Análisis demográfico de las biografías

Daniel Courgeau y Éva Lelièvre

EL COLEGIO DE MÉXICO
EMBAJADA DE FRANCIA EN MÉXICO

CENTRO DE ESTUDIOS DEMOGRÁFICOS
Y DE DESARROLLO URBANO

ANÁLISIS DEMOGRÁFICO DE LAS BIOGRAFÍAS

312.018
C975a

Courgeau, Daniel
Análisis demográfico de las biografías / Daniel Courgeau y Éva Lelièvre ; traducción del francés de Mariela Álvarez. -- México : El Colegio de México, Centro de Estudios Demográficos y de Desarrollo Urbano, 2001.

305 p. ; 22 cm.

ISBN 968-12-0968-0

1. Historia de vida, Análisis de. 2. Método longitudinal. 3. Regresión, Análisis de. I. Lelièvre, Éva, coaut.

Esta obra, editada en el marco del programa de apoyo a la publicación, ha contado con el financiamiento del Ministère des Affaires Etrangères y de la Embajada de Francia en México

Diseño de portada: Irma Eugenia Alva Valencia
Ilustración: *El día ilustrísimo,* Jorge de la Vega, 1964

Primera edición, 2001

Camino al Ajusco 20
Pedregal de Santa Teresa
10740 México, D.F.

ISBN 968-12-0968-0

Impreso en México

ANÁLISIS DEMOGRÁFICO DE LAS BIOGRAFÍAS

Daniel Courgeau y *Éva Lelièvre*

Traducción del francés de Mariela Álvarez

EL COLEGIO DE MÉXICO

ÍNDICE

Agradecimientos 11
Introducción 13

PRIMERA PARTE
GENERALIZACIÓN DEL ANÁLISIS LONGITUDINAL

I. La observación de biografías 21
A) Diversos tipos de encuestas 22
B) La encuesta "Triple biografía" 25
C) El problema de las ponderaciones 28
D) La observación incompleta y errónea de las encuestas 30
E) El problema de los truncamientos 36
F) Conclusión 38
II. Formalización del análisis 41
A) Análisis de una cohorte homogénea que experimenta un solo evento 42
1) Distribución de la ocurrencia de los eventos, notaciones y formalización 42
2) La función de verosimilitud 44
3) Tiempo discreto, tiempo continuo 45
B) Análisis de una cohorte heterogénea y de la interacción entre fenómenos 48
1) Demografía diferencial 49
2) Modelo markoviano o semimarkoviano 50
a) Formalización del modelo 51
b) Relación del modelo con los procesos de conteo poissonianos 53
c) Generalización del modelo 55
C) Hacia un análisis más completo de los comportamientos humanos 58
D) Conclusión 61
III. Métodos de estimación a partir de observaciones truncadas 63
A) Presentación de los truncamientos 63
B) Truncamientos a la derecha 65
1) Estimador de Kaplan-Meier 66
Ejemplos miniatura 70
2) Estimador de Aalen 72

C) Truncamientos a la izquierda 74
1) Un método de corrección 75
2) Resultados obtenidos 77
D) Conclusión 80
IV. Estudio de un evento 81
A) Muestra simple. Un evento único 81
1) Estimación en tiempo discreto 81
2) Estimación actuarial 82
3) Estimación de los cocientes acumulados 84
B) Muestra simple: riesgos múltiples 85
Ejemplos 87
C) Muestras múltiples: pruebas comparativas 87
Ejemplos 90
D) Conclusión 94
V. Estudio recíproco de las interacciones entre dos eventos 95
A) Concepción del análisis 95
1) Dos fenómenos en competencia 95
2) La diferenciación dinámica 96
3) Los diferentes tipos de dependencia 96
4) Dependencia unilateral y causalidad 99
B) Formalización del análisis bivariado 100
1) Estimación actuarial 102
2) Las pruebas posibles 103
3) Las pruebas seleccionadas y la convergencia de las estimaciones 104
C) Análisis práctico 106
1) Simultaneidades e intervalos de tiempo 108
2) Representaciones gráficas 110
D) Conclusión 111
VI. Generalización para situaciones más complejas 113
A) Presentación y límites de la aplicación práctica 113
B) Interacciones entre tres eventos: dos casos de estudio 114
C) Interacciones entre dos eventos con ocurrencias múltiples 118
D) Conclusión 120

SEGUNDA PARTE

GENERALIZACIÓN DE LOS MODELOS DE REGRESIÓN

VII. Formalización estadística del análisis paramétrico 123
A) Algunos modelos paramétricos útiles en demografía 123
1) Distribución exponencial 124
2) Mezcla de distribuciones exponenciales 126
3) Distribución de Gompertz 134

4) Distribución de Weibull 139
5) Distribución Gamma 141
6) Distribución log-normal 143
7) Distribución log-logística 144
8) Distribución de Fisher-Snedecor (F) generalizada 146
9) Comparación de las diversas distribuciones 148
B) Modelos de regresión 149
1) Modelos de riesgos proporcionales 150
a) Distribución exponencial 152
b) Distribución de Weibull 152
c) Distribución de Gompertz 153
d) Distribución log-logística 153
e) Verificación de la validez del modelo 153
2) Modelos de tiempo de ocurrencias aceleradas 156
a) Nexos entre los modelos de tiempo de ocurrencias aceleradas y los modelos de riesgos proporcionales 156
b) Distribución log-logística 157
c) Verificación de la validez del modelo 158
3) Modelos más complejos 159
C) Conclusión 159
VIII. Métodos de estimación de modelos paramétricos 161
A) Cálculo de la verosimilitud cuando ciertos intervalos están abiertos 162
B) Estimación de los parámetros y pruebas de su valor 164
C) Ejemplos de estimación de los parámetros 165
1) Modelo exponencial 165
2) Utilización del método de Newton-Raphson 173
3) Modelo de Weibull 174
4) Modelo de Gompertz 179
5) Modelo log-logístico de ocurrencias aceleradas 184
D) Comparación de modelos paramétricos 190
E) Conclusión 194
IX. Métodos de análisis semiparamétrico 197
A) De las regresiones paramétricas a los modelos de riesgos proporcionales semiparamétricos 197
1) Definición 197
2) Construcción de la verosimilitud 198
B) Métodos de estimación 199
1) Estimación de los parámetros 199
2) Estimación del componente no paramétrico 202
C) Algoritmo de Newton-Raphson 204
D) Selección de un modelo para el análisis de interacciones 205
E) Algunos ejemplos de aplicación 206

F) Conclusión 210
Conclusión general 211
A) El análisis de las interacciones entre fenómenos 211
1) Formalización teórica 212
2) Modelos no paramétricos 214
B) El tratamiento de la heterogeneidad de las poblaciones 216
1) Modelos paramétricos 217
2) Modelos semiparamétricos 218
C) Nuevas líneas de investigación 220
Anexo 1
I. Correspondencia de los términos en inglés, francés y español 223
II. Fórmulas básicas 225
III. Recordatorios y demostraciones 227
Producto integral 227
Prueba de una hipótesis nula, prueba general 228
Normalidad asintótica 229
Relación entre los procesos de conteo y el análisis de regresión de los intervalos 230
Planteo de Cox para la estimación de la verosimilitud parcial 231
Proceso de conteo y martingalas (Aalen) 232
IV. *Softwares* disponibles 233
Anexo 2 241
Anexo 3 279
Bibliografía por temas 291

AGRADECIMIENTOS

Ante todo queremos agradecer aquí a Philippe Antoine (ORSTOM), Didier Blanchet (INED), Catherine Bonvalet (INED), Marco Botai (Instituto de Estadística, Pisa, Italia), Michel Bozon (INED), Arnaud Bringé (INED), Christian Delbès (Fundación Nacional de Gerontología), Jean-Marie Firdion (INED), Fátima Juárez (El Colegio de México, México), Monique Lefebvre (INED), Henri Leridon (INED), Gilles Pison (INED), Denise Pumain (INED), Benoit Riandey (INED), Eric Sammartino (INED), Laurent Toulemon (INED) y Bernard Zarca (CREDOC), participantes en el Seminario de Formación de los Métodos de Análisis Biográfico (INED, 24 a 28 de agosto de 1987), quienes nos recomendaron calurosamente que presentáramos por escrito la exposición teórica y práctica que realizamos en dicho seminario.

De igual manera queremos agradecer a Marc Barbut (EHESS), Bui Trong Lieu (Universidad de París V), Hervé Le Bras (INED) y Denise Pumain (INED), quienes fueron los miembros del jurado de la tesis que presentó Éva Lelièvre en la Escuela de Altos Estudios de Ciencias Sociales el 23 de marzo de 1988, acerca de los "Métodos matemáticos y estadísticos para el análisis de historias de vida".

También ofrecemos nuestro agradecimiento a Jan Hoem (Universidad de Estocolmo, Suecia) por los numerosos intercambios escritos y verbales sobre el tema de esta obra.

Finalmente, reconocemos la labor de François Milan en la captura del manuscrito, a Nicole Berthoux por la elaboración de las figuras y gráficas y a Hella Courgeau por la compilación y revisión del material de este libro.

INTRODUCCIÓN

Hasta el presente, el análisis longitudinal básicamente se ha desarrollado considerando cada fenómeno demográfico por separado. El principal objetivo ha sido aislar cada uno de estos fenómenos en estado puro. En particular se ha considerado necesario separar la influencia del evento estudiado de la de los fenómenos perturbadores: la mortalidad y los movimientos migratorios. Para ello se ha planteado un cierto número de hipótesis, sin que realmente haya sido posible verificarlas con las fuentes existentes (Henry, 1959, 1966).

Hay que considerar que este análisis se ha desarrollado a partir de datos agregados, como las estadísticas del estado civil o, en el mejor de los casos, los datos de los registros de población. Si bien estas fuentes ofrecen la posibilidad de tratar cada evento eliminando el efecto perturbador de los otros, casi no permiten analizar las interacciones entre fenómenos. Por lo tanto, en los manuales de demografía clásica encontramos que los fenómenos están aislados en estado puro, y se tratan por separado en cada uno de los diferentes capítulos: nupcialidad, fecundidad, mortalidad, desplazamientos y migraciones (Pressat, 1961; Henry, 1972).

No obstante, algunos demógrafos han destacado la utilidad de analizar las interacciones entre los fenómenos demográficos. Así, Pressat (1966) hacía énfasis en que "la búsqueda de las correlaciones entre fenómenos demográficos, si bien todavía constituye un dominio inexplorado, debería poder enriquecer considerablemente nuestro conocimiento". Pressat, sin embargo, no indicaba ningún camino a seguir para el análisis de esas correlaciones. Asimismo, cuando Henry (1972) hizo un examen de la nupcialidad, indicaba que "en cuanto a los emigrantes, podríamos vernos tentados a sustituir su nupcialidad en el extranjero por la que habrían tenido si se hubieran quedado, siendo que tal nupcialidad en el extranjero depende de condiciones que pueden ser muy diferentes". Este reconocimiento de la interacción entre los dos fenómenos y el cambio de comportamiento respecto de la nupcialidad una vez que los individuos han emigrado, es de mucha validez. Sin embargo, ante la ausencia de datos, Henry no llevó más lejos su análisis.

Henry (1959) también abordó otro problema importante: la heterogeneidad. Si bien "en el caso de una cohorte homogénea, la historia estadística de los individuos que la componen es la misma que la historia estadística

de la cohorte", cuando trabajamos con una cohorte heterogénea ese resultado ya no se sostiene. Así, por ejemplo, en el caso más simple que es el de una población que puede descomponerse en dos subpoblaciones —cada una con una probabilidad diferente pero constante a lo largo del tiempo en que se experimenta el evento estudiado—, la población en conjunto no tendrá un cociente constante, esto es, la media de los cocientes de las dos subpoblaciones. Si tal es el caso en el instante inicial, al paso del tiempo se produce una selección que elimina de la población que está expuesta al riesgo a la subpoblación que tiene el cociente más elevado, lo que ocasiona que a partir de cierto tiempo el cociente de la población observada tienda hacia el de la subpoblación que tiene el cociente más bajo. Por supuesto que esta heterogeneidad podría ser mucho más compleja, y todo indica que es necesario estudiarla.

"Para saber exactamente cuál es el efecto de la heterogeneidad de los grupos humanos habrá que orientar las investigaciones de la demografía diferencial hacia las características individuales, físicas y psicológicas, preocupándose por estudiar a la vez la dispersión y la correlación de los índices demográficos en el interior de los grupos, bastante suscintos, hasta aquí considerados" (Henry, 1959).

Mientras el demógrafo siga utilizando las estadísticas del estado civil o de los registros de población que se publican habitualmente, no tendrá manera de abordar estos dos problemas fundamentales: el análisis de las interacciones entre los fenómenos demográficos y el análisis de la heterogeneidad de los grupos humanos. Por lo tanto, hay que valerse de otras fuentes que permitan una observación continua de un grupo de individuos a lo largo de toda su existencia, o al menos de una parte de ella, y de la recolección del mayor número de características en cada encuesta.

Vemos entonces que la unidad de análisis ya no será el evento (deceso, matrimonio, nacimiento, migración, etc.) sino la biografía individual, considerada como un proceso complejo. En la actualidad ya no sólo se pretende aislar cada fenómeno en estado puro sino, por el contrario, se intenta ver cómo un evento en una existencia puede influir sobre la continuación de la vida individual y cómo ciertas características pueden empujar a un individuo a que se comporte de manera diferente a otro.

Este cambio de perspectiva nos conduce a reformular las bases del análisis demográfico en términos del examen de procesos estocásticos complejos. Tratemos de ver con más precisión cómo llegar a eso.

Esos procesos no se producen en un espacio-tiempo abstracto, sino que tienen su fuente en una estructura social particular. Un individuo nacido en una tribu Lobi a comienzos del siglo xx tendrá una biografía de un tipo completamente diferente a la de un individuo nacido en la Francia campesina de la misma época, o a la de uno nacido en la Francia urbana actual. Sin

embargo, en cada una de esas estructuras sociales podemos distinguir sistemas de relaciones que están más o menos desarrollados según el grupo o la sociedad considerados (Kimball y Pearsall, 1954): sistemas familiares, económicos, políticos, religiosos, educativos, asociativos e informales. Claro que nada impide que en el futuro aparezcan nuevos tipos de sistemas de relaciones. Nuestro acercamiento no considera a la sociedad como algo cerrado sino, al contrario, en evolución perpetua.

Cada miembro de una sociedad dada está simultáneamente implicado en los diversos sistemas. Así, un individuo que vive en la actualidad en Francia puede estar implicado en el sistema familiar como cohabitante y padre de un niño; en el sistema económico como ingeniero en la industria automotriz; en el sistema político como consejero municipal de su comuna; en el sistema religioso como católico no practicante; en el sistema educativo como alguien que continúa con cursos de perfeccionamiento de su especialidad; en el sistema asociativo como miembro de un equipo de futbol no profesional y, por último, en el sistema informal como participante ocasional en las reuniones de padres de alumnos, con el objeto de resolver los problemas de la educación de su hijo.

Es la interacción entre esos diversos tipos de participaciones lo que va a generar un espacio y un tiempo propios para cada situación. La movilidad espacial o profesional de un soltero quizás sea más frecuente y cubra distancias más grandes que la de un individuo casado, sobre todo si éste tiene uno o varios hijos. En efecto, este último está ligado a su lugar de residencia y a su trabajo por restricciones debidas al lugar de trabajo de su esposa, al lugar donde van a la escuela sus hijos, etcétera.

El análisis de biografías pretende entonces situar esos cambios en el tiempo y el espacio vividos por los individuos en el marco de su sociedad. De lo que se trata es de ver cómo un acontecimiento familiar, económico o de otro tipo que enfrenta un individuo, modificará la probabilidad de que se produzcan otros eventos en su existencia. Se tratará de ver, por ejemplo, cómo su matrimonio puede influir en su carrera profesional, su movilidad espacial, la aparición de otros eventos como el nacimiento de un niño, o la ruptura con su familia de origen, etcétera.

Llegamos así al análisis de las interacciones entre fenómenos demográficos que tanto nos interesa. El área de estudio de este análisis el la de las biografías individuales.

Si tratamos de comprender los comportamientos de un individuo habremos de interesarnos por sus orígenes sociales y toda su historia pasada. Aquí partimos del supuesto de que esos comportamientos no son innatos sino que se modifican a lo largo de la existencia individual, gracias a las experiencias personales y a sucesivas adquisiciones. Así, dos individuos con el mismo origen social pero que hayan seguido caminos completamente

distintos, tendrán un comportamiento frente al matrimonio, la constitución de la familia, la carrera profesional, etc. que variará a lo largo del tiempo.

Con esto llegamos al análisis de la heterogeneidad de las poblaciones, aunque visto en forma dinámica y no estática. Este análisis, que ocupa un lugar distinguido en el estudio de las biografías individuales, no es determinista sino probabilista desde sus inicios.

Así pues, varios individuos colocados en una misma situación conflictiva tendrán desde el principio probabilidades diferentes de encontrar soluciones a esta situación antes de una cierta fecha. Algunos quizás no las encuentren jamás, otros podrán inventar un comportamiento totalmente nuevo que quizás resuelva mejor la situación de conflicto en la que se encontraban. Le dejamos así al individuo un margen de libertad que puede conducir a situaciones completamente nuevas. Este margen de libertad permite una evolución de la humanidad no inscrita en el origen, sino desarrollada de forma irreversible a lo largo del tiempo.

En este sentido nos encontramos muy cerca de Prigogine y Stengers (1988), quienes reconocían que "el evento crea una diferencia entre el pasado y el futuro... es el producto inteligible de un pasado del que, sin embargo, no podría deducirse. Abre hacia un futuro histórico donde se decidirá si sus consecuencias son insignificantes o tienen sentido". Encontramos aquí ese margen de libertad que puede conducir a situaciones completamente nuevas.

¿Podemos entonces formalizar este análisis de las biografías que inicialmente presentamos de manera informal?

Cuando un individuo nace, su existencia puede seguir una gran variedad de trayectorias. A pesar de ello, esas diversas trayectorias distan de ser todas igualmente probables. La biografía de un individuo puede definirse entonces como el resultado de un proceso estocástico complejo que se desarrolla a lo largo del tiempo, pero que se sitúa en condiciones históricas, geográficas, económicas y sociales dadas.

Llamemos Ω_θ al conjunto de las historias de vida o partes de historias de vida que podríamos observar hasta un instante θ. Como ya dijimos, debemos limitarnos a la observación del pasado, pues el futuro introduce situaciones nuevas no deducibles del pasado. Por ejemplo, antes de la aparición del automóvil no se podía prever que surgiera la profesión de mecánico, y quizás en un futuro cercano los descubrimientos genéticos extenderán la duración de la vida humana hasta los 150 años o más. El análisis que se puede realizar en el instante t no toma en cuenta más que los comportamientos pasados, y los proyecta hacia el porvenir sin tomar en cuenta su evolución. Ésta, sin embargo, se produce a un ritmo lo bastante lento como para que el análisis del pasado esclarezca las probabilidades de los diversos eventos del futuro próximo.

Sea ω_n una biografía enteramente observada, donde n precede a θ y ω_θ una biografía observada hasta el instante θ, que no está terminada. Se puede decir que los eventos de una u otra de esas biografías son variables definidas sobre el espacio general Ω_θ. Por ejemplo, la edad en la que se casa uno de esos individuos es una aplicación (medida de probabilidad) de Ω_θ sobre la abscisa positiva $([0,+\infty])$.

Entonces dotemos a Ω_θ de una sigma-álgebra[1] $\mathcal{B}_\theta$ de los eventos que deseamos analizar en una población dada y de una *aplicación* (medida de probabilidad) P_θ de $\mathcal{B}_\theta$ sobre (0, 1), que nos va a indicar las probabilidades de conocer esos diferentes eventos en la población estudiada. En ese caso $(\Omega_\theta, \mathcal{B}_\theta, P_\theta)$ define bien un campo de probabilidad.

Un instante aleatorio T será entonces una función de tiempo en el espacio de probabilidad de Ω_θ, que en este caso se puede extender más allá de θ, suponiendo que las probabilidades definidas antes de θ permanezcan las mismas en el curso del tiempo. Así, por ejemplo, si un individuo no está casado en θ, se supone que su edad al casarse es una variable aleatoria T que sigue la misma distribución que la observada para los individuos ya casados con características semejantes a las de él. Esta hipótesis ni siquiera es necesaria, y podemos contentarnos con trabajar sobre las biografías completamente terminadas, como se hace en la demografía histórica.

El análisis de las biografías que proponemos aquí consistirá, pues, en estimar la distribución de las probabilidades de las trayectorias seguidas por una población dada. Esta distribución puede variar de una subpoblación a otra y depender de ciertas cualidades de los individuos de la subpoblación (características sociales y económicas de los padres y de los abuelos, por ejemplo). Estas trayectorias se identifican por variables aleatorias T_1, T_2, ... T_n, que representan las duraciones de permanencia en los diversos estados que las componen. Claro está que esas variables no son independientes y la distribución de las trayectorias que queremos estimar resulta de su distribución conjunta.

Nuestro acercamiento supone, pues, que el comportamiento de los individuos se puede describir como un proceso estocástico complejo. Una vez admitido ese modelo, se comienza por la estimación estadística de la distribución de las variables anteriormente definidas, que desarrollamos en esta obra. Habiendo ya conocido esas distribuciones, será posible acercarse a la distribución más compleja del conjunto de la trayectoria.

Partiremos ante todo de los métodos para conseguir esas biografías. Por lo general no se puede disponer de una recolección exhaustiva de las biografías, sino de datos recogidos de una muestra de individuos cuya biografía

[1] Conjunto de las partes Ω_θ.

completa no se observa. Por lo tanto, tendremos que evaluar los diversos problemas que nos plantea esta observación incompleta.

A continuación formalizaremos el análisis y los métodos de estimación partiendo de casos simples para irlos complicando progresivamente. Del estudio de un evento pasaremos al estudio recíproco de las interacciones entre dos eventos, antes de generalizar esos modelos no paramétricos hacia situaciones más complejas. Pasaremos luego a los modelos paramétricos, los cuales permiten hacer que intervenga el efecto de un gran número de características sobre la duración de la permanencia en un estado dado. Por último, introduciremos modelos semiparamétricos que combinan los dos acercamientos precedentes.

Estos diversos métodos se ilustrarán mediante su aplicación a situaciones muy diferentes con la intención de mostrar su generalidad. Como anexo se incluyen ciertos programas que permiten realizar esos análisis, a fin de que el lector del libro pueda utilizarlos.

Este manual suministra, a la vez, una presentación teórica detallada de los métodos de análisis de las biografías y una posibilidad de aplicación práctica de esos métodos tanto a los archivos ya existentes, como a aquellos que se creen a partir de encuestas biográficas.

PRIMERA PARTE

GENERALIZACIÓN DEL ANÁLISIS LONGITUDINAL

I. LA OBSERVACIÓN DE BIOGRAFÍAS

En la introducción mencionamos cuáles eran los objetivos teóricos del análisis de las biografías. A fin de alcanzar esos objetivos hay que disponer de un número suficiente de observaciones de biografías, para poder producir resultados seguros. En este caso, la muestra ideal consiste en una observación exhaustiva de la población estudiada, la cual suministrará la biografía detallada de cada uno de sus miembros.

De la observación exhaustiva que constituye el censo sólo puede derivarse un número restringido de preguntas sobre la biografía. Por ejemplo, a partir de 1962 se plantea en Francia una pregunta sobre la residencia al 1 de enero del año anterior al censo. La explotación de tal pregunta, que permite estimar cifras efectivas de migrantes intercensales, desconocerá sin embargo la fecha de esa migración, las migraciones múltiples a lo largo del periodo entre censos, y los retornos, lo que implica una subestimación de la movilidad. Además, cuando se quiere situar las migraciones en relación con los acontecimientos profesionales o familiares, los datos ya no se prestan a ello porque son incompletos o están truncados.

Como ejemplo, vemos que se conoce el estatus matrimonial de cada individuo, pero se desconoce su fecha de matrimonio si está casado, y la del divorcio si está divorciado, etc. Esos datos son prácticamente inutilizables para el análisis de las biografías, pues no recogen sino una parte ínfima de la historia de vida.

Resulta más interesante considerar los datos de estado civil, que también son exhaustivos. Ellos constituyen la fuente esencial para el análisis demográfico clásico de los fenómenos considerados por separado. Los cuadros que publican los institutos de estadística (el INSEE en Francia) no permiten, en cambio, ningún análisis de la interacción entre fenómenos diferentes, de no ser la eliminación del efecto perturbador de la mortalidad. De todas maneras, sólo se refieren a tres fenómenos demográficos: la nupcialidad, la fecundidad y la mortalidad. Por último, si bien registran los eventos correspondientes, no suministran de manera precisa datos sobre las poblaciones a las que se refieren. En efecto, para estimar esa población hay que basarse en los datos de un censo anterior. Las migraciones que se desconocen contribuyen a sesgar esta estimación, sobre todo si no se trabaja sobre el conjunto del país. Una vez más, estos datos resultan insuficientes para el análisis de biografías.

Los datos del panel provienen de una muestra permanente, o casi permanente, de individuos a quienes se interroga con regularidad. Sin embargo, la información que se recoge de esta manera plantea problemas de análisis. En efecto, como a los individuos se les interroga acerca de su situación cuando se realiza cada entrevista, las historias individuales resultan incompletas y esos "huecos" vuelven difícil todo análisis estadístico de las interacciones, sin hipótesis constrictivas (*cf.* capítulo III.C). Si bien esta técnica del panel se opone a la de la encuesta puntual, no siempre resulta fácil explotar el carácter temporal de las informaciones obtenidas en el nivel individual. No obstante, esos datos que se recogen sucesivamente en una misma muestra son extremadamente ricos para la medición de las evoluciones de nivel global. La encuesta sobre el empleo del INSEE es de tipo panel y se realiza cada año sobre una muestra que se renueva parcialmente año con año.

Los datos de los registros de población, exhaustivos al igual que los datos del estado civil, son mucho más satisfactorios cuando se mantienen correctamente. Recordemos que ellos recogen, además de los nacimientos, matrimonios y decesos, las migraciones internas que contribuyen a modificar la población de una unidad administrativa. Lo que se les escapa son los cambios de situación profesional y diversos acontecimientos no registrados, como las uniones temporales. Cabe advertir además que esos registros no existen sino en un pequeño número de países (no los hay en Francia) y que desempeñan un papel esencialmente administrativo. Cuando no están centralizados, la búsqueda de cada paso de la trayectoria de un individuo en todas las comunas donde ha vivido es demasiado laboriosa, de manera que no puede realizarse más que en pequeñas subpoblaciones. Un seguimiento de este tipo fue el de los individuos nacidos en la parroquia sueca de Arnas entre 1896 y 1905, a lo largo de 50 años de su vida (Wendel, 1953), y otro el de 50 parejas nacidas en Bélgica durante el decenio de 1910, durante más de 60 años de su existencia (Duchène, 1985).

Vemos entonces que si bien entre las fuentes citadas hasta el momento la mejor es el registro de la población, no se le puede utilizar de manera exhaustiva a causa de que implica costos prohibitivos. En cambio, el seguimiento de una muestra con la ayuda del registro es totalmente realizable y conduce a datos del mismo tipo que los de una encuesta y sin duda de mejor calidad, pero mucho menos completos (en particular faltan datos sobre la vida profesional).

A) DIVERSOS TIPOS DE ENCUESTAS

En un país que no dispone de registros de población, las encuestas son necesarias para el análisis de las biografías. Resultan igualmente útiles en los países que disponen de registros de población, pues permiten captar los

acontecimientos no registrados. Hay varios tipos de encuestas que se pueden llevar a cabo.

La *encuesta prospectiva* es aquella que sigue a los individuos desde su fecha de entrada a la población sometida a análisis (14 años, por ejemplo, si se estudia el matrimonio, fecha de terminación de los estudios para el ingreso a la carrera profesional, etc.) y representa la opción más favorable. Una *encuesta de entrevistas repetidas* o una *encuesta renovable* es de tipo prospectivo, pero no sigue forzosamente a los individuos desde su fecha de entrada en la población analizada y los pierde si se salen del campo de la muestra. En ese caso es necesario hacer una encuesta retrospectiva cuando se realiza la primera entrevista, para conocer toda la existencia pasada de los encuestados y seguir a los individuos que se salen del campo de la muestra. En estas condiciones, dichas encuestas suministrarán información acerca de todos los eventos acaecidos a los individuos hasta la fecha de la última entrevista. Como por lo general no se sigue a los individuos hasta su deceso, ciertos intervalos observados serán *intervalos abiertos a la derecha*. Así, por ejemplo, ciertas mujeres habrán tenido su primer hijo durante la encuesta y no se sabe si tendrán el segundo después y, en caso de que sí lo tengan, se ignora cuándo será la fecha de ese segundo nacimiento. Más adelante veremos que la información suministrada por una encuesta como ésa es, sin embargo, utilizable.

Si se realiza una encuesta prospectiva hay que prestar especial atención a las salidas de observación de los individuos durante el curso de la encuesta (individuos cuya dirección no se encuentra más a partir de cierta fecha). En este caso, la salida de observación puede efectivamente no ser aleatoria, sino que está relacionada con los individuos cuyo comportamiento es muy diferente del resto de la población. Un medio para verificar esta hipótesis sería tratar por separado a la subpoblación que sale de la observación antes del final de la encuesta, para ver si no tiene ya de por sí un comportamiento diferente del resto de los encuestados mientras sigue siendo observada.

De esta manera A. Monnier obtuvo resultados bastante buenos valiéndose de una encuesta longitudinal en tres entrevistas (1974, 1976 y 1979) llevada a cabo en el INED. Esta encuesta tenía como objetivo confrontar los proyectos de fecundidad con los comportamientos reales. Para realizarla, Monnier volvió a interrogar —casi siempre por correo— a las mujeres de su muestra inicial y logró minimizar el sesgo debido a las salidas de observación: 82% de las encuestadas respondió a la segunda entrevista y 89% a la tercera (Monnier, 1987).

Cuando se realizan encuestas prospectivas es importante, pues, que utilicemos todos los medios para localizar a un encuestado que haya emigrado. El tomar los nombres y direcciones de las personas que lo conocen de cerca ha resultado útil para lograr ese rencuentro.

Una *encuesta retrospectiva* evita ese inconveniente. En efecto, con ésta sólo se interroga una vez a los individuos y se les piden todas las fechas en que se produjeron los acontecimientos estudiados desde que ellos ingresaron al grupo de población sometida a análisis. De nuevo se tendrán intervalos abiertos a la derecha a partir de la fecha de la entrevista, pero como esta encuesta es independiente de las fechas en que se produjeron los diversos acontecimientos, la información podrá utilizarse sin problema.

Una encuesta como ésta presenta, sin embargo, ciertos riesgos de sesgo, sobre los que ahondaremos más adelante. Sin embargo cabe mencionar aquí el caso de los individuos que fallecieron o el de aquellos que emigraron al extranjero antes de la encuesta. Para esos individuos el fallecimiento o la migración pueden no ser independientes del evento estudiado. Así por ejemplo, si se estudia la nupcialidad, es evidente que ciertos padecimientos o ciertas enfermedades disminuyen las posibilidades de casarse y aumentan simultáneamente los riesgos de fallecer.

Es preciso considerar además los problemas de memoria de los encuestados. En efecto, cuando se realiza una encuesta retrospectiva se suele interrogar a los individuos acerca de acontecimientos que ocurrieron 50 o incluso 60 años antes. Es posible pensar que la confiabilidad de la memoria disminuye con el tiempo trascurrido desde el acontecimiento rememorado; que ciertos datos carecerán de certeza, y que incluso se omitirán ciertos eventos. Por lo tanto, es importante verificar la calidad de los datos recogidos de manera retrospectiva: en los países que disponen de registros de población, se puede confrontar los datos de las encuestas retrospectivas con los de los registros. Más adelante presentaremos brevemente ciertos resultados que obtuvimos en Bélgica (Courgeau, 1985) y otros provenientes de Suecia (Lyberg, 1983). Un análisis de esta índole debe realizarse sobre un número importante de elementos, para que podamos estar seguros de la validez de los resultados obtenidos mediante una encuesta retrospectiva. La percepción del tiempo puede ser muy diferente en diversas sociedades, y el mismo problema puede tener varias soluciones.

Fuera de estos dos tipos principales de encuestas existen otros que conducen a situaciones más complejas y con sesgos mucho más importantes; regresaremos a ellos más adelante.

Antes de detallar tales tipos de encuestas presentaremos primero, de manera sucinta, la encuesta sobre la biografía familiar, profesional y migratoria: la "Triple biografía", que realizamos en el INED en 1981, y que proporcionará numerosos ejemplos de aplicación de los diversos métodos de análisis presentados en esta obra.

B) La encuesta "Triple biografía"

Esta encuesta se sitúa dentro de una corriente de experiencias que ya tiene antigüedad en el INED.

La primera encuesta de este tipo la realizó Guy Pourcher en 1961 y se relacionó con el poblamiento de París (Pourcher, 1964). Este autor interrogó a una muestra de individuos presentes en París acerca de las diversas etapas de su existencia pasada, mediante la aplicación de un primer modelo de cuestionario. Sin embargo, su análisis demográfico clásico de las encuestas desaprovechó gran parte de la información recogida (en particular las fechas de los diversos eventos).

Nuestra encuesta "Triple biografía" se propone analizar las interacciones de los diversos aspectos de la vida familiar, profesional y migratoria de los encuestados. Este análisis condicionó la selección del muestreo y las preguntas planteadas.

Se trató de una encuesta estadística biográfica, retrospectiva, que se llevó a cabo en 1981 con carácter nacional sobre los individuos nacidos entre 1911 y 1935. Esos individuos se incluyeron sin someterlos a una selección ligada a su historia de vida. Habría sido preferible aplicar esta encuesta a algunas generaciones, por ejemplo, las nacidas en 1911, 1921 y 1931, lo cual hubiera producido grupos de individuos que habrían experimentado diversos acontecimientos coyunturales (crisis económicas, guerras, etc.) exactamente a las mismas edades. Lamentablemente, es difícil instrumentar dentro del análisis esta restricción en un país que carece de registros de población y donde el acceso a las bases de datos para realizar un muestreo está particularmente restringido. La ley de 1975 sobre la informática y las libertades prohíbe que durante 100 años posteriores a su publicación se utilice un archivo para usos distintos de los de su finalidad original. Por supuesto, recomendamos ampliamente que en los países que disponen de registros de población de donde se puede extraer una base de datos que sea útil para realizar un muestreo, se observen generaciones precisas. Tal fue el caso, por ejemplo, de una encuesta que realizó el Instituto Max-Planck de Berlín bajo la dirección de Karl Ulrich Mayer, en la que fue posible observar tres cohortes nacidas en 1929-1931, 1939-1941 y 1949-1951.

Para realizar esta encuesta en Francia tuvimos que tomar como base del muestreo los archivos de las viviendas censadas en 1975 o existentes a partir de entonces, en las comunas pertenecientes a la "muestra patrón" del INSEE. Como sólo 45% de los hogares incluía a un adulto de entre 45 y 69 años, numerosas visitas habrían sido inútiles y muy desalentadoras para los encuestadores. Una original solución a ese problema fue emplear la misma muestra para otra encuesta sobre "la vida familiar y la vida profesional" de las mujeres que tuvieran uno o varios niños a su cargo. Las dos subpoblaciones

consideradas son distintas por completo: sólo 10% de los hogares comprende, a la vez, un niño de menos de 16 años y un adulto de entre 45 y 69 años. Si éstas hubieran sido totalmente distintas, el diseño de la muestra habría sido perfectamente no informativo acerca de las historias de vida recogidas. Sin embargo, en este caso sólo se trató de un inconveniente menor. Esta solución, en cambio, permitió limitar las visitas inútiles e hizo que la encuesta valiera la pena, y por tal razón finalmente la escogimos.

A partir, entonces, de la misma muestra de 16 500 viviendas, procedimos a la aplicación simultánea de dos encuestas. Los hogares que podían responder a ambas encuestas se repartieron al azar utilizando la paridad del número de la ficha con la dirección del domicilio. Cuando en el hogar había varias personas con edades entre 45 y 69 años la persona por interrogar se escogía utilizando el método de Kish.[1] Esas encuestas se realizaron con la ayuda de las direcciones regionales del INSEE y de sus encuestadores: sólo 11% de las viviendas extraídas condujo a un fracaso (negativa, ausencia de larga duración, imposibilidad de encontrar al ocupante). Por otra parte, las encuestas descartaron 17% de las viviendas vacías o residencias secundarias y 20% de los hogares fuera del alcance (Riandey, 1985). De esa manera se dispuso de 4 602 cuestionarios cuyas respuestas versaban sobre la "biografía familiar, profesional y migratoria". La aplicación de la encuesta duraba un promedio de una hora y diez minutos y algunas entrevistas llegaban a más de dos horas.

Veamos ahora el contenido de los cuestionarios.

Para permitir un análisis de las interacciones de los diversos eventos hay que recoger la fecha precisa de éstos y su localización. Así, en lo que respecta al primer matrimonio de un individuo, a éste se le preguntaba:

— ¿A qué edad se casó usted (mes, año)?
— ¿Cuál fue su lugar de residencia luego de su matrimonio (comuna, provincia, departamento o país)?

Respecto de un periodo de empleo se preguntaba, entre otras cosas:

— ¿Cuál es la fecha de inicio del periodo (mes, año)?
— ¿En qué lugar está el establecimiento (comuna, departamento)?
— ¿Cuál es la actividad precisa del establecimiento?
— ¿De manera muy exacta, cuál es la profesión principal (al inicio del periodo)?
— ¿Cuál es la profesión principal al final del periodo?
— Si este periodo se terminó, ¿en qué fecha (mes, año) ocurrió?

[1] Recordemos que el método de Kish permite seleccionar a una persona entre varias sin dejarle ningún margen al encuestador, gracias a un cuadro donde las dos entradas son el número de los elegibles y un "número Kish" atribuido aleatoriamente al hogar.

Por último, respecto de las sucesivas residencias se preguntaba, entre otras cosas:[2]

— ¿Cuál fue la fecha de la mudanza (mes, año)?
— ¿Cuál fue el lugar de residencia (comuna, departamento)?

Aunque vemos que las fechas se pedían en meses, esa información parece poco confiable pues numerosos cuestionarios no la incluyen. La confrontación con los datos de los registros de población lo revela claramente (Duchêne, 1985).

Para la buena definición de un estado es preciso pedir la mayor cantidad posible de información sobre sus características. Así, en lo que respecta a los empleos, se pregunta sobre doce características que permiten, por ejemplo, reconstruir el código de categorías socioprofesionales del INSEE o definir categorías precisas (se emplea, por ejemplo, en el sector público y en el sector privado). De igual manera, para cada residencia se pide el estatus de ocupación en cinco puestos, lo que permite clasificar a los individuos en forma diferente (propietario o no propietario, por ejemplo).

En un cuestionario de ese tipo no hay que olvidar la información sobre los orígenes familiares, los hermanos y hermanas del encuestado y sus niños, que completará la parte puramente biográfica.

De igual manera, es indispensable que se pueda confrontar la biografía del encuestado con la de su pareja. Inicialmente se deseaba interrogar a los dos esposos para tener información suplementaria sobre los orígenes familiares, la biografía profesional o las migraciones prenupciales del compañero, pero la segunda entrevista provocó numerosos rechazos durante la encuesta piloto, por más que se acortara su duración a media hora. Además, como con frecuencia la segunda entrevista se aplicaba al esposo de una mujer a la que se había interrogado primero en el hogar, se corría el riesgo de introducir sesgos muy importantes. Para evitarlos se renunció a las dos entrevistas en favor de un tiraje Kish. Sin embargo se le pedía al encuestado que proporcionara abundante información sobre los orígenes y la carrera profesional (en el momento del matrimonio y en la actualidad) de su pareja.

Concebido de esa manera, el cuestionario ya ha posibilitado la realización de numerosos análisis de interacción de fenómenos (Courgeau, 1984, 1985a, 1985b, 1987; Courgeau y Lelièvre, 1986, 1988; Lelièvre 1987a, 1987b).

[2] El lector interesado en las preguntas planteadas acerca de los periodos de empleo y de inactividad y sobre las residencias sucesivas, encontrará una reproducción de esta parte del cuestionario en el manual *Méthodes de mesure de la mobilité spatiale* (Courgeau, 1988).

C) El problema de las ponderaciones

Ya vimos que en la encuesta "Triple biografía" sólo se interrogaba a una de las personas del hogar cuya edad estaba entre 45 y 69 años. A priori parecería útil ponderar la muestra en función del número de individuos representados por cada encuestado. Demostraremos que eso no funciona cuando se trata de realizar un análisis según los modelos que presentamos en esta obra, pues el diseño de muestreo no aporta ninguna información sobre las historias de vida recogidas.

Veamos primero, sobre un ejemplo práctico, los resultados obtenidos cuando se utilizan las ponderaciones y cuando no se emplean.

Se trata de apreciar el efecto de la edad al inicio de la *permanencia* o supervivencia en un estado (en este caso residir en cierto domicilio) y el de la duración de esa *permanencia* sobre la probabilidad de efectuar un cambio de residencia para hombres activos de las generaciones nacidas entre 1926 y 1935. Para ello suponemos que un modelo multiplicativo de tipo Gompertz se aplica bien a los datos (tal es el modelo que desarrollaremos en el capítulo VII.A.3). Sólo mencionaremos aquí que si el modelo se verifica el cociente de migración se escribe:

$$h(t; z) = \exp(\alpha z + \beta t)$$

donde z es un vector renglón de características individuales (aquí, la unidad para la constante y las variables binarias son iguales a la unidad si el individuo inicia su permanencia en el grupo de las edades implicadas), α un vector columna de parámetros que indican el efecto de esas características sobre el cociente, t la duración de permanencia, y β un parámetro que indica el efecto de esta duración de permanencia.

En el cuadro 1 presentamos las estimaciones de estos parámetros según sea que se introduzcan o no las ponderaciones. Vemos que esas ponderaciones son importantes, pues casi duplican el número de individuos por tomarse en cuenta. En cambio se observa que tienen un efecto poco importante sobre los parámetros estimados, considerando la desviación estándar de esos parámetros en el modelo sin ponderación. Así por ejemplo, un individuo de menos de 20 años tendrá un cociente de migración igual a 0.110 al cabo de 10 años de *permanencia* conforme al modelo que no hace intervenir las ponderaciones, e igual a 0.113, en el modelo que sí las hace intervenir.

No retomaremos aquí la demostración detallada que proporciona Hoem (1985) y muestra que la ponderación no ha de introducirse bajo condiciones mucho más generales. Indicaremos, sin embargo, cuáles son las condiciones precisas que conducen a no preocuparse del diseño de la muestra cuando se analizan historias de vida.

CUADRO 1
Efecto de la edad y de la duración de permanencia sobre los cambios de residencia de los hombres activos de las generaciones nacidas entre 1926 y 1935

Características individuales	*Modelo sin ponderación*		*Modelo con ponderaciones*
	Efecto estimado	*Desviación estándar*	*Efecto estimado*
Constante	– 2.258	0.0607	– 2.263
Con menos de 20 años	0.612	0.0671	0.627
Entre 20 y 24 años	0.553	0.0686	0.573
Entre 25 y 29 años	0.356	0.0740	0.309
Entre 30 y 34 años	0.250	0.0827	0.216
Duración de permanencia (β)	– 0.056	0.0033	– 0.054
Número de individuos	3 429		6 315

La condición principal para llegar a ese resultado es que el diseño de la muestra sea independiente de las historias de vida de las personas encuestadas. Se dice, además, que ese diseño de muestra es *no informativo*. Así, en la encuesta realizada en Francia se seleccionaba el individuo al que se iba a interrogar mediante el método de Kish, el cual no utiliza ninguna información acerca de la biografía de las personas que se van a escoger, puesto que el "número de Kish" es atribuido aleatoriamente al hogar. Claramente se advierte que éste es el caso de un diseño de muestra *no informativo*.

Otra condición es que la observación de las biografías individuales, que puede ser tan sólo parcial, se realice de manera independiente de un individuo al otro. En particular, esta condición se verifica si la historia de vida se observa habiendo fijado previamente una duración de matrimonio o desempleo para todos los individuos encuestados.

Si esas dos condiciones se cumplen es factible demostrar que *no resulta útil tomar en cuenta las ponderaciones* en el análisis de las biografías. Éstas presentan la misma distribución de probabilidad que si se hubiera observado la población total al hacerse una encuesta exhaustiva. En este caso se dice que hay que ignorar el diseño de la muestra. Dicho resultado es independiente del modelo utilizado y sigue siendo válido incluso si está mal especificado.

Las condiciones que conducen a tal resultado podrían no cumplirse más si se realizara una encuesta retrospectiva. En efecto, una encuesta como ésa sólo interroga a los individuos que sobrevivieron hasta su aplicación. En tal caso cabe pensar que quienes escapan a la encuesta pueden haber tenido un comportamiento demográfico muy diferente del de los sobrevivientes. Si

así fuera, los individuos bajo observación habrían sido sometidos a un proceso de *selección* suministrado por el diseño de muestra informativo. En cambio, si las personas fallecidas o que emigraron al extranjero hubieran tenido un comportamiento idéntico al de los sobrevivientes durante su vida en el país, entonces ese sesgo de selección habría desaparecido por completo. El diseño de muestra se vuelve no informativo (Hoem, 1985).

Sabemos que en Francia y en otros países la mortalidad de los solteros es mucho mayor que la de los casados, a edades iguales. En Francia, por ejemplo, entre 1967 y 1969, el cociente de mortalidad de los hombres solteros de 45 a 49 años es de 612 por cada 10 000, mientras que el cociente de mortalidad de los casados llega sólo a 298 por cada 10 000 (Vallin y Nizard, 1977), o sea, menos de la mitad del de los solteros. Una selección semejante introduce errores en los cocientes que se estiman con la ayuda de una encuesta retrospectiva, cuyo diseño de muestra se vuelve informativo. La encuesta sobreestimará los cocientes de nupcialidad.

En ese caso existe una posibilidad de corregir los errores si se dispone de información sobre el estado civil o, aún mejor, de registros de población. Aquí es posible introducir una ponderación que será diferente de la que se debe al diseño de muestra. Para eso hay que contar con probabilidades de inclusión en la muestra de los diversos tipos de historia de vida.

D) La observación incompleta y errónea de las encuestas

Anteriormente mencionamos algunas de las imperfecciones de las encuestas. Regresemos ahora a considerar en detalle esos problemas, a veces irresolubles.

Resulta útil presentar aquí algunos ejemplos de situaciones complejas debidas al método de recolección de datos. Supongamos, entonces, que interrogamos al individuo acerca de los diversos cambios de residencia que experimentó o los diversos nacimientos de los que tuvo conocimiento en los últimos cinco años. Vemos que si en ciertos casos hay siempre intervalos abiertos a la derecha, además de ellos habrá *intervalos abiertos a la izquierda* e incluso ciertos intervalos de los que no se conocerá ni el comienzo ni el final. Esta información no se puede utilizar más que introduciendo hipótesis muy fuertes acerca de la distribución de los acontecimientos en el curso del tiempo. De esa manera, si se supone que los acontecimientos siguen una distribución exponencial, los intervalos abiertos a la izquierda no introducirán ningún sesgo. En efecto, veremos (capítulo VII.A.1) que la distribución condicional de $(T-t_0)$, sabiendo que T es superior a t_0 (estando este instante situado cinco años antes de la encuesta, por ejemplo), es la misma que la distribución de T. Sin embargo, vemos que esta condición tiene pocas posibilidades de ser verificada, y por lo

tanto no aconsejamos practicar ese tipo de recolección de datos cuando no se dispone de otra información sobre el fenómeno estudiado.

Un medio para evitar un sesgo como ése consiste en plantear una pregunta sobre la fecha de llegada a la vivienda ocupada a partir de los cinco años anteriores. Tal fue la solución que se escogió en la encuesta sobre la movilidad residencial entre 1973 y 1978, que fue complementaria de la encuesta sobre el empleo del INSEE. Sin embargo se desconoce la movilidad anterior a esa fecha, que puede repercutir en la movilidad observada.

De manera similar, las encuestas que preguntan sobre el último evento ocurrido (la última migración, el último cambio de profesión, el último nacimiento, etc.) conducen de nuevo a riegos de sesgos importantes.[3] Así, la observación de las duraciones de permanencia va a implicar intervalos mucho más largos que los que proporcionaría la observación de la vida entera de los individuos. En ausencia de otras informaciones sobre la duración de permanencia, desaconsejamos enfáticamente ese tipo de recolección de datos. En efecto, cuando se interroga a un individuo acerca de su última migración, se sabe que no ha experimentado otra desde entonces hasta la fecha de la encuesta. El diseño de muestra es, por lo tanto, informativo. En términos más generales, se puede decir que esos problemas de selección están ligados al hecho de que el comportamiento de los individuos se observa bajo condiciones que lo afectan. En todos esos casos la observación respecto del comportamiento que se quiere estudiar no es neutra, y eso acarrea el riesgo de provocar sesgos. Para analizar los datos se necesita plantear hipótesis cuya validez hay que probar.

Cuando se realiza una encuesta prospectiva suele ser difícil realizar el seguimiento de los individuos. He ahí la presencia eventual de un sesgo de selección: las no respuestas y las desapariciones del seguimiento se pueden producir de manera preferencial en una categoría específica de la población encuestada. De igual manera se puede estar en presencia de un sesgo de condicionamiento (Deroo y Dussaix, 1980) debido a una modificación eventual de los comportamientos de los individuos seguidos. Sin embargo la existencia de ese sesgo es, hasta cierto punto, controvertida. F. Bribier realizó un seguimiento de este tipo al observar una cohorte de parisinos desde su retiro, en 1972, hasta su muerte.

Otra fuente de error, tanto en las encuestas prospectivas como en las retrospectivas, es la negativa a responder. Un estudio realizado sobre la encuesta sueca de la fecundidad de 1981 da cuenta de esos errores después de comparar dicha encuesta con los datos de los registros de población (Lyberg, 1983). La tasa de no respuesta para la encuesta fue de 13%. Podría pensarse que los individuos que se negaron a responder pertenecían a un

[3] Para más detalles véase J. Hoem, 1985.

grupo muy selectivo con un comportamiento atípico. Sin embargo, los resultados de esta comparación revelan que no fue así: los diversos análisis efectuados sobre los registros y sobre los datos de la encuesta retrospectiva dan resultados muy similares (cuadro 2).

CUADRO 2
Cocientes de fecundidad de rango 1, estimados a partir de la población total (h_p), la población muestreada (h_s) y de quienes respondieron (h_R) por cada 1000 mujeres

Edad	h_p	h_s	h_R	*Estimación del sesgo de no respuesta*
23	138	120 ± 27*	128 ± 30*	8
24	152	192 ± 36	219 ± 42	27
25	165	158 ± 36	163 ± 40	5
26	168	142 ± 37	154 ± 42	12
27	158	145 ± 40	151 ± 45	6
28	142	129 ± 40	133 ± 45	4
29	125	155 ± 47	177 ± 57	22
30	107	107 ± 42	133 ± 53	26

*Intervalo de confianza de 95 por ciento.
Fuente: Lyberg, 1983.
Encuesta de fecundidad sueca de 1981, mujeres nacidas entre 1941 y 1945.

Los errores de memoria pueden ser importantes, especialmente cuando se entrevista a gente de edad avanzada mediante encuestas retrospectivas. Puede ser que las fechas de varios eventos sean erróneas, o que incluso las hayan olvidado. Por lo tanto, es necesario que se pruebe la calidad de esos datos, comparándolos con los obtenidos de los registros de población. En Bélgica, un país donde los datos de los registros de la población hacen posible esa verificación, se realizó esa prueba sobre el cuestionario de la "Triple biografía" (Poulain *et al.*, 1991).

La prueba se llevó a cabo sobre una muestra de 445 parejas,[4] colocando a los entrevistados en las condiciones más desfavorables. Al principio se entrevistó a ambos esposos simultáneamente, en cuartos separados. Luego de la entrevista, ambos esposos confrontaron sus respectivos relatos de los eventos de su vida y los corrigieron, en particular haciendo referencia a cualquier documento disponible (el libro de familia, recibos de renta, etc.).

[4] Ya antes se había realizado una prueba limitada a 50 parejas (Duchêne 1985; Courgeau, 1985a), que condujo a resultados muy similares a los presentados aquí.

Por último se consultaron los registros independientemente de la entrevista, como una cuarta fuente de información acerca de los mismos eventos. La muestra estuvo limitada a parejas en las que uno de los esposos había nacido entre 1933 y 1942 y el otro en un marco temporal mayor, entre 1933 y 1947. El rango de las edades al momento de la entrevista era, por lo tanto, de 41 a 55 años. A estos entrevistados se les pidió que recordaran eventos ocurridos a veces muchos años atrás. La comparación de las fechas dadas por cada uno de los esposos y por la pareja junta, más lo que proporcionó el registro de la población, se presenta en el cuadro 3 (véase Poulain *et al.*, 1991 para más detalles). Esto resulta satisfactorio en lo que respecta a la fecha de matrimonio y el nacimiento o la muerte de un niño, pues éstos son informes de registros, pero lo es menos cuando se trata de las migraciones del hogar. En efecto, si bien constituye una obligación legal declarar el cambio de lugar de residencia dentro de un plazo de ocho días, puede haber algún retraso que, sin embargo, rara vez excede el mes. Además, los registros de población no registran los cambios de residencia en el extranjero, y sólo recogen parcialmente las emigraciones (inmigraciones) hacia (desde) un país extranjero, en particular en el caso de los militares o la gente ocupada en el Servicio Voluntario en el Extranjero.

El cuadro 3 revela una diferencia muy importante entre el matrimonio y el nacimiento de los niños, por una parte (las fechas son exactas con margen de un mes en más de 90% de los casos), y la migración del hogar o la emancipación de los hijos, por la otra (las fechas son exactas con margen de un mes sólo entre 39 y 67% de los casos). Sin embargo estos últimos porcentajes se incrementan si escogemos un intervalo más o menos de un año: respecto de la emancipación de los hijos, éste sube a 69.8% para los hombres, a 76.7% para las mujeres, y a 78.4% cuando los esposos están juntos; en lo que respecta a las migraciones, suben a 87.5% para las mujeres, a 90.2% para los hombres, y a 93.2% para los dos esposos. Ciertamente, la información que proporcionan las mujeres suele ser mejor que la que dan los hombres, y la que proporcionan los dos esposos es mejor que la que dan el esposo o la esposa por separado. Por lo tanto, siempre que sea posible es preferible entrevistar acerca de los sucesos familiares y las migraciones a los dos esposos juntos, o, como una segunda mejor solución, a la mujer.

Sin embargo, estos errores en las fechas apenas han afectado los resultados de los análisis, sean paramétricos, no paramétricos o semiparamétricos, que hemos realizado usando estos datos (Courgeau, 1991). Aquí presentamos los resultados de un análisis paramétrico de las duraciones de permanencia de más de seis meses en los diferentes lugares de residencia.

El cuadro 4 muestra los resultados de este análisis, en el que hemos tomado en cuenta la duración del matrimonio y el tiempo pasado en el lugar de residencia, así como el número de niños nacidos al inicio del periodo. En un

CUADRO 3

Omisiones y errores en las fechas de matrimonio, nacimiento de los hijos, emancipación de los hijos y migraciones del hogar*

	Matrimonio			*Nacimiento de los hijos*			*Emancipación de los hijos*			*Migración del hogar*		
	Hombres	*Mujeres*	*Pareja*	*Hombres*	*Mujeres*	*Pareja*	*Hombres*	*Mujeres*	*Pareja*	*Hombres*	*Mujeres*	*Pareja*
Total de eventos	445	445	445	1 078	1 078	1 078	310	310	310	1 388	1 388	1 388
Eventos fechados de los cuales:	440	445	445	1 076	1 077	1 078	222	228	222	1 169	1 196	1 237
Fecha exacta (en el curso de 1 mes)	414 (93%)	440 (98.9%)	443 (99.6%)	940 (90.8%)	1 038 (97.8%)	1 045 (98.2%)	83 (39.0%)	115 (50.4%)	119 (53.6%)	651 (55.7%)	723 (60.5%)	833 (67.3%)
Fechados antes (por 1 año o menos)	8 (1.8%)	1 (0.2%)	0 (0.0%)	46 (4.4%)	6 (0.5%)	8 (0.7%)	42 (20.0%)	39 (17.1%)	38 (17.1%)	269 (23.0%)	273 (22.8%)	257 (20.8%)
Fechados después (por 1 año o menos)	13 (2.9%)	3 (0.6%)	1 (0.2%)	33 (3.2%)	10 (1.0%)	8 (0.7%)	23 (10.8%)	21 (9.2%)	17 (7.7%)	103 (8.8%)	83 (6.9%)	63 (5.1%)
Fechados antes (por más de 1 año)	2 (0.4%)	0 (0.0%)	0 (0.0%)	7 (0.7%)	3 (0.3%)	3 (0.3%)	40 (18.9%)	36 (15.8%)	34 (15.3%)	100 (8.6%)	83 (6.9%)	53 (4.3%)
Fechados antes (por más de 1 año)	3 (0.7%)	1 (0.2%)	1 (0.2%)	9 (0.9%)	4 (0.4%)	1 (0.1%)	24 (11.3%)	17 (7.5%)	14 (6.3%)	46 (3.9%)	34 (2.9%)	31 (2.5%)

*Las proporciones en relación con eventos implicados aparecen entre paréntesis.

Fuente: Poulain *et al.*, 1991.

CUADRO 4

Análisis de la movilidad espacial: efecto del tiempo transcurrido desde el matrimonio, de la duración de la estadía en años, y del estatus residencial sobre la probabilidad de cambiar de vivienda de acuerdo con la fuente[a]

Variables	*Hombres (1 260 permanencias)*		*Mujeres (1 310 permanencias)*		*Esposos juntos (1 314 permanencias)*		*Registros (1 189 permanencias)*	
	Modelo 1	*Modelo 2*	*Modelo 1*	*Modelo 2*	*Modelo 1*	*Modelo 2*	*Modelo 1*	
Constante	−2.955***	−2.322***	−3.009***	−2.391***	−3.062***	−2.264***	−3.118***	
	(0.174)	(0.193)	(0.179)	(0.193)	(0.183)	(0.192)	(0.187)	
Inicio de la permanencia en el lugar de residencia en el año del matrimonio	1.464***	0.438***	1.569***	0.574***	1.642***	0.575***	1.564***	Generalización del análisis longitudinal
	(0.172)	(0.174)	(0.177)	(0.178)	(0.182)	(0.181)	(0.186)	
Inicio de la permanencia en el lugar de residencia entre 1 y 4 años después del matrimonio	1.117***	0.572***	1.198***	0.622***	1.301***	0.750***	1.199***	
	(0.160)	(0.160)	(0.157)	(0.158)	(0.161)	(0.161)	(0.166)	
Inicio de la permanencia en el lugar de residencia entre 5 y 9 años después del matrimonio	0.641***	0.375**	0.559***	0.286*	0.684***	0.444***	0.707***	
	(0.164)	(0.165)	(0.162)	(0.163)	(0.164)	(0.165)	(0.166)	
Número de niños al comienzo de la permanencia en el lugar de residencia	0.006	−0.016	0.052	0.042	0.051	0.023	0.033	Observación de biografías
	(0.040)	(0.039)	(0.042)	(0.041)	(0.042)	(0.042)	(0.044)	
Alojado por el empleador		0.485***		0.480***		0.321***		
		(0.091)		(0.090)		(0.078)		
Propietario de la vivienda		−2.431***		−2.347***		−2.538***		
		(0.175)		(0.166)		(0.164)		
Duración de la permanencia en el lugar de residencia	−0.113***	−0.056***	−0.116***	−0.058***	−0.119***	−0.064***	−0.104**	
	(0.0077)	(0.0076)	(0.0076)	(0.0075)	(0.0077)	(0.0075)	(0.0076)	

[a] Parámetros estimados: desviaciones estándar colocadas entre paréntesis.

* Resultado con un nivel de significación de 10 por ciento.

** Resultado con un nivel de significación de 5 por ciento.

*** Resultado con un nivel de significación de 1 por ciento.

Fuente: Courgeau, 1991.

segundo modelo también tomamos en cuenta el estatus residencial indicado por los hombres, las mujeres y ambos esposos juntos (esta información no se considera en los registros de población).

El modelo seleccionado implica que el impacto de la duración de permanencia en el lugar de residencia disminuye exponencialmente la tasa de migración, que se escribe como antes en el tiempo t:

$$h(t;z) = \exp(z\alpha + \beta t),$$

donde z es el vector de las covariantes presentadas antes, y α y β son los parámetros por estimar que mostrarán el efecto de las características y de la duración de permanencia.[5] El cuadro 4 presenta los resultados del primer modelo (excluyendo las características del estatus residencial), estimado separadamente usando los datos proporcionados por los hombres, las mujeres, ambos esposos juntos y los registros de población.[6] Todas las características tienen un efecto muy similar, cualquiera que sea la fuente de información. El número de niños al comienzo del periodo no tiene efecto sobre la duración de la permanencia en el lugar de residencia, mientras que tal duración y la del matrimonio al inicio de la permanencia influyen fuertemente sobre la tasa de migración.

El cuadro también presenta los resultados del segundo modelo, en el que, aparte de las características previas, también se tomaba en cuenta el estatus residencial, al principio observamos un efecto muy significativo del estatus residencial —cualquiera que sea la fuente utilizada— en todas las estimaciones en los mismos intervalos de confiabilidad. El impacto de las otras características es limitado, pero sigue siendo similar al que ocurre cuando éstas se consideran independientemente. Este análisis revela que los errores importantes sobre las fechas de migración y la duración de permanencia en el lugar de residencia no producen sesgos significativos en el análisis de las tasas de migración de acuerdo con las diversas características del encuestado al inicio de la permanencia. En muchos casos los resultados son consistentes, cualquiera que sea la fuente de información utilizada: las pocas diferencias que se observan no modifican las principales cifras del análisis.

E) El problema de los truncamientos

Una de las principales dificultades para el análisis de las biografías es la presencia de truncamientos, nombre con el que se denomina a los huecos

[5] Véase las secciones VII.A.1. y VII.A.3. para una presentación más detallada de este tipo de modelo y de los métodos de estimación de los parámetros.

[6] Los parámetros se estimaron usando el programa RATE desarrollado por N. Tuma.

que se presentan en la información siempre que falta una parte del relato biográfico.

