
> DR.07.s <- DR(AC.5a.tse.GT.2007)
> DR.s.T <- DR.03.s+DR.04.s+DR.05.s+DR.06.s+DR.07.s
> DR.s.T
> DR.03.m <- DR(AC.5a.mirr.GT.2003)
> DR.04.m <- DR(AC.5a.mirr.GT.2004)
> DR.05.m <- DR(AC.5a.mirr.GT.2005)
> DR.06.m <- DR(AC.5a.mirr.GT.2006)
> DR.07.m <- DR(AC.5a.mirr.GT.2007)
> DR.m.T <- DR.03.m+DR.04.m+DR.05.m+DR.06.m+DR.07.m
> DR.m.T
> DR.03.me <- DR(AC.5a.mirrexp.GT.2003)
> DR.04.me <- DR(AC.5a.mirrexp.GT.2004)
> DR.05.me <- DR(AC.5a.mirrexp.GT.2005)
> DR.06.me <- DR(AC.5a.mirrexp.GT.2006)
> DR.07.me <- DR(AC.5a.mirrexp.GT.2007)
> DR.me.T <- DR.03.me+DR.04.me+DR.05.me+DR.06.me+DR.07.me
> DR.me.T
> DR.03.mu <- DR(AC.5a.mirrunif.GT.2003)
> DR.04.mu <- DR(AC.5a.mirrunif.GT.2004)
> DR.05.mu <- DR(AC.5a.mirrunif.GT.2005)
```

```
> DR.06.mu <- DR(AC.5a.mirrunif.GT.2006)
> DR.07.mu <- DR(AC.5a.mirrunif.GT.2007)
> DR.mu.T <- DR.03.mu+DR.04.mu+DR.05.mu+DR.06.mu+DR.07.mu
> DR.mu.T
> #KM
> DR.03.sup <- DR(AC.5a.sup.GT.2003)
> DR.04.sup <- DR(AC.5a.sup.GT.2004)
> DR.05.sup <- DR(AC.5a.sup.GT.2005)
> DR.06.sup <- DR(AC.5a.sup.GT.2006)
> DR.07.sup <- DR(AC.5a.sup.GT.2007)
> DR.sup.T <- DR.03.sup+DR.04.sup+DR.05.sup+DR.06.sup+DR.07.sup
> DR.sup.T
> #NA
> DR.03.sup.na <- DR(AC.5a.sup.GT.2003.na)
> DR.04.sup.na <- DR(AC.5a.sup.GT.2004.na)
> DR.05.sup.na <- DR(AC.5a.sup.GT.2005.na)
> DR.06.sup.na <- DR(AC.5a.sup.GT.2006.na)
> DR.07.sup.na <- DR(AC.5a.sup.GT.2007.na)
> DR.sup.T.na <- DR.03.sup.na+DR.04.sup.na+DR.05.sup.na+DR.06.sup.na+
+   DR.07.sup.na
> DR.sup.T.na
> DR.03.texp.ols <- DR(AC.5a.texp.GT.2003.ols)
> DR.04.texp.ols <- DR(AC.5a.texp.GT.2004.ols)
> DR.05.texp.ols <- DR(AC.5a.texp.GT.2005.ols)
> DR.06.texp.ols <- DR(AC.5a.texp.GT.2006.ols)
> DR.07.texp.ols <- DR(AC.5a.texp.GT.2007.ols)
> DR.texp.ols.T <- DR.03.texp.ols+DR.04.texp.ols+DR.05.texp.ols+
+   DR.06.texp.ols+DR.07.texp.ols
> DR.texp.ols.T
> DR.03.texp.ml <- DR(AC.5a.texp.GT.2003.ml)
> DR.04.texp.ml <- DR(AC.5a.texp.GT.2004.ml)
> DR.05.texp.ml <- DR(AC.5a.texp.GT.2005.ml)
> DR.06.texp.ml <- DR(AC.5a.texp.GT.2006.ml)
> DR.07.texp.ml <- DR(AC.5a.texp.GT.2007.ml)
> DR.texp.ml.T <- DR.03.texp.ml+DR.04.texp.ml+DR.05.texp.ml+
+   DR.06.texp.ml+DR.07.texp.ml
> DR.texp.ml.T
> DR.03.wei.ols <- DR(AC.5a.wei.GT.2003.ols)
> DR.04.wei.ols <- DR(AC.5a.wei.GT.2004.ols)
> DR.05.wei.ols <- DR(AC.5a.wei.GT.2005.ols)
> DR.06.wei.ols <- DR(AC.5a.wei.GT.2006.ols)
> DR.07.wei.ols <- DR(AC.5a.wei.GT.2007.ols)
> DR.wei.ols.T <- DR.03.wei.ols+DR.04.wei.ols+DR.05.wei.ols+DR.06.wei.ols+
+   DR.07.wei.ols
> DR.wei.ols.T
> DR.03.wei.ml <- DR(AC.5a.wei.GT.2003.ml)
> DR.04.wei.ml <- DR(AC.5a.wei.GT.2004.ml)
> DR.05.wei.ml <- DR(AC.5a.wei.GT.2005.ml)
> DR.06.wei.ml <- DR(AC.5a.wei.GT.2006.ml)
> DR.07.wei.ml <- DR(AC.5a.wei.GT.2007.ml)
```