Hablamos de truncamientos a la izquierda cuando el relato comienza en un momento del ciclo de vida y no hay ninguna información disponible sobre el periodo anterior. La recolección de datos sobre las migraciones de los residentes de una zona geográfica a partir de una fecha dada no incluye sus desplazamientos anteriores, en particular aquellos que los condujeron a llegar a la zona en cuestión, a menos que la recolección se complete mediante una encuesta retrospectiva anterior. De igual manera, el estudio de las carreras profesionales de los individuos dentro de una empresa no considerará su trayectoria profesional previa. Se razonará entonces aceptando implícitamente la hipótesis de que el recorrido anterior fuera de la empresa no tiene influencia, o que en todo caso es poco significativa sobre la carrera dentro del grupo estudiado, y se dará mayor importancia, por ejemplo, a la capacitación para el ejercicio profesional como factor explicativo.

Por otra parte, nos encontramos en presencia de un truncamiento a la derecha cuando se interrumpe el relato del que se dispone para cada individuo, ya sea porque el individuo desapareció de la muestra o bien porque el relato se detuvo en la fecha de la entrevista. Cualquiera que sea la forma de recolección de los datos: prospectiva o retrospectiva, los informes biográficos comprenderán historias de vida truncadas en la fecha del término de la observación, a menos que la prospectiva se prolongue hasta la muerte de los encuestados. En el caso de una encuesta puramente prospectiva, podemos encontrarnos en presencia de truncamientos tanto a la derecha como a la izquierda. En el caso de una encuesta retrospectiva la fecha de la entrevista es igualmente la del fin del relato de vida, más allá de la cual no se recoge ninguna información.

Como esos truncamientos están obligatoriamente presentes en la información biográfica, uno de los objetivos de los análisis de la duración de permanencia en un estado es tomar en cuenta los datos truncados en forma sistemática y controlada (Kaplan y Meier, 1958).

El demógrafo se encuentra frente a un archivo en donde para cada individuo se registran series temporales relativas a su vida familiar, su carrera profesional, sus migraciones, etc. Dispone, por lo tanto, de secuencias cronológicas enlazadas por eventos de naturaleza diversa cuyo orden de aparición difiere de un individuo al otro y que no están obligatoriamente presentes en todas las biografías.

Si sólo se consideran la partida del domicilio familiar, el matrimonio y el primer empleo, su orden de aparición en la vida de un individuo puede variar de seis maneras diferentes, y el número de posibilidades se incrementa notablemente si consideramos aquellos casos en los que uno de los eventos jamás se produce (solteros, mujeres en el hogar, agricultores que

permanecen siempre con sus padres), o dos de los eventos no ocurren nunca (solteros que viven en la casa de sus padres, solteros sin actividad profesional, casados sin actividad profesional que viven en la casa de sus padres), y finalmente los casos en los que jamás se vivió ninguno de los tres eventos.

Esos itinerarios complejos constituyen el material del análisis. Más que la edad, los eventos en el tiempo constituyen los pasos del individuo de un estatus a otro, son los testimonios de ajustes, de selecciones —objetivas o no— que el individuo ha hecho en el curso de su existencia. Así, esos eventos constituyen marcas de la "edad social" de los individuos, concepto que la literatura sociológica formalizó (Elder, 1978; Foner y Kertzer, 1978; Hogan, 1978), y describió su evolución (Model, Furstemberg y Strong, 1978) y su carácter sexuado (Langevin, 1986). Conviene, por lo tanto, *modelizar* la aparición de esos eventos considerando sólo uno, o uno que se repite (nacimientos sucesivos), o incluso varios en interacciones.

F) Conclusión

Como ya dijimos, la recolección de las historias de vida, tanto mediante los registros de población como valiéndose de encuestas biográficas, plantea cierto número de problemas.

De todas las fuentes la más satisfactoria es, sin duda, el registro de población, al menos en los países donde éste se lleva en forma correcta. Lamentablemente esos países son pocos y la utilización de esa fuente de información es muy costosa. Además, ciertos eventos e informaciones necesarios para el análisis de las historias de vida no se registran allí. Así, para efectuar un análisis detallado de la fecundidad de las mujeres suecas se tuvo que realizar una encuesta retrospectiva, aunque Suecia disponga de registros perfectamente organizados.

Las encuestas prospectivas son también una fuente muy confiable cuando las encuestas sucesivas se realizan con intervalos no muy prolongados (un año, por ejemplo). En caso de que no sea así, si los retrasos en la obtención de datos analizables son muy elevados, se restringe la utilización de esa fuente que, por otra parte, debe combinarse generalmente con una encuesta retrospectiva inicial.

Las encuestas retrospectivas siguen siendo, pues, la fuente más utilizada para el análisis de las biografías, aunque plantean numerosos problemas que aún no han sido completamente resueltos. Así, la selección de los individuos sobrevivientes en el momento de la encuesta puede hacer que el diseño de muestra se vuelva informativo, en cuyo caso hay que medir los sesgos introducidos y verificar que carezcan de importancia. De igual manera, esas encuestas plantean problemas de memoria, sobre todo cuando se entrevista a

personas de edad avanzada. Los errores pueden ser importantes, pero cabe pensar que los sesgos que introducen en el análisis son mucho más limitados. La comparación de los resultados obtenidos mediante las encuestas retrospectivas y los registros de población deberían permitir dar respuesta a estas cuestiones.

Pese a tales inconvenientes, las encuestas restrospectivas constituyen la forma de recolección de datos más fácil y cómoda. Por lo tanto, es importante garantizar los resultados que suministran en aquellos países donde tal cosa es posible.

II. FORMALIZACIÓN DEL ANÁLISIS

Abordaremos ahora la formalización del análisis de biografías. Esta formalización generaliza los métodos desarrollados por biomatemáticos para el análisis de la mortalidad —en presencia de muestras restringidas—, al "análisis de las duraciones de permanencia", que no considerará la ocurrencia de los eventos como inevitable (como en el caso de la muerte). Es muy posible que un matrimonio, el nacimiento de un niño, una migración o un cambio profesional jamás se produzcan en la existencia de un individuo dado.

La observación de las biografías suministra cierto número de eventos, y lo que queremos estudiar es la repartición de las ocurrencias de estos eventos a lo largo del tiempo. Para determinar con precisión esas ocurrencias se deben satisfacer tres condiciones:

- Hay que disponer de un punto de partida común definido sin ambigüedad. Este instante no está necesariamente señalado por la fecha de nacimiento de cada individuo, sino que corresponde a la fecha de ocurrencia de un suceso anterior a los eventos estudiados (la fecha del primer empleo, del matrimonio, etc.) a partir del cual éstos se pueden producir (cambio de empleo, nacimiento de hijos legítimos, etcétera).
- Hay que marcar el tiempo sobre una escala temporal común: la duración transcurrida desde el instante inicial, la edad de los individuos.
- Por último, es necesario que el evento mismo tenga una definición clara, como suele ocurrir con los eventos demográficos cuya fecha se registra. Consideremos, sin embargo, la necesidad de definir bien cuál es el tipo de evento que se toma en cuenta. Así, las migraciones se pueden definir de maneras muy distintas que habrá que precisar muy bien (conjunto de mudanzas o migraciones intercomunales, por ejemplo, que eliminan las mudanzas intracomunales). Lo mismo sucede con los cambios profesionales (cambios de profesión detallados, cambios de estatus profesional, cambios de categoría socioprofesional).

Con afán de claridad partiremos del caso más simple, que es aquel en que se trabaja sobre una población donde todos los individuos tienen la misma probabilidad de experimentar un evento dado, en cada instante. Esta probabilidad podrá, sin embargo, cambiar en el transcurso del tiempo.

A partir de ese caso teórico simple, podemos introducir una heterogeneidad más compleja de la población sobre la que aquí se trabaja y hacer que intervenga un número creciente de eventos diversos para analizarlos

simultáneamente. Se indicará entonces por qué los modelos presentados en esta obra constituyen una generalización de la demografía diferencial y de los modelos markovianos.

A) Análisis de una cohorte homogénea que experimenta un solo evento

Este caso simple, que corresponde al análisis demográfico de un evento, no introduce nada nuevo respecto de la demografía clásica, pues no es más que una especificación generalizable a casos más complejos.

1) Distribución de la ocurrencia de los eventos, notaciones y formalización

En el seno de una población homogénea los individuos son susceptibles de acceder a un evento dado en la fecha T. Entonces T es una variable aleatoria positiva cuya distribución examinaremos.

Esta variable aleatoria se puede especificar de diversas maneras, pero en el marco de nuestro análisis hay tres tipos de especificaciones que se revelan como las más útiles: la función de permanencia o supervivencia en un estado, la función de densidad de probabilidad (densidad) y, por último, la función de densidad de probabilidad condicional de mantenerse el estado inicial (densidad condicional), o función de riesgo que resulta en un cociente instantáneo de ocurrencia.

La función de permanencia, ya sea el tiempo continuo o discreto, se define como la probabilidad de que la fecha de ocurrencia del evento T sea superior o igual a una fecha t dada:

$$S(t) = P(T \geq t) \quad 0 < t < \infty \tag{1}$$

Esta función es con toda claridad no creciente, continua a la izquierda con $S(0) = 1$ (y solamente si se estudia una muestra en el que todos los individuos experimentan el evento: $\lim_{t \to \infty} S(t) = 0$).

Si bien en la práctica no todos los individuos experimentan el evento esta propiedad límite de la función de permanencia suele ser necesaria. En particular, para introducir a los individuos que jamás experimentan el evento se recurre al artificio matemático siguiente: se establece la hipótesis de la existencia de un punto en el infinito tal que $\lim_{t \to \infty} S(t) = 0$.

Generalmente la distribución T puede tener componentes continuos o discretos.[1] Expresemos primero los términos de ésta en cada uno de los casos.

[1] Tal es el objeto de la discusión que desarrollaremos.

• Si la distribución T es continua, se define la función de densidad de probabilidad como el límite cuando $\Delta t \to 0$ de la probabilidad para que la ocurrencia del evento T esté comprendida en el intervalo $[t,\ t+\Delta t]$ dividido entre Δt :

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t} \tag{2}$$

sea además:

$$f(t) = -\frac{dS(t)}{dt} = -S'(t) \tag{3}$$

inversamente entonces, se tendrá:

$$S(t) = \int_t^{\infty} f(s)ds \tag{4}$$

y que todos los individuos experimenten o no el evento:

$$\int_0^{\infty} f(s)ds = \left[-S(s)\right]_0^{\infty} = -0 + 1 = 1$$

La densidad condicional o función de riesgo define en la fecha t el cociente instantáneo de ocurrencia del evento condicionalmente a la permanencia hasta t por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T \ge t)}{\Delta t} \tag{5}$$

sea además:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\dfrac{-dS(t)}{dt}}{S(t)} = -\frac{d\,\mathrm{Log}\,S(t)}{dt} \tag{6}$$

$h(t)$ es una especificación de la distribución de T, pues al integrar, sabiendo que $S(0) = 1$, se obtiene:

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) = \exp\left(-H(t)\right) \tag{7}$$

y

$$f(t) = h(t)\exp\left(-\int_0^t h(s)ds\right) = h(t)\exp\left(-H(t)\right) \tag{8}$$

donde $H(t)$ es la integral de la densidad condicional o función de intensidad acumulada.

• En el caso en que T es una variable aleatoria discreta y toma valores $t_1 < t_2 < \ldots$ con las probabilidades $f(t_i) = P(T = t_i) = f_i$ entonces para permanecer en un estado dado hasta T, $\left(P(T \geq t)\right)$ es necesario y suficiente salir de éste después de la fecha t, de donde:

$$S(t) = \sum_{i|t_i \geq t} f(t_i) = \sum_{i|t_i \geq t} f_i \tag{9}$$

La densidad condicional o función de riesgo h_i a permanecer hasta t_i es la probabilidad condicional de ocurrencia T en t_i, sea:

$$h_i = P\left(T = t_i | T \geq t_i\right) = \frac{f_i}{S(t_i)} \tag{10}$$

Finalmente, una condición necesaria y suficiente es la de no haber partido de allí en ninguna de las fechas (aquí puntos discretos) que preceden a t, de donde:

$$S(t) = \prod_{i|t_i < t} (1 - h_i) \tag{11}$$

para que en el caso discreto se tenga igualmente $S(t) = \exp\left(-H(t)\right)$ se toma por convención:

$$H(t) = -\sum_{i|\, t_i < t} \log(1 - h_i) \tag{12}$$

función de intensidad acumulada y si los h_i son pequeños:

$$H(t) \simeq \sum_{i|t_i < t} h_i \tag{13}$$

2) La función de verosimilitud

Aunque tenemos la intención de regresar en cada caso particular a los métodos de cálculo y de maximización de la verosimilitud total o parcial, en este capítulo de formalización del análisis conviene presentar la función de verosimilitud a la que los mecanismos de truncamiento le han dado una forma especial. Distinguiremos, como lo hicimos antes, el caso en el que la función de permanencia es continua y aquel en que es discreta.

Siendo continua la función de permanencia, una observación de ocurrencia en la fecha t contribuye a la verosimilitud (esto es, la probabilidad de observar las ocurrencias tal y como fueron recolectadas) mediante un término $f(t)$ que indica la función de densidad de probabilidad de ocurrencia

en t. En cambio, si la observación está truncada en τ, su contribución a la verosimilitud corresponde a su probabilidad de permanencia más allá de τ, sea $S(\tau)$. La verosimilitud total para N observaciones individuales independientes indexadas para j es entonces:

$$L = \prod_{t} f(t_j) \prod_{\tau} S(\tau_j) \tag{14}$$

En el caso en que la función de permanencia es discreta, donde a cada punto t_i corresponde la probabilidad $f_i = P(T = t_i)$, la contribución de una observación de ocurrencia es f_i y de una pérdida de observación en τ es por convención:

$$P(T > \tau) = S(\tau^+) = 1 - \sum_{i|t_i \leq \tau} f_i \tag{15}$$

como lo vimos con anterioridad, en ese caso tendremos:

$$S(\tau^+) = \prod_{i|t_i \leq \tau} (1 - h_i) \tag{16}$$

y

$$f_i = h_i \prod_{k<i} (1 - h_k) \tag{17}$$

De ahí resulta que la verosimilitud total se obtendrá contando el número de acontecimientos d_i en cada punto t_i y N_i que es el número de individuos sometidos a riesgo en ese mismo punto. La contribución a la verosimilitud total es pues:

$$L_i = h_i^{d_i} (1 - h_i)^{N_i - d_i} \tag{18}$$

en consecuencia la verosimilitud total es:

$$L = \prod_{i} h_i^{d_i} (1 - h_i)^{N_i - d_i} \tag{19}$$

Esto corresponde a la verosimilitud que se obtendría en presencia de una serie de binomiales independientes con N_i lanzamientos y una probabilidad de éxito igual a h_i.

3) Tiempo discreto, tiempo continuo

Cuando anteriormente formalizamos el análisis en sus términos más generales, pudimos medir la proximidad de los estimadores obtenidos según fuera

T una variable aleatoria discreta o continua, y esto es aún más evidente si se especifica la función de permanencia en el caso de una distribución mixta. En efecto, si llamamos h_c a la densidad condicional o función de riesgo para la parte continua y $x_1 < x_2 < \ldots$ a los puntos de la parte discreta:

$$S(t) = \exp\left(-\int_0^t h_c(s)ds\right) \prod_{j|x_j<t} (1-h_j) \tag{20}$$

de donde

$$h(t)dt = h_c(t)dt + \Sigma h_j \delta(t-x_j)\,dt \tag{21}$$

donde

δ es la función delta de Dirac: $\delta(x) = \begin{cases} \text{si } x = 0 \\ 0 \text{ si no es igual} \end{cases}$

la función de intensidad acumulada es entonces:

$$H(t) = \int_0^t h(s)ds = \int_0^t h_c(s)\,ds + \sum_{j|x_j<t} h_j \tag{22}$$

donde los componentes, funciones delta de Dirac, definen las contribuciones discretas de la integral. La función de permanencia puede entonces escribirse en el caso discreto, continuo o mixto:

$$S(t) = \mathop{\mathcal{P}}_0^t \big(1 - dH(t)\big) \tag{23}$$

donde el producto integral $\mathcal{P}$ se define de manera análoga a una integral de Riemann.[2] El intervalo $[0, t]$ está escindido en n pequeños intervalos $[0, t_1\,[, [t_1, t_2\,[\ldots [t_{n-1}, t[$. Se considera entonces el límite cuando $n \to \infty$ y $\max (t_{j-1} - t_j) \to 0$ de:

$$\mathop{\mathcal{P}}_0^t \big(1 - dH(t)\big) = \lim \prod_1^n \big(1 + H(t_{j-1}) - H(t_j)\big) \tag{24}$$

se puede entonces escribir que:

$$S(t) = \mathop{\mathcal{P}}_0^t \big(1 - h(u)\,du\big) \tag{25}$$

donde una vez más las funciones delta de Dirac toman a su cargo las contribuciones discretas, así que para la parte continua de la densidad condicional o función de riesgo volvemos a encontrar:

[2] Véase anexo 1.III.

$$\prod_0^t (1 - h_c(s)\,ds) = \exp\left(-\int_0^t h(s)\,ds\right) \tag{26}$$

Vemos así con claridad que en los dos casos, tiempo discreto o continuo, nos enfrentamos con el mismo problema.

Los métodos en tiempo continuo son aquellos que suponen que la fecha de ocurrencia se mide con precisión. En efecto, cualquiera que sea la división temporal que se plantee, las fechas de ocurrencia de los eventos se medirán en unidades discretas, y cuando esas unidades sean suficientemente pequeñas,[3] el problema se tratará en tiempo continuo.

El cociente en tiempo discreto mide la probabilidad de que un individuo experimente el acontecimiento estudiado, sabiendo que para él es factible en esa fecha. Este cociente es una variable no observada, si bien controla la ocurrencia y la cronología de los eventos. De ahí la importancia de esta variable en el modelo de análisis de las biografías.

En tiempo discreto, el número de acontecimientos en t se dividirá entre el número de individuos expuestos hasta antes de t.

En tiempo continuo, la definición precedente ya no es válida para caracterizar al cociente instantáneo. En efecto, en tiempo continuo la probabilidad de que un evento se produzca *exactamente* en la fecha t es ínfima. Por esta razón hay que dividir la probabilidad de que un individuo experimente el acontecimiento estudiado durante el intervalo $[t, t + \Delta t[$, sabiendo que él está expuesto en la fecha t, entre la duración Δt.

Aquí se ve claramente que si $\Delta t = 1$ volvemos a encontrar los términos del cociente instantáneo en tiempo discreto. Y como ya lo vimos antes (5) el cociente instantáneo o densidad condicional en tiempo continuo está dado por:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Si bien es común que nos refiramos al cociente instantáneo como a la probabilidad instantánea de ocurrencia del evento estudiado, de hecho ese cociente no tiene límite superior. Una interpretación más estricta es que este cociente representa la intensidad no observada con la que se producen los eventos, lo que intuitivamente ilustra bien la noción de riesgo. Así, si este cociente es igual a 1.5 ($h(t) = 1.5$ cualquiera que sea $t \geq 0$), eso significa que 1.5 es el número promedio esperado de acontecimientos en una unidad de tiempo, e inversamente $1/h(t)$ mide la duración promedio esperada que transcurre hasta la llegada de un evento, aquí 0.667 unidad de tiempo.

[3] Aquí se toma la posible incertidumbre. Como veremos más adelante, a menudo nos situamos en el caso continuo debido a evidentes criterios de simplicidad de cálculo.

Por último, en lo que respecta a la selección entre tiempo discreto o continuo, ambos métodos proporcionan resultados extremadamente cercanos para el análisis de los fenómenos demográficos. En consecuencia, la selección del método depende más de los costos de cálculo.

Ciertos autores proponen la alternativa de tratamiento de los datos en tiempo discreto, que ha sido desarrollada esencialmente por biometristas. Así pues, la vida de cada individuo se puede considerar como una sucesión de experiencias de Bernouilli,[4] las cuales intervienen entre sí en cada intervalo de tiempo hasta la ocurrencia del evento estudiado. Una trayectoria se resume, entonces, en una sucesión de 0 (ceros) durante todo el tiempo que el individuo permanezca en el estado que precede a la ocurrencia del evento, y termina en un 1 (uno), durante el intervalo en el que se produce el evento. Este tipo de acercamiento se puede utilizar con facilidad sobre muestras reducidas donde el periodo de observación no es demasiado largo ni está subdividido en intervalos demasiado pequeños, pues el costo informático se acrecienta con el número de casos de la tabla planteada.

Hay que insistir, sin embargo, en el hecho de que estos métodos casi siempre darán resultados muy similares a los obtenidos utilizando métodos en tiempo continuo (Arjas y Kanjas, 1988), puesto que los modelos en tiempo discreto convergen hacia las ecuaciones de los modelos en tiempo continuo una vez que se reducen los intervalos tomados en cuenta (Allison, 1982).

B) Análisis de una cohorte heterogénea y de la interacción entre fenómenos

No hay ninguna razón *a priori* para que las hipótesis que están en la base del modelo precedente deban ser verificadas. En primer lugar, las cohortes reales no están compuestas por individuos idénticos y las posibilidades de que los comportamientos de éstos sean los mismos son muy pocas. En segundo lugar, las historias de vida anteriores que han experimentado los diversos miembros de la cohorte pueden sin duda influir sobre sus comportamientos en el futuro. La gran variedad de esas trayectorias inducirá nuevos comportamientos diferentes.

Para tomar en cuenta esta heterogeneidad se han propuesto diversos métodos que aquí presentaremos muy rápidamente, para mostrar en qué aspecto son muy restrictivas sus hipótesis, antes de desarrollar un modelo más general —que lamentablemente es demasiado rico para que se pueda estimar en su totalidad—. En esta obra tendremos, pues, que contentarnos

[4] X es una variable cuya distribución se hace según una ley de Bernouilli; sea p un valor comprendido entre 0 y 1: $P(X=0)=1-p$; $p(X=1)=p$.

con presentar aproximaciones parciales, pero que se pueden estimar con nuestros datos.

1) Demografía diferencial

Podemos llamar demografía diferencial al "estudio de las diferencias entre las diversas categorías (étnicas, religiosas, sociales, etc.) de una población" (Henry, 1959). Esa demografía suele descomponer el conjunto de la población observada en subpoblaciones (hombres y mujeres, por ejemplo, o bien individuos clasificados por un nivel educativo creciente). La aplicación de los métodos de análisis demográfico clásico a cada subpoblación permite comparar los comportamientos particulares. Este tipo de análisis ha sido esencialmente transversal.

Lo que aquí nos interesa es la utilización de esos métodos en el análisis longitudinal. Esto es posible si se introduce una heterogeneidad en la cohorte observada cuyo efecto sobre los cocientes o tasas a cada edad se pueda seguir. Así, por ejemplo, se pueden calcular los cocientes de nupcialidad de las mujeres de una cohorte según su nivel educativo. La comparación longitudinal es por lo tanto de gran interés. Esta comparación presenta, sin embargo, ciertas limitaciones que si no se toman en cuenta pueden conducir a resultados erróneos.

En el análisis transversal se hace intervenir la categoría de individuo en el momento de la observación. El análisis de las diferencias de nupcialidad, fecundidad o mortalidad puede hacerse sobre categorías muy variadas como las socioprofesionales, las definidas según el ser propietario o no de la vivienda en la que se reside, etc. en el momento de la observación. En cambio, para el análisis longitudinal se necesitarán categorías definidas de una vez por todas y que no puedan cambiar durante el curso de la existencia individual. Si bien es posible comparar la nupcialidad de una cohorte de hombres y de mujeres, es mucho más delicado comparar la nupcialidad de los individuos según su nivel educativo, pues algunos se casan antes de terminar sus estudios. En cambio, es difícil poner en evidencia el efecto de la profesión de un individuo sobre su nupcialidad o su fecundidad, pues tal profesión puede cambiar sustancialmente a lo largo de su existencia.

Podríamos vernos tentados a resolver ese problema definiendo subpoblaciones según criterios *a posteriori*. Así, se estudiaría la nupcialidad de los individuos según su nivel educativo a los 50 años. Si la cohorte se observa hasta esa edad, es posible definir diversas subpoblaciones y realizar el estudio de su nupcialidad anterior de manera clásica. Este estudio nos parece poco satisfactorio, pues define de manera fija y definitiva a las subpoblaciones. Más que el nivel educativo que se tiene a los 50 años, lo que va a influir sobre

la probabilidad de casarse de un individuo dado es su trayectoria escolar, universitaria, o incluso la de los estudios realizados simultáneamente con el trabajo. Lo que nos permitirá responder a este asunto será una aproximación distinta a la diferencial, siguiendo al individuo a lo largo de su existencia en diversos dominios. Lo mismo sucederá si queremos analizar el efecto de los cambios de profesión de un individuo sobre su nupcialidad o su fecundidad, o el efecto de cualquier otra característica no fija en el tiempo.

Incluso en el caso en el que sea posible diferenciar a la población en función de características bien definidas o estables (orígenes sociales, tamaño de la familia de origen, rango de nacimiento del encuestado, etc.), su desagregación en subpoblaciones puede conducir rápidamente a efectivos sometidos a riesgos muy poco numerosos. Debido a esto, no aparecerán más diferencias significativas, por más que las pruebas sean completamente realizables (véase capítulo IV.C).

Más adelante veremos cómo la utilización de modelos más formalizados (modelos markovianos o modelos de riesgos proporcionales, por ejemplo) ayuda a evitar que la población se desagregue demasiado, produciendo pruebas mucho más satisfactorias, e incluso considerando numerosas características.

2) Modelo markoviano o semimarkoviano

Los modelos markovianos constituyen un primer paso en el análisis simultáneo de las transiciones entre un cierto número de estados. Estos modelos han sido aplicados desde hace mucho tiempo a los fenómenos demográficos: estudio de la mortalidad por causa (Keyfitz,1968); estudio de la evolución demográfica de varias regiones entre las cuales existen migraciones internas (Rogers, 1973a; 1973b). Aquí presentaremos rápidamente las hipótesis sobre las que se basan, así como los principales resultados a que conducen.

Los *procesos de Markov* que se utilizan en demografía se sitúan en un *espacio finito de estados*. Los desplazamientos de los individuos entre esos diversos estados se desarrollan en un *tiempo* que corresponde a la edad o a la duración transcurrida desde un evento de referencia. Ese evento podría ser el matrimonio para el estudio de las migraciones de la pareja. Los estados del espacio se identifican según el estatus en el que se encuentra el individuo. Por ejemplo, si se estudian las migraciones de la pareja en Francia, los diversos estados serán sus regiones de residencia a todo lo largo de la vida matrimonial. Por último, las *transiciones* se caracterizan por el paso de un estado a otro. En el caso estudiado, esas transiciones serán las diversas migraciones entre regiones de Francia.

Los fenómenos que estudiamos se caracterizan entonces por la intensidad de las transiciones de un estado al otro (o a los otros).

En la situación más simple, donde tenemos sólo dos estados, vivo/muerto, el último es evidentemente absorbente. Pero la mayoría de las etapas del ciclo de vida son estados transitorios con una transición siempre posible hacia el estado final o absorbente (la muerte), por lo que el modelo puede complicarse tanto como se desee.

Queremos además hacer énfasis en que en el estudio de los fenómenos humanos la cadena de Markov que sirve para modelizarlos puede considerarse a menudo como jerárquica, en el sentido de que una vez que se deja un estado no es posible regresar a él (luego del primer nacimiento sólo puede producirse un segundo nacimiento). Los estados de esta cadena se ordenan jerárquicamente hacia el estado absorbente final, de manera que cada uno es transitorio y sin regreso posible.

Antes de formalizar el modelo, veamos más en detalle las hipótesis que lo fundamentan. Es posible que este proceso incluya numerosos estados que un mismo individuo puede ocupar en diversos momentos de su existencia. En el caso de las migraciones interregionales en Francia, por ejemplo, el país está dividido en 21 regiones, entre las cuales se pueden estimar 420 flujos posibles. Vemos pues que los individuos pueden seguir caminos muy complejos. Las condiciones en las que éstos se producen si el modelo se verifica son, sin embargo, muy simplificadoras. La probabilidad de migrar de una región a otra depende de la edad del individuo, pero es independiente de la duración de la permanencia del individuo en la región de origen, de las diversas regiones por las que pasó con anterioridad, así como de la duración de su permanencia en cada región. La probabilidad de regresar a una región en la que se vivió antes es, entonces, la misma que si no se hubiera vivido nunca allí.

Este ejemplo nos permite ver mejor cuántas de esas condiciones tienen pocas posibilidades de ser verificadas. Un modelo de Markov de este tipo ofrecerá una visión simplificada y bastante errónea de la realidad. Para acercarse a un modelo realista habrá que resolver algunas de esas condiciones, lo cual complicará fuertemente la formulación del modelo (Courgeau, 1987).

Consideremos también que esos modelos se han aplicado con mucha frecuencia a poblaciones exhaustivas. En estas condiciones los cálculos de tasas y probabilidades conducen a resultados fácilmente interpretables, pues su varianza es muy débil. El cálculo de esta varianza se hace necesario cuando la muestra disminuye, si queremos comparar diveras estimaciones de probabilidades de transición (Hoem y Funck Jensen, 1982).

a) Formalización del modelo

Como vimos anteriormente, el tiempo debe marcarse con precisión sobre una escala común que aquí llamaremos *la duración de permanencia* (que cubre tanto

las duraciones transcurridas desde un evento inicial como la edad calendario). Como disponemos de un espacio de estados finitos, las transiciones de un estado a otro podrán depender de ese tiempo, pero serán independientes de las etapas anteriores y de las duraciones de permanencia en esas etapas.

Sean:

- t y t' duraciones de permanencia;
- I el espacio finito de los estados;
- I_i (.) los indicadores tales que:

  $I_i(x) = 1$ si el individuo está presente en el estado i en la duración x

  $I_i(x) = 0$ si el individuo no está presente en el estado i en la duración x

$P_{ij}(t, t') = P(I_j(t') = 1 | I_i(t) = 1)$ es la probabilidad de transición de la cadena de Markov del estado i al estado j entre las duraciones t y t'. La intensidad de las transiciones entre i y j se mide por el cociente instantáneo de ocurrencia, cuya significación es la misma que al inicio del capítulo pero que para el caso de transiciones denominaremos cociente instantáneo de transición de un estado a otro:

$$h_{ij}(t) = \lim_{t' \to t} P_{ij}(t, t') . \frac{1}{t' - t} \tag{27}$$

que existe cualquiera que sean $i \neq j$.

Por otra parte, se define:

$$h_i(t) = \lim_{t' \to t} \left(1 - P_{ii}(t, t') \frac{1}{t' - t} \right) = \sum_{j \neq i} h_{ij}(t) \tag{28}$$

Para que $\sum_j h_{ij}(t) = 0$, se plantea $h_{ii}(t) = -h_i(t)$

Las probabilidades de transición satisfacen las ecuaciones diferenciales de Kolmogorov, sea:

$$\frac{dP_{ij}(t, t')}{dt'} = -P_{ij}(t, t') h_j(t') + \sum_{k \neq i} P_{ik}(t, t') h_{kj}(t') \tag{29}$$

lo que se traduce igualmente por $S_i(t)$, la probabilidad de permanencia en i hasta la antigüedad t:

$$\frac{dS_i(t)}{dt} = -h_i(t) S_i(t) \tag{30}$$

lo que tiene como solución

$$S_i(t) = \exp - \int_0^t h_i(x)\,dx. \tag{31}$$

Bajo la hipótesis —que siempre será la nuestra— de intensidades constantes por intervalo, los cocientes instantáneos de transición de un estado a otro serán estimados por cocientes instantáneos de ocurrencia que, para cada intervalo de duración con el índice k, son iguales a:

$$q_k = \frac{n_k}{T_k}$$

$-n_k$: número de transiciones observadas

$-T_k$: duración total observada en el estado

Si N, el tamaño de la muestra, crece, entonces $T_k/N \to \tau_k > 0$ para todos los intervalos de antigüedad y los $U_k = \sqrt{N}(q_k - h_k)$ son independientes, normalmente distribuidos, centrados y de varianza $\sigma_k^2 = q_k/\tau_k$ [5] estimada por:

$$\hat{\sigma}_k^2 = N\,q_k/T_k$$

Como vemos, este cálculo de la varianza es simple si el tamaño de la muestra no es demasiado pequeño.

b) Relación del modelo con los procesos de conteo poissonianos

Aquí consideramos una cadena de Markov homogénea en un espacio finito de r estados cuyas probabilidades de transición verifican:

$$dP(t)/dt = Q\,P(t) \quad \text{y} \quad P(0) = I \qquad \text{matriz de identidad,}$$

cuya única solución está dada por:

$$P(t) = e^{tQ}\,;\, t > 0 \tag{32}$$

donde Q es una matriz $r \times r$ cuyos componentes independientes del tiempo verifican

$$-\infty < q_{ii} < 0, q_{ij} \geq 0 \quad \text{para} \quad i \neq j \quad \text{y} \quad \sum_{j=1}^{r} q_{ij} = 0$$

La interpretación es la siguiente:

[5] *Cf.* anexo 1.III.

Matriz Q:

- $-q_{ii}\,dt$ es la probabilidad instantánea de que un individuo en el estado i en t salga de él durante $[t\,,\,t+dt[$
- $q_{ij}\,dt$ es la probabilidad de que un individuo en el estado i pase al estado j durante $[t\,,\,t+dt[$

Matriz $P(t)$: $p_{ij}(t)$ es la probabilidad de que un individuo que salga de i esté en el estado j con una antigüedad igual a t.

Como ya vimos en el párrafo anterior, las q_{ij} se estiman directamente a partir de las transiciones observadas como cocientes instantáneos de ocurrencia, y:

$$\hat{q}_{ii} = -\sum_{j\neq i} \hat{q}_{ij}$$

Para desarrollar un modelo de movilidad como éste en el caso de una población heterogénea, se le da a la matriz Q de las transiciones la forma particular $Q=\lambda(M-I)$ y entonces las probabilidades de transición están gobernadas por:

$$P(t)=e^{tQ}=e^{t\lambda(M-I)} \tag{33}$$

Si se supone que las duraciones de permanencia están distribuidas de manera exponencial con el parámetro λ (lo que significa decir que los acontecimientos se producen según un proceso de Poisson) y que la duración media de permanencia en un estado es constante (igual a $1/\lambda$), entonces un individuo en el estado i permanece allí durante un periodo τ_0 cuya extensión está distribuida exponencialmente según:

$$\text{prob}\,(\tau_0 \geq t)=e^{-\lambda t} \qquad t>0$$

Al final de este periodo, el individuo pasa a j con una probabilidad m_{ij} (esta vez no se supone que $m_{ij}=0$, así un individuo tiene el tiempo disponible para permanecer en el mismo estado), luego él permanece en j durante un periodo τ_1...

Vemos que en ese caso las probabilidades de transición del estado i al estado j son independientes del tiempo y de etapas anteriores recorridas por el individuo.

Entonces tenemos:

$-q_{ii}dt=-\lambda\left(m_{ii}-1\right)dt$ prob. de dejar el estado i durante $(t,\,t+dt)$

y

$-q_{ij}dt=\lambda m_{ij}dt$ prob. de pasar de i a j durante $(t,\,t+dt)$.

Una interpretación del modelo se refiere a la subordinación de un proceso $Y(t)$ a un proceso markoviano $X(t)$ utilizando un conteo poissoniano $T_\lambda(t)$ como reloj intrínseco (Feller, 1968).

Si $(Y(t))_{t>0}$ son las variables aleatorias que describen la biografía de un individuo, entonces éstas se escriben:

$Y(t) = X(T_\lambda(t))$ donde X es una cadena de Markov discreta cuya matriz de transición es M y prob $(Y(t) = j | Y(0) = i) = p_{ij}(t)$ es el componente (i, j) de $P(t)$.

La ecuación se interpreta entonces como:

(i) espera en i hasta el primer salto de tiempo de un proceso de Poisson;
(ii) en este instante se cambia (o se permanece) según una matriz de transición M;
(iii) espera en j hasta un nuevo salto del proceso de Poisson...

Ahora bien, la probabilidad de que exactamente n transiciones tengan lugar durante $(0, t)$ es:

$$\text{prob}\ (T_\lambda(t) = n) = \frac{(\lambda t)^n e^{\lambda t}}{n!} \tag{34}$$

por otra parte, los destinos de las transiciones están gobernados por:

$$\text{prob}\ (X(k+1) = j | X(k) = i) = m_{ij}$$

tenemos pues la representación del estado de la población dada por M, la dinámica la aporta el reloj $T_\lambda(t)$ y así tenemos:

$$P(t) = e^{t\lambda(M-I)} = \sum_{n=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^n}{n!} M^n \tag{35}$$

Este término general permite tomar una construcción del modelo según la cual la propensión a cambiar y la probabilidad de destinos están identificadas separadamente por λ y M.

c) Generalización del modelo

Ya mencionamos hasta qué punto este modelo, no obstante lo complejo de su formulación, constituía una caricatura de los comportamientos humanos. En efecto, si en la modelización de un solo fenómeno parece satisfactorio tomar en cuenta la duración de permanencia desde el inicio de la observación, cuando se modelizan numerosos fenómenos es necesario ir más allá.

Resulta notorio que mientras más tiempo haya pasado el individuo en un estado dado (residencia, ocupación profesional) menor será su propensión a dejar ese estado. En el estudio de las migraciones, numerosos trabajos han presentado modelos estocásticos de movilidad espacial (Courgeau, 1973) en los que la ley de inercia acumulada es una noción central (McGinnis, 1968), y esos trabajos sobre los procesos semimarkovianos han sido retomados en el terreno de las migraciones (Ginsberg, 1971) y también en el de la movilidad profesional (Singer y Spillerman, 1974). La existencia de esta inercia acumulada ya no se adapta al modelo markoviano, puesto que además de la duración de la vida del individuo, ella introduce una duración de permanencia en cada uno de los estados sucesivos; por ello el proceso se hace no markoviano. Efectivamente, si se puede tomar en cuenta el tiempo transcurrido desde el origen en una modelización que entonces se llama cadena de Markov no homogénea, la duración de permanencia en el último estado no se puede simplemente incorporar.

Para volver a situarse en un esquema markoviano, una de las soluciones adoptadas es la de de multiplicar el espacio de los estados. En efecto, para regular el problema de los efectos de la edad o de la generación que se plantea cuando se hace el análisis de una muestra de individuos, la muestra se puede dividir para que el proceso se ejecute en un nuevo espacio de estados que lo hace markoviano. Las probabilidades de transición son así igualmente específicas a la inercia de cada estado considerado.

Otra posibilidad es la de desarrollar un modelo no markoviano que introduzca una heterogeneidad no observada en la población.

La formulación mediante la ayuda del conteo poissoniano es muy útil si se quiere modelizar la heterogeneidad en el seno de una población (Singer y Spillerman, 1974). En efecto, es posible imaginar no un cociente único λ, sino una matriz de coeficientes λ_i, según el estado de partida del individuo.

Si se plantea la hipótesis de que la influencia de los factores no observados (propensión de los individuos a experimentar el acontecimiento) se puede resumir en un vector U de marcadores $\{u_1, \ldots, u_k\}$ cuyas probabilidades asociadas son $\{\lambda_1, \ldots, \lambda_k\}$, y que los λ_i son independientes de las características individuales observadas, Heckman y Singer proponen considerar una influencia proporcional de los factores no observados sobre la función de permanencia. Así, para calcular cuál es la contribución de una observación a la verosimilitud, se "suma" sobre lo inobservado:

$$S(t) = \Sigma S(t \mid u_i)\lambda_i \tag{36}$$

es la contribución de una observación truncada en t.

El más simple de los modelos considerará sólo dos puntos, al ser cada individuo de un tipo o del otro, indexados por $u_1 = 0$ y $u_2 = \theta$, cuyas probabilidades asociadas son λ y $1 - \lambda$.

Así, la modelización de la heterogeneidad realiza la asociación de dos distribuciones de las cuales una puede ser discreta.

La función de permanencia en la duración t es pues:

$$S(t) = \lambda S(t \mid 0) + (1 - \lambda) S(t \mid \theta). \tag{37}$$

El modelo de migrantes-sedentarios constituye un desarrollo de este modelo (véase capítulo VII.A.2).

En ese caso límite, uno de los grupos se considera como jamás sometido a riesgo, por lo que la función de permanencia se vuelve:

$$S(t) = \lambda S(t \mid 0) + (1 - \lambda) \tag{38}$$

y los cocientes instantáneos:

$$h(t) = f(t \mid 0) / \left[\lambda S(t \mid 0) + (1 - \lambda)\right] \tag{39}$$

Este modelo tiene numerosas aplicaciones en demografía, como el estudio del primer matrimonio hecho por Coale y McNeil (1972). En su análisis, el parámetro λ mide la proporción de la población expuesta al evento, y la función de permanencia en el estado de soltería $S(t \mid 0)$ para aquellos que son susceptibles de casarse se construye como una convolución[6] de una distribución normal y de "retrasos" distribuidos exponencialmente.

En un desarrollo más detallado de este modelo introduciremos una heterogeneidad no observada más compleja, en el capítulo VII.A.2.

Para terminar, podemos decir que la introducción de modelos markovianos ha permitido un avance muy fructífero en la formulación de situaciones demográficas complejas, que implican numerosos estados. Su objetivo es esencialmente descriptivo, lo que explica que se utilice principalmente para proyecciones de poblaciones. Estos modelos, en cambio, casi no permiten un examen y una explicación más completa de los comportamientos humanos. Tal es el análisis que ahora nos toca presentar.

[6] Al ser T y U dos variables aleatorias independientes, la ley de su suma es $P_{U+T} = P_U + P_T$; si sus densidades son respectivamente f_U y f_T la densidad de la suma se construye como la convolución de las densidades:

$$f(t) = f_U * f_T(t) = \int f_U(t - v) f_T(v) dv$$

C) Hacia un análisis más completo de los comportamientos humanos

En relación con los modelos markovianos, el objeto del análisis ha cambiado. Ya no nos preocupamos más por prever cómo serán las poblaciones en los diversos estados durante el curso del tiempo, sino por analizar la biografía de un individuo, intentando ver qué etapas anteriores lo han podido conducir al estado actual.

Para hacer eso vamos a considerar que una biografía es el resultado de un proceso estocástico complejo sobre el cual van a influir las diversas características del medio donde vive el individuo.

A lo largo del tiempo el individuo experimentará sucesivamente diversos eventos de tipos diferentes y se encontrará sometido a condiciones exteriores variables. Así, por ejemplo, puede efectuar una migración antes de adquirir un primer empleo, esto puede ir seguido por un matrimonio y una migración simultáneos, más tarde tendrá sucesivamente dos niños antes de realizar una tercera migración, etc. Puede ser que a lo largo de ese mismo tiempo el individuo haya estado sometido primero a condiciones exteriores muy fuertes provenientes de sus padres (orientación escolar y profesional ligada a los deseos *a priori* de los padres, por ejemplo) y que luego, habiendo comenzado a trabajar, tales influencias de los padres se desvanezcan y él tenga que enfrentar las condiciones del mercado de trabajo que orientarán su profesión en una dirección privilegiada; más tarde, su matrimonio podrá ponerlo bajo nuevas condiciones tanto de empleo como de relaciones personales, etcétera.

Vemos, pues, que cada uno de los eventos se puede caracterizar por la fecha en la que se produjo y por el tipo de suceso del que se trata (matrimonio, nacimiento, migración, cambio profesional, etc.). De esa manera se podrán descubrir las interacciones entre fenómenos demográficos y analizarlas en términos correctos. Simultáneamente, en cada instante de la vida de un individuo se define cierto número de características que pueden tener influencia sobre el desarrollo de su existencia.

La existencia de un individuo se representa, entonces, mediante una serie de parejas de variables aleatorias (T_i, J_i), donde T_i es el instante de llegada del $i^{\text{ésimo}}$ evento de tipo J_i, y mediante una serie de variables $z_i(t)$, que representan todas las características del individuo antes del instante t. Los diversos instantes verifican la relación:

$$T_1 \leq T_2 \leq T_3 \ldots \leq T_i \leq \ldots$$

De manera semejante, en el caso donde se observa una población homogénea se puede definir el cociente instantáneo de ocurrencia del $i^{\text{ésimo}}$ evento de tipo j, para un individuo que tiene las características $z_i(t)$, en el instante t:

$$h_{ij}\left(t, z_i(t)\right) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\left(T_i < t + \Delta t, J_i = j \middle| T_i \geq t; (t_1, j_1); \cdots; (t_{i-1}, j_{i-1}); z_i(t)\right) \quad (40)$$

donde t_k representa el instante de llegada del $k^{\text{ésimo}}$ evento de tipo k y donde $z_i(t)$ representa el conjunto:

$$z_i(t) = \left\{z_k(t): (t_k, j_k), k = 1, \ldots, i-1\right\}$$

Igualmente se puede calcular el cociente instantáneo para el $i^{\text{ésimo}}$ evento, cualquiera que sea su tipo, bajo la forma:

$$h_i\left(t, z_i(t)\right) = \sum_{j=1}^{m} h_{ij}\left(t, z_i(t)\right) \quad (41)$$

donde m es el número total de tipos de eventos a los que está sometido el individuo en el instante t.

Cuando $z(t) = z$, es decir, cuando las características del individuo no dependen de t, se puede calcular sin dificultad una función de verosimilitud correspondiente a los diversos eventos de su biografía.

Para el primer periodo observado $[0, t_1]$ debemos introducir la probabilidad que él no haya conocido ningún evento antes de t_1. Al igual que anteriormente, esta probabilidad es igual a:

$$\exp\left(-\int_0^{t_1} h_1(t, z)\, dt\right)$$

En seguida el individuo conoce el evento de tipo j_1 en el instante t_1, cuya probabilidad se escribe:

$$h_{1j_1}(t_1, z)\, dt$$

A partir de la pareja (t_1, j_1), se puede calcular la probabilidad de no conocer ningún evento en el curso del periodo (t_1, t_2), lo que es igual a:

$$\exp\left(-\int_{t_1}^{t_2} h_2(t, z)\, dt\right)$$

Se puede entonces continuar este procedimiento hasta el momento de la encuesta que conduce a la contribución con la verosimilitud siguiente:

$$\prod_{i=1}^{m}\left[\exp\left(-\int_{t_{i-1}}^{t_i} h_i(t, z)\, dt\right)\right]\left[h_{ij_i}(t_i, z)\right]^{\delta_i} \quad (42)$$

donde δ es una variable igual a la unidad salvo para el final del último intervalo abierto, donde ésta es igual a cero. Al calcular el producto de todas las verosimilitudes individuales, se obtiene la verosimilitud de la muestra observada. La aplicación del método de máxima verosimilitud debería conducir entonces a la estimación de todos los parámetros del modelo, si éstos no son demasiado numerosos.

Este método se puede generalizar en el caso en que $z(t)$ dependa del tiempo, utilizando productos integrales (Kalbfleisch y Prentice, 1980, p. 182).

En efecto, dado el muy elevado número de cocientes instantáneos que hay que estimar, las múltiples características que deben considerarse y el número limitado de individuos encuestados, el analista no puede, de momento, estimar esta distribución conjunta como un todo. Deberá contentarse con aproximaciones parciales que, según ciertas hipótesis, permiten la estimación de esos cocientes y la del efecto de diversas características individuales sobre esos mismos cocientes.

Pese a lo anterior el cuadro general que hemos presentado aquí sigue siendo válido para todos esos análisis parciales.

El análisis no paramétrico aporta una primera serie de respuestas acerca de la diferenciación de los comportamientos. Cuando decimos "no paramétricos", nos referimos a métodos que no presuponen la distribución del evento o de los eventos estudiados. No hay ninguna ley conocida que se ajuste a los datos disponibles. En ese sentido este primer tipo de análisis, preliminar en el estudio de un fenómeno, se emparenta con los métodos clásicos de análisis longitudinal en demografía, si bien éste hace intervenir la posibilidad de analizar simultáneamente varios eventos. Una segunda diferencia con los métodos clásicos de análisis consiste en el cálculo sistemático de las varianzas, desviaciones estándar, covarianzas de los estimadores y, en consecuencia, de estadísticas elaboradas que permiten la comparación. En efecto, los datos disponibles para esos análisis, ya se trate de la encuesta "Triple biografía" o de otras muestras, jamás son exhaustivos sobre una población, y las estratificaciones que les son propias hacen necesario el cálculo de intervalos de confianza para las estimaciones.

Aun si esos datos fueran exhaustivos, el número tan elevado de situaciones en las que se puede encontrar un individuo reduciría los efectivos sometidos a riesgo, y haría necesario el cálculo de esas varianzas y covarianzas. Así, en presencia de una muestra importante, una estratificación muy fina de los datos (por estado, por edad y por duración a partir de un evento origen, por tipo, etc.) hace necesarios esos cálculos a fin de razonar con total seguridad.

A continuación haremos intervenir diversas características de los individuos implicados, mediante la aplicación de métodos paramétricos que constituyen una generalización de los métodos de regresión habituales para

las biografías. Por último, los métodos semiparamétricos harán la síntesis entre los métodos no paramétricos y los paramétricos.

D) Conclusión

En este capítulo presentamos el vocabulario del análisis de las duraciones de permanencia así como las notaciones más usuales dentro del cuadro simplificado del análisis de un solo evento, que se presenta en una cohorte homogénea. También se discutió aquí la consideración de un tiempo continuo o discontinuo.

Para hacer intervenir la heterogeneidad de las cohortes y la interacción entre fenómenos, mostramos que los métodos de la demografía diferencial, útiles en lo transversal, se revelan como insuficientes cuando se aborda lo longitudinal. De igual manera los modelos markovianos, útiles para realizar proyecciones de población, no están adaptados al análisis y explicación de las interacciones entre fenómenos demográficos. Presentamos entonces en forma suscinta los métodos de análisis que proponemos en esta obra y que constituyen una generalización de la demografía diferencial y de los modelos markovianos. Lamentablemente, por el momento no es posible aplicar ese modelo de conjunto a los datos de una encuesta, pues es precisa la estimación de un número demasiado elevado de cocientes y parámetros.

Será necesario entonces que restrinjamos el campo de aplicación a eventos menos numerosos, con características individuales en un número más reducido. No obstante, estos análisis son ya suficientemente complejos y permitirán despejar el terreno.

Por el momento se trata de confrontar las herramientas con los datos, comenzando por los análisis menos ambiciosos y teniendo siempre en cuenta los límites impuestos por la naturaleza misma de la información recogida.

III. MÉTODOS DE ESTIMACIÓN A PARTIR DE OBSERVACIONES TRUNCADAS

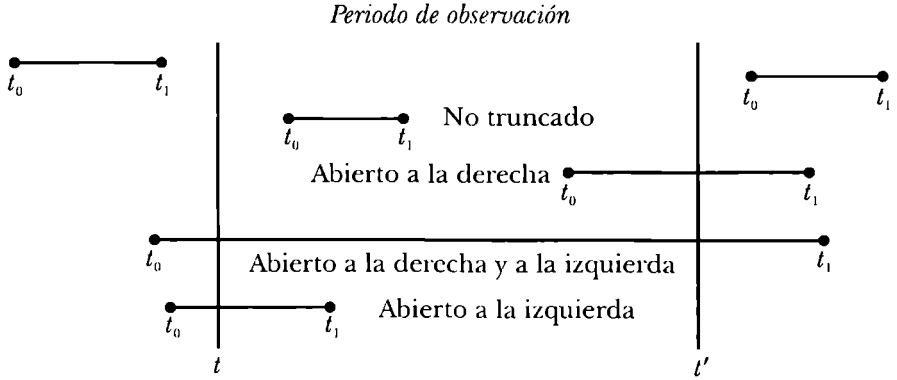
Presentaremos aquí el principio de los métodos de estimación no paramétricos de los cocientes, a partir de diversos tipos de observación de datos que se pueden truncar tanto a la derecha como a la izquierda.

A) Presentación de los truncamientos

Los datos de tipo longitudinal, retrospectivos o prospectivos, según el modo de recolección presentan lagunas de naturaleza diferente. Si el intervalo de tiempo que cubre la observación se extiende desde la fecha t hasta la fecha t' y si se estudia la evolución de un proceso que se inicia en t y se termina en T con $T = t_1 - t_0$, se presentan cuatro situaciones si excluimos los casos donde no se observa ninguna parte del desarrollo del proceso (figura 1):

1) el proceso es enteramente observado, el intervalo no es truncado por la observación;

Figura 1
Diversos tipos de truncamiento de un proceso simple que comienza en t_0 y se termina en t_1, mientras que el periodo de observación comienza en t y se termina en t'



2) t sobreviene después del inicio del proceso, los datos están truncados a la izquierda;

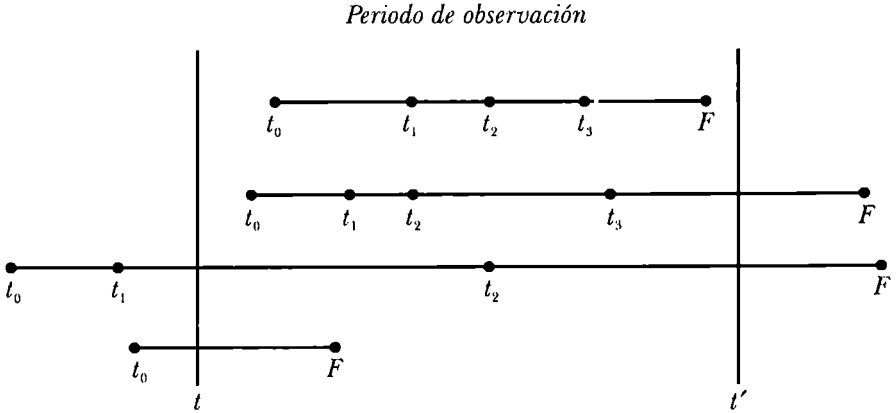
3) t' sobreviene antes del final del proceso, los datos ahora están truncados a la derecha;

4) por último, los datos pueden estar truncados a la vez a la izquierda y a la derecha.

(Asimismo se utilizan los términos siguientes: intervalos abiertos a la derecha e intervalos abiertos a la izquierda).

Se hablará entonces de proceso simple para referirse a un tipo de evento que define un intervalo de manera única, tal como el primer matrimonio, la muerte; pero también hay que considerar procesos de ocurrencias múltiples, como los nacimientos o las migraciones sucesivos. Este último tipo de proceso (figura 2) se desarrolla entre t_0 y F, fecha de finalización, que puede ser anterior a la muerte del individuo o al final del periodo fecundo.

FIGURA 2
Caso de un proceso de ocurrencias múltiples
que se termina en F con el mismo periodo de observación $[t, t']$



El tomar en consideración las observaciones incompletas para el cálculo de los diferentes cocientes instantáneos o duraciones medias de permanencia en un estado es indispensable, sin embargo, presenta dificultades. Según el tipo de truncamiento se proponen respuestas diferentes.

Aquí presentamos dos tipos de soluciones respecto de los datos truncados a la derecha o los truncados a la izquierda.

Como ya expusimos, cualquiera que sea el método de recolección de datos, los intervalos de observación ofrecen prácticamente siempre trun-

camientos a la derecha. Primero demostraremos que éstos no resultan demasiado problemáticos y recordaremos las hipótesis que en ese caso subyacen a las aplicaciones (Feller, 1968).

Después nos referiremos a los intervalos abiertos a la izquierda, que también son muy frecuentes en la práctica. Éstos presentan la particularidad de sesgar los resultados, y hasta el momento no contamos con métodos eficaces para enmendar esos sesgos. Si se percibe que esos truncamientos presentan más problemas a la izquierda del intervalo, puede concluirse entonces que no tenemos capacidad para estimar los efectos del desarrollo pasado del proceso en observación presente o en su previsión. Ante la ausencia de método se solía establecer la hipótesis, clásica aunque insatisfactoria, de que el proceso había comenzado en t —aun cuando se sabe que es anterior—, o bien suponer que el pasado desconocido del proceso no afecta en nada su desarrollo actual, lo cual evidentemente no es muy realista.

B) Truncamientos a la derecha

Es posible hacer una estimación a partir de datos retrospectivos si se toman en cuenta los intervalos truncados a la derecha por la fecha de la encuesta.

La estimación de los tiempos promedio de espera entre los eventos, y la de los cocientes de ocurrencia de esos eventos descansa en parte sobre el hecho de que los tiempos de espera observados están incompletos.

Sin embargo, cuando se examina únicamente el último evento (última migración, último nacimiento, etc., antes de la entrevista), el problema sólo puede resolverse imponiendo condiciones restrictivas.

Situémonos en el caso de eventos que se producen de manera repetida según una ley de Poisson: cambios de residencia, de empleo, nacimientos, etc. Las duraciones de permanencia se distribuyen entonces exponencialmente.

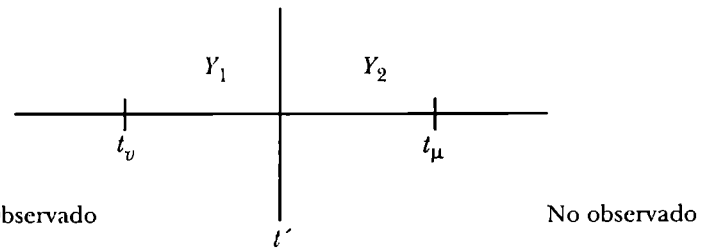
En una encuesta retrospectiva se observan todos los intervalos, de los que sólo el último puede estar truncado; sin embargo, a veces sólo se dispone de la duración transcurrida entre el último evento y la entrevista realizada en el tiempo t': Y_1 (figura 3).

El último evento observado se produjo en la fecha $t_v < t'$. El próximo (no observado) tendrá (quizás) lugar en la fecha $t_\mu > t'$.

Si el próximo evento ocurre en t_μ, la duración $t_\mu - t'$ es aleatoria y debido a las propiedades del proceso de Poisson, Y_2 está distribuido exponencialmente. Además, se demuestra que Y_1 también está distribuido exponencialmente.

Debido a esto, los intervalos truncados corresponden a duraciones de permanencia que en promedio son más largas que las observadas en el resto de la biografía. En efecto, la distribución de los intervalos truncados a la de-

FIGURA 3



recha no es exponencial, sino que tiene una densidad de tipo gamma, y la duración media de los intervalos truncados en la fecha de la entrevista es el doble de la de los otros intervalos entre eventos. Esto demuestra que la interrupción de un proceso estocástico tiende más a producirse durante intervalos largos que durante intervalos cortos, lo que corresponde a la noción intuitiva de que la entrevista tiene mayores posibilidades de interrumpir intervalos más largos entre eventos.

De esta manera, omitir las duraciones truncadas por un final de observación conduce a una fuerte subestimación de las duraciones de permanencia medias, de donde procede una sobrestimación de los cocientes instantáneos en los diversos estados, pues en ese caso se dejarían sistemáticamente de lado las duraciones de permanencia largas.

Finalmente, recordemos que para utilizar esos intervalos abiertos a la derecha se adoptó la hipótesis de independencia entre el evento estudiado y el fin de la observación: los individuos desaparecidos de la muestra se comportarían de manera similar a los que se siguen observando. Esta hipótesis, aun cuando sea realista, puede no ser verificada, en cuyo caso hay que plantear otro tratamiento de los datos truncados (exclusión, estudio aparte, etcétera).

La estimación a partir de datos retrospectivos se hará, entonces, tomando en cuenta los intervalos truncados a la derecha por la fecha de la encuesta. Según el tipo de datos recogidos se utilizarán diversos estimadores. Aquí presentamos el estimador de Kaplan-Meier para los casos en que se observa un solo evento, y el estimador de Aalen que generaliza el anterior a situaciones más complejas.

1) Estimador de Kaplan-Meier

En lo que respecta a la formalización de la estimación no paramétrica, las propuestas de Kaplan y Meier (1958) constituyen siempre una referencia. Describiremos sus técnicas antes de entrar a detallarlas, a fin de presentar de

la mejor manera posible la innovación y el aporte que representaron. En efecto, ellas se preocuparon por primera vez de la estimación de los datos truncados a la derecha y permitieron estimar una función de permanencia que la toma en cuenta.

Tal como vimos, esos truncamientos se deben a la naturaleza de los datos. En efecto, sea cual fuere el tipo de recolección, los datos biográficos se truncan allí donde se detiene el relato de vida. Se dispone, pues, de una muestra de individuos para los que se registra la ocurrencia del evento estudiado si éste tuvo lugar durante el periodo de observación, o si no, la salida de la observación del individuo (por diversas razones: muerte, migración, fecha de la encuesta, etc.), que constituye una pérdida de información en una fecha dada.

El estimador de Kaplan-Meier se establece simplemente a partir de la maximización de la verosimilitud. En efecto, si se observan eventos ocurridos en $t_1 < t_2 < \ldots < t_k$, la verosimilitud de las observaciones se forma para cada antigüedad t_i mediante las contribuciones:

$$L_i = h_i^{d_i}(1-h_i)^{N_i-d_i} \quad \text{(fórmula (18), cap. II.A.2)}$$

donde

d_i es el número de eventos ocurridos en t_i
N_i la población sometida a riesgo hasta antes de t_i
h_i el cociente instantáneo de ocurrencia en t_i

El logaritmo de la verosimilitud total es en consecuencia:

$$\log L = \sum_i \left[d_i \log h_i + (N_i - d_i)\log(1-h_i)\right] \qquad (1)$$

y el estimador del máximo de verosimilitud se obtiene derivando el logaritmo de la verosimilitud, como solución de:

$$d \log L / dh = 0$$

sea para cada $$t_i: \hat{h}_i = d_i/N_i . \qquad (2)$$

El cálculo de la varianza de este estimador se basa en la teoría estándar de los grandes números, lo que puede limitar su utilización cuando las muestras se revelan como demasiado pequeñas.[1]

Asintóticamente, en efecto, $\sqrt{n}\left(\hat{h}_i - h_i\right)$ tendrá una distribución normal de media cero y de matriz de covarianza que se estima por el inverso de la matriz de información de Fisher, sea:

[1] *Cf.* anexo 1.III.

$$\frac{d^2 \log L}{dh_j\, dh_k} = \begin{cases} -\dfrac{N_j}{\hat{h}_j(1-\hat{h}_j)} & \text{si } j = k \\ 0 & \text{si } j \neq k \end{cases}$$

de donde

$$\operatorname{var}(\hat{h}_j) = \frac{d_j(N_j - d_j)}{N_j^3} \tag{3}$$

El mismo resultado se obtendría, además, mediante binomiales independientes.