```
> DR.wei.ml.T <- DR.03.wei.ml+DR.04.wei.ml+DR.05.wei.ml+DR.06.wei.ml+
+   DR.07.wei.ml
> DR.wei.ml.T
> DR.03.tllog.ols <- DR(AC.5a.tllog.GT.2003.ols)
> DR.04.tllog.ols <- DR(AC.5a.tllog.GT.2004.ols)
> DR.05.tllog.ols <- DR(AC.5a.tllog.GT.2005.ols)
> DR.06.tllog.ols <- DR(AC.5a.tllog.GT.2006.ols)
> DR.07.tllog.ols <- DR(AC.5a.tllog.GT.2007.ols)
> DR.tllog.ols.T <- DR.03.tllog.ols+DR.04.tllog.ols+DR.05.tllog.ols+
+   DR.06.tllog.ols+DR.07.tllog.ols
> DR.tllog.ols.T
> DR.03.tllog.ml <- DR(AC.5a.tllog.GT.2003.ml)
> DR.04.tllog.ml <- DR(AC.5a.tllog.GT.2004.ml)
> DR.05.tllog.ml <- DR(AC.5a.tllog.GT.2005.ml)
> DR.06.tllog.ml <- DR(AC.5a.tllog.GT.2006.ml)
> DR.07.tllog.ml <- DR(AC.5a.tllog.GT.2007.ml)
> DR.tllog.ml.T <- DR.03.tllog.ml+DR.04.tllog.ml+DR.05.tllog.ml+
+   DR.06.tllog.ml+DR.07.tllog.ml
> DR.tllog.ml.T
> ####
>
> # Coste en tiempo de cálculo
> set.seed(1234)
> t.tse <- system.time(prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                        in.GT,2003,fit.pop.F,nsim=1000))
> set.seed(1234)
> t.mirr <- system.time(prev.AC.V.sim.mirr(datos.II,in.GT,2003,
+                                          fit.pop.F,nsim=1000))
> set.seed(1234)
> t.mirr.exp <- system.time(prev.AC.V.sim.mirrexp(datos.II,in.GT,2003,
+                                                 fit.pop.F, nsim=1000))
> set.seed(1234)
> t.mirr.unif <- system.time(prev.AC.V.sim.mirrunif(datos.II,in.GT,2003,
+                                                   fit.pop.F,nsim=1000))
> set.seed(1234)
> t.supKM <- system.time(prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                           in.GT,2003,fit.pop.F,nsim=1000))
> set.seed(1234)
> t.supNA <- system.time(prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                           in.GT, 2003,fit.pop.F,'NA',nsim=1000))
> set.seed(1234)
> t.texp.ols <- system.time(prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                              2003,fit.pop.F, nsim=1000))
> set.seed(1234)
> t.texp.tml <- system.time(prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),in.GT,
+                                              2003,fit.pop.F,'ML',nsim=1000))
> set.seed(1234)
> t.wei.ols <- system.time(prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),in.GT,
+                                             2003,fit.pop.F,nsim=1000))
> set.seed(1234)
```