Se obtiene el estimador de la función de permanencia, llamado de Kaplan-Meier, como sigue:

$$\hat{S}(t) = \prod_{t_i<t}(1-\hat{h}_i) = \prod_{t_i<t}(N_i - d_i)N_i^{-1} \tag{4}$$

Ésta es la expresión del estimador para el caso en que ocurren varios eventos al mismo tiempo.

En el caso en que se considera un evento por fecha ($d_i = 1$) el estimador de la función de permanencia se define como sigue:

$$\hat{S}(t) = \prod_{r}\left[(N-r)/(N-r+1)\right] \tag{5}$$

donde N es el tamaño de la muestra;

$t_1 \leq \ldots \leq t_N$ edades ordenadas al momento de las pérdidas o al momento de la ocurrencia de eventos;

r número de pérdidas o de ocurrencia de eventos a la edad $t_r \leq t$.

Si no hay ninguna pérdida de información, el estimador $S(t)$ se reduce al estimador binomial usual: la proporción de sobrevivientes. Este nuevo estimador es consistente y poco sesgado, y es posible obtener una expresión asintótica de su varianza, a la que se conoce bajo el nombre de fórmula de Greenwood.

En efecto, como:

$$\log \hat{S}(t) = \sum_{t_j} \log(1-\hat{h}_j)\,,$$

la varianza asintótica de $S(t)$ se estima como sigue:

$$\operatorname{var}\left\{\log\left(\hat{S}(t)\right)\right\} \simeq \sum_{t_j} \operatorname{var}\left(\log(1-\hat{h}_j)\right)$$

$$\simeq \sum_{t_j} \left(\frac{1}{1-\hat{h}_j} \right)^2 \text{var}\,(\hat{h}_j)$$

$$\simeq \sum_{t_j} \left(\frac{1}{1-\hat{h}_j} \right)^2 \frac{\hat{h}_j(1-\hat{h}_j)}{N_j}$$

$$\simeq \sum_{t_j} \frac{d_j}{N_j(N_j-d_j)}$$

de donde:

$$\text{var}\left\{\hat{S}(t)\right\} = \left(\hat{S}(t)\right)^2 \sum_{t_j} \frac{d_j}{N_j(N_j-d_j)} \tag{6}$$

que se llama fórmula de Greenwood.

Esta varianza podrá utilizarse para comparar la población observada aquí con alguna otra.

A causa de las aproximaciones necesarias para el establecimiento de la fórmula, la función de permanencia puede calcularse para los valores extremos de t. Sin embargo, el intervalo de confianza podría incluir valores exteriores al intervalo (0, 1). Este problema se evitaría utilizando una distribución normal asintótica a $S(t)$ no restringida a este intervalo. Kalbfleisch y Prentice plantean entonces que se calcule la varianza asintótica s^2 de:

$$\log\,(-\log S(t)) = SA(t)$$

estimada mediante:

$$\hat{s}^2(t) = \frac{\sum_{i|\,t_i<t} \frac{d_i}{N_i(N_i-d_i)}}{\left[\sum_{i|\,t_i<t} \log\left(\frac{N_i-d_i}{N_i}\right)\right]^2} \tag{7}$$

Un intervalo de confianza de 95% para $SA(t)$ está dado para $SA(t) \pm 1.96\, s(t)$, lo que corresponde para $S(t)$ a un intervalo de confianza de:

$$S(t)\,\exp\,(\pm\, 1.96\, s(t))$$

que toma entonces sus valores en (0, 1).

Sin embargo, si se considera el proceso de truncamiento en tiempo discreto y las distribuciones en tiempo continuo, es posible realizar otras estimaciones de cocientes instantáneos de ocurrencia.

• Ejemplos miniatura

1. En el caso en que las ocurrencias de los eventos no implican ningún evento simultáneo se aplicará fácilmente la segunda formulación de $S(t)$.

Sean 10 individuos. Se observan fallecimientos de antigüedades 0.8; 1.0; 7.6; 9.2 años, y salidas de la observación en 0.4; 2.1; 4.7; 5.3; 8.6 y 14.0 años.

$$\hat{S}(t_1) = S(0.8) = \frac{8}{9} = 0.888$$
$$\hat{S}(t_2) = S(1.0) = \frac{7}{8} \times \frac{8}{9} = 0.777$$
$$\hat{S}(t_3) = S(7.6) = \frac{3}{4} \times \frac{7}{8} \times \frac{8}{9} = 0.583$$
$$\hat{S}(t_4) = S(9.2) = \frac{1}{2} \times \frac{3}{4} \times \frac{7}{8} \times \frac{8}{9} = 0.292$$

2. Examinemos ahora el caso en que varios individuos experimentaron el evento estudiado con igual antigüedad.

Sea una muestra formada por 250 inquilinos de un centro de asistencia; el evento estudiado es la obtención de empleo. Se observa a los individuos mientras permanecen en el centro; para quienes lo dejan antes de encontrar un empleo sólo se dispone de la fecha de su partida del centro. Los primeros eventos observados se producen de la siguiente manera, bajo la hipótesis de que las partidas ocurren luego de la obtención de empleos:

al cabo de 6 meses → 31 empleos, 12 partidas
al cabo de 7 meses → 11 empleos
al cabo de 9 meses → 10 partidas
al cabo de 10 meses → 15 empleos, 21 partidas
etcétera.

Se puede calcular entonces:

$$\hat{S}(t_1) = 1 - \frac{31}{250} = 0.876$$
$$\hat{S}(t_2) = \left(1 - \frac{31}{250}\right)\left(1 - \frac{11}{207}\right) = 0.829$$

$$\hat{S}(t_3) = \left(1 - \frac{31}{250}\right)\left(1 - \frac{11}{207}\right)\left(1 - \frac{0}{196}\right) = \hat{S}(t_2) = 0.829$$

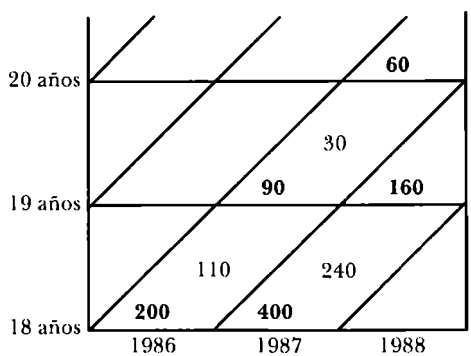
$$\hat{S}(t_4) = \left(1 - \frac{31}{250}\right)\left(1 - \frac{11}{207}\right)\left(1 - \frac{15}{186}\right) = 0.762 \qquad \text{etcétera.}$$

3. Por último, con la ayuda de un ejemplo rudimentario se pueden comparar las estimaciones obtenidas teniendo en cuenta una muestra en su conjunto o estimando por separado las funciones de permanencia para dos subpoblaciones.

De esa manera simplemente se medirá el beneficio que aporta el estimador utilizado de manera clásica en la demografía.

Sea una muestra de estudiantes de 18 años observados durante dos años, cuyo número crece a 400 individuos de 18 años extraídos de la misma población en el segundo año, y cuyas entradas en la vida activa se observan en la figura 4.

FIGURA 4
Diagrama de Lexis



Los estimadores de $S(1)$ calculados para cada subpoblación, respectivamente 0.45 y 0.40, con una desviación estándar de 0.02, no contradicen la hipótesis de la homogeneidad del conjunto de la muestra para la que: $\hat{S}(1) = 0.417$.

A continuación se puede estimar $S(2)$ únicamente con la ayuda de los datos del primer grupo: sea $\hat{S}_r\,(2) = 60/200 = 0.3$, estimador reducido de la función de permanencia en el estado de estudiante. Esto sólo es posible si se conoce todo el detalle de la observación para cada subpoblación de la muestra completa (todos los estudiantes, independientemente de que entren o no a la actividad económica).

Pero se habría podido proceder de otra manera. Calculemos primero la probabilidad de tener dos años de permanencia sin trabajo condicionalmente a un primer año sin entrada en actividad:

$$\hat{S}(2)/\hat{S}(1) = 60/90 = 0.667$$

se calcula así:

$$\hat{S}(2) = [\hat{S}(2) \mid \hat{S}(1)] \times \hat{S}(1) = 0.667 \times 0.417 = 0.278$$

Este cálculo presenta la ventaja de que ignora la diferente observación de las dos subpoblaciones y de que da una estimación de $S(2)$ que toma en cuenta el conjunto de la observación y no sólo una parte de ésta.

En efecto, para calcular $\hat{S}(2)/\hat{S}(1)$ basta con contar a aquellos que continúan expuestos al riesgo después de un año de observación, y para el cálculo de $\hat{S}(1)$ se sabe cuántos permanecen allí después del año, esto es, 250 correspondientes a una muestra inicial de 600.

2) Estimador de Aalen

Posteriormente Aalen (1978) formalizó el análisis al generalizar el cuadro teórico propuesto por Kaplan y Meier (1958) o Nelson (1969).

Presentaremos aquí su propuesta, que engloba las estimaciones planteadas por Kaplan y Meier. Una de las principales innovaciones que aportó Aalen fue la posibilidad de tomar en cuenta al analizar un evento simple (la muerte, por ejemplo) varias modalidades de ese evento (varias causas de fallecimiento, por ejemplo) y de comparar su intensidad en un cuadro teórico que no necesita establecer la hipótesis de independencia de los diferentes riesgos.

En efecto, si nos interesamos por la mortalidad en la muestra sin distinguir sus causas, entonces se impone la estimación de la función de permanencia propuesta por Kaplan y Meier, si se trabaja sobre una muestra pequeña.

Pero cuando se pretende separar los efectos de las diferentes causas de mortalidad, se presenta entonces el problema de las dependencias de éstas. Si se calculan los estimadores de Kaplan y Meier para cada causa, considerando entonces los otros tipos de fallecimiento como pérdidas, lo que de hecho se estima es una hipotética mortalidad debida a una causa específica en ausencia de otros riesgos. La hipótesis de la independencia de los riesgos, que en este caso debemos asumir obligatoriamente, es muy fuerte y sobre todo no está verificada, por lo que sería deseable otra estimación. Podemos entonces reformular el problema en términos de procesos de conteo. Estos procesos permiten estudiar individuos independientes sometidos al riesgo de eventos diversos. El proceso $N(t)$ cuenta

los eventos ocurridos en el curso de $[0, t[$, es univariado si el fenómeno es único y multivariado (N_i (t) $i = 1... k$), si se trata de una colección de k procesos de conteo que pueden ser dependientes.

Para cada uno de los procesos estocásticos N_i (t) definimos un proceso de intensidad $\Lambda_i(t)$ como la probabilidad condicional de ocurrencia del evento i en el intervalo $(t, t+\Delta t)$, conociendo la historia anterior del proceso, $F(t)$. Se escribe entonces:

$$\Lambda_i(t) = \lim_{\Delta t \to 0} \frac{E\left[N_i(t+\Delta t) - N_i(t) \middle| F(t)\right]}{\Delta t} \tag{8}$$

Aalen (1978) introdujo entonces el modelo de intensidad multiplicativa, al escribir el proceso de intensidad bajo la siguiente forma:

$$\Lambda_i(t) = h_i(t) Y_i(t) \tag{9}$$

donde Y_i (t) representa, por ejemplo, la población sometida al riesgo de experimentar el evento i, y h_i (t) representa la intensidad del evento i para un individuo (se trata en este caso de un cociente clásico). Cabe advertir que Y_i se puede definir de manera más compleja, por ejemplo cuando se estudia la difusión de una epidemia, donde Y_i (t) se puede considerar como el producto del número de individuos contagiados por la población sometida a riesgo. La intensidad permanece bien definida y mide el contagio.

La intensidad acumulada está dada entonces por la fórmula (22) del capítulo II.A.3.

$$H_i(t) = \int_0^t h_i(s)\, ds$$

Esta intensidad acumulada se puede estimar fácilmente. En el caso de un cociente de mortalidad por causa, por ejemplo, sea Y (t) la población sometida a riesgo justo antes del instante t. Llamemos $t_{i1} < t_{i2} < ...$ a las fechas observadas del fallecimiento por causa i, entonces se puede escribir el estimador de Nelson de la intensidad acumulada en el instante t.

$$\hat{H}_i(t) = \sum_{t_{ij} \le t} \frac{1}{Y(t_{ij})} \tag{10}$$

y el estimador de su varianza

$$\operatorname{var}\left(\hat{H}_i(t)\right) = \sum_{t_{ij} \le t} \frac{1}{\left[Y(t_{ij})\right]^2} \tag{11}$$

Por supuesto que se pueden observar varios fallecimientos por la causa i en el instante T_{ij}, N_{ij}, que remplazarán al numerador unitario de $\hat{H}_i(t)$ y de $\text{var}\left(\hat{H}_i(t)\right)$. Las intensidades $h_i(t)$ permanecen bien definidas, incluso cuando los riesgos son dependientes y se puede considerar que los $\hat{H}_i(t)$ son procesos casi independientes.

El aporte de Aalen consiste en su aplicación de la teoría de las martingalas a los procesos de conteo. En efecto, el estudio de los procesos puntuales multivariados, cuya teoría estadística se desarrolló a raíz de sus trabajos, requiere el empleo de la teoría de las martingalas y de las integrales estocásticas para establecer las propiedades de los estimadores y de los estadísticos de prueba (*cf.* anexo 1.III).

Por este hecho tal modelo permite que no se le imponga prácticamente ninguna estructura a los procesos N_i de conteo, lo que significa ninguna restricción sobre la interdependencia de los $Y_i(t)$ ni sobre su dependencia respecto del pasado.

Por lo tanto, para cada t, *la única restricción* importante es que los $Y_i(t)$ deben ser función de lo que sucedió antes (con la posibilidad de depender de elementos aleatorios exógenos), pero en ningún caso deben depender del futuro del proceso.

En el marco de una estimación no paramétrica de los $h(t)$ eso significa que el número de individuos sometidos a riesgo se puede modificar en cada instante t y de esa manera se pueden tomar en cuenta las pérdidas de información sobre la base de la experiencia pasada.

En este sentido se advierte claramente el carácter mucho más general de estos modelos, de los que las estimaciones de Kaplan y Meier no son más que un caso particular. En el capítulo IV detallaremos algunas aplicaciones de dichos modelos.

C) Truncamientos a la izquierda

Existen numerosos ejemplos de datos truncados a la izquierda: se conoce el lugar de residencia de un individuo en la fecha t (inicio de la observación), pero no desde cuándo reside allí; se conoce el estatus matrimonial en t, pero no la fecha del matrimonio si ésta es anterior a t, etc. De igual manera, al observar las migraciones de un individuo durante el intervalo $[t, t']$, algunas veces se dispondrá del rango de tales migraciones, pero en el caso más extremo no se registrará ninguna migración durante la observación. Ahora bien, hemos demostrado, por ejemplo, que la frecuencia y las distancias recorridas durante la infancia constituyen factores que inciden de manera muy significativa sobre el comportamiento migratorio de los futuros adultos (Courgeau, 1985; Courgeau, Lelièvre y Wagner, 1986). Así, resulta muy claro

que la hipótesis que ignora la influencia del pasado del proceso sobre su desarrollo futuro, en el caso de las migraciones provoca errores de apreciación que se manifiestan en la interpretación de los fenómenos observados.

Con el afán de poner a prueba los sesgos que introduce este truncamiento a la izquierda, dispondremos de los datos de la encuesta "Triple biografía" a la cual truncaremos artificialmente a la izquierda. De esta manera se podrá examinar la validez del método de estimación comparando los resultados obtenidos sobre el intervalo que permite la observación completa $[t, t']$ y sobre intervalos abiertos arbitrariamente a la izquierda $[t^*, t']$.

La idea central de este método propuesto por B. Turnbull (1974) es estimar el número de individuos que han experimentado el evento antes del inicio de la observación (t) a partir:

1) del número de individuos para quienes la información está truncada, y
2) de los valores de la función de permanencia que se genera iterativamente partiendo de los datos iniciales hasta lograr la obtención de una convergencia conveniente.

Los modelos propuestos dan estimadores de la función de permanencia. Esto no es restrictivo en la medida en que conjuntamente se obtienen otros estimadores (cocientes instantáneos, duraciones medias de permanencia...).

1) Un método de corrección

Supongamos que el proceso estudiado es el nacimiento del primer hijo de las mujeres que pertenecen a un grupo de generaciones dadas. Claro está que el método puede generalizarse a cualquier otro fenómeno estudiado.

Sea T la variable aleatoria que representa la edad de una mujer cuando tiene su primer hijo, contada a partir de su ingreso a la población fecunda. Como no se observa sino el periodo $[t^*, t']$, ciertos nacimientos se pueden producir antes de t^*. Sin embargo, al inicio de la observación, t^*, tenemos un número de mujeres, hayan tenido o no su primer hijo, clasificadas por la generación de nacimiento.

Representaremos entonces mediante:

μ_j el número de madres de la edad a_j, de las que no se tiene la fecha de nacimiento de su primer niño;
r_j el número de mujeres observadas que se hacen madres entre las edades a_{j-1} y a_j;
λ_j el número de mujeres sin hijos, que salen de observación a la edad a_j.

El método que presentamos aquí funciona mediante iteraciones sucesivas:
(i) Esas iteraciones se inician estimando funciones de permanencia iniciales y probabilidades de permanecer sin hijos hasta cada edad,

sobre una parte de la muestra que incluye sólo a las mujeres de quienes conocemos la edad en el momento de nacimiento de su primer hijo.

$$\left\{S_0^0, S_1^0, \ldots S_m^0\right\}$$

donde m es la edad máxima observada en el curso del periodo $[t^*, t']$;

(ii) A partir de esas funciones de permanencia se estiman los valores de los coeficientes α_{ij}^0 siguientes:

$$\alpha_{ij}^0 = \frac{S_{i-1}^0 - S_i^0}{1 - S_j^0} \qquad \text{para } i \leq j \tag{12}$$

que representan la parte de los nacimientos a la edad i de las mujeres que hayan tenido su primer hijo antes de la edad j; sea: $P(a_{i-1} < T \leq a_i \mid T \leq a_j)$, bajo la hipótesis de un comportamiento idéntico de las diversas generaciones. Lo que sigue es una primera estimación del número de mujeres que se hacen madres entre las edades a_{i-1} y a_i:

$$r_i^0 = r_i + \sum_{k=i}^{m} \mu_k \alpha_{ik}^0 \tag{13}$$

obtenido añadiéndole a las mujeres observadas que se hacen madres entre a_{i-1} y a_i, las madres no observadas repartidas según los coeficientes α;

(iii) A partir de estos nuevos valores se pueden estimar nuevas funciones de permanencia mediante:

$$S_i^1 = 1 - \frac{r_i^1}{R_i^1} \tag{14}$$

$$S_j^1 = \left(1 - \frac{r_j^1}{R_j^1}\right) S_{j-1}^1 \qquad \text{donde } j = 2, \ldots m$$

con

$$R_j^1 = \sum_{k=j}^{m} \left(r_k^1 + \lambda_k\right),$$

que es la población sometida a riesgo entre a_{j-1} y a_j.

Se puede entonces retomar (ii) con las nuevas funciones de permanencia S_j^1 que dan nuevos valores α_{ij}^1 y r_i^1, que remplazan a S_j^0, α_{ij}^0 y r_i^0.

Se continúa la iteración hasta que los estimadores S_j^ℓ convergen. Se puede, por ejemplo, hacer una prueba de esta convergencia calculando el valor máximo de $\left| S_j^\ell - S_j^{\ell-1} \right|$ y reiterar el procedimiento mientras este valor siga siendo superior al umbral que fijamos inicialmente.

2) Resultados obtenidos

Las estimaciones se hicieron[2] sobre una muestra de 511 mujeres de la encuesta "Triple biografía" nacidas entre 1926 y 1930. El intervalo que corresponde a la observación completa (ningún truncamiento a la izquierda) para el estudio de los primeros nacimientos es $[t, t']$ = (1942, 1965). El método se aplicó entonces sucesivamente a los intervalos $[t^*, t']$ siguientes: (1943, 1965), (1947, 1965), (1952, 1965).

La figura 5 presenta las diferentes curvas que se obtuvieron en el caso del intervalo amputado de 10 años, según el tipo de consideración sobre los truncamientos. Las curvas que representan la probabilidad de permanecer aún sin niños a las edades señaladas en la abscisa corresponden a las estimaciones obtenidas a partir, respectivamente, del conjunto de los datos completos (función de permanencia realmente observada), de la muestra reducida a los datos no truncados y, finalmente, de los resultados del modelo (datos reconstituidos).

La figura 6 presenta la distribución de los errores en relación con las observaciones para las dos estimaciones, obtenidos mediante el modelo de reconstitución del pasado o a partir de la muestra reducida. Resulta claro que la estimación por reconstitución se revela como una herramienta mejor que la práctica que consiste en seleccionar tan sólo a los individuos cuya observación es completa, cuando se está en presencia de una muestra cuyos datos están en parte truncados a la izquierda.

Cabe hacer, sin embargo, algunos señalamientos. Por una parte, el método se sustenta en la hipótesis de que el proceso estudiado no varía en el tiempo; así, para fenómenos muy marcados por los efectos del periodo, o en los que la evolución de una generación a otra es muy fuerte, se obtendrá una estimación mucho menos satisfactoria. Habría entonces que construir un término α_{ij} de la forma $\alpha_{ij}(t)$.

Por otra parte, este mismo método aplicado a las migraciones da resultados similares aunque menos precisos, lo cual se explica porque las migraciones se encuentran más uniformemente repartidas sobre el intervalo de observación que los primeros nacimientos concentrados a la izquierda del intervalo.

Para la aplicación del modelo a las migraciones (figura 7), luego de haber considerado la edad a la primera migración (después de los 14 años) se toma en cuenta el tiempo transcurrido entre dos migraciones sucesivas. Este indicador tiene la característica de que puede extenderse más allá del intervalo de observación. En la figura 2 encontramos los cuatro casos posibles. Además se puede probar otro modelo de estimación, que consiste en tomar en cuenta

[2] *Cf.* E. Rouy, 1986.

FIGURA 5
Función de permanencia: edad al primer nacimiento.
Datos truncados a la izquierda de 10 años (1952, 1965)

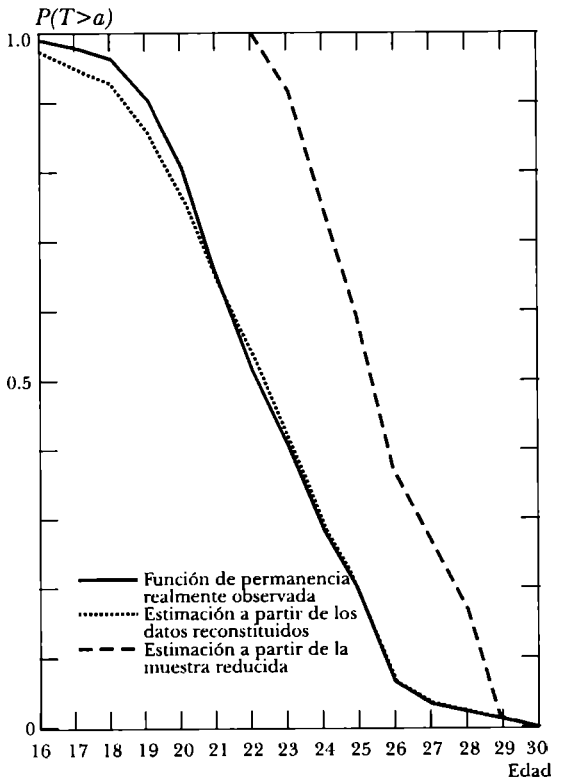


FIGURA 6
Distribución de los errores.
Datos truncados a la izquierda de 10 años (1952, 1965)

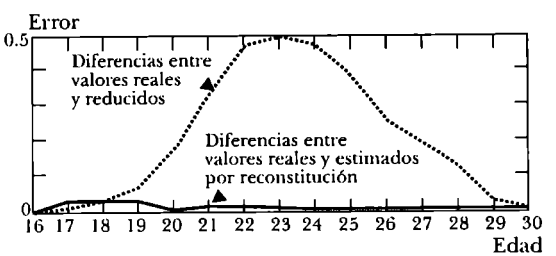
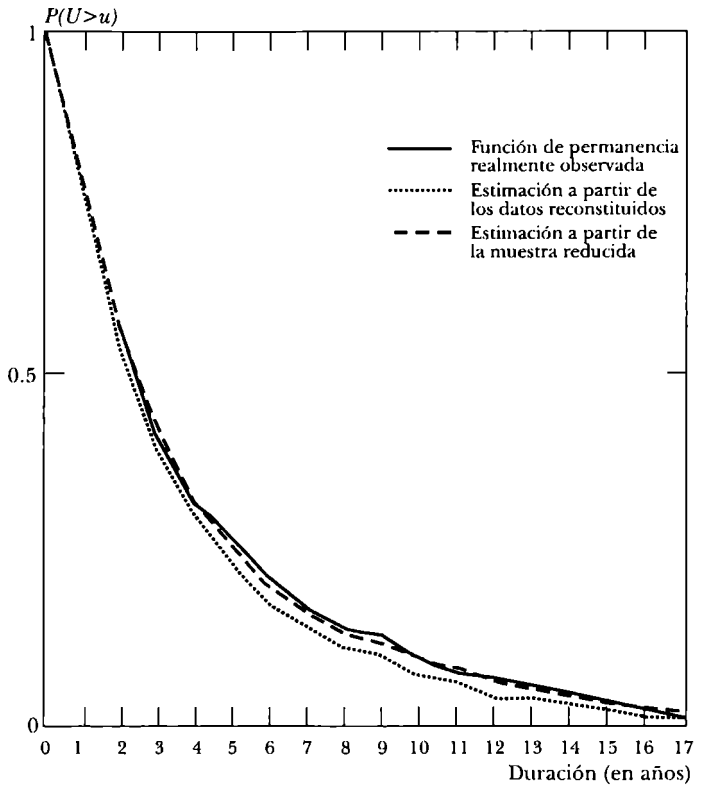


FIGURA 7
Función de permanencia entre la primera y la segunda migraciones.
Datos truncados a la izquierda de 10 años (1952, 1965)



las observaciones truncadas a la izquierda de la misma manera que se hace para las que están a la derecha, pero este modelo necesita complementarse con arreglos más complejos que aquí no presentamos, pues no están acabados.

Hemos propuesto aquí el inicio de una práctica que toma en cuenta los datos truncados. Esta práctica, tal como ya vimos, no exige que se lleven a cabo refinamientos estadísticos costosos en comparación con las mejoras reales para el análisis. Es preciso, sin embargo, contar con hipótesis muy sólidas acerca del comportamiento de la población respecto del fenómeno estudiado, que debe ser estacionario en el curso del tiempo para las diversas generaciones consideradas.

D) Conclusión

Aquí presentamos dos métodos actuariales de estimación no paramétrica a partir de la observación. Ambos fueron desarrollados sucesivamente para responder a la necesidad del cálculo en presencia de informaciones incompletas, truncadas, y luego para ser capaces de tomar en cuenta las fluctuaciones de la población sometida a riesgo.

En efecto, es necesario considerar los datos truncados para no sesgar sistemáticamente los resultados en diversas formas:

- sea subestimando todas las duraciones de permanencia en el caso de truncamientos a la derecha; esta eventualidad se toma ahora en cuenta de manera satisfactoria en todos los modelos de análisis de los datos longitudinales;
- sea limitándose al estudio de subpoblaciones particulares, lo que reduce el campo del análisis;
- sea, sobre todo, ignorando por completo el pasado del proceso, lo que implica importantes sesgos que aún se dominan poco.

Los métodos resultan eficaces para estimar las funciones de permanencia en presencia de truncamientos a la derecha. En cambio, si el intervalo de observación está abierto a la izquierda, proponemos aquí un método iterativo de estimación, el cual se basa, sin embargo, en hipótesis muy restrictivas.

Una vez precisadas las nociones fundamentales que están en la base del análisis de las biografías, vamos a considerar con mayor detalle el análisis de un evento, luego el de dos en interacción, y finalmente el de varios. Todas las estimaciones no paramétricas se desarrollarán gracias a estas nociones.

IV. ESTUDIO DE UN EVENTO

El estudio de un evento se puede hacer en el seno de una misma población o comparando varias poblaciones. Asimismo, puede ser de varios tipos como, por ejemplo, el caso de las migraciones según el destino. A continuación consideraremos el estudio de esos diferentes casos.

A) Muestra simple. Un evento único

La distribución en el curso del tiempo de las salidas de observación y de los eventos que se producen en la población puede ser objeto de diferentes hipótesis que conducen a estimaciones diversas.

1) Estimación en tiempo discreto

Sean $t_1 < t_2 < \ldots t_k$ las fechas de eventos sucesivos observados para una muestra de N observaciones independientes. Supongamos que d_i individuos experimentan el evento en t_i y que m_i individuos desaparecen entre $[t_i, t_{i+1}[$. Por último, sean $N_i = (d + m_i) + \ldots + (d_k + m_k)$ los individuos sometidos a riesgo hasta antes de t_i.

Vemos cómo el estimador de Kaplan y Meier, presentado en el capítulo III.B.1 (fórmulas (1) (3) (4)), se aplica en este caso y conduce a la estimación del cociente instantáneo de ocurrencia:

$$\hat{h}_i = \frac{d_i}{N_i}$$

y de la probabilidad de permanencia:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{N_i}\right) = \prod_{t_i \leq t} (1 - \hat{h}_i)$$

La varianza de $\hat{h}_i$ es asintóticamente igual a:

$$\operatorname{var}\left(\hat{h}_i\right) = \frac{d_i(N_i - d_i)}{N_i^3}$$

Y la de $\hat{S}(t)$ está dada por la fórmula de Greenwood ((6) capítulo III.B.1).

2) Estimación actuarial

Supongamos ahora que los d_i individuos experimentan el evento en el intervalo $[t_{i-1}, t_i[$ y que los m_i individuos salen de la observación en el mismo intervalo. Se plantea la hipótesis de que el cociente instantáneo de ocurrencia permanece constante a todo lo largo del intervalo ($h(t_i) = h_i$), y que el riesgo de truncamiento es también constante ($h_c(t_i) = c_i$) e independiente del cociente instantáneo.

Las contribuciones a la verosimilitud total $L(h, c)$ serán las siguientes para cada intervalo (condicionadas por su permanencia anterior hasta el comienzo del intervalo, con, por convención, $b_i = t_i - t_{i-1}$):

1) Los $N_i - d_i - m_i$ individuos que permanecen más allá de t_i sin desaparecer de la muestra contribuyen mediante:

$$S_{1i}(t) = \exp\left(-b_i(h_i + c_i)\right)$$

2) Los d_i individuos que experimentan el evento durante el intervalo contribuyen mediante:

$$S_{2i}(t) = \int_0^{b_i} h_i e^{-th_i} e^{-tc_i} dt = \frac{h_i}{h_i + c_i}\left[1 - \exp\left(-b_i(h_i + c_i)\right)\right]$$

3) Los m_i individuos que salen de observación durante el intervalo contribuyen mediante:

$$S_{3i}(t) = \int_0^{b_i} c_i e^{-th_i} e^{-tc_i} dt = \frac{c_i}{c_i + h_i}\left[1 - \exp\left(-b_i(h_i + c_i)\right)\right]$$

Así, la contribución para el intervalo $[t_{i-1}, t_i[$ al logaritmo de verosimilitud es:

$$l_i(h_i, c_i) = -(N_i - d_i - m_i)\, b_i (h_i + c_i)$$

$$+ d_i \log\left(\frac{h_i}{h_i + c_i}\right) + m_i \log\left(\frac{c_i}{h_i + c_i}\right)$$

$$+ (d_i + m_i) \log\left\{1 - \exp\left(-b_i(h_i + c_i)\right)\right\}. \qquad (1)$$

Los estimadores h_i y c_i se obtienen derivando y anulando el logaritmo de la verosimilitud.

$$\frac{dl_i}{dh_i} = -(N_i - d_i - m_i)\, b_i + \frac{d_i}{h_i} - \frac{d_i + m_i}{h_i + c_i} + \frac{(d_i + m_i)\, b_i e^{-b_i(h_i + c_i)}}{1 - e^{-b_i(h_i + c_i)}} = 0$$

$$\frac{dl_i}{dc_i} = -(N_i - d_i - m_i)\, b_i + \frac{m_i}{c_i} - \frac{d_i + m_i}{h_i + c_i} + \frac{(d_i + m_i)\, b_i e^{-b_i(h_i + c_i)}}{1 - e^{-b_i(h_i + c_i)}} = 0 \qquad (2)$$

lo que conduce, restando las dos ecuaciones y luego realizando la sustitución, a:

$$\hat{h}_i = -\frac{d_i}{b_i(d_i + m_i)} \log\left(\frac{N_i - d_i - m_i}{N_i}\right)$$

y (3)

$$\hat{c}_i = -\frac{m_i}{b_i(d_i + m_i)} \log\left(\frac{N_i - d_i - m_i}{N_i}\right)$$

Cuando el intervalo $[t_{i-1}, t_i]$ es pequeño (b_i pequeño) entonces $(d_i + m_i)/N_i$ es también pequeño y se puede desarrollar el logaritmo en serie (ignorando los índices i):

$$b\hat{h} = \frac{d}{N} + \frac{1}{2}\frac{d(d+m)}{N^2} + o\left(\frac{d+m}{N}\right)^3 \qquad (4)$$

donde

$$o(\varphi(\cdot)) \text{ significa que } \frac{o(\varphi(\cdot))}{\varphi(\cdot)} \to 0$$

De donde resulta:

$$b\hat{h} = \frac{d}{N - 1/2(d+m)}\left[\frac{N^2 - 1/4\,(d+m)^2}{N^2}\right] + o\left(\frac{d+m}{N}\right)^3$$

de donde

$$b\hat{h} = \frac{d}{N - 1/2(d+m)}\left(1 - \left(\frac{1/2\,(d+m)}{N}\right)^2 + o\left(\frac{d+m}{N}\right)^2\right)$$

$$b\hat{h} = \frac{d}{N - 1/2(d+m)}\left(1 + o\left(\frac{d+m}{N}\right)^2\right) \qquad (5)$$

El estimador del cociente instantáneo en el intervalo $[t_{i-1}, t_i[$ es entonces igual a:

$$\hat{h}_i = \frac{d_i}{N_i - 1/2\,(d_i + m_i)} \qquad \text{cuando } b_i = 1 \qquad (6)$$

bajo la hipótesis de que los eventos y las salidas de observación se producen uniforme e independientemente los unos de los otros sobre el intervalo.

En demografía clásica, se calcula habitualmente un cociente anual como la probabilidad de experimentar el evento en el intervalo en ausencia de salidas de observación:

$$\hat{h}'_i = \frac{d_i}{N_i - 1/2\, m_i} \tag{7}$$

Se hace entonces la hipótesis clásica de independencia: los que salieron de observación entre t_{i-1} y t_i antes de experimentar el evento, habrían conocido este evento si se hubieran quedado, tal como lo hicieron los que siguieron siendo observados en el curso de este periodo.

Las dos fórmulas precedentes permiten ver lo que diferencia a un cociente instantáneo de un cociente anual.

Las gráficas de los cocientes instantáneos, o más exactamente, de la función de probabilidad de densidad condicional, suelen ser útiles para decidir sobre el ajuste de la distribución observada en una familia de distribución conocida. Esto se tratará en detalle en el capítulo VII.

El estimador de la probabilidad de permanecer en el estado inicial es entonces igual a:

$$\hat{S}(t_i) = \prod_{k \le i} (1 - \hat{h}'_i) \tag{8}$$

y su varianza se obtiene de la misma manera que en el caso precedente mediante la fórmula de Greenwood:

$$\operatorname{var}\left(\hat{S}(t_i)\right) = \left(\hat{S}(t_i)\right)^2 \sum_{k=1}^{i} \frac{d_k}{\left(N_k - \frac{1}{2} m_k\right)\left(N_k - \frac{1}{2} m_k - d_k\right)} \tag{9}$$

3) Estimación de los cocientes acumulados

Otra posibilidad consiste en utilizar los cocientes instantáneos acumulados que permiten ver mejor la calidad del ajuste. Esos cocientes acumulados se estiman utilizando la expresión (13) del capítulo II.A.1:

$$\hat{H}(t) = \sum h_i$$

Esta fórmula es aproximativa y no se verifica bien más que cuando los valores de h_i son débiles. Como: $\left(S(t) = \exp(-H(t))\right)$ se obtiene inmediatamente la varianza de $\hat{H}(t)$:

$$\operatorname{var}\left(\hat{H}(t)\right) = \operatorname{var}\left(-\log \hat{S}(t)\right) = \sum_i \frac{d_i}{N_i(N_i - d_i)} \tag{10}$$

Las curvas de los cocientes acumulados presentan la desventaja de que hacen aparecer y destacar las inestabilidades de la distribución, sobre todo en el caso de los datos poco numerosos, aunque su representación gráfica da indicaciones precisas que detallaremos en el capítulo siguiente.

B) Muestra simple: riesgos múltiples

En la presentación que precede no se hizo ninguna distinción entre las diferentes modalidades del evento estudiado. Sólo tomamos en cuenta las pérdidas de información (truncamientos). Supongamos ahora que el evento que se produjo aleatoriamente en el instante T pudiera ser de naturalezas diferentes, identificadas por J. Si, además, cada individuo no puede experimentar más que un solo evento, nos encontramos en presencia de riesgos múltiples. Cuando se distinguen varias causas de fallecimiento, y el individuo no puede morir sino por una de ellas, tenemos una clara aplicación de esta noción. Esta noción de riesgos múltiples interviene todas las veces que el evento estudiado se puede distinguir por tipos.

Consideramos la distribución de la pareja (T, J), donde T mide la duración de permanencia al ocurrir el evento y J la naturaleza del evento, entonces la función de probabilidad de densidad condicional es igual a:

$$h_i(t) = \lim_{\Delta t \to 0} P\left(T < t + \Delta t, J = i \middle| t \leq T\right) \tag{11}$$

por consecuencia:

$$h(t) = \sum_i h_i(t) \tag{12}$$

y sabiendo que las ocurrencias del evento han tenido lugar en t, la probabilidad de que ellas sean de tipo i es:

$$h_i(t) \,/\, h(t)$$

de tal manera que:

$$P(J = i) = \int_0^\infty h_i(t) \exp\left[-\int_0^t h(s)\,ds\right] dt. \tag{13}$$

Un interesante caso particular se presenta si $h_i(t) = \alpha_i h(t)$, lo que implica que T y J son variables independientes.

Tal como vimos, a menos que se consideren unos riesgos independientes de los otros (calculándose entonces las curvas de permanencia a partir de las estimaciones de Kaplan y Meier para cada caso), nos encontramos confrontados con el problema de la interdependencia entre esos diferentes riesgos.

Por otra parte, el hecho de que en el seno de una muestra algunas veces reducida se consideren diferentes causas para la ocurrencia de un mismo evento puede provocar el temor de que las poblaciones sometidas a riesgo se reduzcan y, en consecuencia, resulte imposible estimar de manera confiable las similitudes o divergencias de comportamiento.

En ese caso se advierte que el estimador de los cocientes acumulados propuesto por Aalen (fórmulas (10) y (11), capítulo III.B.2) está totalmente indicado. Si $t_{i1} < t_{i2} < \ldots$ son los instantes en que ocurre el $i^{\text{ésimo}}$ evento en la población observada, si $Y(t_{ij})$ es la población sometida a riesgo hasta antes del instante t_{ij} y si $n(t_{ij})$ individuos experimentan ese evento en el instante t_{ij}, entonces:

$$\hat{H}_i(t) = \sum_{t_{ij} \le t} \frac{n(t_{ij})}{Y(t_{ij})} \qquad \text{estimador de Nelson-Aalen para datos agrupados} \qquad (14)$$

y cuya varianza es

$$\text{var}\left(\hat{H}_i(t)\right) = \sum_{t_{ij} \le t} \frac{n(t_{ij})}{\left[Y(t_{ij})\right]^2} \qquad (15)$$

Aquí $Y(t)$ representa una cantidad de individuos sometidos a riesgo hasta antes de t, que se puede modificar en cada fecha t sin haberse producido necesariamente un evento.

La relación de este estimador con el de Kaplan-Meier resulta clara. Pero aquí los h_i son siempre específicos, incluso cuando los riesgos no son independientes; de esa manera la representación gráfica suministra siempre una información sin ambigüedad a diferencia de las curvas de permanencia de Kaplan-Meier.

La representación gráfica de las curvas $\hat{H}_i(t)$, llamada de Nelson-Aalen, da las indicaciones siguientes:

1) Sus pendientes son las estimaciones del valor de las intensidades acumuladas de cada tipo de evento, que entonces se pueden comparar entre ellas en cada fecha t.
2) Su forma puede utilizarse para probar la distribución paramétrica de los procesos estudiados.

3) El valor de $\hat{H}_i(t)$ en cada punto representa el estimador del número de eventos de modalidad i que se habrían producido si constantemente hubiera estado un individuo sometido a riesgo. Una cantidad que es poco usual.

Ejemplos

1. El siguiente ejemplo presenta las curvas de los cocientes acumulados de mortalidad según tres causas de fallecimiento en el seno de una muestra de 95 individuos.

La figura 1 presenta las curvas de intensidad acumulada. Muestra que las dos causas (1) (3) de mortalidad actúan con una probabilidad constante y comparable al inicio del periodo. Cuando la causa (1) desaparece, la causa (2) aparece. La causa (3) adquiere en ese momento importancia (la pendiente de su intensidad crece bruscamente). Sin embargo, la causa (2) adquiere preponderancia al cabo de 600 días.

2. Como este modelo permite tomar en cuenta de manera multiplicativa la observación $Y(t)$, ésta también se puede estimar mediante el simple conteo de los individuos sometidos a riesgo. Por ejemplo, en un caso de estudio epidemiológico, Aalen propone hacer un modelo de la virulencia de la infección proporcional al número de individos contagiosos $c(t)$ y al número de individuos susceptibles $n(t)$ de contraer la enfermedad. En ese caso se puede escribir el siguiente modelo de intensidad multiplicativa:

$$\Lambda(t) = h(t)\, c(t)\, n(t)$$

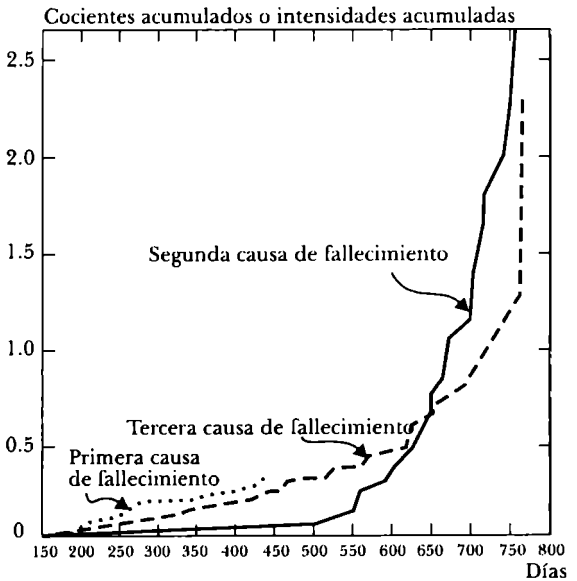
$h(t)$ es una medida del proceso de contagio de un individuo y:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{1}{c(t_i)\, n(t_i)}$$

C) Muestras múltiples: pruebas comparativas

Si hay varias muestras disponibles, a menudo nos vemos conducidos a determinar si la misma distribución gobierna la ocurrencia de los eventos observados en las diversas muestras. La prueba que se utiliza entonces proviene de las técnicas de pruebas de rango. En efecto, nos encontramos en el caso de una comparación de k distribuciones. La hipótesis nula corresponde

FIGURA 1
Mortalidad según tres causas



Causa de fallecimiento	*Porcentaje de individuos*	*Duración de permanencia o supervivencia (en días)*
(1)	23	159 189 191 198 200 207 220 235 245 250 256 261 265 266 280 343 356 383 403 414 428 432
(2)	40	317 318 399 495 525 536 549 552 554 557 558 571 586 594 596 605 612 621 628 631 363 643 647 648 649 661 663 666 670 695 697 700 705 712 713 738 748 753
(3)	37	163 179 206 222 228 249 252 282 324 333 341 366 385 407 420 431 441 461 462 482 517 524 564 567 586 619 620 621 622 647 651 686 761 763

Nota: Este ejemplo ha sido tomado de Aalen (1982). Se trata de una muestra de ratones a los que se atribuyeron tres causas de muerte que no vale la pena precisar aquí.

al hecho de que esas k distribuciones son idénticas y de densidad desconocida. Se dispone entonces de pruebas de rango que utilizan únicamente el número de series de las observaciones cuando éstas[1] se encuentran ordenadas de manera creciente.

[1] Se trata aquí de las duraciones de permanencia.

Sea una muestra que se puede escindir en k poblaciones distintas y $t_1 < t_2 < \ldots < t_n$ las fechas de ocurrencia del evento estudiado. Se supone igualmente que en t_j se registran d_j ocurrencias mientras que N_j individuos están sometidos a riesgo hasta antes de t_j ($j = 1, \ldots, r$). El número de individuos y de ocurrencias del evento correspondientes a las diferentes subpoblaciones será, por otra parte, llamado d_{ij} y N_{ij} con ($i = 1, \ldots, k$).

La distribución de $d_{1j}, \ldots, d_{kj}$ es entonces el producto de binomiales:

$$\prod_{i=1}^{k} \binom{N_{ij}}{d_{ij}} h_j^{d_j} (1-h_j)^{N_j - d_j} \tag{16}$$

donde h_j es el cociente instantáneo común a las k distribuciones.

En consecuencia la distribución de $d_{ij}, \ldots, d_{kj}$ conociendo que d_j, es hipergeométrica:

$$\frac{\prod_{i=1}^{k} \binom{N_{ij}}{d_{ij}}}{\binom{N_j}{d_i}} \tag{17}$$

de donde se puede deducir la media de los d_{ij}:

$$w_{ij} = N_{ij} \frac{d_j}{N_j} \tag{18}$$

la varianza:

$$(V_j)_{ii} = \frac{N_{ij}(N_j - N_{ij})\, d_j (N_j - d_j)}{N_j^2 (N_j - 1)} \tag{19}$$

y la covarianza de d_{ij} y de $d_{\ell j}$ igual a:

$$(V_j)_{i\ell} = \frac{-N_{ij} N_{\ell j} d_j (N_j - d_j)}{N_j^2 (N_j - 1)} \tag{20}$$

La transpuesta del estadístico de prueba v_j es $(d_{1j} - w_{1j}, \ldots, d_{kj} - w_{kj})$ vector con media cero y de matriz de varianza-covarianza igual a V_j.

El vector $v = \sum_{1}^{k} v_j$ que es el de las ocurrencias del evento observadas en cada subpoblación menos el número correspondiente de ocurrencias del evento esperadas, constituye el estadístico de rango.

Las distribuciones son iguales cuando asintóticamente:

$$v' V^{-1} v$$

es un χ^2_{k-1}.

El estadístico de χ^2_{k-1} se forma utilizando solamente $k-1$ elementos, pues si se conocen $k-1$ elementos, el k^e está perfectamente definido. Efectivamente, la suma de los k elementos de v es igual a cero.

Ejemplos

Antes de examinar algunos ejemplos de análisis formulados con los datos de la encuesta "Triple biografía", tomemos a manera de ejercicio (como lo habíamos hecho antes) una aplicación "microscópica".

Tomemos la población de dos inmuebles de la misma categoría situados en dos barrios de una ciudad. Queremos probar la diferencia eventual que existe en la emigración de los habitantes.

Duración	*Primera población*			*Segunda población*			*Conjunto*	
t_j	d_{1j}	N_{1j}	w_{1j}	d_{2j}	N_{2j}	w_{2j}	d_j	N_j
t_1	4	19	2.375	1	21	2.625	5	40
t_2	1	15	0.429		20	0.571	1	35
t_3		14	0.824	2	20	1.176	2	34
t_4	3	14	1.931	1	15	2.069	4	29
t_5	2	10	0.833		14	1.167	2	24
t_6		7	1.667	5	14	3.333	5	21
t_7		6	1.600	4	9	2.400	4	15
t_8	1	6	2.727	4	5	2.273	5	11
t_9		1	0.500	1	1	0.500	1	2
t_{10}	1	1	1.000					

Así, en el instante t_6 sobre el conjunto de las dos poblaciones, el cociente de emigración vale $q_6 = \frac{5}{21} = 0.2381$. Si ese cociente se aplicara por separado a cada inmueble observaríamos 1.667 emigrantes del primero y 3.333 del segundo. Como lo que se observa es respectivamente 0 y 5,

$$\text{el estadístico de prueba } v_6 = \begin{pmatrix} 0 - 1.667 \\ 5 - 3.333 \end{pmatrix} = \begin{pmatrix} -1.667 \\ 1.667 \end{pmatrix}$$

y

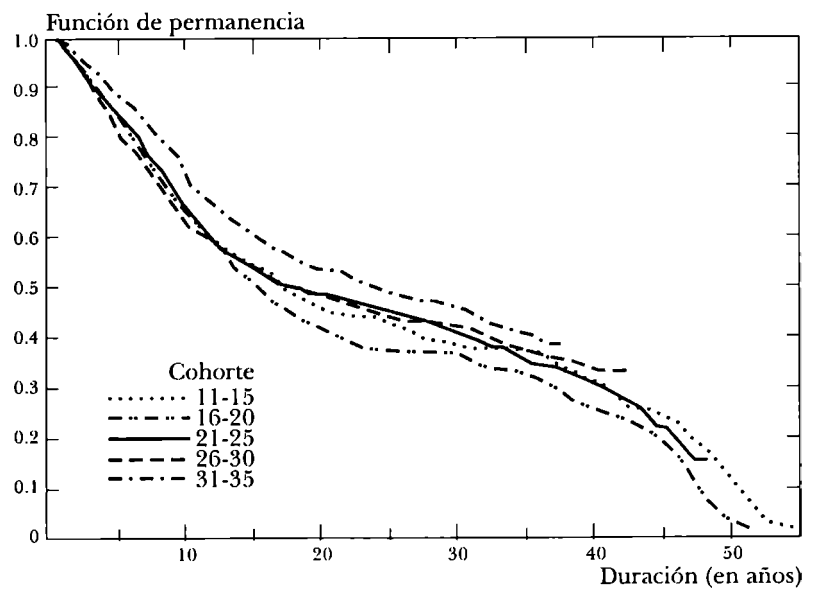
$$v = \sum_{j=1}^{10} v_j = \begin{pmatrix} -1.886 \\ 1.886 \end{pmatrix}$$

de donde la matriz de varianza-covarianza vale:

$$V = \begin{pmatrix} 5.828 & -5.828 \\ -5.828 & 5.828 \end{pmatrix}$$

El estadístico,[2] que se calcula y que se compara con un χ^2_1, tiene como valor $(1.886)^2 (5.828)^{-1} = 0.610$, lo cual no es significativo. Por lo tanto, se puede concluir que no aparece ninguna diferencia en la emigración de los ocupantes de los inmuebles estudiados.

FIGURA 2
Función de permanencia en el primer grupo de categorías socioprofesionales en años. Sexo masculino



Análisis no paramétricos como éstos se pueden realizar mediante los paquetes: Life Tables and Survival Functions 1L de BMDP (P1L) y Lifetest de SAS. Esos dos programas permiten estimar las duraciones de permanencia, los cocientes instantáneos de ocurrencia, la densidad de la distribución y hacer comparaciones para subpoblaciones distintas de la muestra. Por otra parte, permiten disponer de los resultados gráficamente.

[2] Para formar el estadístico de prueba conviene ser prudente y sólo tomar en cuenta que $k - 1$ elementos sea aquí solamente uno.

El ejemplo tratado aquí tiene que ver con el estudio de la vida profesional de los individuos, en particular de su movilidad en la carrera profesional. M.A. Cambois utilizó Lifetest de SAS.

FIGURA 3
Salida de la primera categoría socioprofesional. Sexo masculino

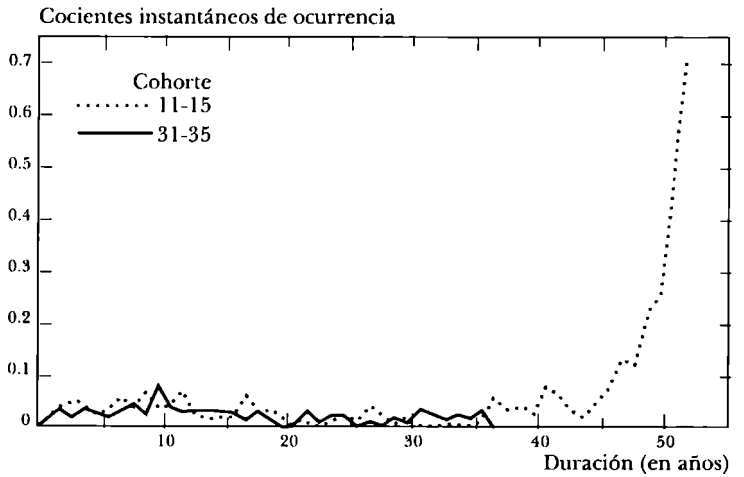


FIGURA 4
Acumulación de los cocientes instantáneos,
salida de la primera categoría socioprofesional. Sexo masculino

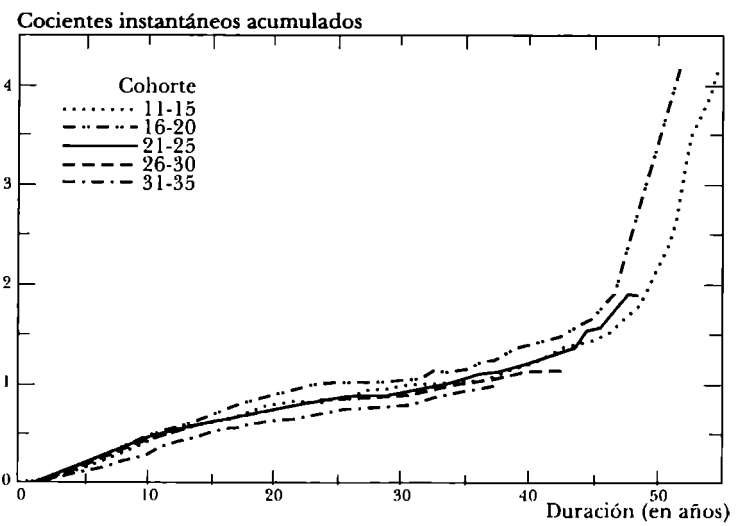
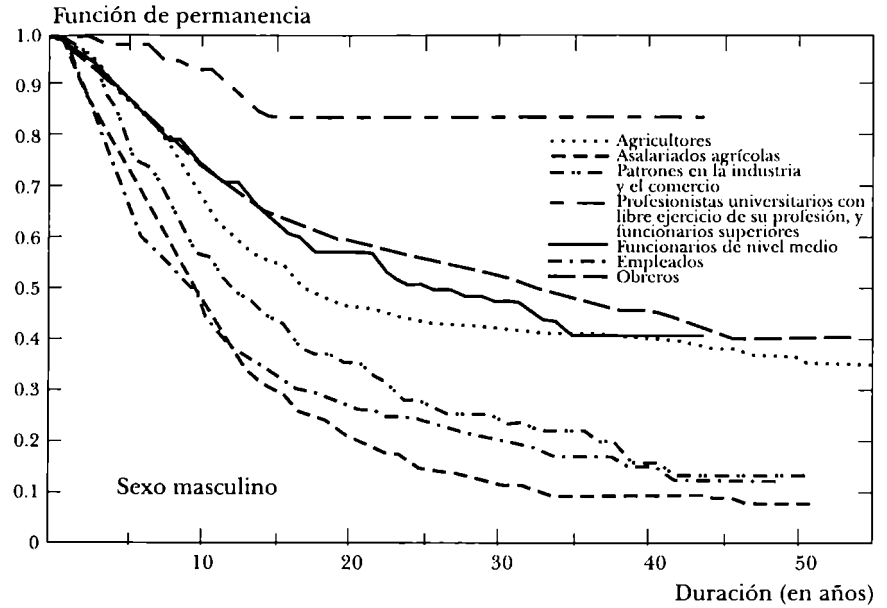
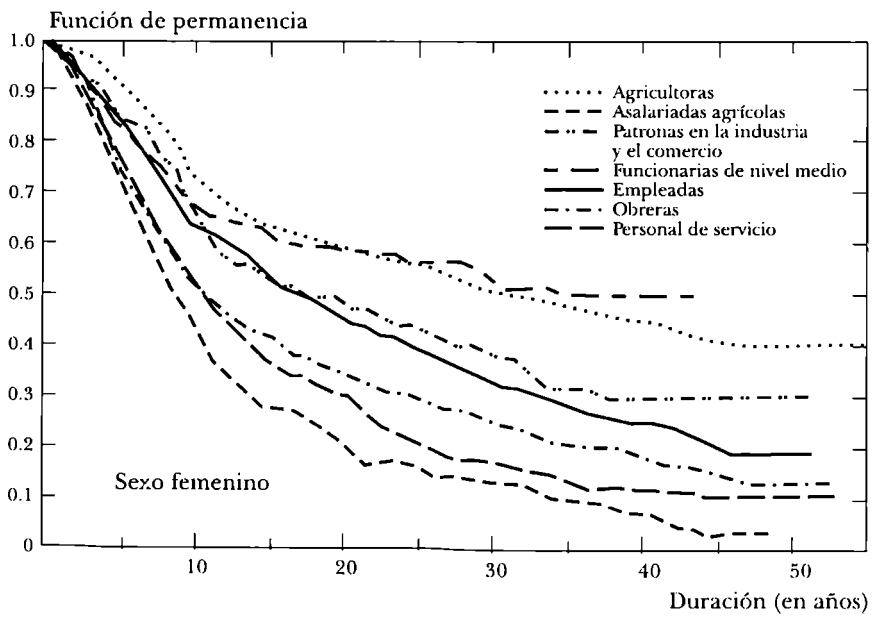


FIGURA 5
Duración de permanencia en el grupo de partida en años





El análisis de las duraciones de permanencia se hace tomando en cuenta los cambios de grupos de categorías socioprofesionales correspondientes a la nomenclatura INSEE1975: agricultores; asalariados agrícolas; patrones de la industria y del comercio; profesiones universitarias de libre ejercicio y funcionarios superiores; funcionarios de nivel medio; empleados; obreros; personal de servicio; artistas; clero; ejército, policía.

En el caso de los hombres, el estudio de la duración de permanencia en el primer grupo socioprofesional (figuras 2 a 4) muestra una estabilidad más grande para las cohortes más recientes. Se presentan aquí tres tipos de curvas para el mismo análisis: la de duración de permanencia, la de los cocientes instantáneos de ocurrencia —sólo para dos cohortes— y finalmente la representación gráfica de los cocientes instantáneos de ocurrencia acumulados. Cabe destacar que los cocientes acumulados superan la unidad (capítulo II.A.3).

La figura 5 presenta las distribuciones para los hombres y las mujeres según la primera categoría socioprofesional. Tanto unos como las otras dejan muy rápidamente un primer empleo de asalariado agrícola; en cambio, si ese primer empleo es una actividad de obrero, el comportamiento es muy distinto según el sexo de los individuos. Las mujeres que iniciaron su vida profesional como obreras abandonaron ese tipo de empleo con igual rapidez que las asalariadas agrícolas o el personal de servicio. En el caso de los hombres, por el contrario, es en la categoría de obrero donde permanecen durante más tiempo, después de la de funcionario superior y profesiones independientes.

D) Conclusión

En este capítulo consideramos los diferentes análisis que se pueden llevar a cabo cuando se estudia el desarrollo de un solo proceso. Los métodos no paramétricos utilizados aquí suministran estimadores de una gran utilidad, en la medida en que se disponga siempre de su varianza. Constantemente se tiene la posibilidad de juzgar la calidad del ajuste de las estimaciones y por esta razón se pueden llevar a cabo comparaciones.

Tal es el caso cuando el proceso que se estudia se expresa según modalidades diferentes, como las causas de mortalidad, o bien cuando se comparan diversas subpoblaciones.

El presente capítulo también nos ha permitido mostrar una diferencia esencial entre este análisis y uno clásico. Aquí podemos prescindir de establecer la molesta hipótesis del estudio del proceso en ausencia de truncamientos.

Cuando estos métodos se aplican a varios procesos abren perspectivas innovadoras. Ahora abordaremos el análisis de las interacciones entre fenómenos demográficos.

V. ESTUDIO RECÍPROCO DE LAS INTERACCIONES ENTRE DOS EVENTOS

Ahora abordaremos el caso bivariado que permite tratar la ocurrencia de dos eventos, como por ejemplo el matrimonio y la salida del mundo agrícola. Presentaremos en primer lugar los conceptos que se encuentran en la base del análisis, antes de pasar a formalizarlo.

A) Concepción del análisis

1) Dos fenómenos en competencia

En el caso univariado se plantean los problemas de riesgos múltiples cuando sólo se observa la ocurrencia de un evento de entre todos los que pueden experimentarse. Sin embargo, el estudio de los comportamientos conduce a considerar las trayectorias individuales tomando en consideración aspectos más complejos. En efecto, tal como vimos en la introducción, los individuos no se enfrentan a elecciones únicas, sino que sus decisiones son más bien el resultado de procesos, objetivos o no, de arbitraje entre sistemas en donde se encuentran implicados y entre los que conviene mantener una coherencia. Así, estudiar un fenómeno demográfico independientemente de lo que rodea a los individuos constituye un análisis reduccionista. Asimismo, no basta con estudiar la influencia de una serie de factores sobre un fenómeno, sino que es necesario tomar en cuenta la distribución de varios fenómenos competitivos en el curso de la vida de un individuo.

En primer lugar abordaremos el caso de dos eventos y estudiaremos la influencia recíproca de uno sobre el otro. Sería simplista esquematizar las etapas del ciclo de vida según una jerarquía que tienda hacia el estado final (la muerte); nuestra aproximación se cuida de caer en este reduccionismo y confronta eventos en diversos dominios. Ciertamente, el ciclo de vida está formado por etapas marcadas por la ocurrencia de eventos de naturaleza variada que no siguen un orden preestablecido. Así, el paso al estatus de adulto se realiza (entre otras cosas) por la partida del domicilio familiar, el matrimonio, el primer empleo, el inicio del compromiso político, etc., es decir, por la ocurrencia de eventos en el dominio de lo familiar y migratorio, profesional o sociopolítico. Por otra parte, cada individuo no habrá de pasar

necesariamente por todas esas etapas ni las experimentará en un orden preestablecido. A cada uno le reconocemos, entonces, diferentes dominios de implicación en cuyo seno se desarrollan sus trayectorias; la coherencia final de estas trayectorias depende de ajustes individuales variados. La exploración de esos ajustes se transforma en el objeto del análisis que determinará las dependencias en juego entre los diferentes sistemas.