```
> t.wei.ml <- system.time(prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),in.GT,
+                                            2003,fit.pop.F,'ML',nsim=1000))
> set.seed(1234)
> t.tllog.ols <- system.time(prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                                in.GT,2003,fit.pop.F,nsim=1000))
> set.seed(1234)
> t.tllog.ml <- system.time(prev.AC.V.sim.tllog(datos.II, c(35,45,55,65,75),
+                                               in.GT,2003,fit.pop.F,'ML',nsim=1000))

> t.tse

   user  system elapsed
 515.70    0.17  516.28

> t.mirr

   user  system elapsed
  40.59    0.01   40.62

> t.mirr.exp

   user  system elapsed
  79.75    0.14   79.72

> t.mirr.unif

   user  system elapsed
  84.93    0.20   84.97

> t.supKM

   user  system elapsed
 439.14    0.20  439.75

> t.supNA

   user  system elapsed
 440.18    0.22  440.95

> t.texp.ols

   user  system elapsed
1404.57    0.31 1406.89

> t.texp.tml

   user  system elapsed
2208.35    0.22 2210.26

> t.wei.ols

   user  system elapsed
1419.58    0.15 1419.97

> t.wei.ml

   user  system elapsed
2245.92    0.16 2247.76

> t.tllog.ols
```

```
   user  system elapsed
1447.85    0.26 1449.45
```

```
> t.tllog.ml
```

```
   user  system elapsed
2218.96    0.22 2221.13
```

B.3. Estimaciones de prevalencia a 5 años en Cataluña

Método Tiempo de Seguimiento Empírico - CAT

```
> set.seed(1234)
> # Prevalencia acumulada a 5 años de 2003 CAT estimada usando distirb de G+T
> AC.5a.tse.CAT.2003 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                         fit.incid.Cat,2003,fit.pop.F,nsim=1000)
> # Prevalencia acumulada a 5 años de 2004 CAT estimada usando distirb de G+T
> AC.5a.tse.CAT.2004 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                         fit.incid.Cat,2004,fit.pop.F,nsim=1000)
> # Prevalencia acumulada a 5 años de 2005 CAT estimada usando distirb de G+T
> AC.5a.tse.CAT.2005 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                         fit.incid.Cat,2005,fit.pop.F,nsim=1000)
> # Prevalencia acumulada a 5 años de 2006 CAT estimada usando distirb de G+T
> AC.5a.tse.CAT.2006 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                         fit.incid.Cat,2006,fit.pop.F,nsim=1000)
> # Prevalencia acumulada a 5 años de 2007 CAT estimada usando distirb de G+T
> AC.5a.tse.CAT.2007 <- prev.AC.V.sim.tse(datos.II, c(35,45,55,65,75),
+                                         fit.incid.Cat,2007,fit.pop.F,nsim=1000)
```

```
> # Fig. DE estimaciones para CATALUNYA TSEmp en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.tse.CAT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tse.CAT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tse.CAT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tse.CAT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tse.CAT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Tiempo de Seguimiento Empírico - Cataluña',
+       outer=TRUE, col.main='steelblue')
```

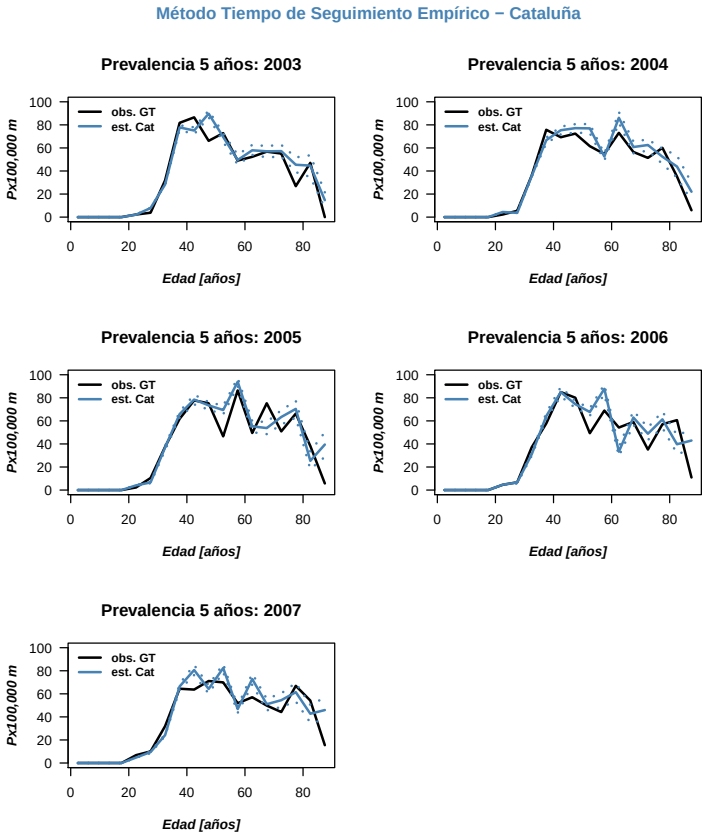

Figura B.2: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 estimada mediante el método de simulación de cohortes "Tiempo de Seguimiento Empírico", basado en la distribución del tiempo de seguimiento observado en seis grupos de edad de la población diagnosticada por esta enfermedad en Girona y Tarragona entre 1999 y 2007, seguida hasta junio de 2010. En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

Método Mirror - CAT

```
> # Prevalencias a 5 años en CATALUÑA por met mirror
> set.seed(1234)
> AC.5a.mirr.CAT.2003 <- prev.AC.V.sim.mirr(datos.II, fit.incid.Cat,
+                                           2003, fit.pop.F, nsim=1000)
> AC.5a.mirr.CAT.2004 <- prev.AC.V.sim.mirr(datos.II, fit.incid.Cat,
+                                           2004, fit.pop.F, nsim=1000)
> AC.5a.mirr.CAT.2005 <- prev.AC.V.sim.mirr(datos.II, fit.incid.Cat,
+                                           2005, fit.pop.F, nsim=1000)
> AC.5a.mirr.CAT.2006 <- prev.AC.V.sim.mirr(datos.II, fit.incid.Cat,
+                                           2006, fit.pop.F, nsim=1000)
> AC.5a.mirr.CAT.2007 <- prev.AC.V.sim.mirr(datos.II, fit.incid.Cat,
+                                           2007, fit.pop.F, nsim=1000)
> #
> # Gráficos CAT estimada por mirror en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirr.CAT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirr.CAT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirr.CAT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirr.CAT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirr.CAT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Mirror - Cataluña', outer=TRUE, col.main='steelblue')
```

Método Mirror-Exponencial - CAT

```
> # Prevalencias a 5 años est. en CAT por met. mirr-exp en base a G+T
> set.seed(1234)
> AC.5a.mirrexp.CAT.2003 <- prev.AC.V.sim.mirrexp(datos.II, fit.incid.Cat,
+                                                 2003, fit.pop.F, nsim=1000)
> AC.5a.mirrexp.CAT.2004 <- prev.AC.V.sim.mirrexp(datos.II, fit.incid.Cat,
+                                                 2004, fit.pop.F, nsim=1000)
> AC.5a.mirrexp.CAT.2005 <- prev.AC.V.sim.mirrexp(datos.II, fit.incid.Cat,
+                                                 2005, fit.pop.F, nsim=1000)
> AC.5a.mirrexp.CAT.2006 <- prev.AC.V.sim.mirrexp(datos.II, fit.incid.Cat,
+                                                 2006, fit.pop.F, nsim=1000)
> AC.5a.mirrexp.CAT.2007 <- prev.AC.V.sim.mirrexp(datos.II, fit.incid.Cat,
+                                                 2007, fit.pop.F, nsim=1000)
```