2) La diferenciación dinámica

Más que referirnos a las correlaciones entre fenómenos, preferimos hablar de interacciones entre éstos, puesto que vamos a considerar formas diversas de dependencia. Pretendemos por eso analizar los comportamientos que son resultado de la confrontación de las diversas exigencias del momento. Así pues, no se trata de estudiar aquí las diferencias entre subpoblaciones designadas por una sola característica o por un conjunto de ellas, sino de identificar cómo se distinguen a lo largo de la vida los miembros de la población de una cohorte homogénea, de acuerdo con los acomodos individuales, obligatorios o voluntarios, que han tenido que hacer.

El estudio de las decisiones preferenciales, de los pasos obligados y de las trayectorias favorables, sienta así las bases de una aproximación a comportamientos que no se pueden aprehender más que dentro de la dinámica de los procesos.

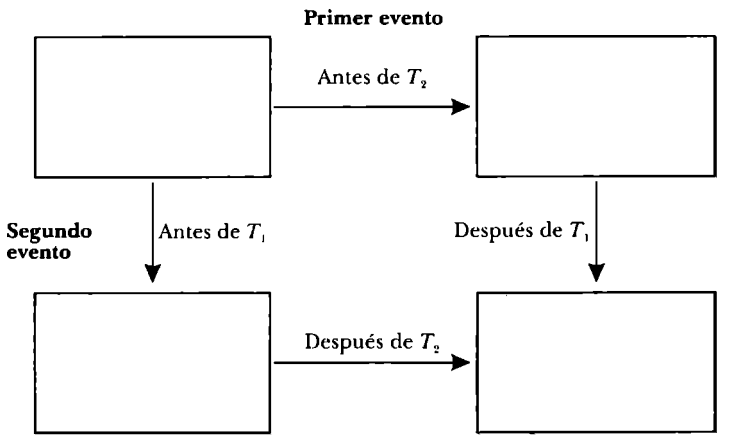
En este caso, al disponer de un grupo supuestamente homogéneo al inicio de un periodo nuestra herramienta nos permitirá explorar la manera en que evoluciona ese grupo y la aparición de la heterogeneidad entre sus miembros, e incluso pondrá en evidencia una heterogeneidad que ha podido existir desde el comienzo. Las diferencias que nos interesan son aquellas que se revelan en el curso del ciclo de vida, en favor de interacciones entre los diferentes dominios en que se encuentran implicados los individuos.

3) Los diferentes tipos de dependencia

Sean T_1 y T_2 las duraciones en las que se producen dos eventos. El estudio consiste en analizar las interacciones, las cuales se miden por el efecto observado sobre la distribución del primer evento cuando se produce el segundo y, recíprocamente, sobre las modificaciones debidas a la aparición del primero sobre la distribución del segundo. Este proceso se esquematiza en la figura 1.

La prueba de igualdad entre los cocientes instantáneos de ocurrencia "antes" y "después" proporciona así una indicación de la dependencia estocástica del primer evento en relación con el segundo, sin que se presuponga

FIGURA 1



la reciprocidad. De esta manera, al disociar las dependencias se puede determinar el sentido de las influencias en términos probabilísticos y no de causalidad determinista.

En las aplicaciones que describiremos en este capítulo, en ninguno de los casos hemos llegado a una constante de *independencia total* entre ambos fenómenos estudiados, lo que correspondería a la verificación de dos igualdades. Vemos así el peligro de tratar los fenómenos demográficos por separado, bajo la hipótesis de una independencia entre ellos.

En cambio, si se verifica sólo una de las dos igualdades ponemos en evidencia una *dependencia unilateral*. En el caso de los hombres que iniciaron su vida profesional en la agricultura, es mucho más frecuente que se casen luego de su partida del sector agrícola, mientras que su cambio de actividad no depende de su estatus matrimonial. En el caso de las mujeres sucede lo contrario, pues el matrimonio las estabiliza fuertemente en el sector agrícola mientras que el abandono de este sector no modifica sus posibilidades de casarse (Courgeau y Lelièvre, 1986). En el caso de las parejas obreras, la adquisición de la primera vivienda no depende del nacimiento del último hijo (de familias completas); sin embargo el hecho de hacerse propietarios vuelve más probable la llegada de un último hijo (Courgeau y Lelièvre, 1988).

Estos resultados ponen claramente en evidencia la importancia relativa que tienen para el grupo estudiado los dos dominios cuya competencia sometemos a prueba. Esas dependencias, que en el análisis no paramétrico aparecen caricaturizadas, serán el objeto de una segunda aproximación semiparamétrica o paramétrica, que permitirá determinar su naturaleza exacta. En particular, en un segundo tiempo se identifica la naturaleza de

los subgrupos que se han distinguido después de examinar las características individuales de sus miembros.

Por último, cuando ninguna de las igualdades sometidas a prueba se verifica, estamos en presencia de una *dependencia recíproca* de dos fenómenos. Así sucede con el matrimonio y la partida del domicilio paterno para las generaciones nacidas entre 1926 y 1935 (Courgeau, Lelièvre y Wagner, 1986). De igual manera, si la probabilidad de migrar hacia una zona fuertemente urbanizada disminuye al nacer el primer hijo, los nacimientos de rango dos y más se reducen después de una migración hacia un centro muy urbanizado (Courgeau, 1987).

Estos análisis nos han permitido trabajar sobre muestras pequeñas o sobre estratificaciones finas. Hemos podido someter a prueba las igualdades precedentes, no sólo sobre el conjunto del periodo estudiado sino también en cada duración o en cada edad.

Puede ser que las dependencias no se observen sino a ciertas edades o durante un periodo dado luego del evento perturbador. La influencia puede también revertirse. Así, cuando se estudian la fecundidad y la actividad femeninas se observa que las mujeres inactivas al casarse y al tener el primer hijo muestran comportamientos de fecundidad fuertemente diferenciados según la edad: las mujeres de menos de 30 años que retoman una actividad son tan fecundas como las que permanecen inactivas; sin embargo, después de los 30 años el hecho de retomar la actividad constituye un freno a un nacimiento suplementario (Lelièvre, 1987).

Es posible poner en evidencia niveles de interpretación más complejos. En el caso del estudio de las interacciones entre fecundidad y migración hacia las zonas fuertemente urbanizadas, se observa una espectacular reducción de la fecundidad de rango superior a 1 entre los migrantes. El problema que se plantea entonces es saber si se trata de un comportamiento de adaptación o de selección, en el caso en que la muestra de las futuras migrantes presente ya un comportamiento fecundo diferente del de las demás mujeres de la región de origen. Los datos biográficos permiten someter a prueba esas diferencias entre futuras migrantes y sedentarias en el seno de la población de origen (Courgeau, 1987).

Efectivamente, las futuras migrantes hacia las metrópolis (para los nacimientos de rangos 2 y 3) tienen ya una fecundidad baja respecto de las sedentarias de las zonas poco urbanizadas. Así se pone en evidencia una *dependencia a priori* de la fecundidad sobre la migración por venir, que se traduce por esta selección en el seno de la población inicial.

Recíprocamente, se observa que la migración hacia las zonas poco urbanizadas tiene un efecto favorable sobre el nacimiento del segundo o tercer hijos. Gracias a una investigación idéntica, esta vez tenemos la posibilidad de poner en evidencia un comportamiento real de adaptación de la

fecundidad de las que migraron fuera de las metrópolis. En efecto, su comportamiento fecundo anterior no difiere en nada del de las citadinas que no dejan las zonas fuertemente urbanizadas.

4) Dependencia unilateral y causalidad

Hemos puesto en evidencia cuatro tipos de dependencia o de independencia entre fenómenos. Podríamos vernos tentados a considerar esas dependencias como las causas de los fenómenos, pero más allá de un razonamiento de causalidad, incluso en el sentido probabilista y no determinista, las analizamos en términos de interacciones y de reciprocidad eventual.

Lo que pretendemos aquí es mostrar en qué difiere el presente análisis de uno de tipo causal. Para situar esta tarea de razonamiento respecto de las prácticas de la explicación causal, examinaremos las teorías probabilísticas. La definición que dan los teóricos de la causalidad probabilística[1] es la siguiente:

> c y e son dos eventos específicos observados,
> c es la causa de e si y sólo si
> *1)* c no se produce después de e;
> *2)* $P(e \mid c) > P(e)$;
> *3)* no existe ningún evento s susceptible de enmascarar a c respecto de e;

s enmascara a c si $P(e \mid s, c) = P(e \mid s) \neq P(e \mid c)$; intuitivamente s es entonces la verdadera causa de e del que c no es más que un indicador.

El estudio de las interacciones, tal como lo proponemos, se caracteriza por la identificación de *dependencias unilaterales*. La relación entre dos fenómenos se analiza minuciosamente mediante una *aproximación recíproca sistemática*. Así pues, nuestra perspectiva jamás excluye la reciprocidad eventual de la influencia y está, como punto de partida, en contradicción con el postulado de antisimetría subyacente en la definición probabilística: si c es la causa de e, es evidente que e no puede ser la causa de c. Aquí nuestro razonamiento claramente se expresa en términos de dependencia y no de causalidad.

Por otra parte, tampoco retenemos el axioma *1)*, común además a todas las definiciones de la relación causal: la anterioridad temporal de la causa. En efecto, la idea de la reciprocidad invalida esta forma de la temporalidad definida en el axioma. Tal como ya vimos, llegamos a determinar comportamientos *a priori*, como por ejemplo los fenómenos de selección que favorecen un tipo particular de migración. Este análisis tiene incluso capacidad para identificar formas complicadas de temporalidad.

[1] Good, Salmon, Suppes, citado por M. Swain, 1987.

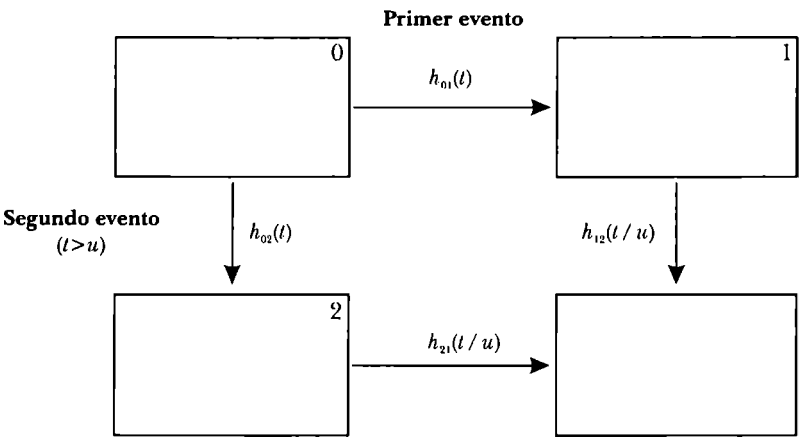
Nos situamos fuera del marco de la explicación causal, en un modelo de análisis de los fenómenos y de sus interacciones. Este análisis nos permite realmente prolongar el análisis demográfico clásico de los fenómenos en "estado puro", al introducir un análisis de los fenómenos "en interacción".

Para la discusión que sigue nos colocaremos en el marco de un estudio bivariado. Se trata, por ejemplo, del análisis de la nupcialidad de los agricultores (Courgeau y Lelièvre, 1986) con un estudio de las interacciones entre el matrimonio y la salida del mundo agrícola, o también del análisis de las interacciones entre la fecundidad y la actividad femenina (Lelièvre, 1987), o entre la fecundidad y las migraciones hacia o fuera de las metrópolis (Courgeau, 1987), o entre el acceso a la propiedad y la llegada del último hijo (Courgeau y Lelièvre, 1988). En cada ocasión estudiamos las interacciones entre dos eventos cuya aparición es resultado de un arbitraje, de una elección de prioridad objetiva o no, y nuestro procedimiento tiende entonces a identificar los diversos tipos de dependencia entre esos dos fenómenos.

B) Formalización del análisis bivariado

Las distribuciones condicionales se formalizan como sigue: Sean T_1 y T_2 las duraciones en las que se producen dos eventos.

Figura 2
Diagrama de los estados en el caso de dos variables



Los cocientes instantáneos (figura 2) se definen por:

$$h_{01}(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_1 < t + \Delta t \,|\, T_1 \geq t, T_2 \geq t), \qquad (1)$$

que es el cociente instantáneo de ocurrencia del primer evento si el segundo no se ha producido antes, y por:

$$h_{21}(t|u) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_1 < t + \Delta t | T_2 = u, T_1 \geq t), \tag{2}$$

el cociente instantáneo de ocurrencia del primer evento si el segundo se ha producido antes de la fecha u.[2]

La prueba de igualdad entre h_{01} y h_{21} así como entre h_{02} y h_{12}, da entonces una indicación de la dependencia estocástica de la variable aleatoria T_1 respecto de T_2:

- dependencia recíproca de dos eventos si $h_{01}(t) \neq h_{21}(t|u)$ y $h_{02}(t) \neq h_{12}(t|u)$,
- dependencia unilateral si $h_{01}(t) \neq h_{21}(t|u)$ y $h_{02}(t) = h_{12}(t|u)$, de la que sabemos al mismo tiempo si se ejerce negativa o favorablemente sobre la ocurrencia de T_1 según que $h_{01}(t) > h_{21}(t|u)$ o $h_{01}(t) < h_{21}(t|u)$, y por último
- independencia total entre los dos eventos si $h_{01}(t) = h_{21}(t|u)$ y $h_{02}(t) = h_{12}(t|u)$.

En consecuencia, en tiempo continuo se puede establecer la densidad de la pareja de variables aleatorias (T_1, T_2) sabiendo que la función de permanencia de T_i es:

$$S_i(t) = \exp\left(-\int_0^t h_{0i}(u)\,du\right) \tag{3}$$

y su densidad de probabilidad:

$$f_i(t) = h_{0i}(t)\, S_i(t) \tag{4}$$

Supongamos que:

$f_2(t_2|t_1)$ función de probabilidad de densidad condicional de T_2 sabiendo que $T_4 = t_1$

$f_1(t_1)$ función de probabilidad de densidad marginal de T_1

se tiene siempre $f(t_1, t_2) = f_2(t_2|t_1) f_1(t_1)$.

De acuerdo con lo anterior, la probabilidad de permanecer en I_1[3] hasta la fecha t se puede escribir en nuestro modelo de la siguiente forma:

[2] Se obtiene simétricamente $h_{02}(t)$ y $h_{12}(t|u) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T_2 < t + \Delta t | T_1 = u, T_2 \geq t)$

[3] Siendo I_1 el estado en el que se encuentran los individuos que han experimentado el primer evento e I el estado inicial.

$$f_2(t_2|t_1) = h_{12}(t_2|t_1)\exp\left(-\int_{t_1}^{t_2}\left(h_{12}(u|t_1)\right)du\right) \tag{5}$$

y la probabilidad de permanecer en I hasta la fecha t_1:

$$f_1(t_1) = h_{01}(t_1)\ \exp\left(-\int_0^{t_1}\left(h_{01}(u)+h_{02}(u)\right)du\right) \tag{6}$$

Si $t_1 \leq t_2$, antes de t_1 no se puede estar más que en I y hay dos maneras de salir de ahí, de donde:

$$f(t_1, t_2) = h_{01}(t_1)\, h_{12}(t_2|t_1) \exp\left[-\int_0^{t_1}\left(h_{01}(u)+h_{02}(u)\right)du - \int_{t_1}^{t_2}\left(h_{12}(u|t_1)\right)du\right] \tag{7}$$

con una expresión similar para $t_2 \leq t_1$.

Ahora bien, T_1 y T_2 son independientes si $f(t_1, t_2) = f_1\,(t_1) f_2\,(t_2)$.

Una condición necesaria y suficiente puede expresarse bajo la forma:

$$h_{21}(t|u) = h_{01}(t) \text{ y } h_{12}(t|u) = h_{02}(t).$$

En el caso en que se verifica la segunda igualdad, la fórmula (7) de densidad de la pareja se vuelve entonces:

$$f(t_1, t_2) = h_{01}(t_1)\exp\left(-\int_0^{t_1}\left(h_{01}(u)+h_{02}(u)\right)du\right) \times h_{02}(t_2)\exp\left(-\int_{t_1}^{t_2} h_{02}(u)du\right) \tag{8}$$

El desequilibrio de la fórmula proviene de la anterioridad de la producción del primer evento en relación a la del segundo.

1) Estimación actuarial

Para estimar estos cocientes instantáneos de ocurrencia vamos a suponerlos constantes a lo largo de todo un año. Esta versión no paramétrica en tiempo discreto permite la estimación, ya que aún no disponemos de algún método satisfactorio puramente no paramétrico (Cox y Oakes, 1984). Se supondrá, de igual manera, que durante esos intervalos los eventos considerados (por ejemplo, matrimonios y salidas de la agricultura) están repartidos uniformemente.

Así, sean: $N_i(t)$ $(i = 0, 1, 2)$ la población en el estado i al comienzo del año t

$n_{ij}(t)$ el número de eventos de tipo j acaecidos en la población del estado i durante el año t.

Los estimadores más simples están dados por:

$$\hat{h}_{0,1}(t) = \frac{n_{0,1}(t)}{N_0(t) - \frac{1}{2}[n_{0,1}(t) + n_{0,2}(t)]} \tag{9}$$

$$\hat{h}_{2,1}(t|u) = \frac{n_{2,1}(t)}{N_2(t) - \frac{1}{2}[n_{2,1}(t) - n_{0,2}(t)]} \tag{10}$$

$$\hat{h}_{0,2}(t) = \frac{n_{0,2}(t)}{N_0(t) - \frac{1}{2}[n_{0,2}(t) + n_{0,1}(t)]} \tag{11}$$

$$\hat{h}_{1,2}(t|u) = \frac{n_{1,2}(t)}{N_1(t) - \frac{1}{2}[n_{1,2}(t) - n_{0,1}(t)]} \tag{12}$$

A fin de eliminar las simultaneidades, en nuestro ejemplo hemos planteado la hipótesis siguiente: los individuos que se casaron en el año de su salida del sector agrícola se remiten a la población de los agricultores casados para realizar el cálculo de los cocientes de salida de la agricultura, y a la población de los solteros salidos de la agricultura para el cálculo de los cocientes de nupcialidad.

Hay otras maneras de tomar en cuenta las simultaneidades, que serán detalladas más adelante.

2) Las pruebas posibles

Uno de los problemas de este análisis es contar con la posibilidad de poner a prueba las hipótesis que determinan las dependencias particulares entre fenómenos. Ahora bien, con frecuencia nos encontramos en presencia de efectivos reducidos, lo que no siempre permite aplicar todas las pruebas posibles.

La prueba más simple es una de rango generalizado propuesta por Peto y Pike (1973) y utilizada por Aalen *et al.* (1980). Las otras pruebas se basan en la diferencia normalizada entre valores observados y valores reales y el cálculo de la matriz de varianza-covarianza de las dos distribuciones.

La prueba, sin embargo, no permite más que una comparación de conjunto entre las tendencias representadas por dos series de cocientes instantáneos h_{oi} y h_{ji}. Se trata de una prueba no paramétrica con dos componentes. Aquí se compara la ocurrencia de uno de los dos eventos según que el segundo se haya producido o no.

Sean:

$t.^1$ las fechas de las transiciones de o a i

$t.^2$ las fechas de las transiciones de j a i

$Y_k\ (t)$ la población sometida a riesgo en el estado k en el instante t.

Se calcula el estadístico de Savage de la siguiente manera:

$$S = \sum_k \frac{Y_j(t_k^1)}{Y_o(t_k^1) + Y_j(t_k^1)} - \sum_k \frac{Y_o(t_k^2)}{Y_o(t_k^2) + Y_j(t_k^2)} \tag{13}$$

y un estimador no sesgado de su varianza

$$V = \sum_k \frac{Y_o(t_k^1)Y_j(t_k^1)}{[Y_o(t_k^1) + Y_j(t_k^1)]^2} + \sum_k \frac{Y_o(t_k^2)Y_j(t_k^2)}{[Y_o(t_k^2) + Y_j(t_k^2)]^2} \tag{14}$$

o bajo la hipótesis de igualdad entre las dos series, el estadístico $SV^{-1/2}$ se distribuye entonces asintóticamente según una ley normal $N(0, 1)$.

Sin embargo, S no puede medir diferencias más que sobre intervalos donde Y_o y Y_j son simultáneamente diferentes de 0.

3) Las pruebas seleccionadas y la convergencia de las estimaciones

Para probar las hipótesis de igualdad de los cocientes instantáneos de ocurrencia preferimos utilizar estadísticos de prueba basados en las desviaciones entre los estimadores de los dos cocientes. En realidad, todos estos estimadores proceden de la teoría asintótica de los cocientes de ocurrencia sobre riesgo, revisados por J. Hoem (1976) para el campo de la demografía.

Efectivamente, cualquiera que sea el modelo de distribución del fenómeno, el estimador actuarial, según la hipótesis de que el cociente permanece constante a todo lo largo del intervalo, es de la forma

$$(h - \hat{h}) / (-d^2 \log (h) / dh^2) \tag{15}$$

que tiene un comportamiento asintótico normal y, por lo tanto, permite el cálculo de intervalos de confianza.

Schou y Vaeth hicieron cálculos de simulación sobre estos estadísticos para determinar cuál era el tamaño mínimo de efectivos que permitía conser-

var su convergencia asintótica. En esta medida, la selección de esos estimadores nos ha parecido la más consistente.

Supongamos que se quiere saber si el cociente instantáneo que corresponde al primer evento permanece sin cambios cuando se produce el segundo. Queremos ahora probar en la fecha t si se verifica la igualdad siguiente:

$$\hat{h}_{01}(t) = \hat{h}_{21}(t)$$

Llamemos $Y_0(t)$ y $Y_2(t)$ a las poblaciones sometidas a riesgo del primer evento, según que éstas aún no hayan experimentado el segundo evento o que, por el contrario, ya lo hayan experimentado.

Cuando la población observada tiende hacia el infinito, tenemos entonces el estadístico siguiente:

$$D(t) = \frac{\hat{h}_{01}(t) - \hat{h}_{21}(t)}{\left(\frac{\hat{h}_{01}(t)}{Y_0(t)} + \frac{\hat{h}_{21}(t)}{Y_2(t)}\right)^{1/2}} \tag{16}$$

donde el denominador es un estimador consistente de la desviación estándar asintótica del numerador teniendo en cuenta el tamaño respectivo de las poblaciones sometidas a riesgo, y es asintóticamente normal $N(0, 1)$ si la igualdad precedente se verifica.

De esa manera tenemos la posibilidad de probar la validez de esta igualdad.

El estadístico $D(t)$ se puede acumular durante una serie de periodos y, en ese caso, el estadístico:

$$D = \sum_t D(t) / \sqrt{m} \tag{17}$$

donde m es el número de periodos considerados, es igualmente normal $N(0, 1)$ si las dos subpoblaciones tienen el mismo comportamiento a todo lo largo del tiempo.

El segundo estimador utiliza el resultado obtenido por Schou y Vaeth (1980), que indica que las variables $\left(\hat{h}_{01}(t)\right)^{1/3}$ y $\left(h_{21}(t)\right)^{1/3}$ tienden más rápidamente que las anteriores hacia su distribución normal. Llamemos $n_{01}(t)$ y $n_{21}(t)$ los números de individuos que experimentaron el primer evento cuando aún no hayan experimentado el segundo o cuando, por el contrario, ya lo hayan experimentado.

Cuando la población observada tiende hacia el infinito, tenemos entonces el estadístico siguiente:

$$SV(t) = \frac{\hat{h}_{01}(t)^{1/3} - \hat{h}_{21}(t)^{1/3}}{\left(\frac{\hat{h}_{01}(t)^{2/3}}{9n_{01}(t)} + \frac{\hat{h}_{21}(t)^{2/3}}{9n_{21}(t)} \right)^{1/2}} \tag{18}$$

donde el denominador es un estimador consistente de la desviación estándar asintótica del numerador y es asintóticamente normal $N(0, 1)$ si se verifica la igualdad de los cocientes instantáneos de ocurrencia.

Este estadístico se puede acumular para una serie de periodos tomados en cuenta de manera idéntica a $D(t)$, sea:

$$SV = \sum_t SV(t) / \sqrt{m} \tag{19}$$

siendo m el número de periodos.

De un estudio de simulación que hicieron Schou y Vaeth sobre la solidez de los estimadores del máximo de verosimilitud en el caso de los datos longitudinales, resulta que los límites de la aplicación de la hipótesis de normalidad siguen siendo válidos mientras que $N\hat{h}_k \geq 10$. Este es un resultado importante para el análisis de las muestras de tamaño pequeño.

La rápida convergencia de estas estimaciones hacia la distribución asintótica normal, demostrada por la simulación de Schou y Vaeth, no es sin embargo verificable si los cocientes instantáneos de ocurrencia estimados son iguales a cero. En ese caso, relativamente frecuente en el análisis de los fenómenos humanos, nos encontramos con la incapacidad de probar diferencias y de establecer intervalos de confianza.

C) Análisis práctico

El análisis de las interacciones entre dos eventos comprende, en la práctica, casos de situaciones muy variadas. Vamos entonces a describir el camino a seguir en situaciones diferentes.

En primer lugar, es necesario tomar nuevamente en cuenta los eventos fechados cronológicamente según una escala idéntica: así, la duración hasta la ocurrencia de algún evento debe medirse respecto de un origen común. Se incorporan los individuos al inicio de su vida profesional, cuando se van de la casa de sus padres, cuando se casan, etcétera.

En segundo lugar, se deben escoger eventos que no forman parte de una serie de pasos en estados jerarquizados; efectivamente, no se pueden estudiar las interacciones entre la primera y la segunda migración puesto que éstas se producen necesariamente en un orden dado, en el que la primera

precede por definición a la segunda. De igual manera, si se consideran el matrimonio y el divorcio nos veremos inevitablemente conducidos a concluir que el matrimonio es la principal causa de divorcio.

Una vez que se han respetado estos dos principios y se han escogido dos eventos, subsisten aún múltiples situaciones:

- ciertos individuos jamás experimentan alguno de los eventos. Así, una parte de la población permanece soltera en forma definitiva, por lo cual es necesario tomar en cuenta los truncamientos a la derecha;
- ciertos eventos son reversibles, lo que significa que los estados a los cuales los conducen son recurrentes; en ese caso hay que tomar en cuenta los retornos. En efecto, se puede dejar la agricultura y regresar a ella, tener un empleo y luego estar desempleado;
- en el estudio de las interacciones entre diferentes dominios de la vida de un individuo, es deseable que se confronte el ciclo de vida familiar (salida del hogar de los padres, cohabitación, matrimonio, nacimientos sucesivos) con un evento único (el primer empleo, por ejemplo). Así se dispone de una serie cronológica de eventos que nos proponemos analizar en interacción con un evento único o con otra serie cronológica.

Por supuesto que se presentan múltiples situaciones que aquí no detallamos.

Vale la pena considerar en detalle el segundo caso que hemos mencionado, el de los eventos demográficos reversibles. Bajo este tema se consideran las idas y vueltas del empleo a la inactividad, de las migraciones hacia las metrópolis, o al contrario, de las metrópolis hacia zonas menos urbanizadas, por ejemplo. Estas transiciones en la modelización corresponden esta vez a los pasos repetidos por estados recurrentes. Entonces se pueden llevar a cabo dos análisis: el de las interacciones entre "la ida" y el segundo evento, y el análisis de las interacciones entre "el regreso" y ese segundo evento. Luego se comparan los cocientes instantáneos de ocurrencia del segundo evento según la situación de partida del individuo.

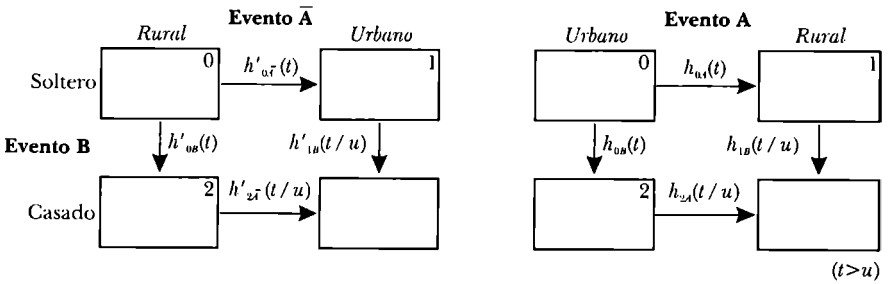
Para el caso en que se estudia la interacción entre las migraciones rural/urbana y la nupcialidad, corresponde el esquema de la figura 3.

Primero se tratan las interacciones entre las migraciones hacia los centros urbanos y los matrimonios: los individuos siguen el recorrido según las flechas correspondientes a los cocientes instantáneos de ocurrencia $h'_{.B}(.)$ y $h'_{.A}(.)$.

Luego se hace el análisis de las interacciones entre las migraciones desde el medio urbano hacia el rural y el matrimonio: los individuos siguen el recorrido según las flechas $h_{.B}(.)$ y $h_{.A}(.)$.

En un segundo tiempo se van a poder comparar de dos en dos los cocientes instantáneos de nupcialidad de los citadinos $h_{0B}(t)$ y $h'_{1B}(t|u)$ según que éstos sean originarios de las metrópolis o que hayan inmigrado

FIGURA 3
Caso bivariado con evento reversible



a ellas anteriormente, siendo originarios del campo. De la misma manera, se podrán comparar los cocientes de nupcialidad de aquellos que viven fuera de las grandes metrópolis $h'_{0B}(t)$ y $h_{1B}(t|u)$ según sean originarios de ellas o no.

Tales aplicaciones no se pueden considerar sin un *software* para ponerlas en funcionamiento. E. Lelièvre inventó ese paquete en el INED. Este *software* llamado Root ha permitido probar las diferentes hipótesis de trabajo que presentamos aquí y ha sido utilizado para todas las aplicaciones que describimos. Su manual de utilización se presenta en el anexo 2.

1) Simultaneidades e intervalos de tiempo

El estudio de las interacciones confronta eventos que se hallan en competencia en un momento del ciclo de vida. Sin embargo, es posible encontrar dificultades para establecer una escala de tiempo que permita el análisis. Efectivamente, según el reloj que escoja, el investigador observa casos de simultaneidad de dos eventos (matrimonio y migración a la misma edad, abandono de actividad y un nacimiento en la misma fecha, etc.). Ahora bien, si desde el punto de vista matemático no se plantea ese problema de los eventos simultáneos —en tiempo continuo siempre se puede detectar la anterioridad de la ocurrencia de un evento sobre otro—, en ciencias sociales el asunto se plantea en forma diferente, pues los eventos que se observan no son más que la realización de decisiones anteriores cuya fecha no registramos.

E incluso, si la recolección de datos se hace con precisión (día, hora, minuto...) y se tiene la posibilidad de disociar en todos los casos la ocurrencia de dos eventos, ¿cuál es la validez de esta distinción? Una vez que el estadístico elimina la eventualidad de ocurrencias simultáneas, ¿qué pueden decir el demógrafo o el sociólogo de dos eventos ocurridos con un mes de diferencia pero vividos como simultáneos?

La distinción de los eventos escogidos como simultáneos sólo puede lograrse si tenemos indicaciones sobre la maduración, la elaboración de las decisiones o de las circunstancias del evento. De no poseer esas informaciones, el demógrafo puede entonces optar por comportarse como un estadístico o decidir que tomará en cuenta esas situaciones de incertidumbre como tales. Así, hemos introducido la noción de "tiempo borroso" (*temps flou* en francés y *fuzzytime* en inglés) (Courgeau y Lelièvre, 1985) que caracteriza al lapso transcurrido entre la maduración (decisión eventual) y la realización (ocurrencia observada) del evento, lapso que es diferente para cada individuo y que el demógrafo desconoce.

La simultaneidad de dos ocurrencias sobreviene cuando éstas se producen en un intervalo de tiempo seleccionado (dos eventos se pueden considerar simultáneos si se producen en el mismo trimestre, en el mismo año...). El investigador, tomando en consideración el "tiempo borroso", no se concede entonces el derecho de decidir acerca de la anterioridad de eventos ocurridos dentro del marco de tiempo elegido.

Este procedimiento controvertido que pretende considerar las simultaneidades por sí mismas no facilita la aproximación estadística al problema, sino que apela directamente a la colaboración multidisciplinaria. En este sentido resulta conveniente caracterizar el conjunto de los movimientos simultáneos considerados como el de las decisiones conjuntas. Ahora bien, sólo la interpretación de los psicólogos y los sociólogos que trabajan sobre la maduración de los proyectos y sobre la lógica de las trayectorias es apta para ser tomada en cuenta.

Aquí queremos destacar el trabajo de E. Klijzing, J. Siegers, N. Keilman y L. Groot presentado en la Conferencia Internacional de Demografía realizada en Jyväskylä en junio de 1987. Estos investigadores, en el marco de un estudio del abandono de la actividad profesional y de los nacimientos, llevan a cabo un experimento para definir el reloj o la escala de tiempo que más se adapta a las tomas de decisión en el caso de la simultaneidad del abandono de la actividad y de los nacimientos.

Su procedimiento consiste en tomar intervalos de tamaño variable (de uno a nueve meses): un nacimiento y un abandono de actividad se considerarían entonces como simultáneos si se producen en este intervalo. En consecuencia, las simultaneidades serán tanto más numerosas cuanto tanto más extenso sea el intervalo. Esos autores proceden entonces a la prueba de igualdad $h_{01}(t) = h_{21}(t|u)$, que en este caso es tener un hijo estando activo o habiendo dejado la actividad (cuadro 1), excluyendo las simultaneidades. Podemos notar que los cocientes permanecen bastante estables, pero sobre todo que el estadístico de prueba permanece prácticamente sin cambio. Durante un intervalo que aquí está restringido a nueve meses como máximo,

CUADRO 1
Pruebas del "tiempo borroso"

Cocientes de fecundidad (dos años)	*Duración de los intervalos de simultaneidad (en meses)*			
	1	*3*	*6*	*9*
h_{01} mujer activa	5.5	5.3	5.1	5.0
h_{21} mujer que deja la actividad	23.2	23.7	22.1	24.1
Estadístico de prueba N (0, 1) diferencia entre h_{01} y h_{21}	–3.311	–3.401	–3.264	–3.495

distinguir o no la fecha de los dos eventos no modifica la estimación de la influencia de uno de los eventos sobre el otro.

En el *software* Root dejamos la opción de tomar en cuenta simultaneidades según ocho opciones, las cuales corresponden a la posibilidad de sustraer del análisis o incorporar a éste los casos en que se producen dos eventos con la misma antigüedad, e incluso de calcular aparte los cocientes instantáneos que les corresponden.

La hipótesis que se conserva en todas las aplicaciones es que aquellos que experimentan un evento en este intervalo están sometidos a riesgo durante un medio-intervalo de tiempo. Esta hipótesis supone que los eventos están repartidos de manera uniforme en el intervalo.

2) Representaciones gráficas

Como ya dijimos en el capítulo anterior, hay varios tipos de representaciones gráficas que el investigador puede utilizar para ilustrar el análisis. Aquí discutiremos las diversas posibilidades, evaluando sus inconvenientes y sus ventajas. Hay varios estadísticos disponibles: los cocientes instantáneos de ocurrencia, los cocientes instantáneos de ocurrencia acumulados y la función de permanencia o supervivencia. Éstos se encuentran representados en las gráficas 2, 3 y 4 del capítulo anterior. En los artículos que hemos publicado escogimos principalmente la representación de los cocientes acumulados. Ésta se presta, sin embargo, a discusión, en la medida en que el "nivel" aparente del fenómeno no corresponde a ninguna noción interesante. En efecto, esta confusión proviene de que el acumulado de los cocientes instantáneos puede superar la unidad puesto que no se trata de la suma de probabilidades sobre una población estable, sino más bien de la suma de riesgos (un riesgo puede ser superior a 1, *cf.* capítulo II. A. 3). De hecho, sólo las desviaciones (las diferencias de pendiente) entre las curvas indican la diver-

gencia de los comportamientos y son éstas a las que hay que considerar. Además, una curva de cocientes acumulados da indicaciones precisas sobre la forma paramétrica que se va a escoger para modelizar el desarrollo de los procesos estudiados.

Por lo que respecta a la representación de la función de permanencia (véase capítulo IV.C), su utilización le resulta menos problemática al lector: la curva decrece a partir de la unidad hasta los valores mínimos correspondientes a la proporción de aquellos que todavía no han experimentado el evento al final de la observación. Sin embargo, esas curvas se obtienen mediante un cálculo realizado a partir de los cocientes instantáneos de ocurrencia y, sobre todo, si los efectivos son reducidos se corre el riesgo de acumular las incertidumbres del cálculo y, por lo tanto, de suavizar demasiado los resultados.

La tercera opción es trazar las curvas de los cocientes instantáneos de ocurrencia, con lo que se obtiene la representación más fiel. Lamentablemente, cuando tenemos que comparar comportamientos los trazos de las curvas de los cocientes aparecen entremezclados y muy caóticos, por lo que no permiten ninguna visualización de la similitud o las diferencias de los comportamientos.

D) Conclusión

En este capítulo hemos presentado el análisis de dos fenómenos en interacción. Al prolongar el análisis demográfico clásico, este procedimiento nos permite poner en evidencia dependencias complejas: dependencias unilaterales cuando se puede identificar un orden o sentido específico, y dependencias recíprocas cuando los dos eventos influyen uno sobre otro.

Cuando se pasa al análisis práctico se presentan los problemas de simultaneidad de las ocurrencias. Tales problemas reflejan cuestiones fundamentales de interpretación que plantean los datos biográficos.

Los modelos presentados permiten una estratificación fina de los datos; sin embargo, tal como veremos en el siguiente capítulo, para la extensión del análisis hacia más de dos eventos será necesario realizar reagrupamientos que permitan llegar a resultados estadísticamente significativos.

VI. GENERALIZACIÓN PARA SITUACIONES MÁS COMPLEJAS

La situación bivariada es el modelo de interacción más simple. Es posible considerar la extensión del marco de análisis para situaciones más complejas, como por ejemplo las trivariadas o las multivariadas. Presentamos aquí la generalización de este análisis.

A) Presentación y límites de la aplicación práctica

En este capítulo abordaremos sucintamente la presentación teórica del problema planteado por la multiplicación de eventos estudiados en el seno de una misma muestra. Veremos, además, los límites de la extensión del problema bivariado que se presentan debido a la multiplicación rápida de los estados.

Al igual que en el caso bivariado, podríamos tratar los otros eventos como variables dependientes del tiempo cuya influencia se mediría sobre los cocientes instantáneos de ocurrencia del evento escogido; sin embargo, más que emprender un estudio de la regresión nos hemos decidido por analizar solamente las interacciones.

Sean $T_i, \ldots, T_k, k$ los diferentes eventos considerados y 0 el estado inicial, tenemos entonces:

$$h_{0i}(t) = \lim_{\Delta t \to 0} \frac{P\left(T_i < t + \Delta t \middle| \bigcap_j \{T_j \geq t\}\right)}{\Delta t}$$

$$h_{ij}(t|u) = \lim_{\Delta t \to 0} \frac{P\left(T_j < t + \Delta t \middle| \{T_i = u\} \bigcap_{\ell \neq i} \{T_\ell \geq t\}\right)}{\Delta t} \quad \text{donde } u < t$$

$$h_{\ell ij}(t|u,v) = \lim_{\Delta t \to 0} \frac{P\left(T_j < t + \Delta t \middle| \{T_i = u\} \cap \{T_\ell = v\} \bigcap_{\substack{n \neq \ell \\ n \neq i}} \{T_n \geq t\}\right)}{\Delta t}$$

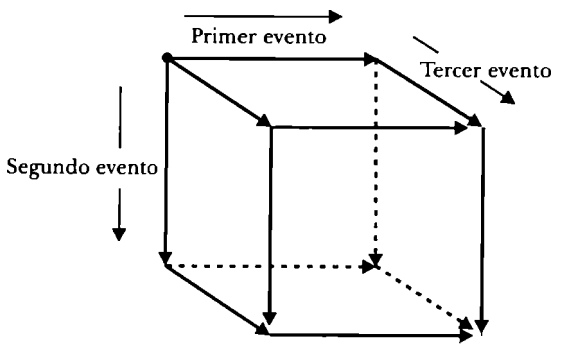
donde $u, v < t$ etcétera.

En el caso bivariado el espacio tiene cuatro estados, en el caso trivariado tendrá ocho, en el multivariado, con k eventos, tendrá 2^k estados, lo que rápidamente se vuelve inconcebible por la reducción de datos a medida que éstos se reparten en el espacio de los estados.

En el caso trivariado donde los individuos inician su trayectoria en 0, aún es posible representar el espacio de los estados y las transiciones simples (figura 1), pues la multitud de eventos simultáneos y la doble o triple simultaneidad haría ilegible el esquema. Cuando se trata de un número más importante de eventos considerados en interacción, ya no se pueden trazar representaciones simples.

Sin embargo en el estudio de los eventos demográficos, el caso de los eventos renovables jerarquizados (la fecundidad en particular) permite considerar una solución a la multiplicidad de estados. En efecto, es posible concebir un estudio de las interacciones entre migraciones sucesivas y nacimientos.

FIGURA 1
Esquema de estudio trivariado

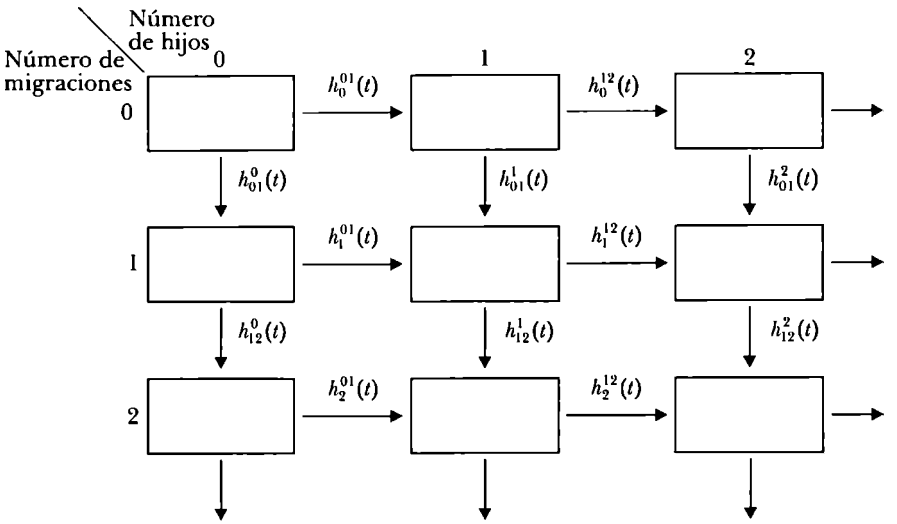


En ese caso se constata que a diferencia de las interacciones múltiples a las que nos referimos antes, nos encontramos aquí un espacio plano con dos dimensiones (figura 2).

B) INTERACCIONES ENTRE TRES EVENTOS: DOS CASOS DE ESTUDIO

A continuación vamos a presentar dos casos de aplicación. Estos dos ejemplos plantean el camino que sigue el investigador cuando enfrenta un problema particular, y un procedimiento del que conviene sacar el mayor partido.

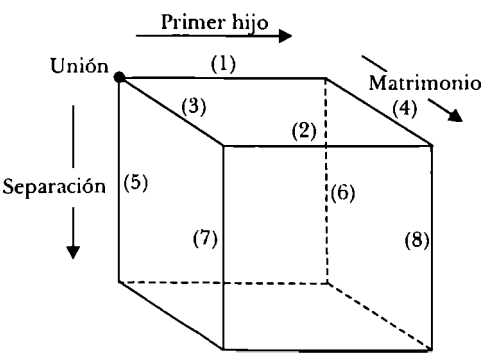
FIGURA 2
Esquema de estudio multivariado en el plano



Fuente: Courgeau y Lelièvre, 1985.

Nos proponemos estudiar los procesos de fecundidad y de separación de diversos tipos de unión (matrimonio o cohabitación). La muestra estudiada es un conjunto de periodos de unión que comienzan o no por un matrimonio y terminan con una separación o un divorcio. El siguiente esquema de un espacio de ocho estados es conveniente para caracterizar la situación; debemos advertir, sin embargo, que ciertas transiciones están prohibidas (las que no llevan numeración) (figura 3).

FIGURA 3
Estadio trivariado de las uniones



El esquema permite así medir las influencias siguientes:

- influencia del matrimonio sobre el nacimiento: comparaciones de (1) y (2) para una pareja de cohabitantes, y recíprocamente;
- influencia de la ocurrencia de un nacimiento sobre el matrimonio: comparación de (3) y (4);

lo que constituye un caso clásico de estudio bivariado.

Además:

- influencia del nacimiento sobre la separación de una pareja que cohabita (5) (6) o está casada (7) (8); esta vez sin recíproco, al contrario de lo que sucede para las influencias siguientes;
- la separación diferencial o no de las parejas sin hijos según estén casadas o cohabiten (7) (5);
- la separación diferencial o no de las parejas que hayan tenido un nacimiento según sean casadas o cohabiten (8) (6).

Sin embargo, este esquema trivariado no permite el análisis de un problema, que es la ocurrencia diferencial de nacimientos según el desenlace del establecimiento de la unión sea un matrimonio o una separación, lo que significa comparar dos tipos de trayectoria (1) según el futuro. De esa manera se prueba una hipótesis de selección, de anticipación: las parejas que se casarán tienen un comportamiento de fecundidad diferente mientras aún están cohabitando; o una hipótesis de adaptación: todos los cohabitantes tienen el mismo comportamiento de fecundidad y sólo cuando se casan intervienen las diferencias (*cf.* capítulo V.A.3).

Para hacer este estudio nos situaremos de nuevo en el marco de un estudio bivariado donde se han distinguido dos subpoblaciones en la muestra de partida. Luego hay que comparar los valores de los cocientes instantáneos de los eventos ocurridos en la trayectoria (1), obtenidos en los análisis sobre las dos subpoblaciones.

El segundo caso de aplicación es el del estudio de las interacciones entre la primera migración después del matrimonio, el primer nacimiento, y el primer cambio (interrupción o regreso) de la actividad profesional después del matrimonio. Esta vez es posible concebir todas las transiciones. Sólo existen idas y venidas entre los estados de actividad e inactividad (figura 4).

Hasta el momento, el estudio ha dividido la muestra de mujeres según su actividad al casarse en activas e inactivas, y se ha efectuado un análisis por separado para los dos grupos.

Con la intención de dar una idea general de la complejidad de los cálculos necesarios, contabilizamos los movimientos efectuados en la duración t: los movimientos simples, dobles o triples (eventos simultáneos) según el estado del cual proceden.

FIGURA 4
Estudio trivariado: migración, nacimiento, actividad profesional

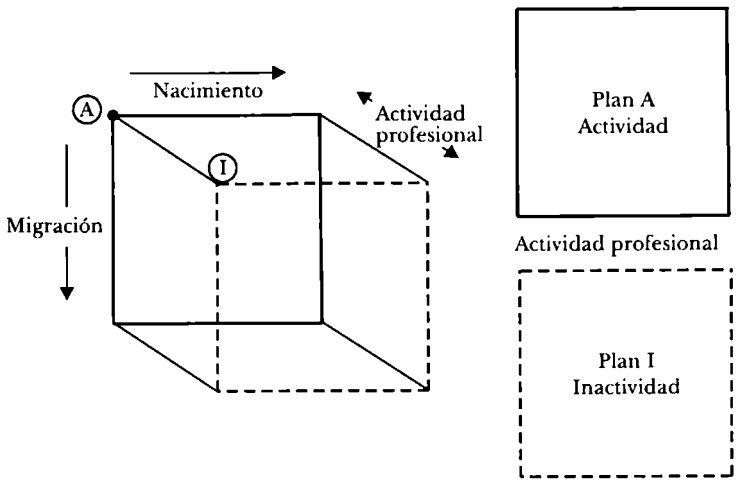


FIGURA 5
Destinos posibles al inicio

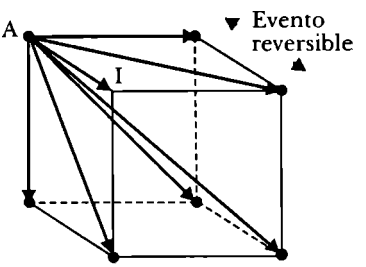


FIGURA 6
Trayectorias posibles luego de haber experimentado un evento irreversible

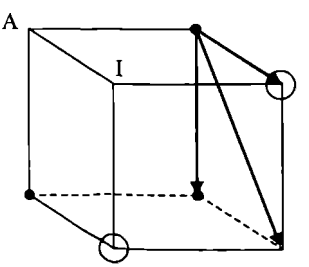
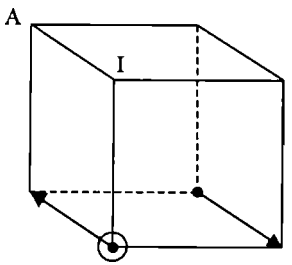


FIGURA 7
Las únicas trayectorias restantes



En efecto, es posible salir de A (respectivamente de I) de siete maneras diferentes (figura 5).

Si ya se experimentó uno de los eventos irreversibles según se esté activo o inactivo en el inicio (cuatro estados posibles simbolizados por O y • en el esquema), se puede salir del estado alcanzado de tres maneras diferentes (figura 6).

Por último, luego de haber experimentado los dos eventos irreversibles no se puede dejar el estado más que de una sola manera: experimentando el evento reversible que conduce hacia el estado final del recorrido (figura 7).

Debido al tamaño de la muestra inicial, esta aplicación ha conducido a comparaciones no significativas desde el punto de vista estadístico. Aquí resulta claro que el análisis puramente no paramétrico debe sustituirse por una modelización más formal (paramétrica o semiparamétrica) que permitirá entonces discriminar ciertos comportamientos en relación con otros, mediante la ayuda del conjunto de la información recogida.

C) INTERACCIONES ENTRE DOS EVENTOS CON OCURRENCIAS MÚLTIPLES

Retomemos el ejemplo que presentamos en la introducción, el de las interacciones entre la fecundidad y las migraciones (Courgeau 1985; Courgeau y Lelièvre, 1987). El espacio de los estados (figura 2) es entonces plano y cada una de las dimensiones es recorrida por las variables aleatorias T_1, T_2 ... T_m ... para las migraciones, y T^1, T^2 ... T^n ... para los nacimientos. Así, cada estado se define por una pareja (m, n) donde m es el número de migraciones anteriores y n el número de hijos ya nacidos. Los cocientes instantáneos de transición entre los diferentes estados son de la forma:

• *para las migraciones*

$$h^n_{i,i+1}(t|u_i,\dots u_1,v_n,\dots v_1) = \lim_{\Delta t\to 0} \frac{P(T_{i+1} < t+\Delta t | T_{i+1} \geq t, T_i = u_i,\dots T_1 = u_1, T^n = v_n,\dots T^1 = v_1)}{\Delta t}$$

• *para los nacimientos*

$$h_m^{j,j+1}(t|u_m,\dots u_1,v_j,\dots v_1) = \lim_{\Delta t\to 0} \frac{P(T^{j+1}<t+\Delta t|T^{j+1}\geq t,T_m=u_m,\dots T_1=u_1,T^j=v_j,\dots T^1=v_1)}{\Delta t}$$

Resulta evidente que no se puede considerar un modelo como éste, que impone una estratificación tan fina de la muestra inicial en $n \times m$ estados en cada duración. Se necesitan hipótesis suplementarias.

En un primer tiempo podemos suponer que la intensidad de la migración no depende sino del rango de ésta, de la duración transcurrida desde la última migración (u), y de la transcurrida desde el último nacimiento (v). Según estas dos hipótesis, la escritura de los cocientes instantáneos sería:

• *para las migraciones*

$$h_{m,m+1}^{n}(t|u,v) = \lim_{\Delta t\to 0} \frac{P(T_{m+1}<t+\Delta t|T_{m+1}\geq t,T_m=u,T^n=v)}{\Delta t}$$

• *para los nacimientos*

$$h_m^{n,n+1}(t|u,v) = \lim_{\Delta t\to 0} \frac{P(T^{n+1}<t+\Delta t|T^{n+1}\geq t,T_m=u,T^n=v)}{\Delta t}$$

Sigue siendo difícil estimar estas expresiones de manera totalmente no paramétrica, lo que obliga a plantear nuevas hipótesis. Admitamos que las intensidades no dependen más de u y v, sino que se puede hacer un estudio de interacciones entre nacimientos y migraciones para el que sólo se retiene la probabilidad de migrar cuando se tienen j hijos:

$$h^j(t|v_j) = \lim_{\Delta t\to 0} P(T_x<t+\Delta t|v_j\leq t,T_x\geq t)$$

y la probabilidad de tener un nuevo hijo cuando ya se migró i veces:

$$h_i(t|u_i) = \lim_{\Delta t\to 0} P(T^x<t+\Delta t|u_i\leq t,T^x\geq t)$$

Bajo tales hipótesis tenemos la posibilidad de responder a las preguntas siguientes: ¿cuáles son los efectos sobre los procesos migratorios de n nacimientos anteriores?, y recíprocamente ¿cuáles son las consecuencias sobre la fecundidad de m migraciones anteriores?

En el caso de las mujeres nacidas entre 1911 y 1925 se pudieron identificar comportamientos diferentes según la edad al casarse. Para las que se casaron jóvenes (entre 15 y 22 años) se evidenció claramente la forma

en que incidía el tamaño de la familia sobre su movilidad: mientras más elevado es el número de nacimientos más móviles son las mujeres. Este efecto es particularmente notorio para los tres primeros nacimientos. Por el contrario, para las mujeres de la misma cohorte casadas más tardíamente (después de los 22 años) este efecto no existe, lo cual plantea la posibilidad de que al inicio estas mujeres dispusieran de una vivienda de tamaño suficiente para la descendencia esperada. Recíprocamente, para ambos grupos se puede descubrir un efecto anticipatorio de la mudanza que precede a un nacimiento.

Estas hipótesis, sin embargo, siguen siendo muy restrictivas, puesto que no permiten tomar en cuenta la última duración de permanencia en una residencia, ni la extensión del último intervalo entre dos nacimientos, lo que evidentemente representa un papel determinante sobre los estimadores, sobre todo si el evento se ha producido con una vecindad temporal bastante próxima. Por lo tanto, resulta necesario recurrir a una modelización más adecuada que tome en cuenta no sólo el rango sino también la duración transcurrida. No obstante esta modelización será más restrictiva, puesto que es paramétrica o semiparamétrica. Ella permite prolongar este primer análisis haciendo intervenir las numerosas características individuales disponibles.

D) Conclusión

Es posible considerar una extensión de las capacidades de los modelos en dos direcciones: sea aumentando el espacio de los estados mediante la confrontación de un número creciente de eventos de dominios distintos, sea estudiando las interacciones entre dos series jerarquizadas de eventos. Esas generalizaciones hacia espacios cada vez más vastos se topan de inmediato en su aplicación con estratificaciones a las que no se puede someter a ninguna muestra, por exhaustiva que ésta sea al inicio. Nos vemos entonces obligados a elaborar hipótesis simplificadoras que, si bien permiten responder a preguntas de primordial interés, sin embargo siguen siendo muy fuertes. Parece casi imposible tratar de relajarlas en el marco de una modelización esencialmente no paramétrica, por lo que es necesario recurrir a aproximaciones más estructurantes (métodos semiparamétricos o paramétricos). En la segunda parte de esta obra abordaremos ese campo extremadamente rico.

SEGUNDA PARTE

GENERALIZACIÓN DE LOS MODELOS DE REGRESIÓN

VII. FORMALIZACIÓN ESTADÍSTICA DEL ANÁLISIS PARAMÉTRICO

Los modelos que se presentaron en la primera parte de esta obra no le imponen ninguna forma *a priori* a los cocientes que se estiman. Por ello, son no paramétricos. Sin embargo, puede resultar útil resumir esta información utilizando un pequeño número de parámetros que permitan encontrar la distribución de conjunto correcta. Además del ahorro que esto aporta, a tales parámetros podría corresponder una interpretación demográfica simple que permitiera esclarecer los procesos observados.

En primer lugar presentaremos algunas distribuciones paramétricas seleccionadas entre las que se utilizan más corrientemente en demografía. Indicaremos simultáneamente los modelos subyacentes que pueden conducir a tales distribuciones y la significación de los parámetros estimados, cuando tengan alguna. Asimismo, daremos ejemplos prácticos provenientes de diversas encuestas, para ilustrar cada tipo de modelo.

Para llevar el análisis más adelante es necesario introducir el efecto de ciertas características individuales que hemos medido y que pueden tener influencia sobre los cocientes. Así, por ejemplo, el nivel educativo de un individuo puede tener un impacto sobre su matrimonio, los nacimientos de sus hijos, sus migraciones o su movilidad profesional; de ahí que sea posible introducir esas diversas características como variables explicativas del comportamiento del individuo, en esos modelos paramétricos. La distribución de la duración de permanencia dependerá, en ese caso, de características a la vez cualitativas y cuantitativas. Discutiremos la forma de esta dependencia y presentaremos diversos tipos de modelos, ilustrándolos siempre con ejemplos prácticos de aplicación.

Esas diversas características permiten introducir la heterogeneidad de las poblaciones sobre las que se trabaja.

A) Algunos modelos paramétricos útiles en demografía

Al igual que antes, nos situamos en una población cuyos individuos, según se supone, experimentaron un evento dado en el instante T. Se trata de una variable aleatoria positiva cuya distribución queremos representar de manera paramétrica.

1) Distribución exponencial

Es la distribución más simple que se obtiene cuando el cociente instantáneo de ocurrencia es una constante, independiente de la duración. Se puede entonces escribir:

$$h(t) = \rho \tag{1}$$

De donde resultan los valores de la función de permanencia:

$$S(t) = \exp\left(-\int_0^t h(s)ds\right) = \exp(-\rho t) \tag{2}$$

y de la función de densidad de probabilidad:

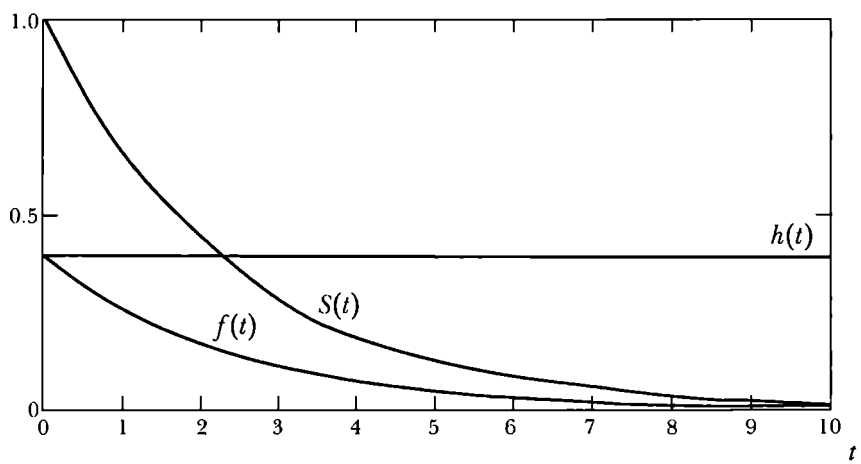
$$f(t) = h(t)S(t) = \rho\exp(-\rho t) \tag{3}$$

que son ambas funciones exponenciales del tiempo. Este modelo no depende más que de un parámetro ρ que siempre es positivo. La duración media de permanencia es igual al inverso de este parámetro, $1/\rho$.

La figura 1 presenta esas tres distribuciones en función del tiempo:

$$S(t), f(t) \text{ y } h(t)$$

FIGURA 1
Función de permanencia, $S(t)$, densidad de probabilidad, $f(t)$, de una distribución exponencial cuyo cociente instantáneo es constante e igual a 0.4, en función de la duración t



Un modelo como éste implica una ausencia completa de memoria entre los individuos sometidos a riesgo; cualquiera que sea la duración transcurrida a partir del instante inicial, el cociente es independiente de ella. Se puede decir entonces que para todo instante t_0 dado, la distribución condicional de $(T - t_0)$, sabiendo que T es superior a t_0, es la misma que la distribución de T.

Una distribución como ésta se observa cuando un individuo está sometido a un gran número de razones para experimentar el evento, y lo experimenta en el momento en que encuentra una de esas razones. Veamos más en detalle cómo representar ese mecanismo.

Sean $T_1, ..., T_n$ instantes aleatorios independientes que tienen la misma distribución en función del tiempo. Esta distribución es tal que cuando $t \to 0$, se puede escribir:

$$\begin{aligned} S(t) &\simeq 1 - \rho t \\ f(t) &\simeq \rho \end{aligned} \tag{4}$$

Si llamamos M_n al mínimo de los instantes aleatorios $T_1, ..., T_n$, la densidad de M_n se puede escribir:

$$f_{M_n}(t) = n\rho(1 - \rho t)^{n-1} \tag{5}$$

pues ese mínimo está cerca del valor 0. De donde resulta que la variable aleatoria $X_n = nM_n$ tiene una densidad:

$$\begin{aligned} f_{X_n}(t) &= \lim_{\Delta t \to 0} \frac{P(t \le nM_n < t + \Delta t)}{\Delta t} \\ &= \frac{1}{n} \lim_{\frac{\Delta t}{n} \to 0} \frac{P\left(\frac{t}{n} \le M_n < \frac{t}{n} + \frac{\Delta t}{n}\right)}{\frac{\Delta t}{n}} = \rho\left(1 - \frac{\rho t}{n}\right)^{n-1} \end{aligned} \tag{6}$$

En ese caso la función de permanencia se escribe:

$$S_{X_n}(t) = \left(1 - \frac{\rho t}{n}\right)^n \tag{7}$$

y el cociente instantáneo:

$$h_{X_n}(t) = \frac{\rho}{1 - \frac{\rho t}{n}} \tag{8}$$

Se ve entonces que cuando $n \to \infty$, $h_{X_n}(t)$ tiende hacia la constante ρ.

Veamos ahora cómo verificar de manera empírica si es factible utilizar esta distribución. Hay diferentes pruebas simples que se pueden llevar a cabo.