```
> # Gráficos CAT estimada por mirror-exp en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.CAT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.CAT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.CAT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.CAT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrexp.CAT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Mirror Exponencial - Cataluña', outer=TRUE,
+       col.main='steelblue')
```

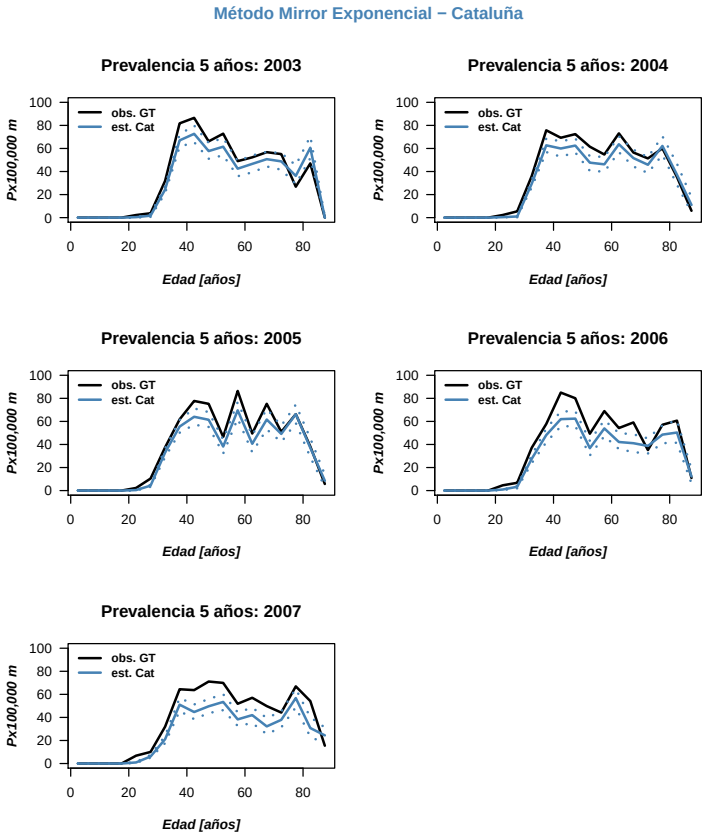

Figura B.3: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 estimada mediante el método de remuestreo sobre la base de Girona y Tarragona según año de incidencia e intervalo de edad, y posterior simulación del tiempo de seguimiento mediante distribuciones exponenciales específicas para el año y la edad, "Método Mirror Exponencial". En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

Método Mirror-Uniforme - CAT

```
> # Prevalencias a 5 años estim en CAT por met. mirror-unif en base a G+T
> set.seed(1234)
> AC.5a.mirrunif.CAT.2003 <- prev.AC.V.sim.mirrunif(datos.II, fit.incid.Cat,
+                                                   2003, fit.pop.F, nsim=1000)
> AC.5a.mirrunif.CAT.2004 <- prev.AC.V.sim.mirrunif(datos.II, fit.incid.Cat, 2004,
+                                                   fit.pop.F, nsim=1000)
> AC.5a.mirrunif.CAT.2005 <- prev.AC.V.sim.mirrunif(datos.II, fit.incid.Cat, 2005,
+                                                   fit.pop.F, nsim=1000)
> AC.5a.mirrunif.CAT.2006 <- prev.AC.V.sim.mirrunif(datos.II, fit.incid.Cat, 2006,
+                                                   fit.pop.F, nsim=1000)
> AC.5a.mirrunif.CAT.2007 <- prev.AC.V.sim.mirrunif(datos.II, fit.incid.Cat, 2007,
+                                                   fit.pop.F, nsim=1000)
```