Si se utiliza el cociente instantáneo de ocurrencia acumulado, en lugar del cociente instantáneo de ocurrencia cuyas variaciones pueden ser importantes, se escribirá:

$$H(t) = \int_0^t h(s)ds = \rho t \tag{9}$$

Vemos que al representar en una gráfica el cociente acumulado en función del tiempo transcurrido se obtiene una recta, si se verifica el modelo exponencial.

Un ejemplo de la utilización de este método lo encontramos en la figura 2. Aquí se examina la probabilidad de hacerse propietario después del nacimiento del último hijo, lo cual fue posible gracias a la encuesta "Triple biografía". Las duraciones se calculan para las familias completas, a partir del nacimiento del último hijo, por lo que son excluidas de esta muestra las mujeres que no hayan tenido un hijo. Los cocientes instantáneos de ocurrencia se estiman con la ayuda de los métodos no paramétricos descritos en la primera parte de este libro. El examen de la figura, que contiene los cocientes instantáneos de ocurrencia acumulados, muestra que para representar esos resultados de manera satisfactoria cabría utilizar una distribución exponencial.

Otro método utiliza el logaritmo de la función de permanencia. Efectivamente, si el modelo exponencial se ajustara a los datos, se escribiría:

$$\log S(t) = -\rho t \tag{10}$$

De esto se deriva que al colocar en una gráfica el logaritmo de la función de permanencia en función del tiempo se obtiene, si el modelo se ajusta a los datos, una recta que pasa por el origen. La figura 3 presenta esos resultados para los mismos datos considerados en la figura 2. De nuevo vemos que éstos se pueden representar mediante una distribución exponencial.

2) Mezcla de distribuciones exponenciales

En el caso anterior se planteó la hipótesis de que la población es homogénea: todos los individuos tienen la misma probabilidad de experimentar el evento. Una hipótesis como ésa es poco verosímil. Podemos suponer que la población se divide en diversas subpoblaciones, cada una con un cociente instantáneo independiente de la duración o el tiempo. El caso más general será aquel

FIGURA 2
Cocientes instantáneos acumulados de las mujeres que se hacen propietarias después del nacimiento de su último hijo, en función de la duración transcurrida después de ese nacimiento

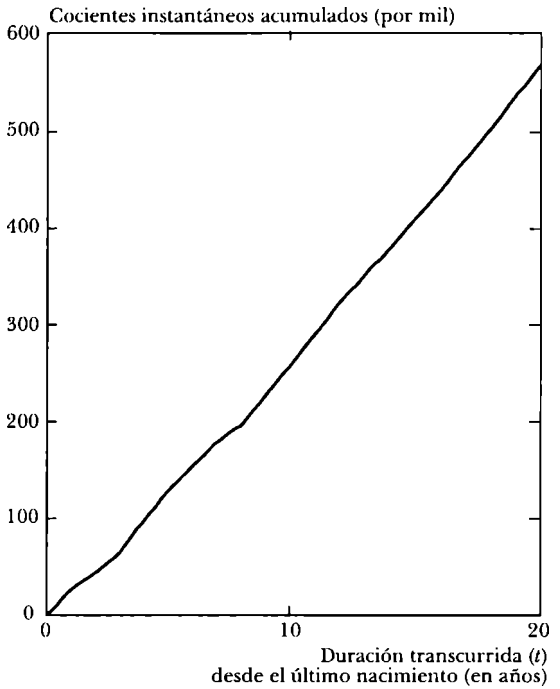
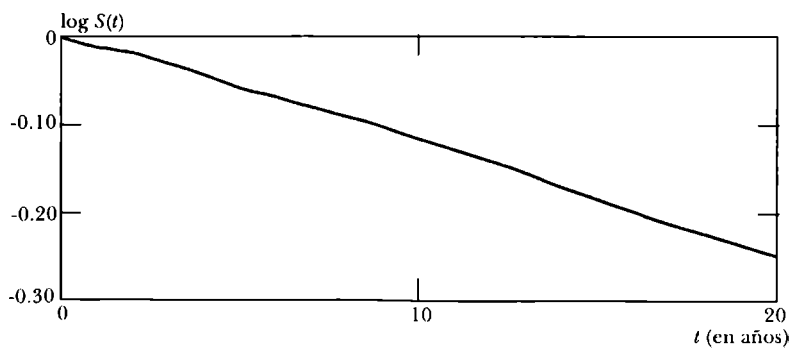


FIGURA 3
Logaritmo de la función de permanencia de las mujeres que aún no son propietarias después del nacimiento de su último hijo, en función de la duración transcurrida desde ese nacimiento



en el que cada individuo tiene un cociente instantáneo diferente del de los otros, pero que siempre es, sin embargo, independiente de la duración o el tiempo.

Partiremos del caso más simple, en el que la población está dividida en dos subpoblaciones, para llegar al más complejo, en el que cada individuo tiene un cociente instantáneo diferente.

En el primer caso, supongamos que la primera subpoblación no está sometida al riesgo de experimentar el evento, mientras que la segunda sí lo está. Caemos entonces en un modelo del tipo migrante-sedentario que se usa con mucha frecuencia en el caso de las migraciones internas (D. Courgeau, 1973) o de la movilidad profesional (Blumen *et al.*, 1955). Ya hicimos una presentación sucinta de este modelo en el capítulo II.B.2.

Llamemos $S(\infty)$ a la probabilidad de no experimentar jamás el evento. La función de permanencia de la segunda subpoblación sometida al riesgo es pues la parte $\big(S(t) - S(\infty)\big)$, en el instante t. Si el cociente instantáneo para esta subpoblación es igual a ρ, se puede escribir bajo la forma diferencial:

$$d\big(S(t) - S(\infty)\big) = -\rho\big(S(t) - S(\infty)\big)dt \tag{11}$$

de donde:

$$S(t) = S(\infty) + \big(1 - S(\infty)\big)\exp(-\rho t) \tag{12}$$

La función de densidad de probabilidad es igual a la inversa de la derivada de $S(t)$:

$$f(t) = \rho\big(1 - S(\infty)\big)\exp(-\rho t) \tag{13}$$

y el cociente instantáneo se escribe:

$$h(t) = \frac{\rho\big(1 - S(\infty)\big)\exp(-\rho t)}{S(\infty) + \big(1 - S(\infty)\big)\exp(-\rho t)} = \frac{\rho}{1 + \dfrac{S(\infty)}{1 - S(\infty)}\exp(\rho t)} \tag{14}$$

Se ve entonces que la población en su conjunto no tiene ya un cociente instantáneo constante, mientras que las dos subpoblaciones que la componen sí lo tienen. Cuando la duración es corta se ve cómo ese cociente se aproxima al valor ρ, y representa el cociente instantáneo de la población móvil. Cuando esta duración aumente ese cociente tenderá hacia cero y representará el cociente instantáneo de la población inmóvil. Será la diferente composición de la población al comienzo y al final de la observación lo que hará variar el cociente instantáneo observado sobre el conjunto de la población.

La duración media de permanencia en el estado inicial es igual a:

$$\int_0^{\infty} \rho t\big(1-S(\infty)\big)\exp(-\rho t)\,dt = \frac{\big(1-S(\infty)\big)}{\rho} \tag{15}$$

Vemos que ésta es igual a la duración media de permanencia de la población móvil, multiplicada por su proporción en la población total.

La figura 4 presenta las diversas distribuciones en función del tiempo cuando el modelo migrante-sedentario se verifica. Todas son uniformemente decrecientes.

Veamos ahora cómo verificar empíricamente si ese modelo es utilizable. Si se trabaja sobre la variación anual de la función de permanencia, se puede escribir:

$$S(t-1)-S(t) = \big(1-S(\infty)\big)(\exp\rho-1)\exp(-\rho t) \tag{16}$$

Eso da, al pasar a los logaritmos:

$$\log \Delta S(t) = \log c - \rho t \tag{17}$$

donde: $\Delta S(t) = S(t-1)-S(t)$ y $c = \big(1-S(\infty)\big)(\exp\rho-1)$

Al colocar en una gráfica semilogarítmica los valores de $\log \Delta S(t)$ en función del tiempo, se obtiene una recta de pendiente $-\rho$, cuando el modelo migrante-sedentario se verifica.

FIGURA 4

Función de permanencia $S(t)$, densidad de probabilidad $f(t)$, y cociente instantáneo $h(t)$, de una distribución del tipo migrante-sedentario, de parámetro $\rho = 0.4$ y $S(\infty) = 0.2$, en función de la duración t

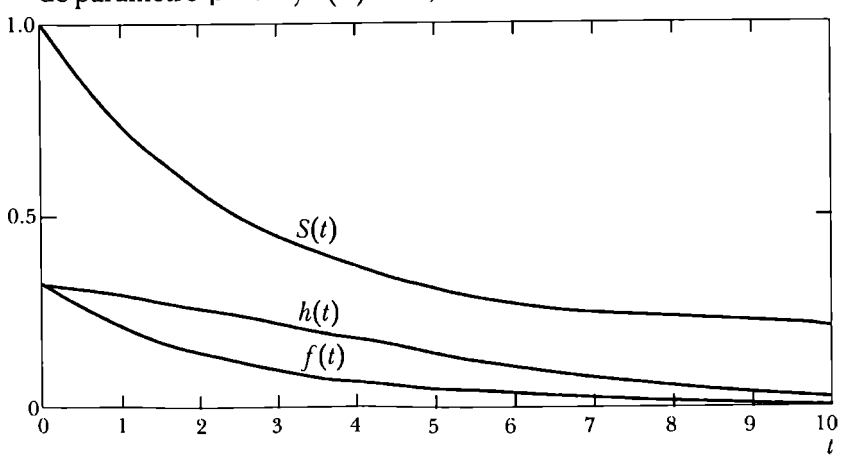
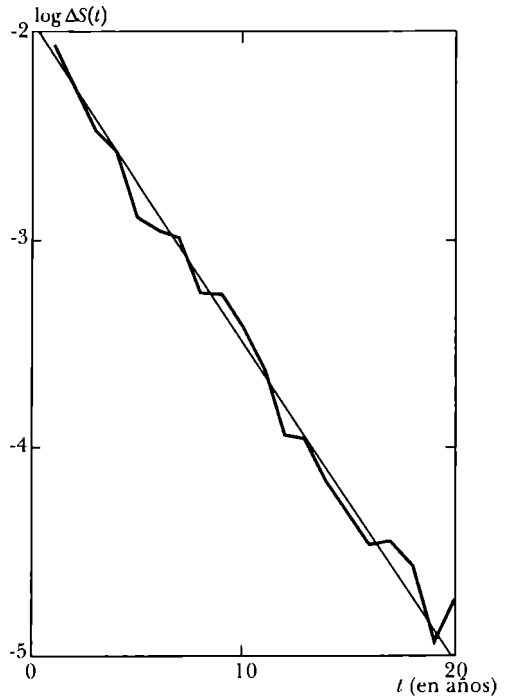


FIGURA 5
Logaritmo de la variación de la función de permanencia en función de la duración t, para los cambios de residencia de las mujeres nacidas entre 1926 y 1936



La figura 5 presenta esos valores para los cambios de residencia de las mujeres nacidas entre 1926 y 1936, en función de la duración de permanencia. Los datos provienen de la encuesta "Triple biografía".

Esos puntos que se apoyan en datos no agregados están más dispersos que en los ejemplos precedentes. Sin embargo se advierte que están correctamente alineados y que en ese caso se verifica el modelo migrante-sedentario. Los migrantes tienen un cociente de movilidad aproximadamente igual a 0.15 y constituyen 86% de la población inicial.

Ese modelo puede hacerse más complejo introduciendo dos probabilidades de migrar distintas de cero y diferentes para cada subpoblación, o introduciendo n subpoblaciones cada una de las cuales tiene una probabilidad de migrar diferente. El tratamiento de esos diversos modelos es similar al del modelo migrante-sedentario.

Veamos aquí el caso más complejo, donde se mezcla una infinidad de distribuciones exponenciales. En ese caso cada individuo tiene siempre un

cociente instantáneo constante en el transcurso del tiempo, pero ese cociente es variable de un individuo al otro. Eso permite introducir una heterogeneidad entre los individuos.

Definamos en ese caso una nueva variable aleatoria P, cuya densidad de probabilidad es $f_P(.)$, que representa la distribución de los diversos cocientes en la población. La densidad de probabilidad de T condicionada por el hecho de que $P = \rho$ para un individuo dado, es como anteriormente:

$$f_{T|P}(t|\rho) = \rho \exp(-\rho t) \tag{18}$$

En ese caso, la densidad no condicionada para P se obtiene sumando en relación a todos los valores posibles de ρ:

$$f_T(t) = \int_0^\infty \rho \exp(-\rho t) f_P(\rho) d\rho \tag{19}$$

La función de permanencia se escribe entonces:

$$S_T(t) = \int_t^\infty f_T(\theta) d\theta = \int_{\theta=t}^\infty \int_{\rho=0}^\infty \rho \exp(-\rho\,\theta) f_P(\rho) d\rho d\,\theta$$

sea:

$$S_T(t) = \int_{\rho=0}^\infty \rho f_P(\rho) d\rho \int_{\theta=t}^\infty \exp(-\rho\theta) d\theta = \int_{\rho=0}^\infty \exp(-\rho t) f_P(\rho) d\rho \tag{20}$$

Se ve que la función de permanencia es la transformada de Laplace, unilateral de la densidad de probabilidad $f_P(\rho)$ de la variable P. En ese caso sabemos que cualquiera que sea la densidad de probabilidad de la variable P, la densidad y la función de permanencia no condicionadas por P serán siempre funciones monótonas de t.

Según la función de densidad de probabilidad de la variable P se llegará a diversos tipos de densidades para la variable T.

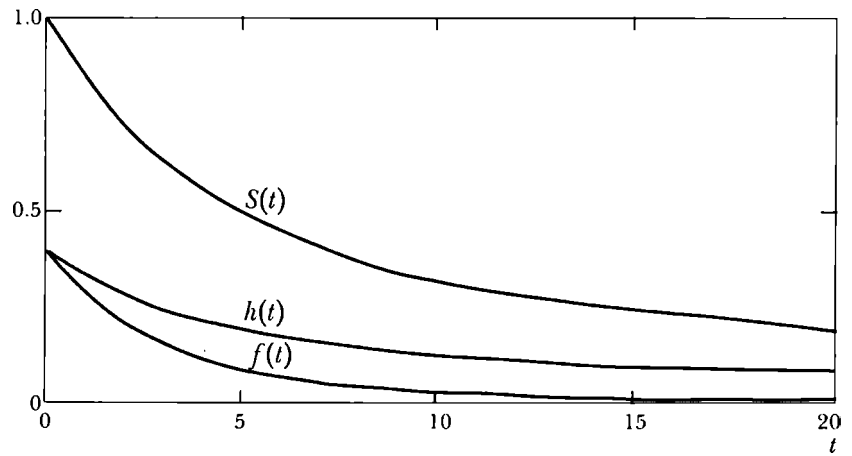
Si, por ejemplo, la variable P está distribuida exponencialmente teniendo como valor promedio ρ_0:

$$f_P(\rho) = \frac{1}{\rho_0} \exp\left(-\frac{\rho}{\rho_0}\right) \tag{21}$$

De donde se sigue que:

$$f_T(t) = \int_0^\infty \frac{\rho}{\rho_0} \exp\left[-\rho\left(t + \frac{1}{\rho_0}\right)\right] d\rho = \frac{1}{\rho_0\left(t + \frac{1}{\rho_0}\right)^2} \tag{22}$$

FIGURA 6
Función de permanencia $S(t)$, función de densidad de probabilidad $f(t)$, y cociente instantáneo $h(t)$, de una distribución de Pareto de parámetro $\rho_0 = 0.4$



de donde

$$S_T(t) = \frac{1}{\rho_0 t + 1} \tag{23}$$

y

$$h_T(t) = \frac{\rho_0}{\rho_0 t + 1} \tag{24}$$

Se ve que la densidad, la función de permanencia, y el cociente instantáneo siguen una distribución de Pareto. En el instante inicial el cociente instantáneo es igual al cociente instantáneo medio de la población, ρ_0. La figura 6 presenta esas diversas distribuciones cuando $\rho_0 = 0.4$.

En forma más general, si la función de densidad de probabilidad de P es una distribución de tipo gamma, de media ρ_0, se puede escribir:

$$f_P(\rho) = \frac{\lambda}{\rho_0}\left(\frac{\lambda\rho}{\rho_0}\right)^{\lambda-1} \frac{\exp\left(-\dfrac{\lambda\rho}{\rho_0}\right)}{\Gamma(\lambda)} \tag{25}$$

donde

$$\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} \exp(-x)\,dx \tag{26}$$

se puede mostrar que la función de densidad de probabilidad sigue todavía una distribución de tipo Pareto:

$$f_T(t) = \lambda \left(\frac{\lambda}{\rho_0} \right)^{\lambda} \left(t + \frac{\lambda}{\rho_0} \right)^{-(\lambda+1)} \tag{27}$$

La función de permanencia es igual a:

$$S_T(t) = \left(\frac{\lambda}{\rho_0} \right)^{\lambda} \left(t + \frac{\lambda}{\rho_0} \right)^{-\lambda} \tag{28}$$

y el cociente instantáneo es igual a:

$$h_T(t) = \frac{\lambda}{t + \frac{\lambda}{\rho_0}} \tag{29}$$

De nuevo esas tres funciones son distribuciones de Pareto. La densidad en el instante inicial es igual a su valor medio para el conjunto de la población, ρ_0. Si $\lambda = 1$ se cae de nuevo en el caso precedente, pues la distribución exponencial es un caso particular de una distribución de tipo gamma.

La distribución de Pareto aparece, pues, en el caso en que se puede considerar que la población estudiada está compuesta por individuos que siguen una distribución exponencial, y que están repartidos según una distribución de tipo gamma.

Es posible verificar empíricamente si una distribución semejante es satisfactoria presentando los valores del inverso del cociente en función del tiempo.

En efecto, las fórmulas (24) y (29) dan:

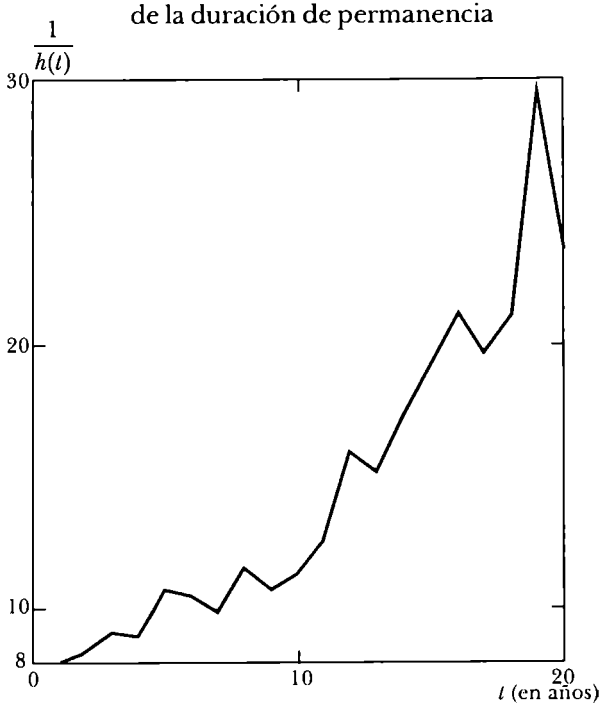
$$\frac{1}{h_T(t)} = \frac{t}{\lambda} + \frac{1}{\rho_0} \tag{30}$$

Si esta relación se verifica, entonces la gráfica es lineal.

La figura 7 presenta esos valores para los cambios de residencia de las mujeres nacidas entre 1926 y 1936, sobre los que ya se probó la validez del modelo migrante-sedentario. Vemos que la relación dista de ser lineal y que no se puede conservar la hipótesis de una distribución de P exponencial o de tipo gamma, mientras que el modelo migrante-sedentario sí se adapta muy bien a esta distribución.

Por supuesto que es posible concebir otros muchos tipos de distribución de P, cada uno de los cuales conducirá a una nueva distribución de las den-

FIGURA 7
Inverso del cociente instantáneo de cambio de residencia de las mujeres nacidas entre 1926 y 1936, en función de la duración de permanencia



sidades, funciones de pemanencia y cocientes instantáneos observados en el conjunto de la población. Recordemos que tales distribuciones siempre serán funciones monótonas del tiempo.

3) Distribución de Gompertz

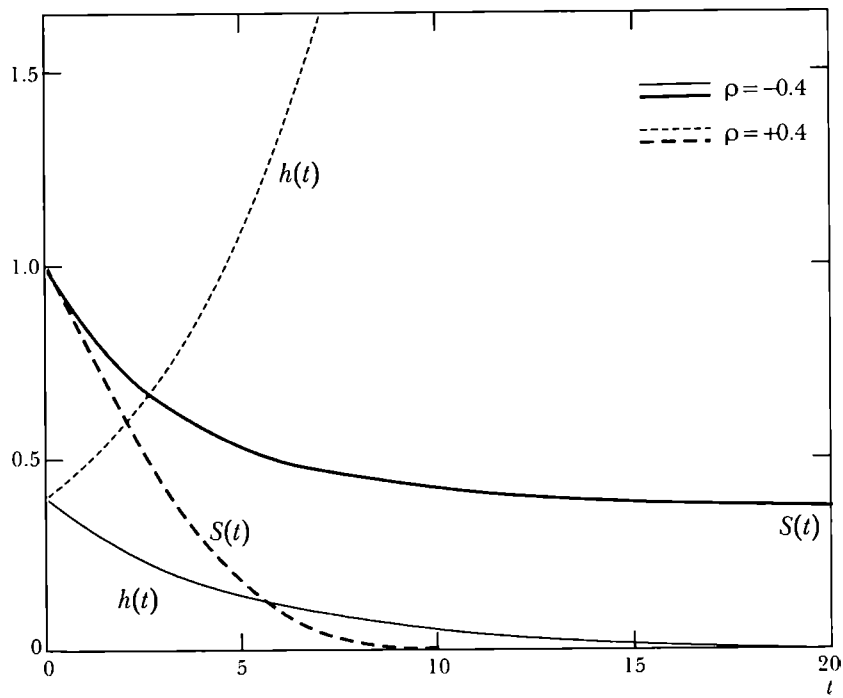
Consideremos nuevamente una población homogénea. Si la variación del cociente instantáneo en el transcurso del tiempo es proporcional a su valor en cada instante, entonces se observa una distribución de tipo Gompertz. Esta condición se puede escribir en la forma siguiente:

$$\frac{dh(t)}{dt} = \rho h(t) \tag{31}$$

que da después de la integración:

$$h(t) = \lambda \rho \exp(\rho t) \tag{32}$$

FIGURA 8
Funciones de permanencia $S(t)$, y cocientes instantáneos $h(t)$, de dos distribuciones de grupos de parámetros $(\rho = -0.4, \lambda = -1)$ y $(\rho = 0.4, \lambda = 1)$



Se ve que según el signo del parámetro ρ, el cociente puede ser uniformemente creciente ($\rho > 0$), o bien uniformemente decreciente ($\rho < 0$). Cuando $\rho = 0$ y $\lambda\rho$ tiende hacia una constante, volvemos a caer en la distribución exponencial precedente. El parámetro λ es siempre del signo ρ, de manera que el cociente sea positivo.

Cuando el coeficiente ρ es positivo se aplica una distribución como ésta a los datos sobre la mortalidad, al menos para las edades superiores a 35 años. Igualmente, tal distribución se ha aplicado a datos sobre la migración o la movilidad profesional, esta vez con un coeficiente negativo. La figura 8 presenta la distribución del cociente en función del tiempo para diversos valores del parámetro ρ, lo que permite distinguir bien los diversos casos.

El cociente instantáneo acumulado puede escribirse:

$$H(t) = \int_0^t \lambda\rho \exp(\rho\theta)\, d\theta = \lambda[\exp(\rho t) - 1] \qquad (33)$$

lo que conduce a:

$$S(t) = \exp\big(\lambda[1 - \exp(\rho t)]\big) \tag{34}$$

Vemos entonces que cuando el parámetro ρ es negativo, se puede interpretar el valor del parámetro λ, escribiendo cuando $t \to \infty$:

$$S(\infty) = \exp\lambda \tag{35}$$

En ese caso tal modelo está cerca del modelo migrante-sedentario, puesto que una parte de la población jamás experimenta el evento. Entonces se puede escribir:

$$S(t) = S(\infty)\exp\big[-\log S(\infty)\exp\rho t\big] \quad \text{si } \rho < 0 \tag{36}$$

lo que después de una diferencia da:

$$d\big(\log S(t) - \log S(\infty)\big) = \rho\big(\log S(t) - \log S(\infty)\big)dt \tag{37}$$

Se ve que la relación precedente (11) se verifica no sobre $S(t)$ sino sobre $\log S(t)$.

Regresando al modelo general, se puede escribir la densidad en la siguiente forma:

$$f(t) = \lambda\rho\exp\big[\rho t + \lambda\big(1 - \exp(\rho t)\big)\big] \tag{38}$$

Esta función de densidad de probabilidad puede escribirse además así:

$$f(t) = \rho\exp(\lambda)\exp\big[\rho t + \log(\lambda) - \exp\big(\rho t + \log(\lambda)\big)\big] \tag{39}$$

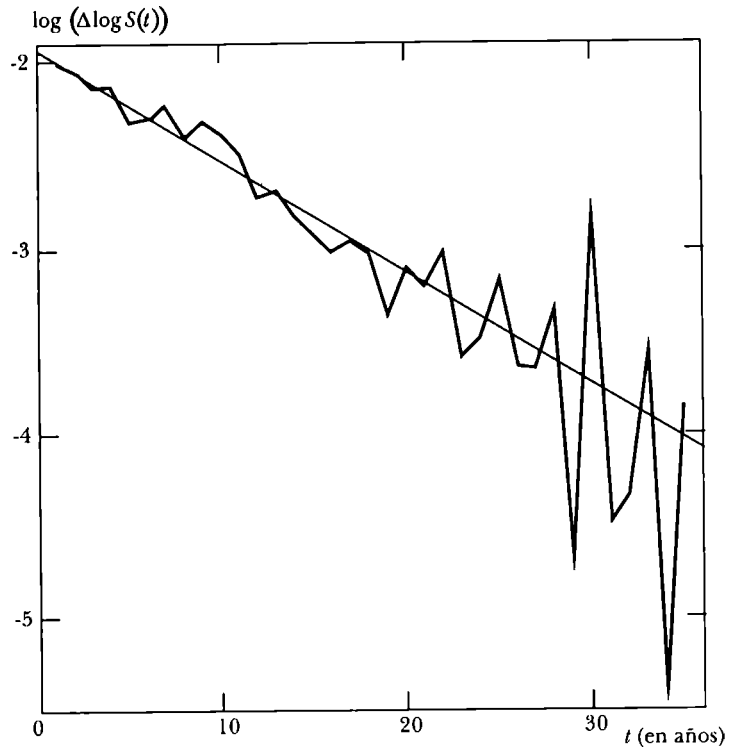
Para verificar empíricamente si se puede utilizar esta distribución, podemos poner el logaritmo de $h(t)$ en función del tiempo. Si se obtuviera aproximadamente una recta, esta distribución sería utilizable.

En la figura 9 presentamos esos valores para los cambios de residencia de las mujeres nacidas entre 1926 y 1936 en función de la duración de permanencia. Los datos son iguales a los utilizados para probar el modelo migrante-sedentario. Vemos que el modelo de Pareto es *a priori* igualmente satisfactorio para esos datos. Para escoger entre los dos modelos habrá que realizar pruebas más precisas.

Además, es posible trabajar sobre la variación anual del logaritmo de la función de permanencia. En efecto, podemos escribir:

$$\log S(t-1) - \log S(t) = \lambda\big(1 - \exp(-\rho)\big)\exp(\rho t) \tag{40}$$

FIGURA 9
Logaritmo del cociente instantáneo en función de la duración de permanencia para los cambios de residencia de mujeres nacidas entre 1926 y 1936



lo que da al pasar a logaritmos:

$$\log\left(\Delta\log S(t)\right) = \log c + \rho t \tag{41}$$

donde:

$$\Delta\log S(t) = \log S(t-1) - \log S(t) \text{ y } c = \log\left[\lambda\left(1-\exp(-\rho)\right)\right]$$

Al colocar en una gráfica los valores de $\log\left(\Delta\log S(t)\right)$ en función del tiempo se obtiene una recta, si se verifica el modelo de Gompertz.

En la figura 10 hemos presentado esos valores para los cambios de residencia de las mujeres nacidas entre 1926 y 1935 en función de la duración de permanencia. Ésta revela una vez más que es posible acercarse a tales migraciones mediante un modelo de Gompertz.

FIGURA 10
Logaritmo de la variación anual del logaritmo de la función de permanencia $S(t)$, en función de la duración de permanencia para los cambios de residencia de las mujeres nacidas entre 1926 y 1936

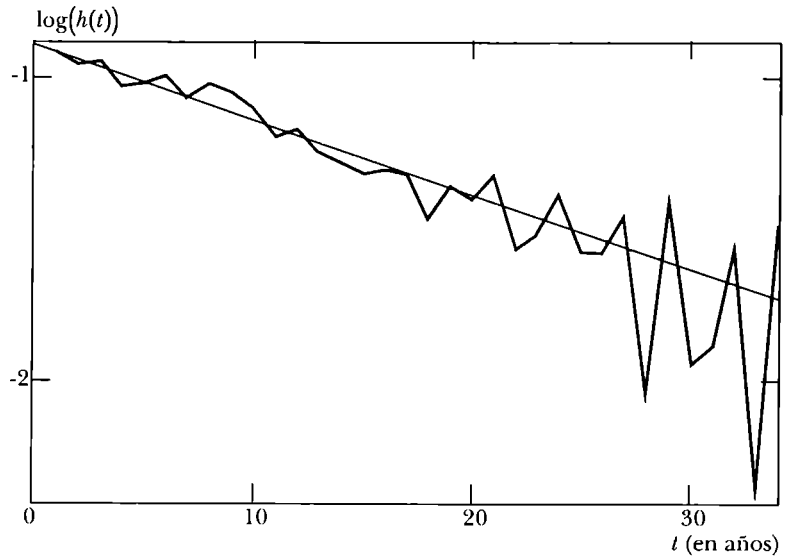
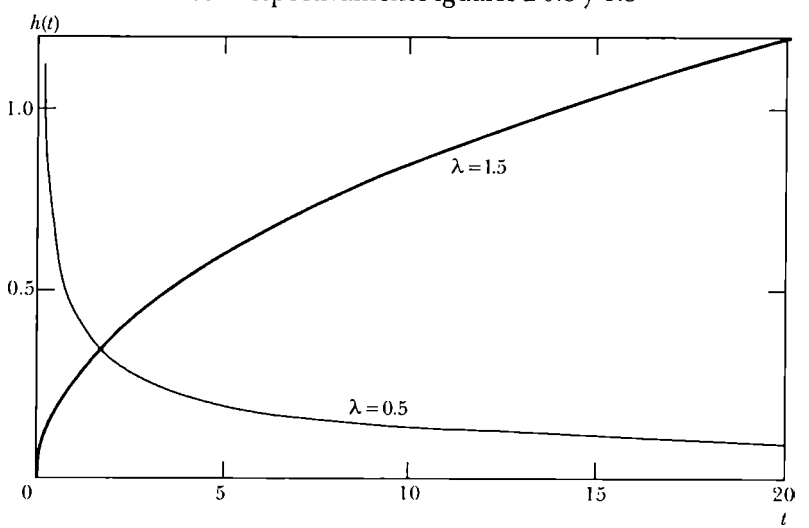


FIGURA 11
Cocientes instantáneos de dos distribuciones de Weibull cuyo parámetro ρ es igual a 0.4 y cuyos parámetros λ son respectivamente iguales a 0.5 y 1.5



Es posible introducir un parámetro más en el modelo. Se obtiene entonces un modelo de tipo Gompertz-Makeham, cuyo cociente instantáneo se escribe:

$$h(t) = \rho_0 + \lambda\rho \exp(\rho t) \tag{42}$$

lo que da:

$$\log S(t) = -\rho_0 t + \lambda[1 - \exp(\rho t)] \tag{43}$$

y:

$$f(t) = \left[\rho_0 + \lambda\rho \exp(\rho t)\right] \exp\left[-\rho_0 t + \lambda\left(1 - \exp(\rho t)\right)\right] \tag{44}$$

Dado que este modelo introduce un parámetro más que el modelo de Gompertz, generalmente se ajusta mejor a los datos empíricos.

4) Distribución de Weibull

La distribución de Weibull proporciona otro tipo de función monótona del tiempo para el cociente. Éste varía como una potencia dada del tiempo, lo que se escribe:

$$h(t) = \lambda\rho(\rho t)^{\lambda-1} \tag{45}$$

donde λ y ρ son parámetros positivos. Se ve que cuando $\lambda = 1$ volvemos a caer sobre una distribución exponencial. Cuando λ es superior a la unidad tenemos un cociente instantáneo que crece uniformemente; cuando λ es inferior a la unidad, decrece uniformemente.

La figura 11 presenta ese cociente en función del tiempo para diversos valores de λ y de ρ.

Al integrar respecto al tiempo, el cociente instantáneo acumulado se escribe:

$$H(t) = \int_0^t \lambda\rho(\rho\theta)^{\lambda-1} d\theta = (\rho t)^{\lambda} \tag{46}$$

de donde:

$$S(t) = \exp[-(\rho t)^{\lambda}] \tag{47}$$

y:

$$f(t) = \lambda\rho(\rho t)^{\lambda-1} \exp[-(\rho t)^{\lambda}] \tag{48}$$

Al igual que la distribución exponencial, esta distribución se observa cuando un individuo está expuesto a un gran número de razones para experimentar el evento observado y lo hace una vez que encuentra una de esas razones.

Sean entonces $T_1, \ldots, T_n$ instantes aleatorios independientes y con la misma distribución en función del tiempo. Esta distribución es tal que cuando $t \to 0$ se puede escribir:

$$S(t) \simeq 1 - (\rho t)^{\lambda} \qquad (49)$$
$$f(t) \simeq \lambda\rho(\rho t)^{\lambda-1}$$

Si llamamos M_n al mínimo de los instantes aleatorios $T_1, \ldots, T_n$, al igual que en el caso exponencial se puede escribir la densidad de M_n:

$$f_{M_n}(t) = n\lambda\rho(\rho t)^{\lambda-1}[1-(\rho t)^{\lambda}]^{n-1} \qquad (50)$$

De lo que resulta que la variable aletoria $X_n = n^{1/\lambda}M_n$ tendrá una densidad igual a:

$$\begin{aligned} f_{X_n}(t) &= \lim_{\Delta t \to 0} \frac{P(t \le n^{1/\lambda}M_n < t + \Delta t)}{\Delta t} \\ &= n^{-\frac{1}{\lambda}} n\lambda\rho(\rho t)^{\lambda-1} n^{\frac{1-\lambda}{\lambda}} [1-(\rho t)^{\lambda} n^{-1}]^{n-1} \\ &= \lambda\rho(\rho t)^{\lambda-1}[1-(\rho t)^{\lambda} n^{-1}]^{n-1} \end{aligned} \qquad (51)$$

De donde resulta:

$$S_{X_n}(t) = [1-(\rho t)^{\lambda} n^{-1}]^n \qquad (52)$$

y:

$$h_{X_n}(t) = \frac{\lambda\rho(\rho t)^{\lambda-1}}{1-(\rho t)^{\lambda} n^{-1}} \qquad (53)$$

Se ve entonces que cuando $n \to \infty$ el cociente tiende hacia el valor

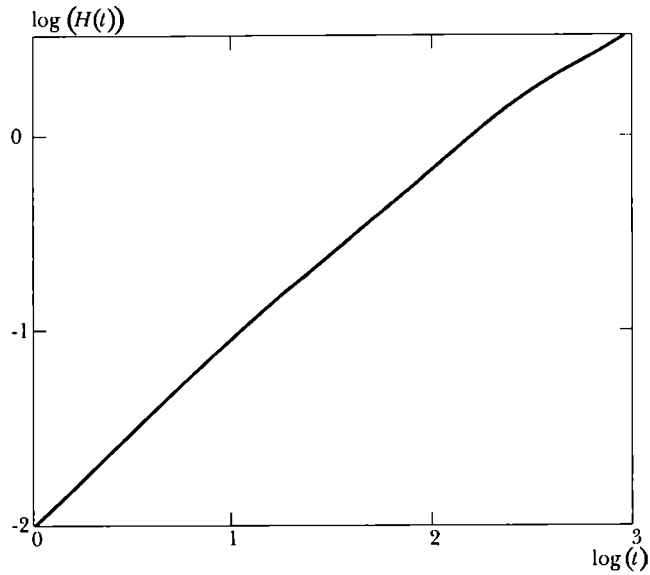
$$h_{X_n}(t) = \lambda\rho(\rho t)^{\lambda-1}$$

que es claramente una distribución de Weibull, dada por la fórmula (45) anterior.

Para verificar empíricamente si se puede utilizar un modelo de Weibull, se prueban los valores de $H(t)$, o de $-\log S(t)$. Se escribe entonces:

$$\log[H(t)] = \log[-\log S(t)] = \lambda(\log\rho + \log t) \qquad (54)$$

FIGURA 12
Logaritmo del cociente instantáneo acumulado $H(T)$, en función del logaritmo de la duración de permanencia en su residencia de mujeres nacidas entre 1926 y 1936



En una gráfica se sitúa el logaritmo del logaritmo inverso de la función de permanencia en relación con el logaritmo del tiempo t. Si el modelo de Weibull es utilizable, se obtiene una recta cuya pendiente es λ y cuya intersección con el eje del tiempo da una estimación de $-\log \rho$.

La figura 12 presenta los resultados obtenidos sobre los cambios de residencia de las mujeres nacidas entre 1926 y 1936. Vemos otra vez que es posible acercarse a esta distribución mediante una distribución de Weibull, para la que el valor aproximado de los parámetros es $\lambda = 0.84$ y $\log \rho = 1.3$. Sin embargo parece que el ajuste es menos bueno que con un modelo migrante-sedentario o de Gompertz. Más adelante veremos cómo escoger entre esas diversas distribuciones.

5) Distribución Gamma

Ésta constituye el último tipo de función monótona del tiempo para el cociente que presentamos aquí.

Esta distribución se obtiene cuando el individuo es sometido a cierto número de riesgos, todos distribuidos de manera exponencial, con el mismo

parámetro ρ. El individuo experimenta finalmente el evento estudiado cuando un número dado, λ, de esas eventualidades ocurre. En ese caso se ve que la función de densidad de probabilidad del evento es:

$$f(t) = \frac{\rho(\rho t)^{\lambda-1} \exp(-\rho t)}{(\lambda-1)!} \tag{55}$$

Esta función de densidad se puede generalizar cuando λ no sea entero. En ese caso se escribe:

$$f(t) = \frac{\rho(\rho t)^{\lambda-1} \exp(-\rho t)}{\Gamma(\lambda)} \tag{56}$$

donde la función $\Gamma(\lambda)$ se define por la relación siguiente:

$$\Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1} \exp(-x) dx \tag{57}$$

donde $\lambda > 0$.

Aunque entre las distribuciones continuas definidas por $t \geq 0$ ésta sea una de las que más se utilizan en estadística, es la más difícil de emplear para el análisis de las biografías debido a la complejidad de la función de permanencia y de los cocientes instantáneos.

Efectivamente, podemos escribir:

$$S(t) = 1 - \frac{\int_0^t x^{\lambda-1} \exp x \, dx}{\Gamma(\lambda)} \tag{58}$$

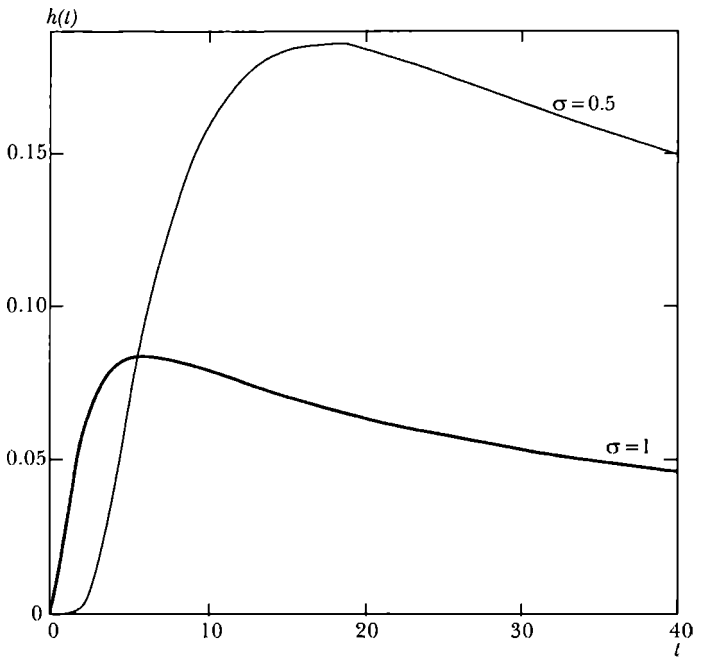
y:

$$h(t) = \frac{\rho(\rho t)^{\lambda-1} \exp(-\rho t)}{\Gamma(\lambda) - \int_0^t x^{\lambda-1} \exp x \, dx} \tag{59}$$

El cociente instantáneo crece del valor 0 al valor λ, cuando $\lambda > 1$, y decrece de infinito a λ cuando $\lambda < 1$. Cuando $\lambda = 1$ volvemos al modelo exponencial.

Dadas las dificultades de estimación de un modelo como éste cuando se trabaja con datos truncados, no lo presentaremos más detalladamente. Remitimos al lector interesado a la obra de Johnson y Kotz (1970).

FIGURA 13
Cocientes instantáneos de dos distribuciones log-normales cuyo parámetro ρ es igual a 0.1 y los parámetros σ son iguales respectivamente a 0.5 y a 1



6) Distribución log-normal

Abordaremos ahora las distribuciones cuyos cocientes ya no son funciones del tiempo que crecen o decrecen uniformemente. Una de las posibilidades es considerar la variable $Y = \log T$ como distribuida normalmente, de media $\log 1/\rho$ y con una desviación estándar σ. En ese caso T tiene una distribución log-normal cuya función de densidad de probabilidad se escribe:

$$f(t) = \frac{1}{\sigma t\sqrt{2\pi}} \exp\left(-\frac{[\log(t\rho)]^2}{2\sigma^2}\right) \tag{60}$$

Para el cálculo de la función de permanencia y del cociente instantáneo se necesita la utilización de la integral normal incompleta:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du \tag{61}$$

que da como valor para la función de permanencia:

$$S(t) = 1 - \phi\left[\frac{1}{\sigma}\log(t\rho)\right] \tag{62}$$

y como cociente instantáneo:

$$h(t) = \frac{1}{\sigma t\left(1 - \phi\left[\frac{1}{\sigma}\log(t\rho)\right]\right)\sqrt{2\pi}} \exp\left(-\frac{[\log(t\rho)]^2}{2\sigma^2}\right) \tag{63}$$

En la figura 13 colocamos los cocientes instantáneos obtenidos para diversos valores de σ. Se advierte que esos cocientes ya no son funciones monótonas de tiempo, sino que pasan por un máximo.

Para la función de permanencia es precisa la utilización de la integral normal incompleta, lo cual implica numerosos problemas de estimación que es difícil resolver simplemente cuando ciertos intervalos están abiertos.

En ese caso preferimos utilizar la distribución log-logística que proporciona una buena aproximación a la distribución log-normal, y en la que la función de permanencia o el cociente instantáneo se expresan con más simplicidad en función del tiempo.

7) Distribución log-logística

Sabemos que la distribución logística tiene una forma muy cercana a la de una distribución normal y que a menudo es posible remplazar la una por la otra sin introducir sesgos importantes.

Al plantear:

$$\log T = Y = -\log\rho + \sigma W \tag{64}$$

se obtiene una distribución log-logística para T cuando W tiene una distribución logística:

$$f_W(w) = \exp w\,[1 + \exp w]^{-2} \tag{65}$$

De esto resulta la función de densidad de probabilidad de T, planteando $\lambda = \frac{1}{\sigma}$:

$$f(t) = \frac{\lambda\rho(\rho t)^{\lambda-1}}{[1 + (\rho t)^{\lambda}]^2} \tag{66}$$

la función de permanencia:

$$S(t) = [1 + (\rho t)^{\lambda}]^{-1} \tag{67}$$

y el cociente instantáneo:

$$h(t) = \lambda\rho(\rho t)^{\lambda-1}[1 + (\rho t)^{\lambda}]^{-1} \tag{68}$$

La ventaja de esta distribución respecto de la log-normal es la forma explícita simple de la función de permanencia y del cociente instantáneo. Cuando los datos implican intervalos abiertos, eso permite una estimación fácil de los parámetros del modelo.

La figura 14 presenta la distribución de los cocientes instantáneos para diversos valores de λ. Si $\lambda > 1$ la curva de los cocientes pasa por un máximo de $t = \dfrac{(\lambda-1)^{1/\lambda}}{\rho}$, y si $\lambda < 1$ dicha curva es constantemente decreciente.

Podemos observar que si la distribución log-logística se adapta a los datos, entonces:

$$\frac{S(t)}{1 - S(t)} = (\rho t)^{-\lambda} \tag{69}$$

El logaritmo de la probabilidad de experimentar el evento después de t respecto de la probabilidad de experimentarlo antes de t es, pues, una función lineal del logaritmo del tiempo t.

Vemos que se puede escribir el cociente instantáneo bajo la forma siguiente:

$$h(t) = \frac{\lambda}{t}[1 - S(t)] \tag{70}$$

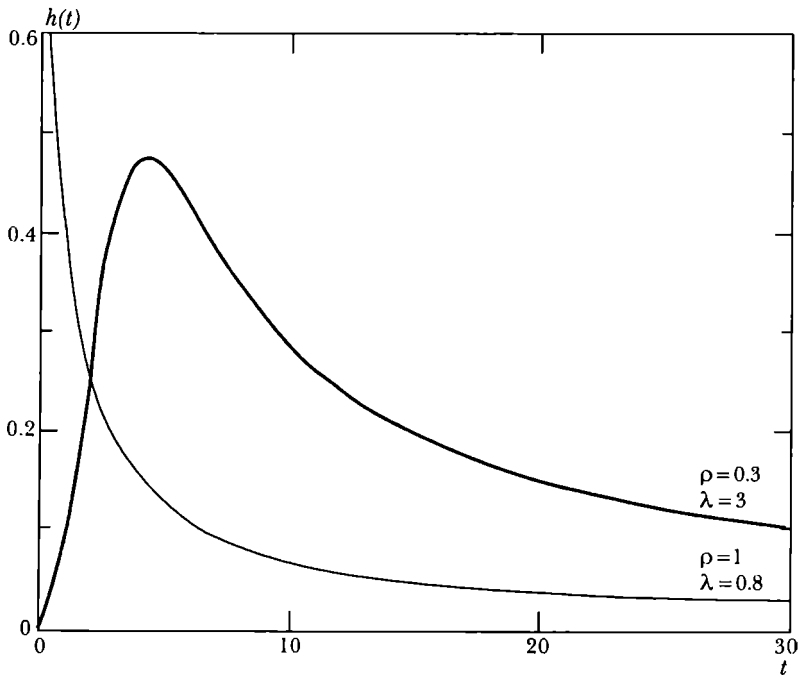
Esta relación significa que el cociente instantáneo es proporcional al número de individuos que ya experimentaron el evento, e inversamente proporcional a la duración transcurrida.

Para probar si un modelo como éste se adapta, podemos escribir:

$$\log\left[\frac{1 - S(t)}{S(t)}\right] = \log(\exp H(t) - 1) = \lambda \log(\rho t) \tag{71}$$

Si el modelo describe convenientemente la observación se obtiene una recta que lleva el logaritmo de la inversa de la función de permanencia menos la unidad, en función del logaritmo del tiempo transcurrido. En la figura 15 presentamos esos valores, al estudiar el hecho de hacerse propietario en fun-

FIGURA 14
Cocientes instantáneos de dos distribuciones log-logísticas cuyos parámetros son respectivamente $(\rho = 1, \quad \lambda = 0.8)$ y $(\rho = 0.3, \quad \lambda = 3)$



ción de la edad de los individuos a partir de los 15 años. Vemos que se puede reconstituir correctamente la distribución de esos datos mediante la ayuda de una distribución log-logística, al menos hasta los 50 años.

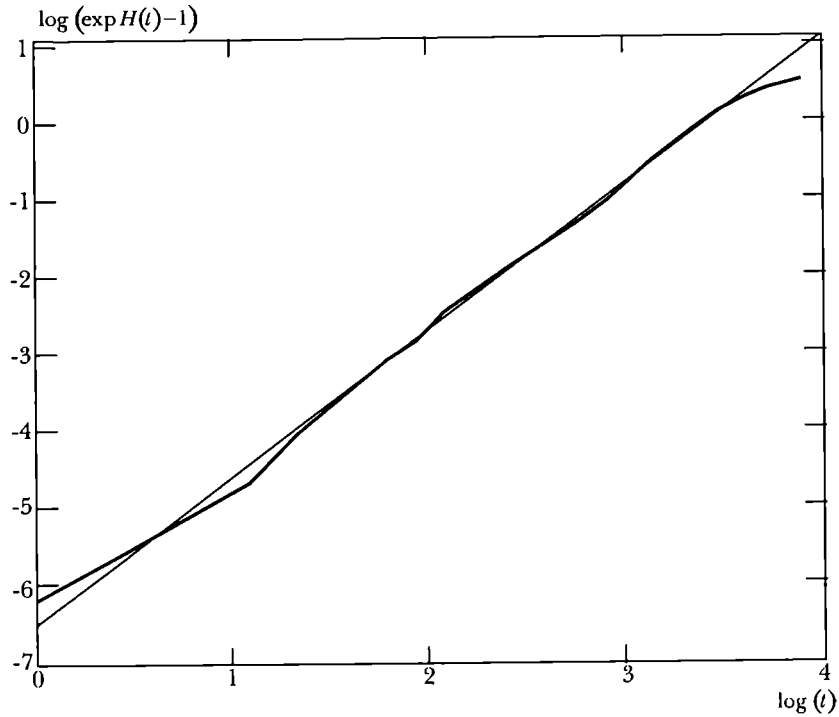
8) Distribución de Fisher-Snedecor (F) generalizada

Este último modelo paramétrico que presentamos aquí incorpora la mayoría de las distribuciones que vimos anteriormente, como casos particulares. Por lo tanto, teóricamente éste permite escoger de entre todas esas distribuciones a aquella que mejor se ajusta a los datos, tal como lo mostraremos en el siguiente capítulo.

Planteemos de nuevo:

$$\log T = Y = -\log \rho + \sigma W \tag{72}$$

FIGURA 15
Logaritmo del inverso de la función de permanencia menos la unidad en función del logaritmo de t, al estudiar el hecho de hacerse propietario en función de la edad de individuos a partir de los 15 años



donde la función de densidad de probabilidad de W es una distribución igual a:

$$f_W(w) = \frac{k_1^{k_1} k_2^{k_2} \Gamma(k_1 + k_2) \exp(k_1 w)}{\Gamma(k_1)\Gamma(k_2)(k_1 + k_2 \exp w)^{k_1 + k_2}} \tag{73}$$

donde $2k_1$ y $2k_2$ son los grados de libertad de esta distribución y $\Gamma(k)$ la función Γ antes definida.

Se ve, por ejemplo, que si $k_1 = k_2 = 1$ el modelo se transforma en una distribución logística:

$$f_W(w) = \frac{\exp(w)}{(1 + \exp w)^2} \tag{74}$$

tenemos entonces una distribución log-logística para T.

Cada uno de los demás tipos de modelos corresponden a valores particulares de k_1 y k_2, pudiendo éstas tender hacia el infinito. Así, por ejemplo, cuando $k_1 = 1$ y $k_2 \to \infty$ se puede escribir:

$$\lim_{k_2 \to \infty,\, k_1 = 1} f_W(w) = \exp(w - \exp w) \tag{75}$$

Vemos entonces que en ese caso se obtiene una distribución de Weibull si $\sigma \neq 1$, y una distribución exponencial si $\sigma = 1$. Respecto de los demás casos, podemos referirnos a Kalbfleisch y Prentice (1980, pp. 28-29).

Si se utiliza la relación (73) se puede calcular la densidad de probabilidad de T, la cual es igual a:

$$f(t) = \frac{k_1^{k_1}\, k_2^{k_2}\, \Gamma(k_1 + k_2)\rho(\rho t)^{\frac{k_1}{\sigma} - 1}}{\Gamma(k_1)\Gamma(k_2)\sigma[k_2 + k_1(\rho t)^{\frac{1}{\sigma}}]^{k_1 + k_2}} \tag{76}$$

De nuevo, si $k_1 = k_2 = 1$ y si se plantea $\lambda = \dfrac{1}{\sigma}$ se tiene una distribución log-logística para T:

$$f(t) = \frac{\lambda\rho(\rho t)^{\lambda - 1}}{\left(1 + (\rho t)^{\lambda}\right)^2} \tag{77}$$

9) Comparación de las diversas distribuciones

El cuadro 1 resume las principales propiedades que permiten optar por una distribución en lugar de otra. Esas propiedades siguen siendo válidas cuando se trabaja con datos que implican intervalos abiertos.

La utilización de los cocientes instantáneos acumulados o de la función de permanencia conducen a trazar curvas empíricas relativamente bien suavizadas, incluso cuando se dispone de pocas observaciones. Podemos, sin embargo, preferir las gráficas que presentan los cocientes instantáneos mismos y que suministran puntos cuyos errores son independientes, si bien en ese caso las curvas son muy desordenadas y suele ser difícil apreciar su forma en conjunto.

También vimos que en numerosos casos los pocos efectivos observados permiten escoger entre diferentes distribuciones posibles. Es preferible entonces escoger la distribución con el menor número de parámetros, e igualmente aquella que representa la forma explícita más simple para la función de permanencia y los cocientes instantáneos. Como veremos más adelante, esto permite una estimación fácil de los parámetros cuando se dispone de

datos que incluyen intervalos abiertos. Así pues, es mejor escoger una distribución log-logística en lugar de una log-normal cuando las dos se ajustan bien a los datos: como vimos antes, la distribución log-logística presenta formas explícitas simples para la función de permanencia y los cocientes instantáneos.

CUADRO 1
Propiedades que permiten escoger entre diversos modelos

Función observada	*Propiedad*	*Modelo*
$h(t)$	independiente de t	Exponencial
$H(t)$	función lineal de t	Exponencial
$\log h(t)$	función lineal de t	Gompertz
$\log\left(\log\left[\Delta S(t)\right]\right)$	función lineal de t	Gompertz
$\log H(t)$	función lineal de $\log t$	Weibull
$\log\left(-\log S(t)\right)$	función lineal de $\log t$	Weibull
$\log\left[\Delta S(t)\right]$	función lineal de t	Migrante-sedentario
$\log\left(\exp H(t) - 1\right)$	función lineal de $\log t$	Log-logístico
$\frac{1}{h(t)}$	función lineal de t	Pareto

B) MODELOS DE REGRESIÓN

Los modelos paramétricos que acabamos de presentar suponen que se trabaja sobre una población homogénea, cuando todos los individuos tienen la misma probabilidad de experimentar el evento en un instante dado, o bien que se trata sobre una población heterogénea, aunque esta heterogeneidad no sea directamente observada.

En efecto, como solemos disponer de diversas características de los individuos encuestados, se puede pensar que éstas influyen sobre la probabilidad de experimentar el evento. Así, el nivel de instrucción, la profesión, los orígenes sociales de un individuo deben ejercer una influencia sobre su matrimonio, el nacimiento de sus hijos, las migraciones que realiza en su vida, etcétera.

Por lo tanto resulta interesante generalizar los modelos precedentes para tomar en cuenta la heterogeneidad observada entre los individuos de la muestra.

Ya que disponemos ahora de diversas características observadas mediante la encuesta para un individuo dado, podemos representarlas bajo la forma de un vector z:

$$z = (z_1, ..., z_s)$$

Tales características pueden ser cuantitativamente variables (el número de migraciones efectuadas por el individuo durante su infancia, el número de sus hermanos y hermanas, por ejemplo) o variables cualitativas que se representan en forma binaria (0 si el individuo es soltero, 1 si está casado al inicio de la permanencia, por ejemplo).

El problema ahora es modelizar el efecto de las diversas características del individuo sobre su duración de permanencia en el estado inicial.

Una primera posibilidad consiste en dividir la población observada en diversos grupos, a fin de hacer un análisis no paramétrico sobre cada uno de esos grupos y luego comparar su comportamiento. Así, es posible comparar los cocientes de migración después de los 15 años para los individuos que hayan experimentado 0, 1, 2, 3, etc. migraciones durante su infancia. En ese caso se utilizan los métodos presentados en los capítulos precedentes.

Se observa, sin embargo, que cuando el número de los grupos aumenta, rápidamente se cae en subpoblaciones cuyo efectivo será demasiado pequeño para llegar a una conclusión segura. Además, ese método no permite la inclusión de características que pueden cambiar en el curso de la permanencia del individuo en el estado inicial. Por ejemplo, se intentaría probar si el matrimonio de un individuo disminuye su probabilidad de migrar.

De allí el interés por utilizar modelos de regresión capaces de permitir que intervenga simultáneamente el efecto de esas diversas características. Claro está que según sea el tipo de modelo utilizado será preciso incluir diversas hipótesis. Resultará útil tratar de probarlas antes de emplear esos modelos.

1) Modelos de riesgos proporcionales

La hipótesis que sustenta estos modelos es la siguiente: las diversas características individuales actúan multiplicativamente sobre una función de riesgo que es la misma para el conjunto de la población, a todo lo largo del tiempo. De donde resulta que los cocientes instantáneos individuales son todos proporcionales entre sí cualquiera que sea la duración transcurrida. Si $h_0(t)$ representa ese cociente inicial, que puede ser de todas las formas paramétricas

presentadas antes, el cociente instantáneo para un individuo que tenga las características z será de la forma:

$$h(t; z) = h_0(t)\exp(z\beta) \tag{78}$$

con $z\beta = z_1\beta_1 + z_2\beta_2 + \ldots + z_n\beta_n$

donde el vector columna β representa los efectos estimados de las diversas características. Resulta fácil ver que cuando todas las variables z son iguales a cero, caemos nuevamente sobre el modelo de base:

$$h\ (t; 0) = h_0\ (t) \tag{79}$$

Si únicamente la variable z_1 es igual a la unidad cuando todas las demás son iguales a cero se ve que:

$$h(t; z_1) = h_0(t)\exp\beta_1 \tag{80}$$

De eso resulta la relación siguiente:

$$\frac{h(t; z_1)}{h(t; 0)} = \exp\beta_1 \tag{81}$$

que es independiente de la forma del cociente inicial $h_0(t)$.

En efecto, de manera formal para dos individuos n_i y n_j, la relación $h_{n_i}(t)/h_{n_j}(t)$ es una constante que depende de z_{n_i} y z_{n_j} pero es independiente de t. Sin embargo esto deja de ser cierto cuando se introducen variables dependientes del tiempo.

Esta relación de dos densidades condicionales generaliza el concepto epidemiológico de los riesgos múltiples, en competencia o concurrentes, para dos grupos distintos. Si bien el parámetro β_i mide el efecto de la variable z_i sobre el cociente instantáneo, a veces es más simple interpretar $\exp(\beta_i)$ como un riesgo relativo.[1]

No obstante, existe la preocupación por ver que los datos satisfagan esta hipótesis de la proporcionalidad. En efecto, esos modelos son extremadamente generales y poco restrictivos (lo que ha determinado su popularidad) y en el caso en que la hipótesis de proporcionalidad no se respeta por completo, las aproximaciones que se suministran suelen ser a menudo satisfactorias. Si no es así resulta útil fragmentar a la población en diversas subpoblaciones

[1] Si z_1 es dicotómica, divide entonces a los individuos en dos subgrupos, y $\exp(\beta_i)$ mide el riesgo relativo de un individuo respecto del de los del grupo de referencia. De no ser así, se puede decir que con una elevación del valor de z_1 de una unidad, el cociente instantáneo de un individuo es multiplicado por $\exp(\beta_i)$.

para las cuales las características que no verifican la hipótesis de proporcionalidad son diferentes (por sexo, por edad al inicio de la observación).[2]

Violar la hipótesis de la proporcionalidad significa, de hecho, que existe una interacción entre la duración (el tiempo) y una o varias de las variables explicativas (lo cual es inmediato si las variables son dependientes del tiempo).

a) Distribución exponencial

En el caso exponencial ese modelo se escribe:

$$h(t; z) = \rho \exp(z\beta) \tag{82}$$

de donde la densidad:

$$f(t; z) = \rho \exp(z\beta) \exp[-\rho t \exp(z\beta)] \tag{83}$$

y la función de permanencia:

$$S(t; z) = \exp[-\rho t \exp(z\beta)]. \tag{84}$$

b) Distribución de Weibull

En el caso de una distribución de Weibull ese modelo se escribe:

$$h(t; z) = \lambda\rho(\rho t)^{\lambda-1} \exp(z\beta) \tag{85}$$

De ahí resulta la función de permanencia siguiente:

$$S(t; z) = \exp[-(\rho t)^{\lambda} \exp(z\beta)] \tag{86}$$

y la función de densidad de probabilidad:

$$f(t; z) = \lambda\rho(\rho t)^{\lambda-1} \exp\left[z\beta 1 - (\rho t)^{\lambda} \exp(z\beta)\right] \tag{87}$$

[2] El análisis no paramétrico, elaborado con anterioridad sobre los datos, da en general una visión muy exhaustiva del comportamiento de las subpoblaciones que se van a distinguir.

c) Distribución de Gompertz

En el caso de una distribución de Gompertz, ese modelo se escribe:

$$h(t; z) = \lambda\rho \exp(z\beta + \rho t). \tag{88}$$

De ahí resulta la función de permanencia siguiente:

$$S(t; z) = \exp[\lambda(1 - \exp \rho t) \exp z\beta] \tag{89}$$

de donde la función de densidad de probabilidad:

$$f(t; z) = \lambda\rho \exp[z\beta + \rho t + \lambda(1 - \exp \rho t) \exp z\beta]. \tag{90}$$

d) Distribución log-logística

El último ejemplo que daremos es respecto de la distribución log-logística que tiene por cociente instantáneo:

$$h(t; z) = \frac{\lambda\rho(\rho t)^{\lambda-1} \exp z\beta}{1 + (\rho t)^{\lambda}} \tag{91}$$

De ahí resulta la función de permanencia:

$$S(t; z) = [1 + (\rho t)^{\lambda}]^{-\exp z\beta} \tag{92}$$

y la función de densidad de probabilidad:

$$f(t; z) = \frac{\lambda\rho(\rho t)^{\lambda-1} \exp z\beta}{[1 + (\rho t)^{\lambda}]^{1+\exp z\beta}} \tag{93}$$

e) Verificación de la validez del modelo

Cuando se trata de las características cualitativas, cada una de ellas define una submuestra que las posee (por ejemplo, la submuestra de los hombres comparados con las mujeres, la submuestra de los agricultores al inicio de la permanencia, etc.). Cuando se trata de características cuantitativas, de nuevo podemos definir submuestras distintas (por ejemplo, la submuestra de los

individuos que tienen dos hermanos, la submuestra de los que inician su permanencia entre 20 y 24 años, etcétera).

Por lo tanto, para todas las subpoblaciones que tengan suficientes efectivos se pueden estimar los cocientes instantáneos y verificar si la hipótesis de proporcionalidad se sostiene con esos resultados.

Así por ejemplo, si el modelo de Weibull se verifica para el conjunto de la población, se pueden poner en una gráfica los valores de $\log\left(-\log S(t; z_i)\right)$ en función de $\log t$. Si el modelo de riesgos proporcionales se verifica, se obtiene entonces una serie de rectas paralelas entre sí.

En términos más generales, se ve que adecuando siempre los valores de $\log\left(-\log S(t; z_i)\right)$ en función del tiempo transcurrido, se obtiene una serie de curvas paralelas entre sí para las diversas características, cuando se verifica el modelo de riesgos proporcionales.

Esto constituye entonces una posibilidad de probar la validez de tal modelo. En la figura 16 colocamos los valores de $\log\left(-\log S(t)\right)$ en función de $\log t$, al estudiar el hecho de hacerse propietario entre los 30 y 35 años, entre las mujeres de la encuesta "Triple biografía". Una de las dos curvas corresponde a las mujeres que no han recibido ningún diploma y la otra a las que han tenido al menos un certificado de estudios. En ese caso se ve que el modelo de riesgos proporcionales es perfectamente válido, y que una distribución de Weibull se aplica correctamente a los datos entre 30 y 45 años (véase a este respecto la fórmula (54)).

Cuando ciertas características no tienen un efecto multiplicativo sobre los cocientes instantáneos, es posible fragmentar la población en varios estratos de acuerdo con esas características, y escribir para el estrato j, por ejemplo:

$$h_j(t; z) = h_{oj}(t)\exp(z\beta) \tag{94}$$

donde los parámetros del modelo de base pueden variar según el estrato considerado. Incluso es posible concebir modelos de tipo diferente según el estrato.

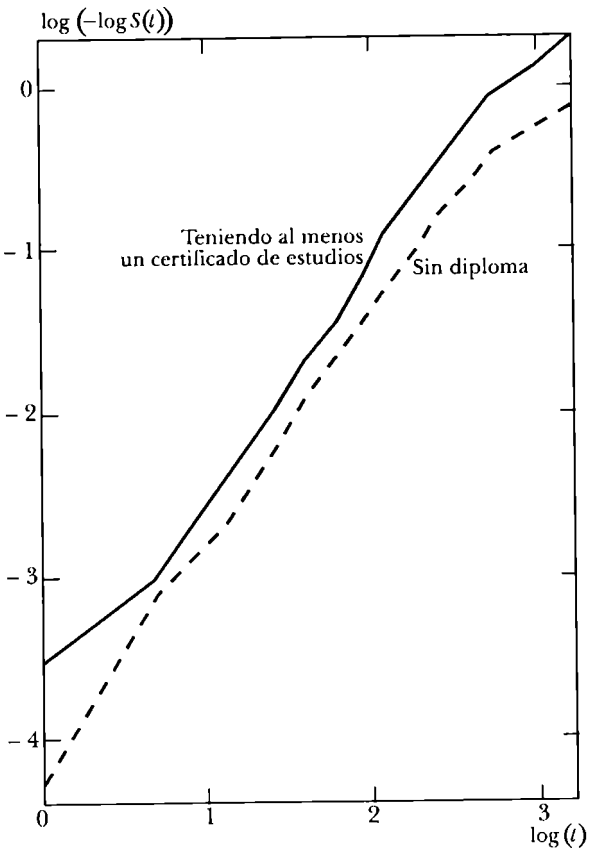
De igual manera, es posible generalizar ese modelo haciendo intervenir una función $\psi(z; \beta)$ y escribiendo:

$$h(t; z) = \psi(z; \beta)h_0(t)\,. \tag{95}$$

Se ve que la función $\psi(z; \beta)$ debe ser igual a la unidad para $z = 0$.
Se puede tomar, en particular:

$$h(t; z) = (1 + z\beta)h_0(t) \tag{96}$$

FIGURA 16
Logaritmo del logaritmo inverso de la función de permanencia en función de t de las mujeres que se hacen propietarias entre los 30 y 55 años, según hayan tenido o no al menos un certificado de estudios



El efecto de las variables es en ese caso aditivo, siendo el cociente instantáneo una función lineal de dichas variables. La utilización de tal modelo puede llevar a estimar cocientes instantáneos negativos si no se impone la condición:

$$1 + z\beta > 0 \text{ para todo } z.$$

Cabe advertir que eso no ocurre cuando se utiliza un modelo como el precedente de tipo multiplicativo, para el cual el cociente instantáneo es siempre positivo cualesquiera que sean los valores de las variables.

2) Modelos de tiempo de ocurrencias aceleradas

Supongamos ahora que el efecto de las características actúa directamente sobre las funciones de permanencia y no sobre los cocientes instantáneos. En el caso multiplicativo, si el individuo tipo —para el cual todas las variables son iguales a cero— tiene una función de permanencia igual a $S_0(t)$, para aquel que tenga las características z, la función de permanencia será:

$$S(t; z) = S_0(t \exp z\beta) \tag{97}$$

donde el vector columna β representa los efectos estimados de las diversas características. En ese caso se ve que la función de densidad de probabilidad se vuelve:

$$f(t; z) = f_0(t \exp z\beta)\exp z\beta \tag{98}$$

y que el cociente instantáneo se escribe:

$$h(t; z) = h_0(t \exp z\beta)\exp z\beta \tag{99}$$

donde $f_0(t)$ y $h_0(t)$ son la función de densidad de probabilidad y el cociente instantáneo del individuo que tiene todas las características iguales a cero.

Entonces es posible escribir ese modelo en términos de variables aleatorias:

$$T = T_0 \exp(-z\beta) \tag{100}$$

donde T_0 es la duración de permanencia de un individuo que tenga todas esas características iguales a cero, y para quien por lo tanto la función de permanencia es $S_0(t)$. En ese caso se ve claramente el efecto multiplicativo sobre esta duración de permanencia, la cual tiene ella misma diversas características. Si se trata de características binarias, por ejemplo, aquellas para las que β_i es positivo acelerarán esta duración de permanencia mientras que las características para las que β_i es negativo la retardarán. Entonces, ese modelo se escribe:

$$\log T = \log T_0 - z\beta \tag{101}$$

a) Nexos entre los modelos de tiempo de ocurrencias aceleradas y los modelos de riesgos proporcionales

Vamos a demostrar que ciertos modelos presentan las dos propiedades precedentes. Éstos deben satisfacer la relación siguiente:

$$h_0^1(t)\exp(z\beta_1) = h_0^2(t\exp z\beta_2)\exp(z\beta_2) \tag{102}$$

donde los índices 1 y 2 representan el modelo de riesgos proporcionales y el modelo tiempo de ocurrencias aceleradas. Dado que esta relación debe cumplirse para todo valor de t y de z, para $z = 0$ se escribe:

$$h_0^1(t) = h_0^2(t) = h_0(t) \tag{103}$$

relación válida para todo valor de t. Sumando en relación a t se obtiene:

$$H_0(t)\exp(z\beta_1) = H_0\left(t\exp(z\beta_2)\right) \tag{104}$$

Se ve con facilidad que la única función $H_0(t)$ que permite verificar esta relación, es la función:

$$H_0(t) = (\rho t)^{\lambda} \tag{105}$$

y que el vector β_1 debe ser proporcional al vector β_2, siendo λ el coeficiente de proporcionalidad. En efecto, se escribe la relación (105) precedente, tomando su logaritmo:

$$\lambda\log(\rho t) + z\beta_1 = \lambda\log(\rho t\exp z\beta_2) = \lambda\log(\rho t) + \lambda z\beta_2 \tag{106}$$

Se ve entonces que la distribución de Weibull o la distribución exponencial, cuando $\lambda = 1$, son las únicas distribuciones que satisfacen a la vez el modelo de riesgos proporcionales y el de tiempo de ocurrencias aceleradas.

b) Distribución log-logística

Un caso interesante por considerar dentro de los modelos de tiempo de ocurrencias aceleradas es el que corresponde a la distribución log-logística. Si:

$$S_0(t) = (1 + (\rho t)^{\lambda})^{-1} \tag{107}$$

De ahí resulta:

$$S(t; z) = (1 + [\rho t\exp z\beta]^{\lambda})^{-1} \tag{108}$$

de donde:

$$f(t; z) = \frac{\lambda\rho[\rho t\exp z\beta]^{\lambda-1}\exp z\beta}{(1 + [\rho t\exp z\beta]^{\lambda})^2} \tag{109}$$

y:

$$h(t;z)=\frac{\lambda\rho[\rho t\exp z\beta]^{\lambda-1}\exp z\beta}{1+[\rho t\exp z\beta]^{\lambda}} \tag{110}$$

Se ve que en ese caso el modelo es diferente del modelo log-logístico de riesgos proporcionales.

El cociente instantáneo se escribe de manera más simple al introducir un primer parámetro β_1 verificando:

$$\exp(\beta_1)=\rho^{\lambda}$$

y remplazando $\lambda\beta$ por β:

$$h(t;z)=\frac{\lambda}{t[1+t^{-\lambda}\exp(-z\beta)]} \tag{111}$$

c) Verificación de la validez del modelo

Al igual que se hizo anteriormente, es posible dividir la muestra total en subpoblaciones que cuenten con un efectivo suficiente y correspondan a las diversas características consideradas. En ese caso la fórmula (101) no muestra que las distribuciones de log T para diversos valores de z se deduzcan la una de la otra por traslación. De ahí resulta, por ejemplo, que la varianza de log T será independiente de la característica que define a la subpoblación. En ese caso basta con calcular la varianza para las diversas subpoblaciones y verificar que sea aproximadamente constante. Para lograrlo es, sin embargo, necesario que no haya intervalos abiertos.

En el caso general, donde hay intervalos abiertos, es necesario considerar por separado cada tipo de distribución para ver qué pruebas se pueden realizar.

Así, por ejemplo, cuando la distribución es log-logística se ve que podemos calcular:

$$\log\left(\frac{1}{S(t)}-1\right)=\lambda\log\rho t+\lambda z\beta \tag{112}$$

En ese caso, si el modelo de tiempo es de ocurrencias aceleradas, $\log(\frac{1}{S(t)}-1)$ será una función lineal de $\log t$, y para los diversos valores de las características cada una de esas rectas tendrá siempre la misma pendiente, λ.

3) Modelos más complejos

El análisis de los diversos parámetros que intervienen en un modelo puede conducir a que éstos dependan en parte, o incluso totalmente, de diversas características.

Aquí presentamos sucintamente el modelo generalizado de Gompertz-Makeham que se puede estimar utilizando el programa Rate (*cf.* anexo 1.VI). Ese modelo se puede escribir:

$$h(t) = \lambda_1 \exp(z_1\beta_1) + \lambda\rho \exp[z_2\beta_2 + \rho(1 + z_3\beta_3)t] \qquad (124)$$

Vemos entonces que el cociente instantáneo depende de tres series de variables, z_1, z_2 y z_3, que pueden ser diferentes unas de otras, pero también contener variables comunes. Ese programa también permite remplazar $\exp(z\beta)$ por $(1 + z\beta)$ y viceversa. Una vez más el modelo lineal implica el riesgo de conducir a cocientes instantáneos estimados negativos para ciertos valores de las características.

C) Conclusión

Este capítulo nos ha permitido mostrar la extrema variedad de funciones paramétricas que podemos escoger para representar diversos tipos de distribución.

En numerosos casos prácticos resultará difícil escoger entre varios tipos de distribuciones cercanas entre sí. Habrá que seleccionar entonces aquella cuyos cocientes instantáneos y la función de permanencia tengan una presentación funcional simple. En el próximo capítulo, que trata sobre los métodos de estimación de los parámetros de esas funciones, veremos con más claridad las razones de esta selección.

El presente capítulo nos ha permitido además introducir diversas características de los individuos observados y modelizar su efecto sobre los cocientes instantáneos, la función de permanencia o la función de densidad de probabilidad. Una vez más será preferible utilizar modelos bastante simples, con efectos aditivos o multiplicativos, para estar en capacidad de estimar los parámetros que intervienen en esos modelos. Sin embargo, antes de aplicar un tipo de modelo dado será necesario verificar que éste se adecue correctamente a las características estudiadas.

VIII. MÉTODOS DE ESTIMACIÓN DE MODELOS PARAMÉTRICOS

En el capítulo anterior presentamos diversos tipos de modelos paramétricos y mostramos los análisis preliminares que permiten escoger un modelo que se adapte a los datos. Para hacerlo resulta particularmente útil dividir la población en subpoblaciones en donde están presentes o ausentes diversas características.

Supongamos ahora que ya hemos optado por uno de esos tipos de modelos. Debemos estimar a continuación los diversos parámetros de ese modelo, así como los coeficientes que se van a aplicar a sus características para representar lo mejor posible el conjunto de los datos. Asimismo se requiere una estimación de la varianza de esos parámetros o coeficientes estimados y de sus covarianzas a fin de efectuar cierto número de pruebas. Por ejemplo, en un modelo multiplicativo será útil ver si el efecto de una característica es significativamente diferente de la unidad, y así observar si esta característica influye o no sobre el fenómeno estudiado. También será útil comparar los efectos de varias características y, en caso de que éstos sean idénticos, se les remplazará por una característica única. Por ejemplo, si el hecho de estar casado, viudo o divorciado influye de manera idéntica sobre la probabilidad de migrar, esas tres características se podrán remplazar por una sola: el hecho de no ser soltero.

Los datos de las encuestas presentan generalmente intervalos abiertos y resulta absolutamente indispensable tomarlos en cuenta en los métodos de estimación. Ya mencionamos los diversos tipos de truncamiento que se encuentran en las encuestas. Por tanto, tratamos primero sobre la manera de calcular la verosimilitud de una observación según contenga un intervalo cerrado o abierto.

A continuación presentaremos los métodos generales de estimación de los diversos parámetros o coeficientes, cuando se dispone de muestras de suficiente tamaño. Esos métodos los desarrollaremos en diversos casos particulares que ilustraremos mediante ejemplos precisos.

Finalmente regresaremos sobre los problemas derivados de la posibilidad de elegir entre diversos modelos paramétricos, lo cual nos conducirá al capítulo siguiente, que trata sobre los modelos semiparamétricos.

A) Cálculo de la verosimilitud cuando ciertos intervalos están abiertos

Ya presentamos el método del máximo de verosimilitud cuando se trata de estimar modelos no paramétricos (capítulo II). Ahora lo abordaremos en el marco de los modelos paramétricos.

La encuesta observa a n individuos y recoge informaciones sobre los eventos acaecidos en el curso de un cierto intervalo de tiempo.

Situémonos primero en el caso en que este intervalo contenga la fecha de entrada de la población sometida a riesgo, pero pueda no contener la fecha en que el individuo experimentó el evento estudiado. En esta última eventualidad se observan *intervalos abiertos a la derecha,* que sin embargo no están desprovistos de información, pues se sabe que antes de cierta fecha el individuo no había experimentado el evento. Representemos el conjunto de las informaciones recogidas bajo la forma de una tríada:

$$(t_i^0, \delta_i, z_i) \quad \text{donde } i = 1, ..., n$$

donde t_i^0 es la duración de observación, ya sea hasta que se produce el evento si éste es observado $(\delta_i = 1)$, ya sea hasta la fecha en el que el individuo se sale de la observación antes de que se produzca el evento $(\delta_i = 0)$; z_i es el vector de las características del individuo al inicio de la permanencia. Vemos, en consecuencia, que δ_i es una variable binaria que indica la ocurrencia o no del evento antes de la salida de la observación.

Esta salida de la observación puede producirse de diversas maneras.

En el caso general, la fecha de salida de la observación es absolutamente independiente de que el individuo haya experimentado o no el evento. Tenemos entonces *salidas de la observación aleatorias* respecto del evento que se estudia, de ahí que podamos introducir una variable aleatoria, T^s, cuya función de permanencia es $O_i(t)$ y la densidad de probabilidad es $q_i(t)$. Esta variable es independiente de la fecha de ocurrencia del evento T, cuya función de permanencia es $S(t; z_i, \beta)$, y la densidad de probabilidad $f(t; z_i, \beta)$, donde β representa el conjunto de los parámetros y coeficientes por estimar. La variable que se observa efectivamente se escribe:

$$T^0 = \min\left(T^s, T\right) \tag{1}$$

Con estas notaciones podemos calcular la probabilidad siguiente:

$$P(t \leq T_i^0 < t + dt, \delta_i = 1; z_i, \beta) = O_i(t) f(t; z_i, \beta) dt \tag{2}$$

En efecto, en ese caso el individuo ha experimentado el evento en el instante t, observado $(\delta_i = 1)$ y, por lo tanto, no ha salido de la observación

en esta fecha. Dado que los dos eventos son independientes, hay que calcular el producto de la probabilidad de que el individuo no haya salido de la observación en el instante t, por $(O_i(t))$ veces la probabilidad de que él haya experimentado el evento en el mismo instante $(f(t; z_i, \beta))$.

De manera similar, para los individuos encuestados que aún no han experimentado el evento se puede escribir:

$$P(t \leq T_i^0 < t + dt, \delta_i = 0; z_i, \beta) = q_i(t)\, S(t; z_i, \beta) dt \qquad (3)$$

Como ni $O_i(t)$, ni $q_i(t)$ proporcionan información sobre los parámetros y coeficientes por estimar, β, la verosimilitud de los datos se puede considerar entonces como proporcional a la cantidad:

$$L(\beta) = \prod_{i=1}^{n} f(t_i; z_i, \beta)^{\delta i} S(t_i; z_i, \beta)^{1-\delta i} = \prod_{i=1}^{n} h(t_i; z_i, \beta)^{\delta i}\, S(t_i; z_i, \beta) \qquad (4)$$

donde $h(t; z, \beta)$ es el cociente instantáneo de ocurrencia del evento.

Se advierte que en ese caso basta con disponer de la distribución de los cocientes instantáneos y de las funciones de permanencia para que sea factible estimar mediante el método del máximo de verosimilitud los valores de los diversos parámetros.

Esta estimación se facilita mucho cuando esas dos distribuciones tienen una forma explícita simple. En cambio, una forma más compleja de una de las dos implica la ejecución de cálculos muy pesados, que rápidamente se pueden volver irrealizables.

Esta verosimilitud sigue siendo válida en otras situaciones donde las salidas de observación ya no son aleatorias. En particular, basta con que tales salidas de observación se produzcan a cada instante de manera independiente de los riesgos a los que están sometidos los individuos, para que dicha verosimilitud sea siempre aplicable. En ese caso, las reglas de salida de la observación pueden depender de la historia pasada del fenómeno estudiado, pero no deben eliminar a individuos expuestos a un riesgo elevado, o por el contrario reducido, de experimentar el evento.

Se habla entonces de un sistema de *salidas de observación independientes.*[1] Nos encontramos en ese caso cuando se detiene la observación luego de que se ha producido el $r^{\text{ésimo}}$ evento, por ejemplo, o cuando se saca de observación a una fracción dada de los individuos sometidos a riesgo luego de la ocurrencia de eventos de determinado rango.

[1] Para una demostración de la validez de la verosimilitud en ese caso, véase Kalbfleisch y Prentice, 1980, pp. 119-122.

B) Estimación de los parámetros y pruebas de su valor

Al reescribir la verosimilitud de las observaciones expuestas en la parte precedente bajo forma logarítmica, obtenemos:

$$\log L(\beta) = \sum_{i=1}^{n} [\delta_i \log h(t_i; z_i, \beta) + \log S(t_i; z_i, \beta)] \tag{5}$$

El método consistirá en dar a los parámetros β los valores que maximizan esta verosimilitud. Para hacerlo se puede trabajar sobre la primera derivada de log $L(\beta)$ respecto de β. Las soluciones de la ecuación que igualan esta derivada a cero proporcionan esos valores.

Cuando el número de las observaciones es suficientemente elevado se demuestra que, bajo condiciones simples generalmente alcanzables ($L(\beta)$ debe ser tres veces diferenciable y ciertas condiciones límite sobre esta tercera derivada deben ser verificadas),[2] esta ecuación tiene una solución. La estimación de los parámetros así obtenida tiene una media asintótica igual a cero con una varianza mínima. La distribución asintótica de este estimador obedece a una ley normal para tantas variables como parámetros se pretenda estimar. Dicha ley tiene como media el verdadero valor de β. Si se calcula la matriz inversa de la segunda derivada de log $L(\beta)$, aún llamada matriz de información de Fisher, se demuestra que el inverso de esta matriz constituye una estimación de la matriz de las varianzas y covarianzas de los parámetros β (*cf.* anexo I.III y capítulo III.B.1).

Ahora es posible efectuar diversas pruebas sobre los parámetros β estimados. Así por ejemplo, para probar si un parámetro β_i se puede considerar como significativamente diferente de cero, se calcula la cantidad $(\hat{\beta}_i^2 / \operatorname{var} \hat{\beta}_i)$. Si β_i no es significativamente diferente de cero este estadístico es una χ^2 con un grado de libertad. De igual manera se puede probar, en términos más generales, si los parámetros β estimados son diferentes de los valores β_0 que se habían elegido *a priori*. Si la matriz de las varianzas y covarianzas se designa como $V(\beta)$, se puede escribir una vez más el estadístico:

$$(\hat{\beta} - \beta_0)^T V(\hat{\beta})^{-1} (\hat{\beta} - \beta_0) \tag{6}$$

que cuando los parámetros β no son significativamente diferentes de β_0, es una χ^2 con tantos grados de libertad como parámetros considerados haya.

Es igualmente posible utilizar directamente la verosimilitud calculando la relación:

$$R(\beta_0) = \frac{L(\beta_0)}{L(\hat{\beta})} \tag{7}$$

[2] Véase a ese respecto Cox y Hinkley, 1974, pp. 281 y ss.

Si deseamos probar la hipótesis de que los parámetros β estimados no son diferentes de los valores dados β_0, entonces la distribución asintótica de $(-2\log R(\beta_0))$ es una χ^2 con tantos grados de libertad como parámetros considerados haya. Este método se generaliza sin problema cuando se quiere probar el valor de un número cualquiera de los parámetros.

Por último, es posible utilizar la primera derivada de la verosimilitud:

$$U(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} \tag{8}$$

Cuando $\beta = \beta_0$ el estadístico $U(\beta_0)$ es asintóticamente normal, de media 0 y de varianza $V(\beta_0)$. Bajo esas condiciones el estadístico:

$$U'(\beta_0)\, V(\beta_0)^{-1}\, U(\beta_0) \tag{9}$$

tiene una distribución asintótica de χ^2 con tantos grados de libertad como parámetros considerados haya. Entonces se puede verificar si el valor β_0 de los parámetros propuestos se puede considerar como un valor satisfactorio, teniendo en cuenta las observaciones.

No avanzaremos más en esta teoría de la estimación y remitimos al lector a la obra de Cox y Hinkley (1974) para conocer más precisiones sobre el método. A continuación presentaremos la estimación en casos precisos.

C) Ejemplos de estimación de los parámetros

A fin de mostrar con mayor claridad las operaciones que habrán de llevarse a cabo para efectuar esta estimación, vamos a considerar diferentes modelos y dar ejemplos prácticos de estimación.

1) Modelo exponencial

El caso más simple es el del modelo exponencial, sin la intervención de variables explicativas. Supongamos que se observan n individuos. Sea d el número de eventos producidos en las fechas t_i, $i = 1, \ldots d$. Cada individuo no experimenta más que un solo evento. En ese caso se observan $(n - d)$ intervalos abiertos cuyos finales de observación se sitúan en las fechas t_j, $j = (d + 1), \ldots n$. Si el cociente instantáneo es igual a ρ, que es el único parámetro por estimar, entonces la aplicación de la fórmula (5) conduce al siguiente logaritmo de la verosimilitud:

$$\log L(\rho) = \sum_{i=1}^{d} \log \rho - \sum_{i=1}^{n} \rho t_i = d \log \rho - \rho \sum_{i=1}^{n} t_i \tag{10}$$

En ese caso se puede decir que:

$$\sum_{i=1}^{n} t_i$$

es el tiempo total transcurrido para los individuos sometidos a riesgo. Entonces se puede escribir:

$$U(\rho) = \frac{\partial \log L(\rho)}{\partial \rho} = \frac{d}{\rho} - \sum_{i=1}^{n} t_i \tag{11}$$

y

$$V(\rho)^{-1} = -\frac{\partial^2 \log L(\rho)}{\partial \rho^2} = \frac{d}{\rho^2} \tag{12}$$

El estimador del máximo de verosimilitud $\hat{\rho}$ de ρ es la solución de la ecuación:

$$U(\rho) = \frac{d}{\rho} - \sum_{i=1}^{n} t_i = 0 \tag{13}$$

lo que da:

$$\hat{\rho} = \frac{d}{\sum_{i=1}^{n} t_i} \tag{14}$$

Se puede entonces decir que se trata de la relación del número de eventos producidos en el tiempo total transcurrido para los individuos sometidos a riesgo.

La relación (12) nos da la varianza estimada de ρ:

$$V(\hat{\rho}) = \frac{d}{\left(\sum_{i=1}^{n} t_i\right)^2} \tag{15}$$

Entonces se puede construir el siguiente intervalo de confianza de 95% para ρ:

$$\frac{d - 1.96\sqrt{d}}{\sum_{i=1}^{n} t_i} < \rho < \frac{d + 1.96\sqrt{d}}{\sum_{i=1}^{n} t_i} \tag{16}$$

Asimismo se puede utilizar el logaritmo de la relación de verosimilitud:

$$-2 \log R(\rho) = 2\left(\rho \sum_{i=1}^{n} t_i - d \log \rho + d\left[\log \frac{d}{\sum_{i=1}^{n} t_i} - 1\right]\right) \tag{17}$$

que tiene una distribución asintótica de χ^2 con un grado de libertad. Se dispone así de un intervalo de confianza de 95%, por ejemplo, para los valores de ρ para los cuales esta función tiene un valor inferior a 3.84.

Por último, se puede utilizar la primera derivada de la verosimilitud, que conduce al estadístico:

$$\left(\frac{d}{\rho}-\sum_{i=1}^{n} t_i\right)^2 \frac{\rho^2}{d} = d + \frac{\rho^2}{d}\left(\sum_{i=1}^{n} t_i\right)^2 - 2\rho \sum_{i=1}^{n} t_i \qquad (18)$$

que tiene una distribución asintótica de χ^2 con un grado de libertad. De nuevo disponemos de un intervalo de confianza de 95% cuando esta función tiene un valor inferior a 3.84. Se deduce que la raíz cuadrada de este estadístico:

$$\left(\frac{d}{\rho}-\sum_{i=1}^{n} t_i\right)\frac{\rho}{\sqrt{d}} = \sqrt{d} - \frac{\rho}{\sqrt{d}} \sum_{i=1}^{n} t_i \qquad (19)$$

tiene una distribución normal que conduce a un intervalo de confianza de 95%, para ρ, que es idéntico al dado por la fórmula (16). Los dos métodos teóricamente diferentes conducen sin embargo a un procedimiento idéntico.

Apliquemos esos diversos métodos a la probabilidad de hacerse propietarias después del nacimiento del último hijo de las mujeres cuya pareja es un obrero especializado. Tales datos provienen de la encuesta "Triple biografía". Entre esas 380 mujeres, 137 se hicieron propietarias antes de la encuesta, lo que nos da: $d = 137$. Esas 380 mujeres han pasado 8 307 años como no propietarias de su vivienda, lo que nos da

$$\sum_{i=1}^{n} t_i = 8\,307$$

A partir de esas cifras tenemos el estimador del parámetro ρ, para el método del máximo de verosimilitud:

$$\hat{\rho} = \frac{137}{8\,307} = 0.0165$$

La varianza y la desviación estándar de este estimador son:

$$V(\hat{\rho}) = \frac{137}{(8\,307)^2} = 1.9853 \times 10^{-6} \quad \text{y} \quad \sigma(\hat{\rho}) = 1.409 \times 10^{-3}$$

De esto resulta el siguiente intervalo de confianza de 95% para ρ, cuando $\sqrt{d} = 11.705$:

$$\frac{137 - 1.96 \times 11.705}{8\,307} < \rho < \frac{137 + 1.96 \times 11.705}{8\,307} \qquad \text{sea} \quad 0.0137 < \rho < 0.0192.$$

El máximo del logaritmo de la verosimilitud es igual a:

$$\log L(\hat{\rho}) = -699.3346$$

Se puede entonces trazar la figura 1, que presenta el logaritmo de esta verosimilitud alrededor de ρ, cuya ecuación es:

$$-2\log R(\rho) = 2(8\,307\rho - 137\log\rho - 699.3346)$$

El intervalo de confianza de 95% se obtiene trazando una paralela al eje de las abscisas, en el punto 3.841 de la ordenada, valor para el cual sólo hay cinco posibilidades sobre 100 de que el valor de ρ esté fuera de este intervalo. De ahí resulta el intervalo de confianza siguiente:

$$0.0139 < \rho < 0.0194$$

Se observa que este intervalo es ligeramente diferente del anterior, debido a que esta curva no es simétrica, pero los resultados obtenidos por los diversos métodos son totalmente coherentes entre sí.

Introduzcamos ahora el efecto de diversas características dadas bajo la forma de un vector. Es posible escribir el cociente instantáneo para el individuo i:

$$h(t; z_i) = \exp(z_i\,\beta) \tag{20}$$

A cada individuo le corresponde el vector $z_i = (z_{i1}, \ldots, z_{is})$ cuyo primer elemento es igual a la unidad de manera que $\rho = \exp(\beta_1)$, esto es, el cociente instantáneo correspondiente a los individuos para quienes las otras características son iguales a cero. Se puede entonces escribir:

$$\log L(\beta) = \sum_{i=1}^{n} z_i\,\beta - \sum_{i=1}^{n} t_i \exp(z_i\,\beta) \tag{21}$$

De esto resulta para la j^e característica:

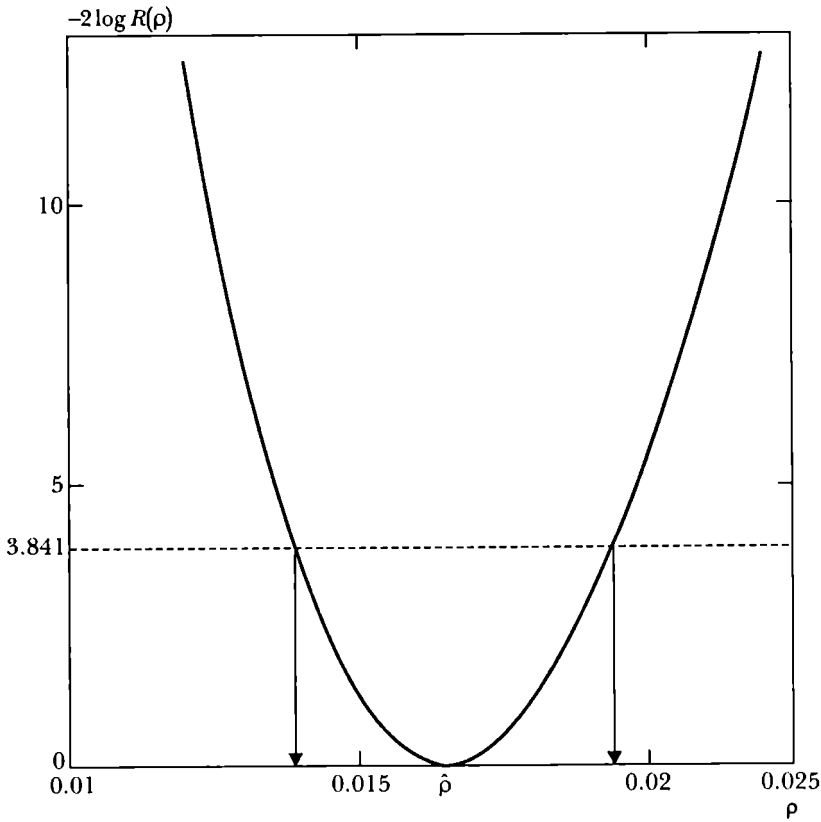
$$U_j(\beta) = \frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{d} z_{ij} - \sum_{i=1}^{n} z_{ij}\, t_i \exp(z_i\,\beta) \tag{22}$$

donde z_{ij} es la j^e característica del individuo i.

De igual manera, la matriz inversa de las segundas derivadas tiene por elemento de la línea j^e y de la columna k^e:

$$-\frac{\partial^2 \log L(\beta)}{\partial \beta_j\, \partial \beta_k} = \sum_{i=1}^{n} z_{ij}\, z_{ik}\, t_i \exp(z_i\beta) \tag{23}$$

FIGURA 1
Logaritmo de la verosimilitud en función del parámetro ρ, y cálculo del intervalo de confianza alrededor de la estimación $\hat{\rho}$, para la probabilidad que tienen los obreros especializados de hacerse propietarios después del nacimiento del último hijo



El método del máximo de verosimilitud conduce a tomar la solución del sistema de ecuaciones:

$$\sum_{i=1}^{d} z_{ij} - \sum_{i=1}^{n} z_{ij}\, t_i \exp(z_i\, \beta) = 0 \quad j = 1, \ldots n \tag{24}$$

En el apartado siguiente veremos un método de resolución de este sistema en el caso general. Situémonos aquí en el caso simple donde no interviene más que una variable binaria. El vector $z_i = (1, z_{i2})$ y el logaritmo de la verosimilitud se escribe:

$$\log L(\beta) = \sum_{i=1}^{d}(\beta_1 + z_{i2}\,\beta_2) - \sum_{i=1}^{n} t_i \exp(\beta_1 + z_{i2}\,\beta_2) \tag{25}$$

De donde resultan las dos derivadas:

$$U_1(\beta) = \frac{\partial \log L(\beta)}{\partial \beta_1} = d - \sum_{i=1}^{n} t_i \exp(\beta_1 + z_{i2}\,\beta_2) = 0 \tag{26}$$

$$U_2(\beta) = \frac{\partial \log L(\beta)}{\partial \beta_2} = \sum_{i=1}^{d} z_{i2} - \sum_{i=1}^{n} t_i\, z_{i2} \exp(\beta_1 + z_{i2}\,\beta_2) = 0$$

Este sistema se puede escribir de manera más simple introduciendo el número de eventos observados cuando $z_{i2} = 0$, (d_0); el número de eventos observados cuando $z_{i2} = 1$, (d_1), y N_0 y N_1 que son los tiempos totales transcurridos para los individuos sometidos al riesgo en las dos categorías:

$$U_1(\beta) = d_0 + d_1 - N_0 \exp\beta_1 - N_1 \exp(\beta_1 + \beta_2) = 0 \tag{27}$$

$$U_2(\beta) = d_1 - N_1 \exp(\beta_1 + \beta_2) = 0$$

De donde resulta:

$$\hat{\beta}_1 = \log \frac{d_0}{N_0}$$

$$\hat{\beta}_2 = \log \frac{d_1 N_0}{d_0 N_1} \tag{28}$$

En ese caso, la matriz inversa de las segundas derivadas es igual a:

$$\begin{pmatrix} N_0 \exp\beta_1 + N_1 \exp(\beta_1 + \beta_2) & N_1 \exp(\beta_1 + \beta_2) \\ N_1 \exp(\beta_1 + \beta_2) & N_1 \exp(\beta_1 + \beta_2) \end{pmatrix} = \begin{pmatrix} d_0 + d_1 & d_1 \\ d_1 & d_1 \end{pmatrix} \tag{29}$$

Lo que da la matriz de las varianzas y covarianzas que es independiente de N_1 y N_2:

$$\frac{1}{d_0 d_1} \begin{pmatrix} d_1 & -d_1 \\ -d_1 & d_0 + d_1 \end{pmatrix} \tag{30}$$

Así, si la varianza de β_1 es de d_0^{-1}, la de β_2 es $(d_0^{-1} + d_1^{-1})$, la covarianza de esos dos parámetros es negativa e igual a $-d_0^{-1}$. Se puede entonces probar un valor *a priori* de los dos parámetros con la ayuda de los diversos métodos que presentamos anteriormente.

Estos resultados se generalizan fácilmente al caso en el que la variable z_2 toma valores enteros iguales a 0, 1, 2, etcétera.

Para ilustrar ese caso, retomemos la probabilidad de hacerse propietarias después del nacimiento del último hijo de las mujeres cuya pareja es un obrero especializado. Ahora haremos que intervenga su nivel educativo para ver si tiene influencia sobre esa probabilidad. Una primera solución consiste en desagregar la muestra en submuestras de niveles educativos dados y estimar, al igual que antes, su parámetro ρ. Hemos hecho diferenciaciones entre los individuos considerando que sus parejas no tengan ningún diploma (0), tengan la constancia de estudios primarios (1) y el certificado de capacidad profesional o el título de estudios primarios (2). De los obreros especializados sólo tres contaban con un grado educativo superior y los dejamos aparte.

Se estimó el logaritmo de ese parámetro para las tres subpoblaciones:

$$\log \hat{\rho}_0 = -4.4257$$

$$\log \hat{\rho}_1 = -3.8863$$

$$\log \hat{\rho}_2 = -3.3779$$

Se ve entonces que el pasar de la primera población a la segunda se debe añadir + 0.5394 al logaritmo de $\hat{\rho}_0$ y que al pasar de la segunda a la tercera hay que añadir + 0.5084 al logaritmo de $\hat{\rho}_1$. Entonces es posible utilizar una variable sintética, tomando los valores 0, 1, 2 según los diplomas que tenga el individuo. Si d_0, d_1, d_2 son los efectivos de mujeres de obreros especializados con los diplomas 0, 1, 2 que se hicieron propietarias antes de la encuesta, y en el caso de que el conjunto de esas mujeres haya pasado N_0, N_1, N_2 años como no propietarias de su vivienda, se demuestra, como anteriormente que las derivadas del logaritmo de la verosimilitud respecto de β_1 y β_2 son:

$$\begin{cases} U_1(\beta) = d_0 + d_1 + d_2 - N_0 \exp\beta_1 + N_1 \exp(\beta_1 + \beta_2) - N_2 \exp(\beta_1 + 2\beta_2) = 0 \\ U_2(\beta) = d_1 + 2d_2 - N_1 \exp(\beta_1 + \beta_2) - 2N_2 \exp(\beta_1 + 2\beta_2) = 0. \end{cases} \tag{31}$$

En el caso que tratamos aquí:

$$d_0 = 54, d_1 = 70, d_2 = 13$$

$$N = 4\,513, N_1 = 3\,412, N_2 = 381.$$

El sistema de las dos ecuaciones (31) se resuelve entonces fácilmente y conduce a las estimaciones:

$$\hat{\beta}_1 = -4.422$$

$$\hat{\beta}_2 = 0.5299.$$

La matriz opuesta de las segundas derivadas se escribe en ese caso:

$$\begin{pmatrix} N_0 \exp\beta_1 + N_1 \exp(\beta_1+\beta_2) + N_2 \exp(\beta_1+2\beta_2) & N_1 \exp(\beta_1+\beta_2) + 2N_2 \exp(\beta_1+2\beta_2) \\ N_1 \exp(\beta_1+\beta_2) + 2N_2 \exp(\beta_1+2\beta_2) & N_1 \exp(\beta_1+\beta_2) + 4N_2 \exp(\beta_1+\beta_2) \end{pmatrix} \tag{32}$$

que es igual a:

$$\begin{pmatrix} d_0 + d_1 + d_2 & d_1 + 2d_2 \\ d_1 + 2d_2 & d_1 + 2d_2 + 2N_2 \exp(\beta_1+2\beta_2) \end{pmatrix} \tag{33}$$

Lo que da la matriz de las varianzas y covarianzas, que esta vez depende de los valores de N:

$$\frac{1}{(d_0+d_1+d_2)\left[d_1+2d_2+2N_2\exp(\beta_1+2\beta_2)\right]-(d_1+2d_2)^2} \times \begin{pmatrix} d_1+2d_2+2N_2\exp(\beta_1+2\beta_2) - (d_1+2d_2) & \\ -(d_1+2d_2) & d_0+d_1+d_2 \end{pmatrix} \tag{34}$$

En nuestro ejemplo, esta matriz se escribe:

$$\begin{pmatrix} 0.0162 & -0.0127 \\ -0.0127 & 0.0181 \end{pmatrix}$$

Se puede entonces ver, por ejemplo, si el efecto del nivel educativo desempeña un papel significativo sobre la probabilidad de hacerse propietario, calculando el estadístico de χ^2, con un grado de libertad

$$\frac{\hat{\beta}_2^2}{V(\hat{\beta}_2)} = \frac{0.281}{0.0181} = 15.52$$

Este efecto es totalmente significativo: mientras mayor sea el nivel educativo mayor será la probabilidad de hacerse propietario.

2) Utilización del método de Newton-Raphson

Ya vimos que para el caso de un modelo exponencial, la resolución de los sistemas de ecuaciones $U_j(\beta) = 0$ se vuelve muy pesada cuando el número de parámetros por estimar es superior a dos.

Resulta útil entonces un acercamiento mediante iteraciones sucesivas, incluso si en ciertos casos esto pudiera conducir a soluciones incorrectas, como lo mostraremos más adelante.

El método de Newton-Raphson es el que se utiliza más comúnmente en este caso. Consideremos la verosimilitud de las observaciones, $L(\beta)$, que va a depender de parámetros β. Esos parámetros son estimados, al igual que aquellos que maximizan esta verosimilitud y que por tanto anulan su derivada, también llamados estadísticos de *score*.[3]

El método de Newton-Raphson se basa pues en el desarrollo limitado al primer orden de las series de Taylor de:

$$U(\beta) = \frac{d \log L(\beta)}{d\beta}$$

Dado un valor β_0, el desarrollo limitado al primer orden es el siguiente:

$$U(\hat{\beta}) = U(\beta_0) - I(\beta^*)(\hat{\beta} - \beta_0) \qquad (35)$$

donde β^* es un valor comprendido entre $\hat{\beta}$ y β_0 e $I(\beta) = -\dfrac{d^2 \log L(\beta)}{d\beta^2}$, matriz de información de Fisher (*cf.* capítulo III.B.1) cuya inversa es el estimador de la matriz de las varianzas y covarianzas para las coordenadas del vector $\hat{\beta}$.

Si β_0 está cerca de $\hat{\beta}$ entonces $I(\beta^*) \simeq I(\beta_0)$ y sabiendo que $U(\hat{\beta}) = 0$, la fórmula (35) se vuelve:

$$\hat{\beta} = \beta_0 + I(\beta_0)^{-1} U(\beta_0) \qquad (36)$$

El miembro de la derecha da el nuevo valor para β y el proceso se reitera hasta que la estimación de β, $\hat{\beta}$ converge hacia una solución aceptable de $U(\hat{\beta}) = 0$.

Este método da por lo general resultados correctos cuando la verosimilitud es unimodal, e incluso si el valor inicial β_0 está lejos del valor que se quiere estimar $\hat{\beta}$. En cambio, si esta distribución es multimodal, el método puede conducir a un máximo relativo que no es el máximo real, y dar resultados completamente incorrectos. Para probar si la solución a la que llegamos es correcta, es útil partir de diversos valores iniciales β_0 y verificar

[3] Véase el anexo I.III.

si se obtiene siempre la misma estimación $\hat{\beta}$. De no ser así, hay que tomar aquella cuya verosimilitud sea máxima.

3) Modelo de Weibull

Consideremos primeramente el modelo de Weibull sin la intervención de variables explicativas.

El logaritmo de la verosimilitud —siempre con d eventos observados sobre n individuos para los que el término de su observación (evento o salida de la muestra) se produce en las fechas t_i— se escribe:

$$\log L(\lambda, \rho) = \sum_{i=1}^{d}\left[\log\lambda + \log\rho + (\lambda - 1)\log\rho t_i\right] - \sum_{i=1}^{n}(\rho t_i)^{\lambda} \tag{37}$$

Las derivadas respecto de los dos parámetros de esta función se escriben

$$\begin{cases} \dfrac{\partial \log L(\lambda, \rho)}{\partial\lambda} = \dfrac{d}{\lambda} + d\log\rho + \sum\limits_{i=1}^{d}\log t_i - \rho^{\lambda}\sum\limits_{i=1}^{n} t_i^{\lambda}\log(\rho t_i) = 0 \\ \dfrac{\partial \log L(\lambda, \rho)}{\partial\lambda} = \dfrac{\lambda d}{\rho} - \lambda\rho^{\lambda-1}\sum\limits_{i=1}^{n} t_i^{\lambda} = 0 \end{cases} \tag{38}$$

La primera ecuación nos da ρ en función de λ:

$$\rho = \left(\frac{d}{\sum\limits_{i=1}^{n} t_i^{\lambda}}\right)^{\frac{1}{\lambda}} \tag{39}$$

Al colocar este valor de ρ en la segunda ecuación, se obtiene:

$$f(\lambda) = \frac{d}{\lambda} + \sum_{i=1}^{d}\log t_i - d\frac{\sum\limits_{i=1}^{n} t_i^{\lambda}\log(t_i)}{\sum\limits_{i=1}^{n} t_i^{\lambda}} = 0 \tag{40}$$

Esta ecuación puede resolverse mediante aproximaciones sucesivas utilizando su derivada respecto de λ:

$$f'(\lambda) = -\frac{d}{\lambda^2} - d\frac{\left[\sum\limits_{i=1}^{n} t_i^{\lambda}(\log t_i)^2\right]\left[\sum\limits_{i=1}^{n} t_i^{\lambda}\right] - \left[\sum\limits_{i=1}^{n} t_i^{\lambda}\log t_i\right]^2}{\left[\sum\limits_{i=1}^{n} t_i^{\lambda}\right]^2} \tag{41}$$

Partiendo de un valor λ_0 se estima un nuevo valor λ_1 corregido de la relación $f(\lambda_0)/f'(\lambda_0)$. Se continúa hasta la obtención de un valor $f(\lambda_0)$ lo suficientemente cercano a cero, en el umbral que se habrá fijado con antelación. Así se obtiene $\hat{\lambda}$ que llevado a la fórmula (39) da igualmente $\hat{\rho}$.

Veamos ahora las segundas derivadas de $\log L(\lambda,\rho)$, que denominaremos l, bajo forma simplificada

$$\begin{cases} -\dfrac{\partial^2 l}{\partial\lambda^2} = \dfrac{d}{\lambda^2} + \rho^{\lambda} \sum_{i=1}^{n} t_i^{\lambda} [\log(\rho t_i)]^2 \\ -\dfrac{\partial^2 l}{\partial\rho^2} = \dfrac{\lambda d}{\rho^2} + \lambda(\lambda-1)\rho^{\lambda-2} \sum_{i=1}^{n} t_i^{\lambda} \\ -\dfrac{\partial^2 l}{\partial\lambda\partial\rho} = -\dfrac{d}{\rho} + \rho^{\lambda-1}(1+\lambda\log\rho) \sum_{i=1}^{n} t_i^{\lambda} + \lambda\rho^{\lambda-1} \sum_{i=1}^{n} t_i^{\lambda} \log t_i. \end{cases} \tag{42}$$

Al disponerse de $\hat{\lambda}$ y de $\hat{\rho}$ se puede calcular la matriz de las varianzas y covarianzas de esos parámetros con la ayuda de esas tres fórmulas y probar diversos valores supuestos para esos parámetros.

Apliquemos ahora ese modelo a la probabilidad de hacerse propietarias luego del nacimiento del último hijo de las mujeres cuya pareja es un obrero.

La estimación de los parámetros mediante el método indicado aquí conduce a:

$$\hat{\lambda} = 1.0200$$
$$\hat{\rho} = 0.0168$$

El parámetro $\hat{\rho}$ estimado mediante este modelo está muy cerca del parámetro $\hat{\rho}$ estimado con el modelo exponencial ($\hat{\rho} = 0.0165$) y el valor del parámetro $\hat{\lambda}$ está cerca de la unidad. Estos resultados son coherentes con el hecho de que el modelo de Weibull se hace idéntico al modelo exponencial cuando $\lambda = 1$.

Entonces resulta útil probar si se puede considerar $\hat{\lambda}$ como diferente de la unidad o no. Un primer método consiste en utilizar las varianzas estimadas mediante las ecuaciones (42):

$$V(\hat{\lambda}) = 6.29840 \times 10^{-3}$$
$$V(\hat{\rho}) = 3.22399 \times 10^{-6}$$
$$\operatorname{cov}(\hat{\lambda}, \hat{\rho}) = 8.86654 \times 10^{-5}$$

La variable $(\hat{\lambda} - 1)$ es, si se verifica la hipótesis de igualdad, una ley normal de media igual a cero y de desviación estándar $\sigma(\hat{\lambda}) = 7.936 \times 10^{-2}$

Vemos que el valor 0.02 verifica perfectamente la hipótesis de igualdad,y que el modelo exponencial se adapta bien a esos datos.

Otro método consiste en calcular $f(1)$, primera derivada del logaritmo de la verosimilitud para $\lambda = 1$ y:

$$\hat{\rho}_1 = \frac{d}{\sum_{i=1}^{n} t_i} = 0.0165 :$$

$$f(1) = d + \sum_{i=1}^{d} \log t_i - d \frac{\sum_{i=1}^{n} t_i \log t_i}{\sum_{i=1}^{n} t_i} = 3.2326 \tag{43}$$

La matriz opuesta de las segundas derivadas en el punto $(1, \hat{\rho}_1)$ se expresa simplemente bajo la forma:

$$\begin{cases} -\dfrac{\partial^2 l}{\partial \lambda^2} = d + \hat{\rho}_1 \sum_{i=1}^{n} t_i [\log (\hat{\rho}_1 t_i)]^2 \\ -\dfrac{\partial^2 l}{\partial \rho^2} = \dfrac{d}{\hat{\rho}_1^2} \\ -\dfrac{\partial^2 l}{\partial \lambda \, \partial \rho} = \sum_{i=1}^{n} t_i \log (\hat{\rho}_1 t_i). \end{cases} \tag{44}$$

Podemos, por lo tanto, estimar los términos de la matriz inversa, que se escribe en nuestro caso:

$$\begin{pmatrix} 6.0522 \times 10^{-3} & 8.791 \times 10^{-5} \\ 8.791 \times 10^{-5} & 3.258 \times 10^{-6} \end{pmatrix}$$

Si la hipótesis se verifica, la cantidad:

$$3.2326 \times \sqrt{6.0522 \cdot 10^{-3}} = 0.252$$

se puede considerar correctamente como extraída del proceso de una ley normal.

Ahora introduciremos las características z de diversos individuos. El modelo de Weibull, en el caso de un modelo multiplicativo, se escribe:

$$h(t, z) = \lambda \rho (\rho t)^{\lambda - 1} \exp z\beta \tag{45}$$

El logaritmo de la verosimilitud es entonces:

$$\log L(\lambda, \rho, \beta) = \sum_{i=1}^{d} [\log\lambda + \log\rho + (\lambda - 1)\log\rho t_i + z_i\beta] + \sum_{i=1}^{n} -(\rho t_i)^{\lambda} \exp z_i\, \beta \tag{46}$$

Los parámetros de λ, ρ y β se estiman resolviendo el sistema de ecuaciones siguiente:

$$\begin{cases} \dfrac{\partial \log L}{\partial \lambda} = \dfrac{d}{\lambda} + d\log\rho + \sum_{i=1}^{d} \log t_i - \rho^{\lambda} \sum_{i=1}^{n} t_i^{\lambda} \log(\rho t_i) \exp z_i\, \beta = 0 \\ \dfrac{\partial \log L}{\partial \rho} = \dfrac{\lambda d}{\rho} - \lambda\rho^{\lambda-1} \sum_{i=1}^{n} t_i^{\lambda} \exp z_i\, \beta = 0 \\ \dfrac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{d} z_{ij} - \sum_{i=1}^{n} z_{ij} (\rho t_i)^{\lambda} \exp z_i\, \beta = 0 \end{cases} \tag{47}$$

En este caso vemos que no se puede llegar a una estimación $\hat{\lambda}$, $\hat{\rho}$ y $\hat{\beta}$ más que utilizando el método Newton-Raphson que necesita el cálculo de las segundas derivadas del logaritmo de la verosimilitud. En consecuencia, hay que calcular la matriz cuyos términos son los siguientes:

$$\begin{cases} -\dfrac{\partial^2 \log L}{\partial \lambda^2} = \dfrac{d}{\lambda^2} + \sum_{i=1}^{n} (\rho t_i)^{\lambda} [\log(\rho t_i)]^2 \exp z_i\, \beta \\ -\dfrac{\partial^2 \log L}{\partial \rho^2} = \dfrac{\lambda d}{\rho^2} + \lambda(\lambda - 1)\rho^{\lambda-2} \sum_{i=1}^{n} t_i^{\lambda} \exp z_i\, \beta \\ -\dfrac{\partial^2 \log L}{\partial \beta_j^2} = \sum_{i=1}^{n} z_{ij}^2 (\rho t_i)^{\lambda} \exp z_i \beta \\ -\dfrac{\partial^2 \log L}{\partial \lambda\, \partial \rho} = -\dfrac{d}{\rho} + \dfrac{1}{\rho} \sum_{i=1}^{n} [1 + \lambda \log(\rho t_i)](\rho t_i)^{\lambda} \exp z_i\, \beta \end{cases} \tag{48}$$

$$
\begin{cases}
-\dfrac{\partial^2 \log L}{\partial \lambda \, \partial \beta_j} = \displaystyle\sum_{i=1}^{n} z_{ij} (\rho t_i)^{\lambda} \log(\rho t_i) \exp z_i \, \beta \\[2ex]
-\dfrac{\partial^2 \log L}{\partial \rho \, \partial \beta_j} = \dfrac{\lambda}{\rho} \displaystyle\sum_{i=1}^{n} z_{ij} (\rho t_i)^{\lambda} \exp z_i \, \beta \\[2ex]
-\dfrac{\partial^2 \log L}{\partial \beta_j \, \partial \beta_k} = \displaystyle\sum_{i=1}^{n} z_{ij} \, z_{ik} (\rho t_i)^{\lambda} \exp z_i \, \beta
\end{cases}
$$

Considerando nuevamente el caso que ya presentamos en el capítulo VII B.1.e, aquí estudiamos el hecho de hacerse propietarias entre los 30 y 45 años, para las mujeres de la encuesta "Triple biografía", que carecen de un diploma o tienen al menos una constancia de estudios. Esta estimación conduce a los parámetros siguientes:

$$\hat{\lambda} = 1.4816$$
$$\hat{\rho} = 0.0475$$
$$\hat{\beta}_1 = 0.352$$

El parámetro $\hat{\beta}_1$ positivo revela claramente que hay mayores probabilidades de hacerse propietaria a partir de los 30 años cuando se tiene un diploma. Para ver si su efecto se puede considerar como significativamente diferente de cero, hay que utilizar la matriz de las varianzas y covarianzas que estimamos simultáneamente con esos parámetros. En ese caso, la matriz se escribe:

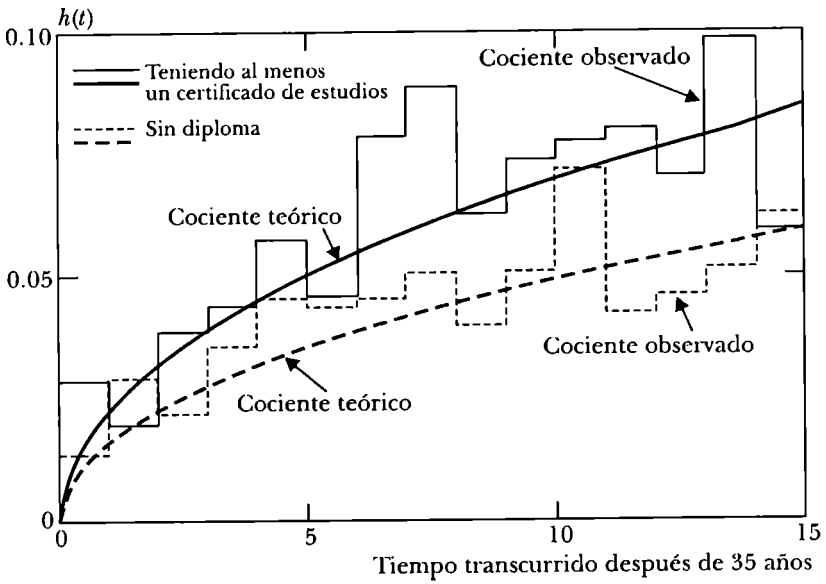
$$10^{-6} \begin{pmatrix} 1812.60 & 23.07 & 99.90 \\ 23.07 & 2.55 & -69.27 \\ 99.90 & -69.27 & 4019.74 \end{pmatrix}$$

Se puede calcular entonces un estadístico de χ^2:

$$\frac{\hat{\beta}_1^2}{V(\hat{\beta}_1)} = \frac{0.124}{4019.74 \times 10^{-6}} = 30.865$$

que muestra claramente que el efecto del nivel educativo es totalmente significativo para la probabilidad de hacerse propietaria después de los 30 años.

FIGURA 2
Cocientes instantáneos estimados de manera no paramétrica o con un modelo de Weibull, respecto al hecho de que las mujeres se hagan propietarias entre los 30 y 45 años, según carezcan de un diploma o tengan al menos un certificado de estudios



La figura 2 presenta los cocientes instantáneos observados y estimados mediante el modelo de Weibull cuando se tiene o no un diploma.

4) Modelo de Gompertz

Presentamos de nuevo la estimación de los parámetros de un modelo de Gompertz antes de hacer que intervengan diversas características individuales. En ese caso, el logaritmo de la verosimilitud se escribe:

$$\text{Log}L(\lambda, \rho) = \sum_{i=1}^{d} (\log\lambda + \log\rho + \rho t_i) + \sum_{i=1}^{n} \lambda[1 - \exp(\rho t_i)] \qquad (49)$$

Las primeras derivadas respecto de los parámetros, se escriben simplemente:

$$\begin{cases} \dfrac{\partial \log L}{\partial \lambda} = \dfrac{d}{\lambda} + \sum_{i=1}^{n} (1 - \exp \rho t_i) = \dfrac{d}{\lambda} + n - \sum_{i=1}^{n} \exp \rho t_i = 0 \end{cases} \qquad (50)$$

$$\left\{\frac{\partial \log L}{\partial \rho}=\frac{d}{\rho}+\sum_{i=1}^{d} t_i-\lambda \sum_{i=1}^{n} t_i \exp \rho t_i=0.\right. \tag{51}$$

Este sistema de ecuaciones se resuelve fácilmente por sustitución. Expresemos λ en función de ρ de la primera ecuación:

$$\lambda=\frac{d}{\sum_{i=1}^{n} \exp\ \rho t_i-1} \tag{52}$$

Este valor llevado a la segunda, nos da una ecuación donde no figura más que ρ:

$$f(\rho)=\frac{d}{\rho}+\rho \sum_{i=1}^{d} t_i-d \frac{\sum_{i=1}^{n} t_i \exp\ \rho t_i}{\sum_{i=1}^{n} \exp\ \rho t_i-1}=0. \tag{53}$$

Esta ecuación se puede resolver mediante aproximaciones sucesivas utilizando su derivada respecto de ρ:

$$f'(\rho)=-\frac{d}{\rho^2}+\sum_{i=1}^{d} t_i-d \frac{\left(\sum_{i=1}^{n} t_i^2 \exp \rho t_i\right)\left(\sum_{i=1}^{n} \exp \rho t_i-1\right)-\left(\sum_{i=1}^{n} t_i \exp \rho t_i\right)^2}{\left(\sum_{i=1}^{n} \exp \rho t_i-1\right)^2}. \tag{54}$$

El método de estimación es idéntico al presentado en el caso del modelo de Weibull y conduce a las estimaciones $\hat{\rho}$ y $\hat{\lambda}$.