```
> # Gráficos CAT estimada por mirror-unif en base a G+T
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.CAT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.CAT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.CAT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.CAT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.mirrunif.CAT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Mirror Uniforme - Cataluña', outer=TRUE,
+       col.main='steelblue')
```

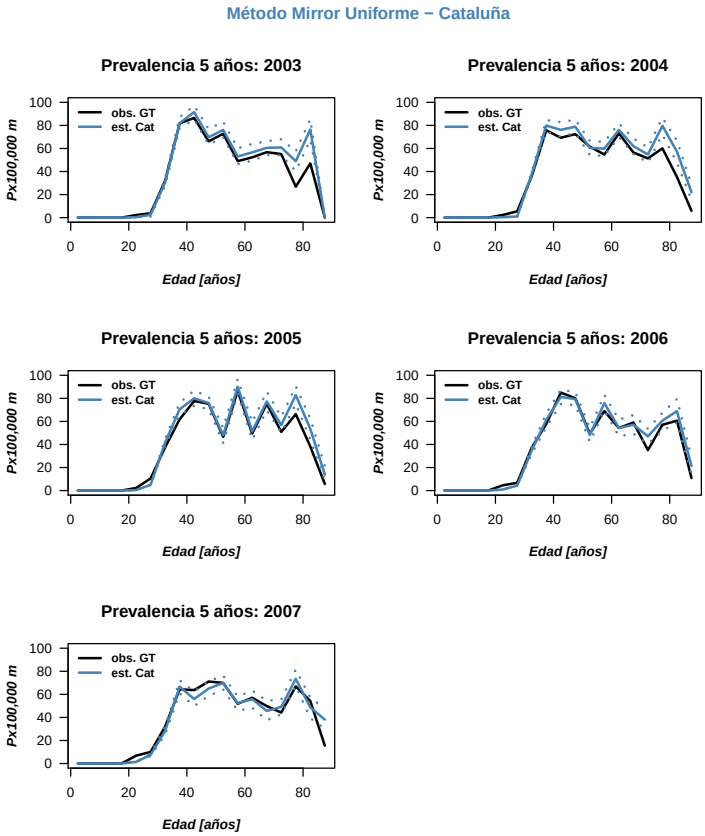

Figura B.4: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 estimada mediante el método de remuestreo sobre la base de Girona y Tarragona según año de incidencia e intervalo de edad, y posterior simulación del tiempo de seguimiento mediante distribuciones uniformes específicas del año y la edad, "Método Mirror Uniforme".

Método Supervivencia Empírica KM - CAT

```
> # Estim. prev. 5 años Catalunya con supervivencia empíricas KM de G+T
> set.seed(1234)
> AC.5a.sup.CAT.2003 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                  fit.incid.Cat, 2003,fit.pop.F,nsim=1000)
> AC.5a.sup.CAT.2004 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                  fit.incid.Cat, 2004,fit.pop.F,nsim=1000)
> AC.5a.sup.CAT.2005 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                  fit.incid.Cat, 2005,fit.pop.F,nsim=1000)
> AC.5a.sup.CAT.2006 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                  fit.incid.Cat, 2006,fit.pop.F,nsim=1000)
> AC.5a.sup.CAT.2007 <- prev.AC.V.sim.tsup(datos.II, c(35,45,55,65,75),
+                                  fit.incid.Cat, 2007,fit.pop.F,nsim=1000)
> #
> # Gráficos prev. a 5 años CAT estimada por superv. emp. KM en G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.sup.CAT.2003, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.sup.CAT.2004, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.sup.CAT.2005, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.sup.CAT.2006, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.sup.CAT.2007, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Supervivencia empírica (Kaplan-Meier) - Cataluña',
+       outer=TRUE, , col.main='steelblue')
```

Método Tiempo de Muerte Exponencial - CAT

```
> # Prev. 5 años Catalunya estim. por modelo superv. exp. ajustado en G+T
> # parámetros estimados por método gráfico
> set.seed(1234)
> AC.5a.texp.CAT.2003 <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2003,fit.pop.F,nsim=1000)
> AC.5a.texp.CAT.2004 <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2004,fit.pop.F,nsim=1000)
> AC.5a.texp.CAT.2005 <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2005,fit.pop.F,nsim=1000)
> AC.5a.texp.CAT.2006 <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2006,fit.pop.F,nsim=1000)
> AC.5a.texp.CAT.2007 <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2007,fit.pop.F,nsim=1000)
> # parámetros estimados por 'ML'
> set.seed(1234)
> AC.5a.texp.CAT.2003.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                     fit.incid.Cat, 2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.CAT.2004.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
```

```
+                                    fit.incid.Cat, 2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.CAT.2005.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                    fit.incid.Cat, 2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.CAT.2006.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                    fit.incid.Cat, 2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.texp.CAT.2007.ml <- prev.AC.V.sim.texp(datos.II, c(35,45,55,65,75),
+                                    fit.incid.Cat, 2007,fit.pop.F,'ML',nsim=1000)
```