Veamos ahora las segundas derivadas del inverso del logaritmo de la verosimilitud:

$$\left\{\begin{aligned} -\frac{\partial^2 \log L}{\partial \lambda^2}&=\frac{d}{\lambda^2} && (55)\\ -\frac{\partial^2 \log L}{\partial \rho^2}&=\frac{d}{\rho^2}+\lambda \sum_{i=1}^{n} t_i^2 \exp\ \rho t_i && (56)\\ -\frac{\partial^2 \log L}{\partial \lambda\, \partial \rho}&=\sum_{i=1}^{n} t_i \exp\ \rho t_i. && (57)\end{aligned}\right.$$

Al disponer de las estimaciones de $\hat{\lambda}$ y de $\hat{\rho}$ es posible estimar la matriz de las varianzas y covarianzas de estos parámetros y probar diversos valores esperados para ellos. La matriz de las varianzas y covarianzas se puede escribir:

$$\frac{\hat{\lambda}^2\hat{\rho}^2}{d^2 + d\,\hat{\lambda}\,\hat{\rho}^2 \sum_{i=1}^{n} t_i^2 \exp\rho t_i - \hat{\lambda}^2\hat{\rho}^2\left[\sum_{i=1}^{n} t_i \exp\rho t_i\right]^2} \times \begin{pmatrix} \dfrac{d}{\hat{\rho}^2} + \lambda \sum_{i=1}^{n} t_i^2 \exp\rho t_i - \sum_{i=1}^{n} t_i \exp\rho t_i & \\ -\sum_{t=1}^{n} t_i \exp\rho t_i & \dfrac{d}{\hat{\lambda}^2} \end{pmatrix} \tag{58}$$

Introduzcamos ahora el efecto de diversas características. En el modelo de riesgos proporcionales teníamos:

$$h(t; z) = \lambda\,\rho \exp(\rho t + z\beta). \tag{59}$$

El logaritmo de la verosimilitud se escribe entonces:

$$\log L(\lambda, \rho, \beta) = \sum_{i=1}^{d}(\log\lambda + \log\rho + \rho\, t_i + z_i\beta) + \sum_{i=1}^{n} \lambda[1 - \exp\rho\, t_i]\exp z_i\beta \tag{60}$$

Para simplificar, las derivadas respecto de los parámetros, se escriben utilizando la siguiente notación $\log L(\lambda, \rho, \beta) = \ell$:

$$\begin{cases} \dfrac{\partial l}{\partial \lambda} = \dfrac{d}{\lambda} + \sum_{i=1}^{n}(1 - \exp\rho\, t_i)\exp z_i\,\beta = 0 \\ \dfrac{\partial l}{\partial \rho} = \dfrac{d}{\rho} + \sum_{i=1}^{d} t_i - \lambda \sum_{i=1}^{n} t_i(\exp\rho\, t_i)\exp z_i\,\beta = 0 \\ \dfrac{\partial l}{\partial \beta_j} = \sum_{i=1}^{d} z_{ij} + \lambda \sum_{i=1}^{n} z_{ij}(1 - \exp\rho\, t_i)\exp z_i\,\beta = 0 \end{cases} \tag{61}$$

Este sistema se resuelve utilizando el método de Newton-Raphson que necesita el cálculo del inverso de las segundas derivadas:

$$
\begin{cases}
-\dfrac{\partial^2 l}{\partial \lambda^2} = \dfrac{d}{\lambda^2} \\[2ex]
-\dfrac{\partial^2 l}{\partial \rho^2} = \dfrac{d}{\rho^2} + \lambda \sum_{i=1}^{n} t_i^2 (\exp \rho t_i) \exp z_i \beta \\[2ex]
-\dfrac{\partial^2 l}{\partial \beta_j^2} = -\lambda \sum_{i=1}^{n} z_{ij}^2 (1 - \exp \rho t_i) \exp z_i \beta \\[2ex]
-\dfrac{\partial^2 l}{\partial \lambda \partial \rho} = \sum_{i=1}^{n} t_i (\exp \rho t_i) \exp z_i \beta \\[2ex]
-\dfrac{\partial^2 l}{\partial \lambda \partial \beta_j} = -\sum_{i=1}^{n} z_{ij} (1 - \exp \rho t_i) \exp z_i \beta \\[2ex]
-\dfrac{\partial^2 l}{\partial \rho \partial \beta_j} = \lambda \sum_{i=1}^{n} t_i \, z_{ij} (\exp \rho t_i) \exp z_i \beta \\[2ex]
-\dfrac{\partial^2 l}{\partial \beta_j \, \partial \beta_k} = -\lambda \sum_{i=1}^{n} z_{ij} \, z_{ik} (1 - \exp \rho t_i) \exp z_i \beta
\end{cases}
\tag{62}
$$

El programa Rate elaborado por N. Tuma y D. Pasta permite la estimación de todos esos parámetros. En ese caso, el modelo se escribe de manera ligeramente distinta

$$h(t; z) = \exp(\rho t + z\beta) \tag{63}$$

El primer vector z_1 siempre tiene, de hecho, el valor 1, lo que hace que:

$$\exp(\beta_1) = \lambda\rho \tag{64}$$

Todos los otros parámetros son idénticos a los del modelo aquí presentado.

A manera de ejemplo daremos los resultados obtenidos cuando se estudia la permanencia en una residencia de los hombres nacidos entre 1931 y 1935. Esos datos fueron también extraídos de la encuesta "Triple bio-

grafía".[4] La característica que aquí se considera es la edad del individuo al inicio de la permanencia, que se hace intervenir bajo la forma de una serie de variables binarias: menos de 20 años, de 20 a 24 años, de 25 a 29 años, de 30 a 34 años, de 35 a 39 años, y de 40 a 44 años. Para que la estimación sea posible no se debe incluir el grupo de 45 años y más, ya que éste se define cuando todas las variables binarias precedentes son iguales a cero.

El número de las duraciones de permanencia observado es de 2 523, del cual 493 seguían transcurriendo en el momento de la encuesta.

Como término de comparación se toma el modelo exponencial sin hacer intervenir ninguna variable. Un modelo como éste conduce a un cociente instantáneo constante estimado en 0.1237 teniendo como valor máximo del logaritmo de la verosimilitud –6273.15.

Cuando se hace que intervengan los diversos grupos de edad en un modelo de Gompertz, se obtiene un nuevo máximo del logaritmo de la verosimilitud igual a –5991.67. Al utilizar el logaritmo de la relación de verosimilitud del segundo modelo comparado con el primero, se llega a $-2 \log R = 562.96$, que tiene una distribución χ^2 con siete grados de libertad, si el segundo modelo no aporta nada respecto del primero. Como la probabilidad de alcanzar este valor es muy débil, ese segundo modelo resulta por lo tanto mucho más satisfactorio que el primero.

La estimación de los diversos parámetros β y ρ se presenta en el cuadro 1, con sus desviaciones estándar, los valores del cuadrado de la relación de los parámetros con su desviación estándar (prueba de nulidad de su valor), que son los de una χ^2 con un grado de libertad si la variable z_i no tiene ningún efecto sobre la probabilidad de migrar, y finalmente los valores exp β_i. La interpretación de esos valores es simple en el caso de las variables binarias: ellos indican el aumento o la disminución del cociente cuando el individuo tiene esta característica.

El parámetro $\hat{\rho}$ es igual a –0.0629, con una desviación estándar de 4.485×10^{-3}, lo que conduce a una prueba χ^2 con un grado de libertad igual a 196.784. Esto revela un fuerte efecto de la duración de permanencia sobre la probabilidad de migrar, de manera que luego de una duración de 10 años, esta probabilidad se reduce a casi la mitad:

$$\exp(-0.629) = 0.525.$$

En el umbral de 5%, el efecto de la edad es altamente significativo antes de los 35 años. Más allá, la probabilidad de migrar se puede considerar como constante e igual a 0.0654, valor igual a la mitad del cociente calculado sobre el conjunto de las migraciones. Este efecto de la edad llega al máximo

[4] Para más detalles sobre este análisis véase D. Courgeau (1985a y 1985b).

CUADRO 1
Estimación de los parámetros β de la desviación estándar, prueba de χ^2 con un grado de libertad sobre la nulidad de su valor y estimación de exp (β_i)

Características consideradas	$\hat{\beta}_i$	$\sigma(\hat{\beta}_i)$	$\left[\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right]^2$	$exp\,\beta_i$
Constante	− 2.727	0.2880	89.187	0.0654
Menos de 20 años	1.131	0.2920	15.008	3.100
20-24 años	1.410	0.2912	23.460	4.097
25-29 años	0.892	0.2940	9.203	2.440
30-34 años	0.626	0.2978	4.417	1.870
35-39 años	0.135	0.3064	0.194	1.145
40-44 años	− 0.177	0.3292	0.127	0.889

entre los 20 y 24 años, cuando la probabilidad de migrar es más de cuatro veces superior a la de los individuos con edades de 35 años y más.

En el cuadro 2 se presenta la matriz de las varianzas y covarianzas de los parámetros β y ρ (última línea). Esas varianzas y covarianzas se pueden usar para llevar a cabo diversas pruebas. Por ejemplo, si queremos probar si la probabilidad de migrar es diferente a los 35-39 años y a los 40-44 años, se puede construir el estadístico $\beta'_1 C_1^{-1} \beta_1$ donde $\beta'_1 = (0.135 - 0.177)$, β_1 es el vector columna que le corresponde y C_1 es la siguiente matriz de las varianzas y covarianzas:

$$C_1 = \begin{pmatrix} 9.3911 \times 10^{-2} & 8.3419 \times 10^{-2} \\ 8.3419 \times 10^{-2} & 1.0838 \times 10^{-1} \end{pmatrix}$$

Si los dos parámetros no se pueden considerar como diferentes, este estadístico es una χ^2 con dos grados de libertad. En este caso encontramos 0.289, lo que muestra que podemos reagrupar esas dos clases de edades en una sola.

5) Modelo log-logístico de ocurrencias aceleradas

Daremos este último ejemplo que corresponde al caso en el que el cociente instantáneo puede pasar por un valor máximo antes de decrecer.

Partimos directamente del modelo que hace intervenir diversas características de los individuos.

CUADRO 2
Matriz de las varianzas y covarianzas estimadas
de los diversos parámetros del modelo de Gompertz

	Constante β_1	*Menos de 20 años*	*De 20 a 24 años*	*De 25 a 29 años*	*De 30 a 34 años*	*De 35 a 39 años*	*De 40 a 44 años*	*Constante* ρ
Constante β_1	8.3387×10^{-2}							
Menos de 20 años	-8.3240×10^{-2}	8.5282×10^{-2}						
De 20 a 24 años	-8.3242×10^{-2}	8.3491×10^{-2}	8.4786×10^{-2}					
De 25 a 29 años	-8.3205×10^{-2}	8.3556×10^{-2}	8.3550×10^{-2}	8.6439×10^{-2}				
De 30 a 34 años	-8.3219×10^{-2}	8.3533×10^{-2}	8.3527×10^{-2}	8.3606×10^{-2}	8.8706×10^{-2}			
De 35 a 39 años	-8.3240×10^{-2}	8.3495×10^{-2}	8.3490×10^{-2}	8.3555×10^{-2}	8.3531×10^{-2}	9.3911×10^{-2}		
De 40 a 44 años	-8.3284×10^{-2}	8.3420×10^{-2}	8.3417×10^{-2}	8.3452×10^{-2}	8.3439×10^{-2}	8.3419×10^{-2}	1.0838×10^{-1}	
Constante ρ	-3.2884×10^{-5}	-5.7207×10^{-5}	-5.5561×10^{-5}	-7.8316×10^{-5}	-7.0116×10^{-5}	-5.6831×10^{-5}	-3.0415×10^{-5}	2.0112×10^{-5}

En el capítulo anterior mostramos (111) que el cociente instantáneo se escribe:

$$h(t;\lambda,z)=\frac{\lambda}{t\left[1+t^{-\lambda}\exp(-z\beta)\right]} \tag{65}$$

donde el parámetro ρ se ha introducido mediante una primera variable z_1 igual a la unidad para todos los individuos encuestados, lo que significa que:

$$\exp(\beta_1)=\rho^{\lambda}. \tag{66}$$

En ese caso, la función de permanencia se escribe:

$$S(t;\lambda,z)=[1+t^{\lambda}\exp z\beta]^{-1}. \tag{67}$$

El logaritmo de la verosimilitud se vuelve entonces:

$$\begin{aligned}\log L(\lambda,z)=&\sum_{i=1}^{d}\left[\log\lambda-\log t_i-\log\left(1+t_i^{-\lambda}\exp(-z_i\,\beta)\right)\right]\\&-\sum_{i=1}^{n}\log\left[1+t_i^{\lambda}\exp z_i\,\beta\right].\end{aligned} \tag{68}$$

Las derivadas respecto de los parámetros se escriben:

$$\begin{cases}\dfrac{\partial l}{\partial\lambda}=\dfrac{d}{\lambda}+\displaystyle\sum_{i=1}^{d}\frac{\log t_i}{1+t_i^{\lambda}\exp z_i\,\beta}-\sum_{i=1}^{n}\frac{\log t_i}{1+t_i^{-\lambda}\exp(-z_i\,\beta)}=0\\[2ex]\dfrac{\partial l}{\partial\beta_j}=\displaystyle\sum_{i=1}^{d}\frac{z_{ij}}{1+t_i^{\lambda}\exp z_i\,\beta}-\sum_{i=1}^{n}\frac{z_{ij}}{1+t_i^{-\lambda}\exp(-z_i\,\beta)}=0\end{cases} \tag{69}$$

De nuevo para resolver este sistema se puede utilizar el método de Newton-Raphson.

El cálculo del inverso de las segundas derivadas en relación con los parámetros da:

$$\begin{cases}-\dfrac{\partial^2 l}{\partial\lambda^2}=\dfrac{d}{\lambda^2}+\displaystyle\sum_{i=1}^{n}\frac{2^{\delta_i}(\log t_i)^2}{\left(1+t_i^{-\lambda}\exp(-z_i\,\beta)\right)\left(1+t_i^{\lambda}\exp z_i\,\beta\right)}\\[2ex]-\dfrac{\partial^2 l}{\partial\beta_j^2}=\displaystyle\sum_{i=1}^{n}\frac{2^{\delta_i}z_{ij}^2}{\left(1+t_i^{-\lambda}\exp(-z_i\,\beta)\right)\left(1+t_i^{\lambda}\exp z_i\,\beta\right)}\end{cases} \tag{70}$$

$$\begin{cases} -\dfrac{\partial^2 l}{\partial\lambda\,\partial\beta_j} = \displaystyle\sum_{i=1}^{n} \dfrac{2^{\delta_i} z_{ij} \log t_i}{\left(1+t_i^{-\lambda}\exp(-z_i\,\beta)\right)\left(1+t_i^{\lambda}\exp z_i\,\beta\right)} \\ -\dfrac{\partial^2 l}{\partial\beta_j\,\partial\beta_k} = \displaystyle\sum_{i=1}^{n} \dfrac{2^{\delta_i} z_{ij}\, z_{ik}}{\left(1+t_i^{-\lambda}\exp(-z_i\,\beta)\right)\left(1+t_i^{\lambda}\exp z_i\,\beta\right)} \end{cases}$$

donde $\delta_i = 1$ cuando el evento se ha producido antes de la salida de la observación del $i^{\text{ésimo}}$ individuo y $\delta_i = 0$ en el caso contrario. La matriz inversa de información nos permite estimar $\hat{\lambda}$ y los valores de $\hat{\beta}_0$. Para esos valores de los parámetros proporciona igualmente una estimación de la matriz, de sus varianzas y covarianzas.

Apliquemos este método al estudio de el hacerse propietaria, en función de la edad de las mujeres, a partir de los 15 años. Al principio no hacemos intervenir ninguna otra característica. Partiendo de un valor $\lambda^* = 2$ y $\beta_1^* = -6$, se obtiene el máximo del logaritmo de la verosimilitud al cabo de cinco iteraciones, con un valor de derivadas inferior a 10^{-3}. Los valores obtenidos son:

$$\hat{\lambda} = 1.922$$

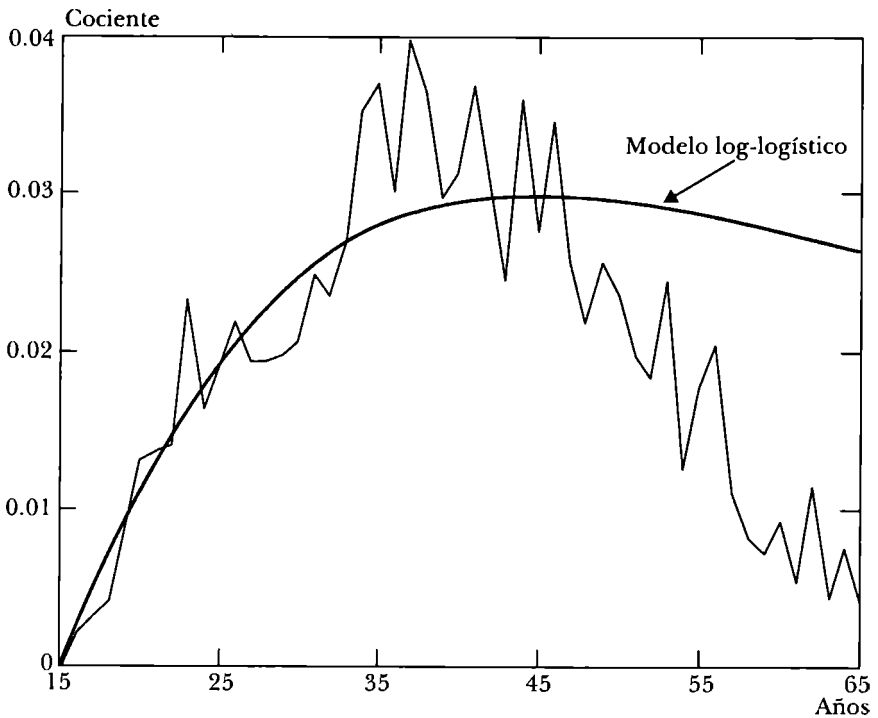
$$\hat{\beta}_1 = 6.556 \text{ o sea } \rho = 0.0341$$

El logaritmo de la verosimilitud, que era igual a –7 463.34 para los valores iniciales es ahora igual a –7 146.79. La matriz de las varianzas y covarianzas se estima en

$$10^{-4}\begin{pmatrix} 17.47 & -57.24 \\ -57.24 & 200.99 \end{pmatrix}$$

La figura 3 presenta los cocientes instantáneos calculados por el método no paramétrico, y los obtenidos con el modelo log-logístico. Se puede ver que si entre los 15 y 35 años el modelo se ajusta bien a las observaciones, entre 35 y 45 años los subestima y más tarde, por el contrario, los sobreestima. Igualmente hemos colocado en la figura 4 las funciones de permanencia en el estado de no propietaria, calculadas de manera no paramétrica y utilizando el modelo log-logístico. Para la estimación no paramétrica se ve en qué medida el hecho de trabajar sobre la función de permanencia borra las diferencias aleatorias que aparecen con los cocientes instantáneos. Observemos, sin em-

FIGURA 3
Cocientes instantáneos correspondientes al hecho de que las mujeres se vuelvan propietarias, a partir de los 15 años, estimados de manera no paramétrica o con un modelo log-logístico



bargo, que la función de permanencia no paramétrica está por debajo de la del modelo log-logístico a partir de los 35 años, pero la supera de nuevo después de los 55 años.

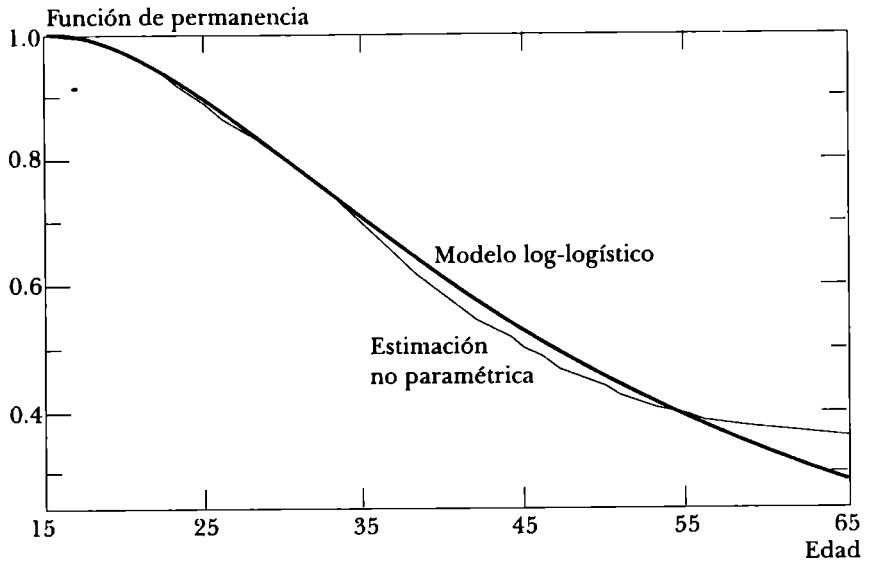
Hagamos ahora intervenir el nivel educativo de las encuestadas bajo la forma de una variable binaria z_2 igual a 0 cuando la mujer no tiene ningún diploma, e igual a 1 cuando ha obtenido al menos una constancia de estudios. La aplicación del método de Newton-Raphson conduce a la siguiente estimación de los parámetros:

$$\hat{\lambda} = 1.928$$

$$\hat{\beta}_1 = -6.732, \text{ o sea } \rho = 0.0019$$

$$\hat{\beta}_2 = 0.256$$

FIGURA 4
Función de permanencia para el hecho de que las mujeres se vuelvan propietarias a partir de los 15 años, calculada de manera no paramétrica o con un modelo log-logístico



El parámetro $\hat{\beta}_2$ positivo muestra de nuevo que hay mayor probabilidad de hacerse propietario a partir de los 15 años cuando se está diplomado. Para ver si su efecto se puede considerar como significativamente diferente de cero hay que utilizar la matriz de las varianzas y covarianzas, que se ha estimado simultáneamente con esos parámetros. En ese caso, la matriz se escribe:

$$10^{-4}\begin{pmatrix} 17.58 & -58.60 & 1.57 \\ -58.60 & 230.75 & -40.69 \\ 1.57 & -40.69 & 57.22 \end{pmatrix}$$

Se puede así calcular un estadístico de χ^2:

$$\frac{\hat{\beta}_2^2}{V(\hat{\beta}_2)} = \frac{0.0655}{57.22 \times 10^{-4}} = 11.45$$

que muestra que el efecto del nivel educativo es totalmente significativo sobre la probabilidad de hacerse propietario en todas las edades, tanto como lo es para las mujeres de más de 30 años. Eso confirma las diferencias que habíamos mostrado en la figura 16 del capítulo VII.

D) Comparación de modelos paramétricos

En primer lugar, es posible introducir en esos modelos un número muy grande de variables. Como los datos utilizados generalmente provienen de encuestas biográficas, disponemos de numerosos elementos de la vida de los encuestados que vamos a ligar al fenómeno estudiado. Es importante, por lo tanto, que entre los diversos modelos que hacen intervenir a esas variables se escoja a aquel que permita la mejor explicación del fenómeno.

Retomemos, a manera de ejemplo, los resultados obtenidos cuando se estudian las duraciones de permanencia en una residencia de los hombres nacidos entre 1931 y 1935. Más arriba habíamos presentado los resultados obtenidos al hacer variar la edad del individuo al inicio de la permanencia y su duración de permanencia. De hecho la encuesta "Triple biografía" proporciona numerosos ejemplos más sobre la biografía y las características de los encuestados. En ese caso, es importante hacerlos intervenir y ver si la calidad del modelo mejora. Así, las variables sobre la etapa del ciclo de vida familiar en la que se encuentra el encuestado pueden incidir sobre su probabilidad de migrar: ¿un individuo soltero al inicio de la permanencia migra de manera distinta a como lo hace un individuo casado? Ser propietario o no de la vivienda en que se reside puede igualmente influir sobre esta movilidad. Asimismo, las etapas de la vida económica del individuo pueden estabilizarlo o, al contrario, conducirlo a migrar. Además, los eventos de carácter político (guerras, servicio militar, etc.) pueden provocar movimientos migratorios particulares. Por último, hay elementos de origen familiar (movilidad durante la infancia, número de hermanos y hermanas, etc.) que pueden influir sobre la movilidad.[5]

A manera de ejemplo, hemos escogido hacer que esas variables intervengan de manera acumulativa, partiendo de un modelo exponencial y encaminándonos hacia un modelo de Gompertz. Esto permite ver si la adición de nuevas variables mejora la calidad del modelo. Medimos esa calidad mediante el logaritmo de la relación de verosimilitud entre modelos sucesivos. El cuadro 3 proporciona el resultado de esas diversas etapas.

Todas las χ^2 son significativas y muestran que la adición de otras variables aporta nuevos elementos para explicar la migración. Se puede pensar que ciertas variables introducidas a lo largo del proceso y que están correlacionadas con el grupo de edades explican mejor el comportamiento migratorio.

En ese caso, el efecto de la edad se encontraría reducido en los modelos donde se hace intervenir más precisamente el estatus matrimonial, el ser propietario o no de la vivienda en que se reside, etc. Efectivamente, se verifica que un individuo casado ve reducida su movilidad a 80% de la que presenta

[5] Para más detalles véase D. Courgeau (1985a y 1985b).

CUADRO 3

Efecto de la adición de nuevas variables sobre la calidad del modelo para estudiar los cambios de residencia de los hombres nacidos entre 1930 y 1935

Tipo de modelo	*Nuevas variables añadidas*	*Núm. de variables añadidas*	*Máximo de logaritmo de semejanza*	*Diferencia de χ^2 con el modelo anterior*
Exponencial	Constante	1	− 6273.15	
Exponencial	Grupos de edad	6	− 6113.38	319.53
Exponencial	Duración de permanencia	1	− 5991.67	243.43
Gompertz	Caract. familiares	5	− 5963.17	57.00
Gompertz	Estatus ocupación del alojamiento	3	− 5755.45	415.44
Gompertz	Caract. profesional	10	− 5685.59	139.72
Gompertz	Eventos políticos	3	− 5642.93	85.31
Gompertz	Orígenes familiares	3	− 5637.67	10.53

un soltero, y que un propietario ve reducida su movilidad a 20% de la que presenta un inquilino, etcétera.

Ahora bien, eso es lo que se observa cuando en el modelo se introducen todas las variables. El cuadro 4 presenta el efecto de la pertenencia a los diversos grupos de edad, una vez que se ha tomado en cuenta el efecto de todas las demás características. El cuadro muestra claramente que esta pertenencia ya no incide sobre la probabilidad de migrar. En cambio, el efecto de características precisas, como el estatus matrimonial, el hecho de ser propietario, el haber tenido padres muy móviles durante la infancia, etc. explican mucho mejor los cambios de comportamiento migratorio del individuo.

Otra posibilidad de comparar diversos modelos y de escoger el más adecuado es utilizar un modelo lo bastante general como para que englobe el mayor número de modelos de los presentados aquí, así como de casos particulares. En el capítulo anterior presentamos la distribución de Fisher-Snedecor generalizada que abre esta posibilidad.

El logaritmo de la verosimilitud de esta distribución se escribe, cuando todos los eventos son observados:

$$\log L(\rho, \sigma, k_1, k_2) = \sum_{i=1}^{n} \Bigg(k_1 \log k_1 + k_2 \log k_2 + \log \Gamma(k_1 + k_2) - \log \Gamma(k_1) - \log \Gamma(k_2) + \frac{k_1}{\sigma} \log \rho + \left(\frac{k_1}{\sigma} - 1 \right) \log t_i - \log \sigma - (k_1 + k_2) \log \left(k_2 + k_1 (\rho t_i)^{\frac{1}{\sigma}} \right) \Bigg) \qquad (71)$$

CUADRO 4
Efecto del grupo de edad al inicio de la permanencia cuando todas las variables intervienen simultáneamente en el modelo

Grupos de edad	$\hat{\beta}_i$	$\sigma(\hat{\beta}_i)$	$\left[\frac{\hat{\beta}_i}{\sigma(\hat{\beta}_i)}\right]^2$	$\exp\beta_i$
Menos de 20 años	0.234	0.298	0.619	1.264
20-24 años	0.201	0.295	0.466	1.223
25-29 años	0.014	0.294	0.002	1.014
30-34 años	- 0.160	0.297	0.293	0.852
35-39 años	- 0.399	0.303	1.741	0.671

De donde resultan las derivadas respecto de los diversos parámetros:

$$
\begin{cases}
\dfrac{\partial \ell}{\partial \rho} = \displaystyle\sum_{i=1}^{n} \frac{k_1}{\sigma\rho}\left[1 - \frac{k_1 + k_2}{\sigma\rho\left[k_2(\rho t_i)^{-\frac{1}{\sigma}} + k_1\right]}\right] = 0 \\
\dfrac{\partial \ell}{\partial \sigma} = \displaystyle\sum_{i=1}^{n}\left(-\frac{k_1}{\sigma^2}\log(\rho t_i) - \frac{1}{\sigma} + (k_1 + k_2)\frac{k_1 \log \rho t_i}{\sigma^2[k_2(\rho t_i)^{-\frac{1}{\sigma}} + k_1]}\right) = 0 \\
\dfrac{\partial \ell}{\partial k_1} = \displaystyle\sum_{i=1}^{n}\Big(\log k_1 + 1 + \psi(k_1 + k_2) - \psi(k_1) \\
\qquad + \frac{\log \rho t_i}{\sigma} - \log\left(k_2 + k_1(\rho t_i)^{\frac{1}{\sigma}}\right) - \frac{k_1 + k_2}{k_2(\rho t_i)^{-\frac{1}{\sigma}} + k_1}\Big) = 0 \\
\dfrac{\partial \ell}{\partial k_2} = \displaystyle\sum_{i=1}^{n}\left(\begin{array}{l}\log k_2 + 1 + \psi(k_1 + k_2) - \psi(k_2) \\ -\log\left(k_2 + k_1(\rho t_i)^{\frac{1}{\sigma}}\right) - \dfrac{k_1 + k_2}{k_2 + k_1(\rho t_i)^{\frac{1}{\sigma}}}\end{array}\right) = 0
\end{cases}
\tag{72}
$$

donde la función $\psi(k) = \dfrac{\partial \log \Gamma(k)}{\partial k}$ es la derivada de la función Γ. De manera semejante, derivando nuevamente esas expresiones es posible obtener la matriz de información[6] y utilizar el método de Newton-Raphson para estimar los cuatro parámetros del modelo.

Sin embargo, para poder estimar los valores paramétricos del modelo cuando k, o k_2 tienden hacia el infinito y probar esos valores, es útil remplazar k_1 y k_2 por dos nuevos parámetros q_1 y q_2 iguales a:

$$q_1 = \frac{k_2 - k_1}{\sqrt{k_1 k_2 (k_1 + k_2)}}$$

$$q_2 = \frac{2}{k_1 + k_2} \tag{73}$$

lo que da para k_1 y k_2 en función de esos nuevos parámetros

$$k_1 = \frac{2}{q_1^2 + 2q_2 + q_1\sqrt{q_1^2 + 2q_2}} \tag{74}$$

$$k_2 = \frac{2}{q_1^2 + 2q_2 - q_1\sqrt{q_1^2 + q_2}} \tag{75}$$

Con esos nuevos parámetros, la función del logaritmo de la verosimilitud será finita cuando k_1 o k_2 tiendan hacia el infinito, y el sistema de ecuaciones (72) no tendrá derivadas iguales a cero. Por ejemplo, si

$$k_1 \to \infty, \text{ vemos que } q_1 \to -\frac{1}{\sqrt{k_2}} \text{ y } q_2 \to 0 .$$

Se ve que el modelo log-normal corresponde al punto (0.0), el modelo de Weibull corresponde al punto (1.0), el modelo log-logístico al punto (0.1). Hay otros tipos de modelos que entran además en esta familia. Por lo tanto resulta posible probar en ese caso general la validez de los diversos tipos de modelos.[7]

[6] Para más detalles respecto a esta estimación y acerca de las pruebas posibles, véase Prentice (1975).

[7] Para más detalles respecto a esta estimación y acerca de las pruebas posibles véase Prentice (1975).

Si introducimos ahora la posibilidad de salir de la observación antes de experimentar el evento, los cálculos se vuelven complejos rápidamente, pues no se dispone de una expresión simple para los cocientes instantáneos o la función de permanencia. Sin embargo, si las salidas de observación son poco numerosas siempre es posible acercarse a la verosimilitud mediante la relación (71) que proporcionará una estimación cercana a los parámetros del modelo.

E) Conclusión

Este capítulo nos ha permitido dar métodos de estimación precisos de los parámetros de los diversos tipos de modelos presentados en el capítulo anterior. De igual manera ha permitido medir el efecto de las características que pueden influir sobre la probabilidad de experimentar el evento estudiado. En cada uno de los casos hemos proporcionado las ecuaciones que posibilitan esta estimación, y los mejores medios para resolverlas.

Esos métodos proporcionan simultáneamente una estimación de la matriz de las varianzas y covarianzas de los parámetros estimados. Su utilización ofrece la posibilidad de efectuar todas las pruebas que se quieran, las cuales muestran en particular cuáles de las características tienen un efecto significativo sobre la duración de permanencia.

En todo este capítulo hemos trabajado sobre distribuciones que son funciones continuas del tiempo. Sin embargo se puede disponer, en el caso de ciertas encuestas, de datos reagrupados sobre periodos dados (trimestrales o anuales, por ejemplo). A partir de los modelos presentados aquí, resulta fácil generar los modelos de tiempo discontinuo correspondientes al introducir un reagrupamiento respecto del tiempo. Por ejemplo, si los datos son anuales se puede introducir un tiempo continuo subyacente T', y el tiempo medido T representa la parte entera de ese tiempo subyacente. En ese caso se puede escribir:

$$f(t) = P(T = t) = P(t \leq T' < t+1) = S(t) - S(t+1) \tag{76}$$

$$\lambda(t) = P(T' < t+1 | T' \geq t) = \frac{S(t) - S(t+1)}{S(t)} \tag{77}$$

Con esas notaciones, la verosimilitud de las observaciones se puede expresar de manera semejante a la que se tenía en el caso continuo (véase la fórmula 4):

$$\log L(\beta) = \sum_{i=1}^{n} \left[\delta_i \log \left(1 - \frac{S(t_i + 1; z_i, \beta)}{S(t_i, z_i, \beta)} \right) + \log S(t_i, z_i, \beta) \right]. \qquad (78)$$

Los métodos de estimación presentados son válidos en este caso. Notemos, sin embargo, que cuando el periodo de reagrupamiento es corto respecto del conjunto de la duración de la observación, el uso de un modelo de tiempo discreto da resultados prácticamente idénticos a los de un modelo de tiempo continuo.

Se supone que todas las características que hicimos intervenir fueron definidas al inicio de la permanencia. En efecto, esas características se pueden modificar antes de que se produzca el evento estudiado y cambiar su probabilidad de aparición. De esa manera, si se estudia la probabilidad de hacerse propietario, el hecho de casarse puede modificarla. Por lo tanto, resulta útil que pueda hacerse que intervengan características que dependen del tiempo. Así, la variable estatus matrimonial puede tomar el valor 0 hasta que ocurre el matrimonio, cuando se vuelve igual a 1. Los métodos que presentamos aquí se generalizan sin problema en ese caso, si bien los cálculos son mucho más laboriosos. En efecto, se puede escribir el logaritmo de la verosimilitud (5) cuando las variables z dependen del tiempo, bajo la siguiente forma:

$$\log L(\beta) = \sum_{i=1}^{n} \left[\delta_i \log h\left(t_i; z_i(t_i), \beta\right) + \int_0^{t_i} h\left(t; z_i(t), \beta\right) dt \right] \qquad (79)$$

la búsqueda de los valores β de que maximicen esa función se hace en una forma idéntica a la que presentamos en este capítulo.

Es igualmente posible hacer que intervenga una heterogeneidad no observada, si se tiene una idea de su distribución entre los individuos encuestados. En el capítulo VII presentamos cierto número de distribuciones que fueron obtenidas haciendo intervenir una heterogeneidad no observada, en el caso en el que el modelo de conjunto es exponencial. Una vez más, esos modelos se estiman con los métodos presentados en dicho capítulo. Cabe advertir, sin embargo, que si bien en las ciencias sociales ya se ha podido evidenciar el efecto de numerosas características observadas sobre los encuestados, aún no se dispone sino de pocos elementos sobre las diferencias de comportamiento que no están relacionadas con esas características. Por lo tanto, resulta muy peligroso modelizar esta heterogeneidad no observada, sin tener información seria. Heckman y Richards (1985) han demostrado que según la distribución que se elija, el efecto de ciertas características puede cambiar de manera importante. Trussel y Richards (1985) han mostrado que las conclusiones del análisis dependen no sólo de las distribuciones elegidas sino también del tipo de dependencia del tiempo de las variables introducidas.

Se puede pensar que esta heterogeneidad no observada se reduce fuertemente cuando se hace intervenir un máximo de características de los individuos encuestados. Eso permite evitar una modelización incorrecta de la heterogeneidad no observada.

Indiquemos, por último, que la utilización de modelos paramétricos está sometida a hipótesis muy fuertes sobre la distribución de los cocientes instantáneos y que con frecuencia disponemos de muy pocos encuestados, lo que impide verificarlas con precisión. Para evitar este inconveniente existe la posibilidad de considerar modelos más generales, no haciendo intervenir bajo la forma paramétrica más que el efecto de las características, dejando una estimación no paramétrica para los cocientes instantáneos de riesgo inicial. En particular, los modelos de riesgos proporcionales permiten hacerlo. Desarrollaremos este análisis semiparamétrico en el capítulo siguiente.

IX. MÉTODOS DE ANÁLISIS SEMIPARAMÉTRICO

Numerosos investigadores prefieren los métodos semiparamétricos a los paramétricos para el análisis de biografías. Ciertamente, los primeros permiten que los cocientes instantáneos dependan de las características individuales, sin imponer una formalización del efecto de duración. A continuación presentaremos en detalle el aporte de esos métodos.

A) De las regresiones paramétricas a los modelos de riesgos proporcionales semiparamétricos

Para analizar las relaciones entre variables explicativas y la ocurrencia de un evento hemos presentado numerosas formalizaciones paramétricas de distribución del evento. Éstas son las que generalmente se utilizan, en particular la distribución exponencial, la de Weibull, o incluso la de Gompertz, que difieren por la manera en que consideran el tiempo, así como otras distribuciones log-normales o gamma, que son más difíciles de estimar en la práctica.

Tal como vimos, esos modelos hacen que las variables intervengan de manera multiplicativa sobre los cocientes (propiedad de los *modelos de riesgos proporcionales*). Asimismo, éstos definen otra clase de modelos log-lineales llamados *modelos de ocurrencias aceleradas* por el efecto multiplicativo que tienen las *variables* sobre *T*.

1) Definición

Los modelos semiparamétricos, que fueron introducidos por Cox (1972), modelizan los cocientes instantáneos de la manera siguiente:

$$h(t; z) = h_0(t) \exp(\beta z) \qquad (1)$$

pero esta vez $h_0(t)$ es una función desconocida arbitraria de t, denominada cociente instantáneo inicial.

Estos modelos constituyen una generalización de los modelos precedentes, que encontramos si le damos una forma a $h_0(t)$: Si $h_0(t) = \exp(a)$, el modelo es exponencial.

En la interpretación $h_0(t)$ será lo más a menudo el cociente instantáneo para el individuo estándar ($z = 0$). Sin embargo, como h_0 no está definida, el modelo está parcialmente no especificado y, en consecuencia, se denomina semiparamétrico. Por otra parte, su denominación como modelo de riesgos proporcionales se deriva, dado que las variables incorporadas son fijas, de la relación que existe entre las densidades condicionales de dos individuos. Como la relación de las densidades de dos individuos es constante cualquiera que sea t, sus cocientes instantáneos son por lo tanto proporcionales (*cf.* capítulo VII.B.1).

2) Construción de la verosimilitud

Sea una muestra de n individuos. El riesgo en el instante t se mide por el cociente instantáneo $h(t; z)$ y se busca estimar los parámetros β desconocidos que miden la influencia de z sobre $h(.)$. Para cada individuo i se dispone de t_i, c_i, z_i, donde t_i es la fecha en que ocurre el evento si $\delta_i = 1$, y donde c es la fecha de truncamiento (el individuo desaparece de la observación) si $\delta_i = 0$. La verosimilitud se forma de la manera siguiente:

$$L = \prod_{i=1}^{n} [f(t_i; z_i)^{\delta_i} S(t_i; z_i)^{1-\delta_i}] \tag{2}$$

que después del desarrollo se escribe:

$$L = \prod_{i=1}^{n} h(t_i; z_i)^{\delta_i} \exp\left[-\int_0^{\infty} \sum_{\ell \in R_{t_i}} h(u; z_\ell) du\right] \tag{3}$$

donde R_t es la población sometida a riesgo en $t - 0$, expresión semejante a la (4) dada en el capítulo VIII.A.

Como una parte de la expresión de los cocientes instantáneos no se especifica (el cociente de riesgo inicial $h_0(t)$), la estimación de los diversos parámetros no puede hacerse maximizando directamente la verosimilitud anterior. En efecto, ante la ausencia de restricciones sobre $h_0(t)$, no podemos encontrar un máximo a la expresión de la verosimilitud. Se recurrirá entonces a métodos de estimación específicos.

Efectivamente, es necesario estimar por una parte el valor de los parámetros β, los cuales actúan de manera paramétrica sobre el cociente instantáneo $h(t, z)$. Esta estimación se mide mediante el sesgo de la maximización de una forma parcial de la verosimilitud. Posteriormente, conociendo el valor $\hat{\beta}$, se estimará el cociente de riesgo inicial $h_0(t)$ de manera no paramétrica.

B) Métodos de estimación

1) Estimación de los parámetros

Mostremos primero las estrechas relaciones que existen entre la verosimilitud marginal de rango y la verosimilitud parcial que vamos a maximizar para obtener los estimadores de los parámetros.

La expresión de verosimilitud marginal de rango resulta de su elaboración en relación con la distribución marginal de rangos y las técnicas de pruebas de rango (Kalbfleisch y Prentice, 1980).

Sean las fechas de ocurrencia $t_1, ..., t_n$ para los n individuos de la muestra con los vectores de variables $z_1, ..., z_n$. Al formar $t_{(1)} < t_{(2)} < ... < t_{(n)}$ como la serie de fechas ordenadas, obtenemos dos nuevos estadísticos:

$$\theta(t) = \left[t_{(1)}, \ldots, t_{(n)}\right] \text{ el estadístico de orden y}$$

$$R(t) = \left[(1), \ldots, (n)\right] \text{ el estadístico de rango.}$$

El primero contiene las fechas ordenadas y el segundo los rangos correspondientes a las fechas ordenadas. Por ejemplo, si $n = 4$ y se observa que $t_1 = 12$, $t_2 = 5$, $t_3 = 9$, $t_4 = 17$, entonces $\theta(t) = (5, 9, 12, 17)$ y $R(t) = (2, 3, 1, 4)$.

Sea G el grupo de las transformaciones de $\mathbb{R}^+$ en $\mathbb{R}^+$ estrictamente creciente y diferenciable. El problema de la estimación de los parámetros β en la expresión de los cocientes instantáneos $h(t; z) = h_0(t) \exp(\beta z)$ no varía en relación con el grupo G de las transformaciones aplicadas a t. Además, la acción de las transformaciones sobre $\theta(t)$ deja a $R(t)$ sin variación. Para el ejemplo anterior, si la transformación se define por $u = 3 \times t$ resulta que:

$$\theta(u) = [15, 27, 36, 51] \text{ y } R(u)[2, 3, 1, 4] = R(t)$$

Se puede entonces decir que el problema de estimación de los parámetros β es el mismo cualquiera que sea la transformación impuesta a $\theta(t)$ y que sólo $R(t)$ contiene la información sobre los parámetros β cuando $h_0(t)$ es totalmente desconocido. Bajo esas condiciones se dice que el estadístico de rango $R(t)$ es marginalmente exhaustivo para la estimación de los parámetros β en ausencia de una especificación de $h_0(t)$.

Así, el orden de esos eventos importa más que las fechas exactas de su ocurrencia.

Se forma, pues, una verosimilitud marginal que es proporcional a la probabilidad de observar la cronología tal como la recogimos.

Esto nos remite al cálculo de la verosimilitud parcial definida por Cox.[1] Condicionalmente a la población sometida a riesgo y al hecho de que la ocurrencia del evento se produzca en t_i, la probabilidad de que el individuo (i) experimente ese evento es igual a:

$$\frac{h_0(t_i)\exp\left(z_i(t_i)\beta\right)}{\sum_{\ell\in R_i} h_0(t_i)\exp\left(z_\ell(t_i)\beta\right)} \tag{4}$$

donde R_i es el conjunto de las etiquetas de los individuos sometidos a riesgo en $t_i - 0$. Esta expresión se reduce a:

$$\frac{\exp\left(z_i(t_i)\beta\right)}{\sum_{\ell\in R_i}\exp\left(z_\ell(t_i)\beta\right)} \tag{5}$$

De esa manera se ignoran los intervalos donde no se ha producido ningún evento o para los que no se dispone de ninguna información sobre los z. La verosimilitud parcial se forma entonces tomando el producto sobre todas las fechas:

$$PL(\beta) = \prod_{i=1}^{n}\frac{\exp\left(z_i(t_i)\beta\right)}{\sum_{\ell\in R_i}\exp\left(z_\ell(t_i)\beta\right)} \tag{6}$$

Esta verosimilitud se escribe de manera más simple:

$$PL(\beta) = \frac{\exp\sum_{i=1}^{n} z_i(t_i)\beta}{\prod_{i=1}^{n}\left[\sum_{\ell\in R_i}\exp\left(z_\ell(t_i)\beta\right)\right]} \tag{7}$$

En la práctica, varios eventos tienen lugar en la misma fecha en el seno de la muestra. Por lo tanto, es conveniente que se tome en cuenta esta eventualidad.

En ese caso se observan n individuos, pero las fechas de ocurrencia ordenadas son $t_1 < \ldots < t_k$. Denominamos d_i al número de individuos para quienes la fecha de ocurrencia del evento es t_i y, sabiendo que los rangos así definidos no son afectados por el valor de los d_i, se tiene entonces:

$$PL(\beta) = \prod_{i=1}^{k}\frac{\exp(s_i\beta)}{\left[\sum_{\ell\in R_i}\exp(z_\ell\beta)\right]^{d_i}} \tag{8}$$

donde $s_i = \sum z_j(t_i)$ es la suma de las variables explicativas de los d_i individuos que experimentan el evento (o el truncamiento) en t_i.

[1] *Cf.* anexo 1.III.

El estimador β del máximo de verosimilitud se obtiene entonces como solución de las ecuaciones habituales:

$$\frac{d \log PL(\beta)}{d\beta_i} = 0 \tag{9}$$

es decir:

$$s_{ji} - d_i A_{ji}(\beta) = 0 \tag{10}$$

donde s_{ji} es el $j^{\text{ésimo}}$ componente del vector s_i
y:

$$A_{ji}(\beta) = \frac{\sum_{\ell \in R_i} z_{j\ell} \exp(z_\ell \beta)}{\sum_{\ell \in R_i} \exp(z_\ell \beta)} \tag{11}$$

Mostremos ahora las relaciones entre esta verosimilitud parcial y la verosimilitud (total) utilizada habitualmente.

El método de estimación de la verosimilitud parcial propuesto por Cox es una contribución determinante, en la medida en que su puesta en práctica es muy similar a la de las estimaciones ordinarias del máximo de verosimilitud, en circunstancias donde las densidades de probabilidades asociadas a las distribuciones estudiadas son complejas.

Sin regresar a los detalles formales de la demostración de Cox (anexo 1.III) podemos, sin embargo, describir su fundamento respecto de la técnica del máximo de verosimilitud.

El método de verosimilitud parcial es muy cercano al del máximo de verosimilitud en que, de igual manera, procede en dos etapas:

1) construir una verosimilitud a partir de la observación, concerniente a parámetros desconocidos;
2) encontrar los valores que maximizan la función construida.

Este método difiere sólo en que la verosimilitud parcial es el producto de las contribuciones de cada evento observado, mientras que la verosimilitud total o usual es el producto de las contribuciones de cada individuo de la muestra.

Para construir la verosimilitud parcial no se conserva más que uno de los factores de la verosimilitud (total). En efecto, dentro del marco de la modelización de riesgos proporcionales, la verosimilitud (total) usual está formada por dos factores: uno que contiene la información concerniente al parámetro β y el otro que contiene la información concerniente a β e igualmente a $h_0(t)$. La verosimilitud parcial sólo retiene, pues, el primer término, al que trata entonces como una verosimilitud total. Este primer factor depende sólo del orden de ocurrencia de los eventos (lo que permite demostrar que

la verosimilitud parcial es de hecho una verosimilitud marginal de rango) y no de la fecha exacta en que ocurren.

Los estimadores obtenidos son asintóticamente no sesgados, distribuidos normalmente. No obstante éstos no son consistentes, puesto que la información sobre las fechas exactas no se utiliza en la estimación.

2) Estimación del componente no paramétrico

Luego de haber detallado la estimación de los parámetros asociados a las variables explicativas, hay una parte más delicada que consiste en estimar el componente no paramétrico de los cocientes instantáneos de ocurrencia.

En efecto, recordemos que éstos se expresan bajo la forma siguiente:

$$h(t; z) = h_0(t)\exp(z\beta)$$

La función de permanencia o de supervivencia en una modelización semiparamétrica se expresa bajo la forma:

$$S(t; z) = \exp -\int_0^t h_0(u)e^{z\beta}du \tag{12}$$

sea además:

$$S(t; z) = \left[\exp -\int_0^t h_0(u)du\right]^{\exp(z\beta)} \tag{13}$$

y finalmente:

$$S(t; z) = S_0(t)^{\exp(z\beta)} \tag{14}$$

El método de verosimilitud parcial ha permitido estimar los parámetros β sin hacer una hipótesis sobre los $h_0(.)$.

Ahora estimaremos los $h_0(.)$ sirviéndonos de los parámetros β estimados, pero aplicando un método de estimación prácticamente no paramétrico.

Sean $t_1 < \ldots < t_k$, las fechas de ocurrencia de los eventos observados, y supongamos que durante el intervalo $[t_i, t_{i+1}[$ las salidas de la observación se producen en t_ℓ que pertenece a ese intervalo. La verosimilitud está así constituida por las contribuciones de los individuos que experimentan el evento al inicio del intervalo; sea:

$$S_0(t_i)^{\exp(z\beta)} - S_0(t_i + 0)^{\exp(z\beta)} \tag{15}$$

donde T_i es el conjunto de sus etiquetas.

La contribución de los individuos que se salen de la observación en t_ℓ es:

$$S_0(t_\ell + 0)^{\exp(z\beta)} \tag{16}$$

donde M_i es el conjunto de sus etiquetas.

De donde resulta que la verosimilitud es igual a:

$$L = \prod_{i=1}^{k} \left\{ \prod_{\ell \in T_i} \left(S_0(t_i)^{\exp(z_\ell \beta)} - S_0(t_i + 0)^{\exp(z_\ell \beta)} \right) \prod_{\ell \in Mi} S_0(t_i + 0)^{\exp(z_\ell \beta)} \right\} \tag{17}$$

Al igual que en el caso del estimador de Kaplan-Meier, haremos la hipótesis de que los eventos se producen al inicio del intervalo:

$$S_0(t) = S_0(t_i + 0) \text{ para } t_i < t \leq t_{i+1}$$

en un modelo de tiempo discreto donde los cocientes instantáneos se expresan bajo la forma $h_0(t_i) = 1 - \alpha_i$.

La verosimilitud, con esta nueva notación, se vuelve:

$$L = \prod_{i=1}^{k} \left(\prod_{\ell \in Ti} \left(1 - \alpha_i^{\exp(z_\ell \beta)} \right) \prod_{\ell \in R_i - T_i} \alpha_i^{\exp(z_\ell \beta)} \right) \tag{18}$$

Tomando para β los valores $\hat{\beta}$ obtenidos al estimar la verosimilitud marginal o parcial y al derivar el logaritmo de la verosimilitud anterior, se obtienen los α_i como soluciones de:

$$\sum_{j \in T} \frac{\exp(z_j \hat{\beta})}{1 - \alpha_i^{\exp(z_j \hat{\beta})}} = \sum_{\ell \in R_i} \exp(z_\ell \hat{\beta}). \tag{19}$$

Si en t_i se produce un solo evento, encontramos analíticamente:

$$\hat{\alpha}_i = \left(1 - \frac{\exp(z_i \hat{\beta})}{\sum_{\ell \in R_i} \exp(z_\ell \hat{\beta})} \right)^{\exp z_i \hat{\beta}} \tag{20}$$

Si no se produce, es necesaria una solución iterativa con un valor inicial aconsejado α_{i0}, tal como:

$$\log \alpha_{i0} = \frac{-d_i}{\sum_{\ell \in R_i} \exp(z_\ell \hat{\beta})} \tag{21}$$

que se obtiene sustituyendo el estimador de α_i siguiente en (19):

$$\hat{\alpha}_i^{\exp(z_j\hat{\beta})} = \exp(e^{z_j\hat{\beta}} \log \hat{\alpha}_i) \tag{22}$$

que es poco diferente de:

$$\hat{\alpha}_i^{\exp(z_j\hat{\beta})} \simeq 1 + e^{z_j\hat{\beta}} \log \hat{\alpha}_i \tag{23}$$

Consideremos ahora los algoritmos de maximización utilizados.

C) Algoritmo de Newton-Raphson

Ya presentamos (capítulo VIII.C.2) el método de estimación que utiliza el algoritmo de Newton-Raphson. Recordemos que la expresión de la verosimilitud parcial es de la forma siguiente:

$$PL(\beta) = \prod_{i=1}^{k} \frac{\exp(s_i\beta)}{\left[\sum_{\ell \in R_i} \exp(z_\ell \beta)^{d_i}\right]} \tag{24}$$

El estimador del máximo de verosimilitud $\hat{\beta}$ se obtiene entonces como solución $U(\hat{\beta}) = 0$, sea:

$$\sum_{i=1}^{k} \left(s_{ji} - d_i A_{ji}(\beta)\right) = 0 \tag{25}$$

donde s_{ji} es la $j^{\text{ésima}}$ componente de:

$$s_i = \sum z(t_i)$$

y $A_{ji}(\beta)$ está dado por

$$A_{ji}(\beta) = \frac{\sum_{\ell \in R_i} z_{j\ell} \exp(z_\ell \beta)}{\sum_{\ell \in R_i} \exp(z_\ell \beta)} \tag{26}$$

A fin de utilizar una iteración de Newton-Raphson para obtener β que maximice $PL(\beta)$ se calcula:

$$I_{hj}(\beta) = -\frac{d^2 \log PL(\beta)}{d\beta_h d\beta_j} = \sum_{i=1}^{d} d_i \, C_{hji} \tag{27}$$

donde:

$$C_{hji} = \frac{\sum_{\ell \in R_i} z_{h\,\ell}\, z_{j\ell} e^{z_\ell \beta}}{\sum_{\ell \in R_i} \exp(z_\ell \beta)} - A_{hi}(\beta) A_{ji}(\beta) \tag{28}$$

En nuestro caso la matriz de varianza-covarianza estimada está dada por $I_{h_j}(\hat{\beta})^{-1}$ y el vector de los parámetros estimados es $A_{ji}(\hat{\beta})$, de donde:

$$\hat{\hat{\beta}} = [I_{hj}(\hat{\beta})]^{-1} A_{ji}(\hat{\beta}) . \tag{29}$$

D) Selección de un modelo para el análisis de interacciones

Nuestro enfoque de las interacciones plantea privilegiar el instante de ocurrencia del evento perturbador si éste se produce antes del evento estudiado, como punto de inflexión del comportamiento de los individuos. De tal manera que se podría modelizar una relación única entre el cociente instantáneo de ocurrencia y las variables individuales de las cuales un indicador señala la ocurrencia del evento perturbador tal que:

$$h(t) = h_0(t) \exp(z_t \beta) \tag{30}$$

donde z_t es un vector de características que incluye una coordenada z_{it} igual a 0 si el segundo evento aún no ha tenido lugar en la fecha t, e igual a 1 si sucede lo contrario.

Tan sólo para privilegiar esta perturbación hemos escogido no tratarla como una simple variable explicativa dependiente de la duración, sino más bien medir las eventuales modificaciones debidas a su ocurrencia bajo la influencia de las diversas variables introducidas. Así, siguiendo a Crowley y Hu (1977), utilizamos una modelización de riesgos propocionales que distingue la formulación de los cocientes instantáneos antes y después de la perturbación:

$$h..(t|u) = h^*(t) \exp[z\beta_1 + H(t-u)(\beta_0 + z'\beta_2)] \tag{31}$$

donde

$$H(x) = \begin{cases} 0 \text{ si } x < 0 \\ 1 \text{ si } x \geq 0 \end{cases}$$

y u es la fecha de la perturbación.

Retomando las notaciones de la figura 2 (capítulo V) también se puede escribir:

$$h_{01}(t) = h^*(t) \exp(z\beta_1) \tag{32}$$

$$h_{21}(t) = h^*(t) \exp(z\beta_1 + \beta_0 + z'\beta_2) \tag{33}$$

donde $h^*(t)$ es el cociente inicial de riesgo no paramétrico, β_1, β_0 y β_2 son los coeficientes que se van a estimar a partir de las variables z (introducidas antes del fenómeno perturbador) y z' (introducidas después del fenómeno perturbador, y pueden entonces ser las mismas que las precedentes u otras adquiridas al producirse la perturbación). Este modelo supone que las diversas variables tienen un efecto multiplicativo sobre el cociente estimado.

Para cada característica se obtiene, pues, su efecto principal (antes de la perturbación u ocurrencia del evento secundario) que puede encontrarse eventualmente modificado al producirse la perturbación o evento secundario por los valores de los parámetros β_0 y β_2. Esta formalización del modelo supone que el cociente de riesgo inicial no paramétrico sigue siendo el mismo antes y después de la ocurrencia del evento perturbador.

El modelo que hemos privilegiado no se puede llevar a la práctica más que mediante un solo paquete comercial (BMDP) que requiere, sin embargo, una programación secundaria pesada (anexo 1.III). Por ello hemos desarrollado un paquete particular, EVACOV (anexo 1.III), cuyo manual de uso se incluye en el anexo 3. Los resultados que presentamos a continuación se han obtenido gracias a ese medio informático.

E) Algunos ejemplos de aplicación

Con el objeto de ilustrar las posibilidades de análisis que ofrece el modelo elegido, retomamos aquí dos ejemplos tratados a partir de los datos de la encuesta "Triple biografía".

En el estudio de las interacciones entre nupcialidad y actividad profesional agrícola (Courgeau y Lelièvre, 1986), los hombres y las mujeres que proceden del mundo campesino tienen comportamientos muy diferentes. El análisis semiparamétrico permitirá identificar y caracterizar con precisión los comportamientos puestos en evidencia. El modelo escogido (31) se aplica en un primer tiempo al análisis del efecto que tiene cada una de las características por separado, sobre la nupcialidad y sobre el abandono de la agricultura.

En los efectos perturbadores del matrimonio (respecto del abandono de la agricultura) reencontramos las influencias unilaterales de interacción: el matrimonio retiene a las mujeres en el trabajo agrícola mientras que el abandono de la agricultura favorece el matrimonio de los hombres. Por otra parte, las variables que caracterizan el entorno familiar y la situación del individuo en su seno tienen un efecto mucho más débil sobre el comportamiento de los hombres que sobre el de las mujeres, y esto tanto para su matrimonio como para su abandono de la agricultura (cuadro 1).

CUADRO 1
Valores de los parámetros en la expresión de los cocientes instantáneos de nupcialidad y de los cocientes instantáneos de abandono de la agricultura

	Mujeres			*Hombres*		
	Efecto principal	*Perturbación*	*Interacción*	*Efecto principal*	*Perturbación*	*Interacción*
	β_1	β_0	β_2	β_1	β_0	β_2
			Nupcialidad			
Hijo(a) mayor	0.151	0.098	−0.483**	−0.171*	0.344**	0.144
Padre agricultor	−0.208**	−0.124	0.067	0.182	0.383**	0.032
Abandono de la agricultura		−0.034			0.384**	
			Abandono de la agricultura			
Núm. hermanos	0.017**	−0.804**	−0.002	0.003	−0.170	0.000
Hijo(a) mayor	−0.352**	−0.884**	0.238	−0.067	−0.161	−0.046
Padre agricultor	−0.949**	−1.274**	0.658**	−0.563**	−0.483**	0.490**
Matrimonio		−0.806**			−0.125	

*Resultado con un nivel de significación de 10 por ciento.
** Resultado con un nivel de significación de 5 por ciento.

Cuando el conjunto de variables interviene simultáneamente en el modelo, se pueden identificar verdaderas estrategias.

Tomemos el ejemplo de las mujeres, presentado en el cuadro 2. Éste nos permite responder a la pregunta de cuáles son las mujeres que, una vez casadas, permanecen en el mundo agrícola:

- Antes de su matrimonio, la hija mayor de un agricultor con sólo dos hijos, tiene un riesgo relativo[2] igual a $\exp\big(0.012+(-0.320)+(-0.928)\big) = 0.290$, mientras que la segunda hija de una familia de obreros agrícolas con cinco hijos tiene un riesgo relativo de abandono de la agricultura igual a $\exp\big(4(0.012)\big) = 1.048$. El abandono del mundo agrícola de esta última es 3.85 veces más probable que el de la primera.
- Una vez que la mujer se ha casado estando en el medio agrícola podemos observar que las características adquiridas en ese matrimonio tienen un efecto desfavorable sobre el abandono de la agricultura. Las mujeres que se quedan están, de hecho, muy tipificadas. Se trata de hijas mayores

[2] El cociente instantáneo se calcula como producto del cociente instantáneo de riesgo inicial y del riesgo relativo que caracteriza a cada individuo, puesto que se calcula a partir del valor de las características individuales.

de un grupo de hermanos no muy numeroso, hijas de agricultor, casadas con un agricultor: mujeres que "hacen carrera" en la agricultura.

CUADRO 2
Modelo óptimo de abandono de la agricultura de las mujeres (valores de los parámetros)

Conjunto de las variables	*Efecto principal* β_1	*Perturbación* β_0	*Interacción* β_2
Núm. hermanos	0.012**		0.000
Hija mayor	−0.320**		0.296
Padre agricultor	−0.928**		0.806*
Matrimonio		−0.228	
Agricultora al casarse			−1.040
Marido agricultor			−0.359**
Suegro agricultor			−0.126

* Resultado con un nivel de significación de 10 por ciento.
** Resultado con un nivel de significación de 5 por ciento.

Como segundo ejemplo tomemos el de la adquisición de la primera vivienda cuando se termina de constituir la familia (Courgeau y Lelièvre, 1988).

En efecto, en el caso de las mujeres casadas de la muestra a las que se observa más allá de su vida fecunda, se llega a obtener información de cuándo se termina de constituir la familia: 66% de esas parejas se hace propietario por primera vez luego de este acontecimiento (cohortes nacidas entre 1911 y 1935). Calculamos entonces los efectos de las variables tomadas por separado, para cinco grupos profesionales a los que pertenecen sus esposos.

La adquisición de la primera vivienda durante la vida fecunda es siempre menos probable para las parejas en que el marido es un obrero especializado (cuadros 3 y 4) proveniente de una familia numerosa o cuya descendencia es grande. Una vez que la familia se ha constituido, las restricciones puramente familiares se atenúan hasta no desempeñar más un papel significativo. Además los diplomas, que son indicadores de la carrera de los esposos, aparecen como claramente favorables, mientras que no tienen un papel primario sobre las adquisiciones precoces, que están más determinadas por un origen rural.