```
> # Fig. estimaciones prev 5 años cATALUNYA (TMExp-ML en base a G+T)
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.texp.CAT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.texp.CAT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.texp.CAT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.texp.CAT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.texp.CAT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Tiempo de Muerte Exponencial (param. ML) - Cataluña',
+       outer=TRUE, col.main='steelblue')
```

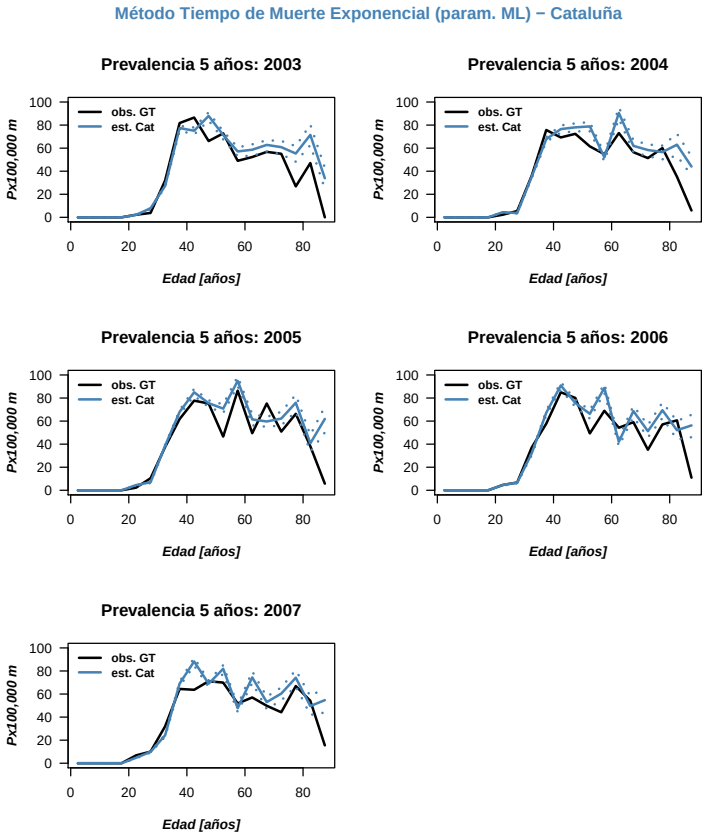

Figura B.5: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 realizadas en base a modelos de supervivencia con distribución Exponencial ajustados (mediante ML) en seis grupos de edad de la población de Girona y Tarragona, método "Tiempo de Muerte Exponencial". En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

Método Tiempo de Muerte Weibull - CAT

```
> # Prev. 5 años estim para Catalunya con weibull ajustadas en G+T
> # parámetros de weibull ajustados desde método gráfico
> set.seed(1234)
> AC.5a.wei.CAT.2003 <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                fit.incid.Cat, 2003,fit.pop.F,nsim=1000)
> AC.5a.wei.CAT.2004 <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                fit.incid.Cat, 2004,fit.pop.F,nsim=1000)
> AC.5a.wei.CAT.2005 <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                fit.incid.Cat, 2005,fit.pop.F,nsim=1000)
> AC.5a.wei.CAT.2006 <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                fit.incid.Cat, 2006,fit.pop.F,nsim=1000)
> AC.5a.wei.CAT.2007 <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                                fit.incid.Cat, 2007,fit.pop.F,nsim=1000)
> # parámetros de weibull ajustados por ML
> set.seed(1234)
> AC.5a.wei.CAT.2003.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                               fit.incid.Cat, 2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.CAT.2004.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                               fit.incid.Cat, 2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.CAT.2005.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                               fit.incid.Cat, 2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.CAT.2006.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                               fit.incid.Cat, 2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.wei.CAT.2007.ml <- prev.AC.V.sim.twei(datos.II, c(35,45,55,65,75),
+                               fit.incid.Cat, 2007,fit.pop.F,'ML',nsim=1000)
```

```
> # Fig. DE prev 5 años PARA cATALUNYA (TMWeibull-ML en base a G+T)
> par(mfrow=c(3,2), omi=c(0,0,0.3,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.wei.CAT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.wei.CAT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.wei.CAT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.wei.CAT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.wei.CAT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Tiempo de Muerte Weibull (param. ML) - Cataluña',
+       outer=TRUE, col.main='steelblue')
```

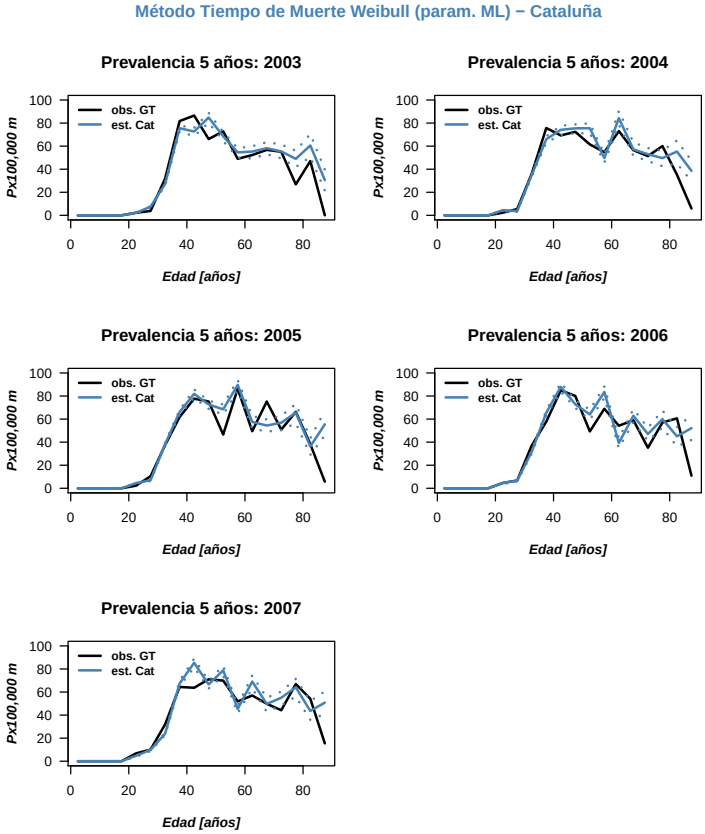

Figura B.6: Prevalencia a 5 años de cáncer de cuello de útero en Cataluña para cada año entre 2003 y 2007 realizadas en base a modelos de supervivencia con distribución Weibull ajustados (mediante ML) en seis grupos de edad de la población de Girona y Tarragona, método "Tiempo de Muerte Weibull". En negro se muestra la prevalencia observada en la población de Girona y Tarragona, y en azul la prevalencia estimada para Cataluña (mediana en línea continua y percentiles 2,5 y 97.5 % en línea punteada).

Método Tiempo de Muerte log-Logístico - CAT

```
> # Prevalencia 5 años estimada para Catalunya usando superv log-L ajustada en G+T
> # estimacion de parámetros de modelo log-L por métodos gráficos
> set.seed(1234)
> AC.5a.tllog.CAT.2003 <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2003,fit.pop.F,nsim=1000)
> AC.5a.tllog.CAT.2004 <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2004,fit.pop.F,nsim=1000)
> AC.5a.tllog.CAT.2005 <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2005,fit.pop.F,nsim=1000)
> AC.5a.tllog.CAT.2006 <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2006,fit.pop.F,nsim=1000)
> AC.5a.tllog.CAT.2007 <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2007,fit.pop.F,nsim=1000)
> # estimacion de parámetros de modelo log-L por ML
> set.seed(1234)
> AC.5a.tllog.CAT.2003.ml <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2003,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.CAT.2004.ml <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2004,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.CAT.2005.ml <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2005,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.CAT.2006.ml <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2006,fit.pop.F,'ML',nsim=1000)
> AC.5a.tllog.CAT.2007.ml <- prev.AC.V.sim.tllog(datos.II,c(35,45,55,65,75),
+                                  fit.incid.Cat, 2007,fit.pop.F,'ML',nsim=1000)
> #
> # Fig. estimacion PARA cATALUNYA en base a mod ajust (ML) en G+T
> par(mfrow=c(3,2), omi=c(0,0,0.5,0), font=2, font.lab=4, las=1)
> plot.prev.sim(data.tmp=AC.5a.tllog.CAT.2003.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2003",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tllog.CAT.2004.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2004",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tllog.CAT.2005.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2005",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tllog.CAT.2006.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2006",leg=2)
> plot.prev.sim(data.tmp=AC.5a.tllog.CAT.2007.ml, ylim.t=c(0,100),
+               title="Prevalencia 5 años: 2007",leg=2)
> title('Método Tiempo de Muerte log-Logístico (param. ML) - Cataluña',
+       outer=TRUE, col.main='steelblue')
```