Resulta igualmente posible introducir variables dependientes del tiempo en los modelos semiparamétricos. M. Murphy, al analizar en Inglaterra la entrada a una vivienda de tipo social luego del matrimonio de los individuos, introdujo este tipo de variable para los nacimientos sucesivos. Por ejemplo, mientras no se ha producido el primer nacimiento esa variable es igual a

CUADRO 3

Valores de los parámetros en la expresión de los cocientes instantáneos de adquisición durante la vida fecunda de las parejas en las que el marido es un obrero especializado

Variables	*Efecto principal* β_1	*Perturbación* (a) β_0	*Interacción* β_2
Diploma	-0.022	-0.347**	0.273*
Núm. hermanos y hermanas	-0.057**	-0.164	-0.003
Diplomas del marido		-0.375**	0.316**
Núm. hermanos y hermanas del marido		0.095	-0.076**
Descendencia final		0.304*	-0.206**

(a) Aquí se trata del nacimiento del último hijo.
* Resultado con un nivel de significación de 10 por ciento.
** Resultado con un nivel de significación de 5 por ciento.

CUADRO 4

Valores de los parámetros en la expresión de los cocientes instantáneos de adquisición una vez constituida la familia (parejas en las que el marido es un obrero especializado)

Variables	*Efecto principal* β_1
Diploma	0.538**
Núm. hermanos y hermanas	0.005
Diploma del marido	0.594**
Núm. hermanos y hermanas del marido	0.013
Descendencia final	-0.002

** Resultado con un nivel de significación de 5 por ciento.

cero, y se vuelve igual a la unidad después del primer nacimiento. El cuadro 5 presenta esos resultados y revela un efecto altamente significativo del primero y tercer nacimientos. Una vez que éstos se han producido, la probabilidad de tener una vivienda de tipo social se multiplica por dos después del primer nacimiento y por 2.37 después del tercero.

CUADRO 5
Parámetros estimados para la entrada a una vivienda de tipo social en Inglaterra después del matrimonio en 1961-1965, según el número de nacimientos anteriores (variable dependiente del tiempo)

Variable	*Parámetro estimado*
Primer nacimiento	0.720***
Segundo nacimiento	0.142
Tercer nacimiento	0.863***
Log de la verosimilitud	−1669.56

*** Resultado con un nivel de significación de uno por mil.
Fuente: M. Murphy, 1984.

En estos ejemplos se advierte que el análisis permite dar precisiones que ayudan a explicar los comportamientos observados. Ciertas hipótesis planteadas al efectuarse estudios más cualitativos pueden entonces confirmarse o invalidarse, y se plantean otras que se examinarán mediante la colaboración de otras disciplinas (sociología cualitativa o psicología, por ejemplo). Esos análisis muy elaborados proporcionan un material de interacción pluridisciplinaria. Los resultados obtenidos demandan el conocimiento de otras disciplinas o lo llaman a explorar un campo de análisis común.

F) CONCLUSIÓN

Este análisis, más flexible que una modelización puramente paramétrica, ofrece una alternativa interesante cuando se quiere medir la influencia de las características individuales sobre los cocientes instantáneos estimados.

Además, cuando no queremos tomar en cuenta explícitamente la heterogeneidad no observada de los comportamientos individuales descritos, el componente no paramétrico "recupera" esa varianza sin que se le imponga nada a su distribución.[3] Los *software* convencionales (BMDP, RATE, GLIM) no siempre toman en consideración esta opción de manera explícita, pero con un pequeño suplemento de programación esto resulta posible. No siempre es posible tomar en cuenta las características dependientes del tiempo.

[3] Durante el seminario de la UIESP sobre el análisis de las biografías (París, marzo de 1988), de los seis trabajos solicitados para comparar los métodos, cuatro equipos escogieron un modelo semiparamétrico de análisis.

CONCLUSIÓN GENERAL

En la introducción de este libro nos referimos a sus dos principales objetivos: en primer lugar, desarrollar los métodos que permiten el análisis de las *interacciones* entre fenómenos demográficos y, en segundo lugar, abordar y tratar la *heterogeneidad* de las poblaciones observadas. Si bien ambos problemas se habían planteado desde hacía mucho tiempo (Henry, 1959; Pressat, 1966), hasta el momento no se les había dado ninguna solución satisfactoria. Muy a menudo bastaba con eliminar el efecto perturbador de un fenómeno sobre otro, sin analizar las interacciones más complejas entre los fenómenos. Asimismo, las poblaciones eran consideradas como homogéneas o, en el mejor de los casos, se usaban criterios simples para desagregarlas en subpoblaciones a las que se trataba por separado.

La recolección de las biografías individuales, obtenidas cada vez más a menudo mediante encuestas retrospectivas, nos ha ofrecido la posibilidad de aportar respuestas nuevas a esos dos problemas. En efecto, los métodos que hemos presentado a lo largo de este libro analizan las interacciones entre fenómenos demográficos, sociales y económicos haciendo intervenir simultáneamente la heterogeneidad de las poblaciones consideradas. Su utilización cada vez más frecuente por parte de los demógrafos, y los numerosos resultados que ya han contribuido a evidenciar, hacen que en la actualidad constituyan una herramienta privilegiada del análisis longitudinal.

Al finalizar este libro sintetizaremos las respuestas de estos métodos a los dos problemas inicialmente planteados y trataremos de identificar nuevas vías de investigación abiertas por ellos.

A) El análisis de las interacciones entre fenómenos

Este primer acercamiento se basa en la recolección de diversos eventos que afectan la existencia de un individuo, superando así el análisis de los fenómenos puramente demográficos, pues esos eventos pueden ser de tipos muy diversos.

Puede tratarse, pues, de eventos del mundo puramente *físico* que afectarán tanto la existencia de algunos individuos como la de millones de seres humanos. Esos eventos por lo general se registran fuera de las biografías individuales y se pueden agregar a éstas con facilidad. Un sismo, una erupción

volcánica, un invierno riguroso, etc., son eventos completamente independientes de las sociedades que afectan; sin embargo, habrán de tocar y modificar el curso de la vida de los seres humanos sometidos a su efecto.

Asimismo, se puede tratar de eventos de orden *biológico* como la pubertad, la menopausia, el parto, etc. En ese caso, la aparición de uno de ellos puede ser diferente según la sociedad en donde vive el individuo. Pero sobre todo, desde otra perspectiva, es posible estudiar cómo pueden modificar la existencia de quienes los experimentan.

Otros eventos de tipo *social, económico y político* van a tener consecuencias importantes en la continuación de la vida del individuo. Algunos dependen de la vida pasada de los individuos. Así, el hecho de hacerse ingeniero en la vida profesional está ligado a los diferentes diplomas que el individuo obtuvo durante su vida escolar y universitaria. Por otra parte, hacerse ingeniero va a influir sobre la ocurrencia de los eventos que vendrán. De igual manera, un evento de tipo político, como la participación en un movimiento de huelga, podrá influir sobre la carrera profesional de un individuo.

Abordamos, por último, eventos más complejos de orden *psicológico*. El inicio de una amistad o el apego a un lugar pueden modificar de manera importante las actitudes futuras de un individuo. Resulta fácil ver la común dificultad de captar esos eventos en una encuesta de tipo demográfico. Sin embargo no hay que subestimar su importancia, que puede ser muy grande.

Todos esos eventos se pueden localizar, con mayor o menor precisión, en la vida de los individuos. Para la demografía ellos constituyen el equivalente de lo que son las partículas para los físicos. Así pues, el número de tipos de eventos por considerar no está fijado de una vez por todas, sino que puede variar de un periodo al siguiente y de una sociedad a otra. De manera semejante podemos concentrarnos en el estudio de la interacción entre dos, tres, etc. tipos de eventos tratando de eliminar el efecto perturbador de los otros.

A partir del momento en que concentramos la atención sobre las fechas de aparición sucesiva de los eventos es posible formalizar de manera teórica el análisis que se va a realizar.

1) Formalización teórica

Aquí no retomaremos en detalle la formalización que hemos presentado en la introducción y en el capítulo II, sino que haremos precisiones sobre las hipótesis que la fundamentan.

Antes de analizar la interacción entre fenómenos debemos plantear la hipótesis de que cada evento tiene una probabilidad inicial de producirse y esta probabilidad cambia con el tiempo y se modifica cuando otros fenómenos se realizan antes de su ocurrencia.

La observación de una biografía individual no permite estimar esas probabilidades, pues no disponemos sino de una sola realización del proceso. En cambio, cuando observamos una muestra de individuos tenemos la posibilidad de estimar esas probabilidades.

Sin embargo, para que esas probabilidades logren una significación clara se necesitan otras hipótesis. En particular hay que trabajar sobre una subpoblación lo bastante homogénea como para que las interacciones estudiadas se pongan en evidencia. Así, hemos aislado la subpoblación de los solteros que inician su actividad económica en la agricultura para poner en evidencia los lazos entre la nupcialidad y el abandono del mundo agrícola. Por supuesto que en un segundo tiempo trataremos de ver si existen otras fuentes de heterogeneidad en la subpoblación aislada. Ése será el objeto del tratamiento de la heterogeneidad.

La observación de una población homogénea implica igualmente la observación de los miembros de una generación o de una cohorte bien acotada. En efecto, no es posible plantear preliminarmente la hipótesis de estabilidad de comportamientos de una generación a la otra, y de hecho ésta tiene muy pocas posibilidades de verificarse. Más bien lo que aquí proponemos estudiar es la evolución de los comportamientos de una generación a la siguiente.

La utilización de los datos de las encuestas retrospectivas para poner en evidencia esta evolución de los comportamientos nos va a llevar a plantear nuevas hipótesis. En efecto, con este modo de observación no registramos toda la existencia de las personas encuestadas, sino sólo los eventos que se produjeron antes de la encuesta. Disponemos así de biografías truncadas a la derecha, y para tener una estimación correcta de las probabilidades tomamos la hipótesis implícita de que los individuos que no experimentaron los eventos estudiados estaban sometidos a la misma probabilidad de conocerlos durante la observación que la de aquellos que realmente los vivieron. De nuevo se trata de una condición de homogeneidad de la subpoblación sobre la que se trabaja.

Vemos así que el análisis de las interacciones entre fenómenos se justifica perfectamente cuando se observa durante un periodo bastante largo a una subpoblación suficientemente homogénea. Hay que percibir con claridad que esta hipótesis, hecha aquí sobre una subpoblación, es idéntica a la que se plantea en demografía clásica cuando se trabaja sobre el conjunto de la población de un país. El hecho de distinguir subgrupos en ese conjunto no puede más que introducir una homogeneidad más grande, y desde el punto de vista metodológico constituye una generalización de los métodos del análisis demográfico clásico.

2) Modelos no paramétricos

Una vez planteadas esas hipótesis, hemos demostrado que los modelos no paramétricos permiten estimar las probabilidades de transición o de paso de un estado a otro, sin requerir hipótesis suplementarias.

Resulta evidente que mientras más estados diferentes hagan intervenir esos modelos, mayor será su precisión. El modelo univariado, que se utiliza en la demografía clásica, resulta en ese caso insuficiente: mezcla los comportamientos de los individuos en situaciones muy diferentes respecto de otros fenómenos que interfieren con el que se estudia. Sin embargo, ése fue el modelo que se usó principalmente hasta periodos recientes.

Ahora conviene emplear un modelo bivariado, que permite analizar de manera muy fina las interferencias entre dos fenómenos. Tales interferencias, que ya se han revelado como muy complejas, permiten una mejor comprensión de los comportamientos humanos.

También hemos puesto en evidencia dependencias de diversos tipos. El nivel más débil de dependencia es aquel en donde el primer fenómeno es independiente del segundo y éste lo es a su vez del primero. Esta *independencia total* entre varios fenómenos hasta el momento jamás ha sido observada en el análisis de la encuesta "Triple biografía", y resulta muy rara en las poblaciones humanas, lo cual revela que los diversos aspectos sociológicos, económicos, políticos, etc. de los fenómenos humanos están estrechamente relacionados entre sí.

Resulta mucho más interesante observar una *dependencia unilateral*. En ella se advierte que el haber experimentado uno de los eventos modifica la probabilidad de experimentar el segundo. Pero que, a la inversa, el haber experimentado el segundo no incide para nada en la probabilidad de experimentar el primero. Tal dependencia unilateral ha aparecido con mucha frecuencia en los análisis que hemos realizado a partir de los datos de la encuesta "Triple biografía". Así, en el caso de las mujeres, no hemos descubierto ninguna influencia del abandono del mundo agrícola sobre su nupcialidad, mientras que una vez casadas en el mundo agrícola, permanecerán allí mucho más tiempo que las solteras. En el caso de los hombres, se reveló una dependencia unilateral opuesta a la de las mujeres. Las posibilidades de casarse de ellos se duplican cuando salen del mundo agrícola (Courgeau y Lelièvre, 1986).

Por último, con frecuencia se ha observado también una *dependencia recíproca* entre dos fenómenos. Cada vez que uno se produce, se modifican las probabilidades de aparición del otro. Así, si la probabilidad de migrar hacia una zona fuertemente urbanizada disminuye después de cada nacimiento, los nacimientos de rango dos o más se reducen después de una migración hacia una metrópoli (Courgeau, 1987).

Puede ser que las dependencias sólo sean observadas en ciertas edades, o durante un periodo dado después del evento perturbador, o que incluso la influencia se invierta. De esa manera, cuando se estudian los nexos entre la fecundidad y la actividad femenina se observa que las mujeres inactivas al casarse y al ocurrir el primer nacimiento tienen una fecundidad fuertemente diferenciada según la edad: si a edades tempranas aquellas que desempeñan una actividad son tan fecundas como las que permanecen inactivas, después de los 30 años, el hecho de estar activa constituye un freno a un nacimiento suplementario (Lelièvre, 1987).

Finalmente, es posible poner en evidencia otros niveles de interpretación más complejos. Así, en el estudio de la interacción entre la fecundidad y la migración entre una metrópoli y una zona poco urbanizada se reveló una modificación importante de la fecundidad de rango dos y más. Se plantea, sin embargo, el problema de saber si se trata de un comportamiento de adaptación o de selección: adaptación si es la migración la que induce una modificación de la fecundidad de los migrantes; selección si en la zona de partida se observa un comportamiento de fecundidad diferente entre los futuros migrantes y los sedentarios. Una vez más los métodos que se presentan en este libro permiten probar las diferencias entre los futuros migrantes y los sedentarios definitivos en la población de partida. Es posible demostrar que en Francia la hipótesis de la *selección* se verifica en el caso de la migración hacia las metrópolis. Efectivamente, los futuros migrantes ya tienen una fecundidad débil respecto a la de los sedentarios de las zonas poco urbanizadas, siendo dicha fecundidad semejante a la de las mujeres que ya migraron (Courgeau, 1987).

De esta manera se manifiesta una *dependencia a priori* de la fecundidad sobre la migración que vendrá, lo que se traduce en una selección en el seno de la población inicial. Por otra parte, se observa un aumento de la fecundidad de las mujeres que migran hacia zonas menos urbanizadas. Gracias a una investigación idéntica a la anterior, esta vez pudimos percatarnos de un comportamiento de *adaptación* de la fecundidad de las mujeres que migran fuera de las metrópolis. Su comportamiento fecundo anterior no difiere en nada del de las citadinas que permanecen definitivamente en esas metrópolis (Courgeau, 1987).

Esos métodos bivariados se pueden completar mediante métodos trivariados o multivariados, que también hemos presentado en este libro. Si bien la estimación de esos modelos no plantea problemas teóricos complicados, en la práctica los efectivos observados generalmente no permiten que el análisis no paramétrico llegue muy lejos por esta vía. En efecto, el número de cocientes instantáneos por estimar crecerá rápidamente a medida que aumente el número de situaciones posibles. Las poblaciones sometidas a riesgos muy pronto van a resultar insuficientemente numerosas como para permitir una estimación precisa.

Cuando se pretende introducir la heterogeneidad de las poblaciones observadas, se plantea un problema similar. Para continuar utilizando métodos no paramétricos es necesario fragmentar la población en estudio en subpoblaciones suficientemente homogéneas respecto de las diversas características cuyo efecto queremos revelar: orígenes sociales, número de hermanos y hermanas, rango de nacimiento, diplomas, etc. Una vez más, los efectivos de las subpoblaciones observadas van a disminuir rápidamente al aumentar el número de características que se toman en cuenta. Cuando esos grupos se reducen y desaparecen gradualmente resulta imposible extraer conclusiones.

Por lo tanto, es preciso utilizar otros métodos para analizar esa heterogeneidad.

B) El tratamiento de la heterogeneidad de las poblaciones

Los eventos que surgen en el curso de la vida de un individuo no son los únicos constituyentes de su trayectoria personal. Numerosas características más son propias del individuo desde su nacimiento o éste las va adquiriendo durante su infancia; son elementos importantes que pueden actuar de manera diferente sobre el curso de su existencia.

Los orígenes familiares del individuo desempeñan así un papel relevante. El hecho de tener un padre agricultor, obrero o funcionario superior sitúa a un individuo desde su nacimiento en determinado medio, cuya influencia sobre su carrera futura es evidente. De igual manera, ser el hijo mayor o el menor, ser hijo único o tener muchos hermanos y hermanas, ser niño o niña, nacer en un medio rural o en una ciudad, etc., son algunas de las tantas características que van a ejercer influencia sobre la vida futura del individuo.

Esta heterogeneidad puede, además, generalizarse al caso en el que se estudia el devenir de un individuo a partir de un instante inicial que no es su nacimiento. De esa manera, cuando se estudia la sucesión de las migraciones, ese instante inicial corresponde al momento de instalarse en una nueva residencia. A partir de entonces, todas las características del individuo al inicio de su permanencia: edad, estado matrimonial, número de hijos, profesión, etc. se podrán asociar con la duración de la permanencia (Courgeau, 1985). Por supuesto que sus orígenes familiares pueden intervenir igualmente entre todas esas características.

Los métodos paramétricos se han revelado muy útiles para tratar una heterogenedidad como la arriba mencionada.

1) Modelos paramétricos

Esos modelos constituyen la generalización en el análisis de las duraciones de permanencia de los modelos de regresión que se utilizan, por ejemplo, en econometría. Éstos necesitan hipótesis más restrictivas que las de los modelos no paramétricos presentados antes. En efecto, hay que modelizar no sólo la duración de permanencia sino también el efecto de diversas características sobre la ocurrencia del evento estudiado.

En el capítulo VIII presentamos una gran variedad de modelos paramétricos que permiten resumir la distribución de las duraciones de permanencia mediante un pequeño número de parámetros. Esos modelos proporcionan cocientes instantáneos uniformemente crecientes, decrecientes o constantes en el curso del tiempo, así como cocientes instantáneos que pasan por un máximo antes de decrecer. En demografía se observan todos esos tipos de distribución. Las distribuciones multimodales, además, se pueden construir combinando varias de las distribuciones precedentes.

Cuando varias de esas distribuciones se adaptan con la misma precisión a una distribución observada, es preferible escoger aquella cuyo cociente instantáneo y función de permanencia tienen formas explícitas y simples. Esto facilita grandemente la estimación de los parámetros en función de los datos, cuando éstos están en parte truncados.

Una vez modelizada la duración de permanencia se hará intervenir el efecto de las diversas características individuales sobre los cocientes instantáneos observados, para lo cual hemos presentado dos tipos principales de modelos.

El modelo de riesgos proporcionales supone que los cocientes instantáneos de los individuos que tienen una característica dada son proporcionales a los de quienes no la tienen, siendo ese coeficiente de proporcionalidad el mismo para todas las duraciones. Vemos que se trata de una hipótesis muy fuerte que deberá probarse sobre todas las características consideradas. Cuando no se verifica es necesario dividir las poblaciones en subpoblaciones, sobre las que se estiman dos modelos de riesgos proporcionales para todas las demás características. Igualmente, se determinará si otros modelos se adaptan mejor a esos datos. Podemos dar como ejemplo nuestro análisis de los cambios de residencia en función de más de 30 características de los individuos al inicio de la permanencia. Debido a que según el sexo del individuo se obtenían cocientes instantáneos que no se podían considerar proporcionales, tuvimos que separar la muestra en dos submuestras, distinguiendo a los hombres de las mujeres (Courgeau, 1985).

Por su parte, el modelo de tiempo de ocurrencias aceleradas supone que las características actúan directamente sobre el tiempo vivido por los individuos. Así, quienes tengan una característica dada van a vivir el evento

estudiado de manera acelerada o retardada respecto de aquellos que no la tengan.

Estos dos tipos de modelos parecen adaptarse mejor a los comportamientos humanos, y se han podido utilizar exitosamente en numerosos estudios. Por supuesto que se pueden utilizar otros tipos de modelos, y en el futuro quizás algunos se revelarán como más completos que el modelo de riesgos proporcionales o el de ocurrencias aceleradas. En ocasiones hemos empleado un modelo lineal que agrega una constante positiva o negativa al cociente instantáneo cada vez que el individuo presenta una característica dada. Al contrario del modelo de riesgos proporcionales, en ese caso un individuo puede tener un cociente instantáneo estimado negativo. Tal posibilidad revela una desventaja del modelo lineal.

En todos los casos, esos modelos paramétricos sólo pueden hacer que intervengan las características que se observaron en la encuesta. Ahora bien, es posible pensar que otras características más difíciles de observar o de medir en una encuesta, o incluso algunas que el investigador no piensa que podrían influir sobre el evento estudiado, tengan un efecto no despreciable sobre los cocientes instantáneos. Así pues, se corre el riesgo de que esta *heterogeneidad no observada* afecte los parámetros que miden el efecto de las características observadas. De ahí la idea de introducir una distribución paramétrica o incluso no paramétrica de esta heterogeneidad no observada, que actúa de manera multiplicativa sobre la distribución de la duración de permanencia. Podemos demostrar que, en esta condición, es posible estimar los nuevos valores de los parámetros que corresponden a las características observadas y que toman en cuenta la heterogeneidad no observada. Parecería que según la distribución supuesta de la heterogeneidad no observada (Heckman y Singer, 1985), o incluso según la distribución paramétrica que se toma para estimar el efecto de las características (Trusell y Richard, 1987), los parámetros estimados pueden variar enormemente e incluso ser de signo contrario.

Estos resultados conducen a privilegiar la aproximación semiparamétrica que desarrollaremos ahora, y para la que se dispone de resultados más precisos en cuanto al efecto de la heterogeneidad no observada sobre la estimación de los parámetros. Asimismo, esta aproximación permite tomar en cuenta simultáneamente la heterogeneidad de las poblaciones y las interacciones entre los fenómenos.

2) Modelos semiparamétricos

Estos modelos son mucho más flexibles que los precedentes, pues no modelizan en forma paramétrica el efecto de la duración de permanencia sobre

el cociente instantáneo. En cambio, introducen un efecto semejante al de las diversas características observadas. Tendremos entonces, en forma similar, modelos de riesgos proporcionales, modelos de tiempo de ocurrencias aceleradas o cualquier otro tipo de nexo entre la duración de permanencia y las características observadas.

En el capítulo IX mostramos cómo estimar el efecto de las características mediante el método de verosimilitud parcial, y el efecto no paramétrico de la duración de permanencia sobre el cociente instantáneo.

Esos modelos ofrecen cierto número de ventajas sobre los precedentes.

En primer lugar ha sido posible determinar de manera teórica hasta qué punto la omisión de características en un modelo como ése afecta los parámetros estimados de las características observadas (Bretagnolle y Hubert-Carol, 1988). Se demostró que tal omisión no afecta el signo de los parámetros estimados, pero sí implica una reducción del valor absoluto de esos parámetros. Esto significa que si el efecto de una característica parecía importante cuando se omitía el de otras independientes, la introducción de éstas en el modelo semiparamétrico sólo reforzaba el efecto de la primera característica. En cambio, ciertas características que parecían no tener ningún efecto significativo podían volverse completamente significativas cuando se introducían las no observadas inicialmente.

Esos resultados son muy importantes, pues nos proporcionan seguridad sobre el sentido de los efectos observados, aun cuando no sepamos si hemos introducido en el modelo todas las características que afectan la duración de permanencia.

En segundo lugar, ese modelo simplifica la introducción fácil y simultánea de las interacciones entre los fenómenos demográficos y la heterogeneidad de las poblaciones observadas. De esa manera, en el modelo de tipo bivariado se pueden introducir cocientes iniciales de riesgo distintos para los diversos tipos de eventos, o cocientes instantáneos proporcionales entre ellos, cuando el análisis no paramétrico ha mostrado que ello se ha verificado convenientemente. Hemos optado por esta última solución al analizar los nexos entre la nupcialidad, el abandono de la agricultura y las diversas características individuales (Courgeau y Lelièvre, 1986).

También existe otra posibilidad para facilitar que intervengan interacciones más complejas: aquella en que las características se modifican según la duración de permanencia. Cabe mencionar aquí que la utilización de un modelo puramente paramétrico permite introducir de manera semejante algunas características dependientes del tiempo transcurrido.

C) Nuevas líneas de investigación

La recolección de biografías y su análisis con los métodos presentados aquí abren un vasto campo de investigación que apenas hemos empezado a explorar. Hay que tomar en cuenta, sin embargo, que ese campo está situado dentro de una corriente mucho más general que abarca todas las ciencias humanas. En la actualidad encontramos entre los antropólogos, los psicólogos, los sociólogos, etc., investigadores que analizan historias de vida, aplicando cada uno hipótesis y métodos muy diferentes que son propios de cada disciplina. Incluso el material de base, la biografía, a veces se recoge de maneras tan distintas que a un investigador le costaría mucho esfuerzo utilizar las biografías recolectadas por un colega de otra disciplina.

No obstante, nos parece importante que cada una de las ciencias se abra hacia las otras. Uno de los objetivos de este libro es mostrar con claridad el aporte original de la demografía sobre ese tema.

En primer lugar, varias veces hemos indicado la dificultad e incluso la imposibilidad que tiene un demógrafo de recolectar y tratar características psicológicas del individuo o características sociológicas de un grupo. Eso introduce una heterogeneidad no observada, que, tal como vimos, plantea numerosos problemas. Consideramos que es preciso realizar una investigación común sobre ese tema, pues resultaría muy enriquecedora para las ciencias humanas.

Una segunda línea de investigación consistiría en emprender una reflexión conjunta sobre las bases teóricas de nuestros planteamientos. Sería útil averiguar si existe un tronco teórico común a las diversas ciencias humanas que analizan las historias de vida, y tratar de delimitarlo. Podemos preguntarnos qué es, de manera precisa, lo que cada una de estas ciencias explora y por qué parece difícil enlazar esas diversas perspectivas acerca de un mismo fenómeno. El antropólogo, el demógrafo, el psicólogo, el sociólogo, etc., interrogan al mismo individuo pero desde puntos de vista y con objetivos diferentes. Una vez que recogen su imagen, cada uno parece trabajar sobre un objeto distinto. Lo que importa analizar a partir de ahora son los nexos entre esos objetos.

También resulta interesante ver si se pueden generalizar los métodos de análisis biográfico a las estructuras sociales más complejas. La historia de una familia, de una empresa, de una nación, etc., ¿se puede analizar con los mismos métodos aplicados a una historia de vida individual? Si bien entre esas historias existen similitudes —los eventos individuales afectan tanto a una familia como a una empresa—, la unidad de análisis es mucho más compleja, pues una familia o una empresa están compuestas por numerosos individuos cuyas relaciones y cambios en esas relaciones forman parte de los eventos que se van a analizar. Éste es un dominio cuya exploración es importante.

Estas conclusiones revelan, finalmente, que los análisis que hemos desarrollado a lo largo de este libro trascienden lo meramente individual, pues tienden hacia una comprensión más profunda de las sociedades humanas. En efecto, esas biografías individuales comprenden una información muy rica acerca de la sociedad que las conforma, pero que es igualmente modificada por ellas. Aquí intentamos explorar al máximo esta información individual, para permitir el esclarecimiento de los comportamientos humanos de una manera nueva.

ANEXO 1

I. Correspondencia de los términos en inglés, francés y español

1) *Accelerated failure time model*
Modèle à sorties accélérées
Modelo de ocurrencias aceleradas

2) *Baseline hazard underlying rate*
Quotient instantané sous-jacent
Cociente instantáneo de riesgo inicial

3) *Censored data*
Données tronquées
Datos truncados

4) *Censored intervals*
Intervalles ouverts, intervalles tronqués
Intervalos abiertos, intervalos truncados

5) *Censoring time*
Sortie d'observation
Salida de la observación

6) *Censorship*
Troncature
Truncamiento

7) *Competing risks*
Risques multiples, concurrents
Riesgos múltiples, concurrentes o en competencia

8) *Density probability function*
Densité de probabilité
Función de densidad de probabilidad

9) *Event history analysis*
Analyse de biographies
Análisis de biografías

10) *Failure time distribution*
Distribution des temps de sortie
Distribución de las fechas de ocurrencia de los eventos

11) *Hazard function*
Densité conditionnelle, fonction d'intensité
Densidad condicional, función de intensidad, función de riesgo, función de probabilidad de densidad condicional de mantenerse en el estado inicial

12) *Hazard rate/instantaneous failure rate*
Quotient instantané (d'occurrence ou de passage)
Cociente instantáneo (de ocurrencia del evento o de la transición de un estado a otro), probabilidad instantánea de ocurrencia del evento

13) *Integrated or comulative hazard function*
Quotients cumulés
Función de intensidad acumulada, integral de la densidad condicional, cocientes acumulados, cocientes instantáneos de ocurrencia acumulados

14) *Life tables*
Tables de séjour (de survie)
Tablas de permanencia o supervivencia en un estado

15) *Partial likelihood*
Vraisemblance partielle
Verosimilitud parcial

16) *Product limit estimate*
Estimateur de Kaplan-Meier
Estimador de Kaplan-Meier

17) *Proportional hazard model*
Modèle à risques proportionnels
Modelo de riesgos proporcionales

18) *Proportional hazard regression model*
Modèle de regression à risques proportionnels
Modelo de regresión de riegos proporcionales

19) *Right (left) censored intervals*
Intervalles ouverts à droite (à gauche)
Intervalos abiertos a la derecha (a la izquierda)

20) *Survival time*
Durée de séjour
Duración de permanencia o supervivencia (en un estado)

21) *Survivor function s(t)*
Fonction de séjour (de survie)
Función de permanencia o supervivencia (en un estado)

II. Fórmulas básicas

Función de permanencia

- En tiempo continuo

$$S(t) = P(T \geq t) \quad S(0) = 1$$

$$S(t) = \int_t^{\infty} f(s)ds$$

$$S(t) = \exp\left(-\int_0^t h(s)\,ds\right) = \exp\left(-H(t)\right)$$

- En tiempo discreto

$$S(t) = \sum_{i|t_i \geq t} P(T = t_i)$$

$$S(t) = \prod_{t_i < t} (1 - h_i)$$

Función de densidad de probabilidad

- En tiempo continuo

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

$$f(t) = -\frac{dS(t)}{dt} = -S'(t)$$

$$f(t) = h(t)\exp\left(-\int_0^t h(s)ds\right) = h(t)\exp\left(-H(t)\right)$$

- En tiempo discreto

$$f(t_i) = P(T = t_i)$$

Cocientes instantáneos de ocurrencia, función de intensidad, función de densidad de probabilidad condicional

- En tiempo continuo

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t P(T \geq t)}$$

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{-dS(t)}{dt}}{S(t)} = -\frac{d\text{Log}S(t)}{dt}$$

- En tiempo discreto

$$h(t) = \sum_i h_i \delta(t - t_i)$$

donde $\delta(0) = 1$

$$\delta(x) = 0, \forall x \neq 0$$

y $\quad h_i = f(t_i) / [f(t_i) + f(t_{i+1})) + \ldots]$

tenemos igualmente el estimador de la *intensidad del proceso*

$$H(t) = \sum_{i|t_i < t} \text{Log}(1 - h_i)$$

si los h_i son pequeños

$$H(t) \simeq \sum_{i|t_i > t} h_i$$

Caso bivariado

Sean T_1 y T_2 las duraciones en que se producen dos eventos.

- Cocientes instantáneos

$$h_{0i}(t) = \lim_{\Delta t \to 0} \frac{P(T_i < t + \Delta t | T_i \geq t, T_j \geq t)}{\Delta t}$$

$$h_{ji}(t|u) = \lim_{\Delta t \to 0} \frac{P(T_i < t + \Delta t | T_j = u, T_i \geq t)}{\Delta t} \qquad u \leq t$$

- Función de densidad de probabilidad

$$f_i(t_i) = h_{0i}(t_i) \exp - \int_0^{t_i} \left(h_{0i}(u) + h_{0j}(u)\right) du$$

$$f_j\left(t_j|t_i\right) = h_{ij}\left(t_j|t_i\right) \exp\left(-\int_{t_i}^{t_j} h_{ij}\left(u|t_i\right) du\right) \qquad t_i \leq t_j$$

de donde la densidad conjunta:

$$f\left(t_i, t_j\right) = h_{0i}(t_i) h_{ij}\left(t_j|t_i\right) \exp\left[-\int_0^{t_i} \left(h_{0i}(u) + h_{0j}(u)\right) du - \int_{t_i}^{t_j} h_{ij}\left(u|t_i\right) du\right] \qquad t_i \leq t_j$$

III. Recordatorios y demostraciones

Producto integral

El producto integral se define a la derecha de manera análoga a una integral de Riemann: el intervalo [0, t] está subdividido en pequeños subintervalos $[0 = x_0, x_1[; [x_1, x_2[; \ldots; [x_{n-1}, x_n = t[$

sea $\xi_j \in [x_j, x_{j+1}[$ se considera como producto integral el límite

$$\lim_{n \to \infty} \prod^{n} \left(1 - h\left(\xi_j\right)\left(x_{j+1} - x_j\right)\right) \text{ con } \max\left(x_{j+1} - x_j\right) \to 0$$

donde $h\left(\xi_j\right)\left(x_{j+1} - x_j\right)$ es h_k si $[x_j, x_{j+1}[$ contiene un punto de apoyo de la función.

Prueba de una hipótesis nula, prueba general

Si bien existen pruebas muy concluyentes para problemas simples asociados con las distribuciones exponenciales, algunas veces en casos muy particulares donde se necesitan métodos más generales, es preciso hacer estimaciones.

Sea $f_Y(y;\theta)$ la densidad de la variable aleatoria Y, y la hipótesis nula: $\theta=\theta_0$ con sus alternativas $\theta>\theta_0$.

Si se considera la alternativa particular $\theta_A=\theta_0+\delta$ δ pequeño, entonces la relación de verosimilitud se define:

$$r_{A0}=f_Y(y;\theta_A)/f_Y(y;\theta_0)$$

Recordatorio: bajo $H_0:\theta=\theta_0$, se considera que $r_{A0}(Y)$ es una variable aleatoria continua tal que, cualquiera que sea $\alpha, 0<\alpha<1$, existe un único tal c_α que:

$$P(r_{A0}(Y)\geq c_\alpha; H_0)=\alpha$$

El Lema de Neyman Pearson indica que para un α dado, la región crítica definida por $r_{A0}(y)\geq c_\alpha$ es la mejor.

El logaritmo de la relación de verosimilitud se define igualmente

$$\log r_{A0}=\log\left(f_Y(y);\theta_0+\delta\right)/f_Y(y;\theta_0)$$

$$\log r_{A0}=\delta\, d\log\left[f_Y(y);\theta_0\right]/d\theta_0$$

conforme a la hipótesis de que f_Y es continua y derivable.

Para δ suficientemente pequeño, se obtiene entonces la región crítica de la relación de verosimilitud para valores importantes de

$$U.(\theta_0)=d\log\left[f_Y(y);\theta_0\right]/d\theta_0$$

llamado el *score* de eficacia para Y.

En las aplicaciones, si los componentes $Y_1 \ldots Y_n$ son independientes entonces el logaritmo de la verosimilitud es la suma de las contribuciones de cada una de las variables aleatorias, y se puede escribir:

$$U.(\theta_0)=\Sigma U_i(\theta_0) \text{ donde } U_i=d\,\mathrm{Log}\left[f_{Y_i}(y_i);\theta_0\right]/d\theta_0$$

El espacio en cuyo seno se formaliza el problema debe ser regular: se puede diferenciar e integrar, pues no existe discontinuidad en la distribución inicial. Se muestra entonces que:

$$E\left(U.(\theta_0);\theta_0\right)=0 \text{ de donde}$$

$$\text{var}\left(U.(\theta_0);\theta_0\right)=E\left(U^2.(\theta_0);\theta_0\right)=I.(\theta_0)$$

donde $I.(\theta_0)$ es la matriz de información de Fisher, información sobre θ contenida en Y.

Si las Y_i son independientes entonces $I.(\theta_0)=\sum I_i(\theta_0)$ y

$$I_i(\theta_0)=\text{var}\left(I_i(\theta_0);\theta_0\right)=E\left[-d^2\log\left(f_{Y_i}(y_i);\theta_0\right)/d\theta_0^2;\beta_0\right]$$

Normalidad asintótica

Sean $Y_1 \ldots Y_n$ variables aleatorias independientes equidistribuidas de densidad $f(y)$, se tiene entonces

$$E\left(U(\theta);\theta\right)=0$$

$$E\left(U^2(\theta);\theta\right)=E\left(U'(\theta);\theta\right)=I(\theta)>0 \qquad (1)$$

Vamos a demostrar que es $\sqrt{n}(\hat{\theta}-\theta)$ es $N\left(0,1/I(\theta)\right)$

$\hat{\theta}$ es un estimador convergente ("consistente") hacia θ, el desarrollo en serie de Taylor de $U(\theta)$ respecto del valor exacto de θ es:

$$U.(\hat{\theta})=U.(\theta)+(\hat{\theta}-\theta)\,U.(\theta)+1/2(\hat{\theta}-\theta)^2U''.(\theta^*) \qquad (2)$$

donde

$$\left|\theta^*-\theta\right|<\left|\hat{\theta}-\theta\right|$$

y

$$U.(\hat{\theta})=d\log f(y;\hat{\theta})/d\hat{\theta}=0$$

pues $\hat{\theta}$ es el máximo de verosimilitud.

La fórmula (2) se vuelve entonces

$$1/\left(\sqrt{n}I(\theta)\right)\sum U_i(\theta)=\sqrt{n}(\hat{\theta}-\theta)\left[-\sum U_i'(\theta)/nI(\theta)+e_n\right] \qquad (3)$$

donde $\left|e_n\right|\le\left|\hat{\theta}-\theta\right|g(y)/I(\theta)=Op(1)g(y)$ es una función integrable y es una medida de $nU''.(\theta^*)$ próxima a $\hat{\theta}=\theta$.

La ley débil de los grandes números se aplica a la relación que tiende entonces hacia $1+Op(1)$, (3) se vuelve

$$1/\left(\sqrt{n}I(\theta)\right)\Sigma U_i(\theta) = \sqrt{n}(\hat{\theta}-\theta)\,[1+Op(1)]$$

la ley fuerte de los grandes números se aplica a la izquierda y encontramos entonces que $\sqrt{n}(\hat{\theta}-\theta)$ sigue una $N\left(0, 1/I(\theta)\right)$ por el hecho de (1).

Relación entre los procesos de conteo y el análisis de regresión de los intervalos

• Ley de probabilidad que gobierna el número de eventos observados

Sea N_t el número de eventos que se producen en un intervalo de tiempo de cualquier duración t, N_t sigue una ley de Poisson de parámetro λt:

$$P(N_t = n) = e^{-\lambda t}(\lambda t)^n / n! \qquad n \in \mathbb{N}$$

La función generadora de los momentos correspondientes se escribe:

$$g_{N_t}(s) = \Sigma s^n P(N_t = n) = \exp\left(\lambda t(s-1)\right)$$

en consecuencia

$$E(N_t) = g'_{N_t}(1) = \lambda t$$

$$\text{var}(N_t) = g''_{N_t}(1) + g'_{N_t}(1) - [g'_{N_t}(1)]^2 = E(N_t) = \lambda t$$

por tanto N_t/t converge en probabilidad hacia λt, lo que justifica la denominación de cociente instantáneo de ocurrencia.

• Relación con el análisis de las duraciones de permanencia o supervivencia

El parámetro λt del proceso de Poisson se transformará, a fin de que evolucione conforme a una tendencia que se va a estimar.

Si la variable aleatoria X representa el intervalo que separa el origen del primer evento $P(X > x) = P(N_x = 0) = \exp(-\lambda x)$

de función de repartición $F_x = 1 - \exp(-\lambda x) \qquad x \geq 0$

de densidad $f_x(x) = \lambda \exp(-\lambda x)$ exponencial de parámetro λ.

Para representar una tendencia se va a considerar una forma particular del parámetro; sea $\lambda(t) = \exp(a + bt)$ lo que implica que $\lambda(t)$ no puede ser negativo, así la probabilidad de que partiendo de t_i el evento siguiente se produzca en el intervalo $[t_{i+1}, t_{i+1} + \Delta_{i+1}[$ es

$$\lambda(t_{i+1})\exp\left[-\int_{t_i}^{t_{i+1}}\lambda(u)du\right]\Delta t+0(\Delta t)$$

Si los eventos se producen en las fechas $(t_1, \ldots t_n)$ entonces la verosimilitud es de la forma:

$$L=\lambda(t_1)\exp\left(-\int_0^{t_1}\lambda(u)du\right)\cdot\lambda(t_2)\exp\left(-\int_{t_1}^{t_2}\lambda(u\right)du)\ldots$$

$$L=\left(\prod_{i=1}^{n}\lambda(t_i)\right)\exp\left(-\int_0^{tn}\lambda(u)du\right)$$

Planteo de Cox para la estimación de la verosimilitud parcial

• Verosimilitud

Las duraciones de permanencia de n individuos están distribuidas independientemente según

$$h(t)=h_0(t)\exp\left(\beta z_i(t)\right)$$

donde $h_0(.)$ y β son desconocidos, y $z_i(t)\in\mathbb{R}^k$ es el vector de las características del $i^{\text{ésimo}}$ individuo en la fecha t.

La verosimilitud se vuelve entonces:

$$L=\prod_i\left[\exp\left(\beta z_i(t_i)\right)h_0(t_i)\exp\left(-\int_0^{t_i}\exp\left(\beta z_i(u)\right)h_0(u)du\right)\right]$$

definimos $Y_i^{\beta}=\exp\left(\beta z_i(u)\right)1_{\{u\le t_i\}}$ y $Y^{\beta}=\sum_i Y_i^{\beta}$

entonces

$$L=\left[\prod_i Y_i^{\beta}(t_i)h_0(t_i)\right]\exp-\int_0^{\infty}Y^{\beta}(u)h_0(u)du$$

$$L=\prod_i\frac{Y_i^{\beta}(t_i)}{Y^{\beta}(t_i)}\left[\prod_i Y^{\beta}(t_i)h_0(t_i)\right]\exp-\int_0^{\infty}Y^{\beta}(u)h_0(u)du$$

$\exp-\int_0^{\infty}Y^{\beta}(u)h_0(u)du$ es igual a 1 si no queda ningún individuo sometido a riesgo después de la última ocurrencia del evento.

Cox plantea usar la primera parte de la verosimilitud para estimar β

$$PL(\beta) = \prod_i \frac{Y_i^{\beta}(t_i)}{Y^{\beta}(t_i)} = \prod_i \left[\frac{\exp(\beta z_i(t_i))}{\sum_{j|\ t_j \le t_i} \exp(\beta z_j(t_i)} \right]$$

Esto implica no considerar los términos que contienen la información sobre el lapso entre cada ocurrencia del evento sucesiva, de allí el nombre de verosimilitud parcial y su nexo estrecho con la verosimilitud marginal de los rangos.

La verosimilitud parcial puede entonces ser interpretada como un producto de verosimilitudes, no para todos los individuos como en el caso de la verosimilitud total o habitual, sino para todas las ocurrencias del evento.

La cuestión, entonces, es la siguiente: si se sabe que un evento se produce en la fecha t_i ¿cuál es la probabilidad (la verosimilitud) de que le concierna más al individuo r que a otro?

Es decir,

$$PL_r = \frac{h_r(t_i)}{h_r(t_i) + h_{r+1}(t_i) + \ldots + h_n(t_i)}$$

el valor de cada PL_r va a depender así del rango de la fecha más que del valor de las diferentes duraciones de permanencia.

Se va a maximizar a PL_r usando técnicas similares a las que se utilizan para maximizar las verosimilitudes totales o usuales (Newton-Raphson, por ejemplo).

Proceso de conteo y martingalas (Aalen)

El proceso $N(t)$ cuenta los eventos ocurridos en el curso de $[0, t[$. Es univariado si el fenómeno es único, y multivariado $(N_i)\, i = 1 \ldots k$ si se trata de una colección de k procesos de conteo, que pueden ser independientes. Si $\{F_t\}$ la σ-álgebra[1] es el conjunto de todos los eventos observados durante $[0, t[$; es la historia completa del proceso hasta t.

La intesidad del proceso $N(t)$ está dada por:

$$\Lambda_i(t) = \lim_{h \to 0} \frac{1}{h} E\left(N_i(t+h) - N_i(t) \mid \{F_t\}\right)$$

En el marco de la modelización markoviana donde el proceso $N_i(t)$ enumera las transiciones fuera de un estado i en el curso de $[0, t]$, si se

[1] $\{F_t\}$ es una familia de partes de Ω (conjunto de cronologías posibles) estables para la unión y la reunión así como para la complementación.

establece $Y_i(t)$, como el número total de individuos en i justo antes de t, el proceso de intensidad se expresa en función de $Y_i(t)$ y de un componente paramétrico a estimar $h(t)$ dependiente del tiempo[2] según un modelo multiplicativo:

$$\Lambda_i(t) = h_i(t)Y_i(t)$$

Justificando la aproximación como aplicación de la teoría basada en las martingalas e integrales estocásticas (Aalen, 1975), se pueden en efecto considerar los aumentos siguientes:

$$dN_i(t) = N_i(t+dt) - N_i(t) \text{ de esperanza } \Lambda_i(t)dt;$$

entonces $dM_i(t) = dN_i(t) - \Lambda_i(t)dt$ es un proceso estocástico de media igual a cero, de donde los

$$M_i(t) = N_i(t) - \int_0^t \Lambda_i(u)du$$

son martingalas integrales, ortogonales ($\in \mathcal{M}^2$). En efecto, existe un proceso único A_i tal que:

$$M_i = N_i - A_i \in \mathcal{M}^2$$

y $-dA_i(t)/dt$ se llama la intensidad de n en $\{F_t\}$

El número de eventos $N_i(t)$ durante $[0, t]$ se estima mediante

$$N_i(t) = \int_0^t \Lambda_i(s)ds$$

llamado compensador del proceso de conteo. $N_i(t)$ es pues la suma del compensador y de un ruido aleatorio que es una martingala.

IV. *Softwares* disponibles

Nuestra intención no es realizar aquí una revisión sistemática de todos los *softwares* existentes en este campo del análisis. En efecto, un gran número de ellos —utilizado regularmente por nuestros colegas extranjeros— no está disponible en Francia.[3] Por lo tanto, esta presentación dista de ser exhaustiva.

[2] El parámetro del tiempo está constituido por la edad o cualquier otra variable de la duración.

[3] Citemos, por ejemplo, Survreg de Preston y Clarkson (1983) que estima modelos paramétricos y de riesgos proporcionales; PHGLM, un programa disponible en SAS (1983) pero que

Esencialmente consideraremos cinco *softwares* disponibles en todo el mundo: Lifetest (SAS), Lifereg (SAS), GLIM (NAG), PL1(BMDP) y PL2 (BMDP), y dos *software* desarrollados en el INED y disponibles en STATA (Bocquier, 1996; Lelièvre y Bringi, 1998).

Dos softwares *de análisis no paramétrico*

Lifetest (SAS) y PL1 (BMDP) ofrecen casi las mismas facilidades. En el capítulo IV.C vimos los resultados de un análisis hecho con Lifetest, así que no detallaremos los cálculos que se efectúan con esos *softwares* puesto que ya los presentamos ampliamente en ese capítulo. Daremos aquí el ejemplo de la llamada al procedimiento y algunas instrucciones.

PROC LIFETEST	→	llama al procedimiento
TIME	→	designa las variables que miden duraciones de permanencia e índices de truncamiento
STRATA	→	divide la muestra en subgrupos
TEST	→	llama a diferentes pruebas estadísticas en cada estrato
FREQ	→	indica la frecuencia de la aparición de cada variable
BY	→	sirve igualmente para dividir la muestra en subgrupos, pero no hará ninguna prueba sobre la homogeneidad de éstos.

Para obtener las curvas de permanencia de los ejemplos del capítulo IV se dieron las siguientes instrucciones:

```
DATA don1 ; SET survie.grp ;
PROC SORT DATA = don1 ; BY sexe grouprof;
PROC LIFETEST ;
TIME dur1*cens1 ;
STRATA grouprof ;
BY sexe ;
```

También es muy sencillo utilizar el procedimiento PL1 de BMDP, ya que no hay necesidad de preparar específicamente los datos.

---

no ha sido ni sustentado ni documentado por ellos, pues fue elaborado por los usuarios. El *software* hace la estimación de la verosimilitud parcial en el caso de modelos de riesgos proporcionales y parece ser muy fácil de emplear. Loglin, documentado en *Loglin 1.9 User's Guide* (1976) de Olivier y Neff (Harvard University Public Health Science Computer Facility) es ampliamente utilizado por J. Hoem. Por último, Censor, de Meeker y Duke (1981), estima sólo modelos paramétricos.

El software *para el análisis de las interacciones entre dos eventos: ROOT (INED)*

Los *softwares* clásicos que hemos utilizado para el estudio de un solo evento (SAS, BMDP) no proponen un estudio de las interacciones entre dos eventos; por lo tanto, en el INED hemos desarrollado un *software* de análisis no paramétrico de las interacciones que permite llevar a cabo este estudio.

Al confrontar lo concerniente a las pruebas y a los casos de simultaneidad hemos tratado de preservar un máximo de opciones con la intención de que el investigador disponga de una herramienta flexible capaz de facilitarle una experimentación completa sobre los datos. Ese *software* documentado ya se ha aprovechado en numerosas aplicaciones. La Asociación Nacional del Software (Association Nationale du Logiciel), cuya función, entre otras, es apoyar la difusión de los *softwares* elaborados por investigadores de la comunidad científica, aseguró su difusión a partir de octubre de 1987.

En el anexo 2 presentamos los detalles de dicho *software*.

Software *de análisis paramétrico y semiparamétrico*

• La tarea de estimar modelos paramétricos con GLIM sigue una lógica particular (MCullagh y Nelder, 1983). La expresión más general de los cocientes por estimar es la siguiente:

$$h(t, z) = h_0(t) \exp(\beta z)$$

los que se explora es:

$$p_i(z) = P(T < t_i \mid T \geq t_{i-1}; z)$$

En GLIM todos los modelos propuestos especifican la forma de $p_i(z)$, ya que las probabilidades que investigamos son soluciones estándar que pertenecen al intervalo [0,1] y los modelos inicialmente podrían dar soluciones fuera de este intervalo. El nexo sistemático de igualdad entre los $p_i(z)$ y una función lineal de las variables explicativas es remplazado entonces por una transformación de la forma $F^{-1}(p)$ donde F^{-1} es una función continua monótona creciente de [0, 1] en $\mathbb{R}$. Generalmente se toma F^{-1} como inverso de una función de repartición conocida. F^{-1} se llama *link function*.

En GLIM disponemos de $F^{-1}(p) = \log p$ por *default* y también de:

$F^{-1}(p) = \log(p/1-p)$ nexo logit

$F^{-1}(p) = \phi^{-1}(p)$ ϕ función de reparto normal

$F^{-1}(p) = \log(-\log(1-p))$ complementaria log-log

La distribución complementaria log-log es la que mejor se adapta si pretendemos estimar una expresión semiparamétrica en el caso de un modelo de riesgos proporcionales (Diamond, McDonald y Shah, 1986). En efecto, si se plantea $q_i(z) = 1 - p_i(z)$, la probabilidad condicional de permanencia se puede escribir entonces:

$$q_i(z) = \exp -h(t; z)\,dt \quad \text{que se desarrolla como sigue}$$

$$q_i(z) = \exp -h_0(t)\exp(\beta z)\,dt$$

$$q_i(z) = \exp[-\exp(\beta z)h_0(t)\,dt]$$

de donde $$\log\left(-\log(1 - p_i(z))\right) = c_i + \Sigma\beta_j z_j$$

$$\text{donde } c_i = \log \int_{t_{i-1}}^{t_i} h_0(t)dt$$

Éste es el método desviado que permite hacer un análisis semiparamétrico, mediante un procedimiento que es inmediato en el caso del análisis paramétrico.

• Lifereg de SAS no comprende más que dos opciones de funciones además de la exponencial: Weibull y Gamma. Sin embargo, esas posibilidades son comparables a las de GLIM, así que no las detallaremos aquí.

• PL2 de BMDP es el único de los cuatro *softwares* que puede tomar en cuenta variables dependientes del tiempo; así pues, daremos aquí un ejemplo de programación posible. Sea el estudio de la influencia de diversas variables individuales sobre los cocientes instantáneos de reincorporación de un desempleado a un empleo de larga duración luego de una etapa de búsqueda de empleo. Escogemos un modelo simple del siguiente tipo:

$$\log h(t) = c(t) + \beta z + \beta' z'(t)$$

Las variables explicativas son la edad en que se produjo la primera interrupción de actividad, la edad en el momento de la encuesta, la región, el estatus matrimonial, los diplomas, la experiencia profesional, y la nacionalidad. Sean siete variables fijas y una dependiente del tiempo, que está constituida por la observación semana tras semana de las posibilidades ofrecidas al individuo (haya habido propuestas o no).

Ejemplo de programación:

```
\INPUT UNIT 6. CODE = CHOMEURS                                    (1)
\FORM TIME = WEEK. STATUS = OBS. RESPONSE = 1                     (2)
\REGRESSION
 COVARIATE = AGE1, AGE, REGION, MAR, DIPL, TRAV, NAT.             (3)
```

```
ADD = PROPSEM                                                    (4)
AUXILIARY = PROP1 TO PROP52                                      (5)
```

Subrutina Fortran que define la relación de la variable con el tiempo y que se debe compilar con el programa principal.

```
SUBROUTINE P2LFUN(Z, ZT, AUX, TIME, NFXCOV, NADD, NAUX, ISUB, X)  (6)
DIMENSION Z(7), ZT(1), AUX(52)                                   (7)
ZT(1) = AUX(TIME)                                                (8)
RETURN
END
```

(1) Leer el archivo CHOMEURS en la unidad 6.
(2) Declaracion de la unidad de tiempo WEEK, el indicador de truncamiento OBS; la observación se hace realmente si es igual a 1.
(3) Se especifica el modelo con siete variables fijas.
(4) Se añade la variable PROPSEM que depende del tiempo y aún no está definida.
(5) Declaración de 52 variables dicotómicas que valen 1 si el individuo ha tenido una propuesta y 0 si no. Esas variables se llaman AUX en la subrutina que sirve para formar la variable dependiente del tiempo.
(6) Declaración tipo de la subrutina, la variable *X* no documentada en el manual de BMDP corresponde al indicador de los valores faltantes no utilizado aquí.
(7) Z(7) corresponde a las variables fijas declaradas en COVARIATE, ZT(1) corresponde a la variable dependiente del tiempo, AUX(52) corresponde a PROP1 → PROP2.
(8) Se define la variable dependiente del tiempo PROPSEM como provista del valor de PROP en TIME que aquí es WEEK.

• Rate es un *software* que desarrolló N. Tuma. Fue concebido para estimar los modelos paramétricos o semiparamétricos del tipo siguiente.

1) El cociente es independiente de la duración, t, pero depende de características dadas bajo la forma de un vector z. Ese cociente se escribe entonces:

$$h(t) = \exp(z\beta)$$

donde la primera componente de z es igual a la unidad.

En cambio, puede haber varios tipos de estados de partida y llegada. Este modelo también se puede aplicar a los datos de panel.

2) El cociente es siempre independiente de la duración t, y depende de características dadas bajo la forma de un vector z. Ese modelo introduce una heterogeneidad no observada ε, cuya distribución es de tipo gamma. Puede

presentarse bajo forma aditiva o multiplicativa. En el primer caso, el cociente instantáneo se escribe:

$$h_{T|E}(t|\varepsilon) = \varepsilon(z\beta)$$

donde la función de densidad de probabilidad ε es:

$$f(\varepsilon) = \frac{\lambda}{\Gamma(\lambda)}(\lambda\varepsilon)^{\lambda-1}\exp(-\lambda\varepsilon) \text{ con } \Gamma(\lambda) = \int_0^{\infty} x^{\lambda-1}\, e^{-\lambda} dx$$

En el segundo caso tenemos:

$$h(t|\varepsilon = \varepsilon\exp(z\beta)$$

Puede haber varios tipos de estados de partida y llegada y el tiempo puede dividirse en periodos entre los cuales pueden variar los parámetros.

3) Se trata ahora de un modelo semiparamétrico cuyo cociente depende de las características. Se puede escribir:

$$h(t) = h_0(t)\exp(z\beta)$$

El programa Rate no estima el cociente $h_0(t)$, sino sólo el efecto de diversas características z sobre el cociente $h(t)$.

4) El cociente depende de la duración y de las características bajo la forma de un modelo de Gompertz-Makeham. Se puede escribir:

$$h(t) = \exp(z\beta) + \exp(z'\beta' + (z''\beta'')t)$$

Vemos que el cociente depende entonces de tres series de variables z, z' y z'', que pueden ser diferentes unas de otras, pero también contener variables comunes. La primera variable de z, z' y z'' es igual a la unidad. El programa permite igualmente estimar un modelo de tipo:

$$h(t) = z\beta + z'\beta'\exp((z''\beta'')t)$$

En todos los casos, el programa Rate proporciona la estimación de los parámetros β y de su matriz de las varianzas y covarianzas. Por tanto, eso permite efectuar todas las pruebas posibles sobre el valor de esos parámetros.

Software *de análisis semiparamétrico:* EVACOV

A fin de realizar un análisis semiparamétrico según la formulación descrita anteriormente, E. Lelièvre desarrolló en el INED un *software* específico. EVACOV es un programa escrito en Fortran que pone en marcha el análisis semiparamétrico. Este programa está inspirado en el de A. Chang, P. Wang y A. McIntosh (Kalbfleisch y Prentice, 1980) y permite tomar en cuenta datos truncados a la derecha y hacer que intervengan variables dependientes del evento perturbador.

Se trata, pues, de un modelo que toma en cuenta variables fijas[4] y variables dependientes del tiempo en un punto, que corresponde a la fecha del evento perturbador.

Al disponer de una muestra que contiene numerosas variables explicativas tenemos la posibilidad de tomarlas en cuenta una por una o en asociaciones diversas que el usuario escoge combinándolas entre sí a su antojo.

En efecto, este análisis prolonga el estudio no paramétrico, en el sentido en que se va a calcular la influencia de las características sobre los cocientes instantáneos antes y después del incidente constituido por el evento perturbador. En un análisis de la nupcialidad y de la salida del domicilio de los padres, en un primer tiempo estimaremos la influencia de las características individuales sobre los cocientes instantáneos de nupcialidad. Con la influencia estimada antes de la partida y después de una eventual salida de la casa de los padres para las variables fijas, se tendrá pues el efecto principal y el efecto de interacción de una variable fija del individuo[5] sobre su matrimonio. Luego, una vez que el individuo ha salido del hogar familiar, pueden intervenir variables asociadas: su nuevo lugar de residencia y su estatus profesional pueden entrar en el análisis.

En el caso contrario, cuando se estudia la partida del hogar de los padres, el matrimonio será considerado como el evento que interfiere, el evento perturbador, y se podrán introducir las variables que están asociadas con él.

El número de características que se pueden hacer intervenir está limitado a ocho variables fijas y 14 variables asociadas al evento que se considera como perturbador. Las ocho variables fijas producirán así 16 estimadores a los cuales habrá que añadir los 14 estimadores correspondientes a las variables asociadas, lo que da la estimación posible de 30 parámetros. Cabe señalar que cuando se hace el análisis de muestras pequeñas, la introducción de un número demasiado grande de características conduce a estimaciones poco significativas o que es imposible realizar.

[4] "Fijo" significa que esas características no son modificadas (modificables) por los eventos cuyas interacciones estudiamos.

[5] Por ejemplo, la categoría socioprofesional del padre es una característica no susceptible de ser modificada por los dos eventos estudiados.

El *software* calcula así edad por edad y para cada variable que se toma en cuenta:

$$\Sigma \exp z_\ell \beta,\ \sum_\ell z_{jl} \exp(z_\ell \beta),\ \sum_\ell z_{hl} z_{jl} \exp(z_\ell \beta)$$

a partir de lo cual se calculan la log-verosimilitud, su derivada, y la matriz de varianza-covarianza que sirven para estimar parámetros β en las iteraciones del método de Newton-Raphson (*cf.* capítulo IX).

Luego de haber obtenido los parámetros β se calcula la matriz de varianza-covarianza; después el valor de la parte no paramétrica de los cocientes, para la que coexisten dos formas de cálculo. Si en la fecha t se produce un solo evento, o en el caso de ocurrencias múltiples, se emplea un algoritmo de Newton-Raphson.

Este *software,* que está disponible junto con su manual de uso, ya ha servido para numerosas aplicaciones. Su difusión ha sido asegurada desde octubre de 1987, a través de la red de la Asociación Nacional de Softwares (ANL, Association Nationale de Logiciel).

En el Anexo 3 presentamos ese *software* en detalle.

ANEXO 2

MANUAL DE USO DE ROOT.RAT*

Mayo de 1987
ÉVA LELIÈVRE, INED

* Todos estos softwares también están disponibles en Fortran y en STATA.

ÍNDICE

I. Análisis no paramétrico: el modelo bivariado 245
1) Presentación del modelo 246
2) Estimación de los cocientes y estadísticos de prueba 247
3) Ejemplo 1 248
4) Ejemplo 2 248
5) Ejemplo 3 250
6) Ejemplo 4 250
II. Puesta en práctica del análisis 251
A) Estructura de los datos 251
B) Selecciones sobre los datos 253
1) Los criterios 253
2) Las pruebas posibles y su código 253
3) Las edades mínimas y máximas para el análisis 254
C) La selección de los tratamientos 254
1) Las series de edad, sean cronólogicas o exclusivas 254
2) Tratamiento de los eventos concurrentes o en competencia 256
3) Ejemplos de resultados de los diversos procedimientos de tratamiento de las simultaneidades 257
4) Caso de reversibilidad de los eventos 260
5) Comparación de eventos en secuencias de desenlace variable 262
III. Los *softwares*. Puesta en práctica de los procedimientos informáticos 263
A) Actividad femenina y fecundidad 263
1) La formación de los comandos necesarios 264
2) Escritura del *software* del cálculo 267
3) Los archivos creados por el ejecutable 268
4) Las comparaciones ulteriores 269
B) Primera migración después del matrimonio y primer nacimiento 271
1) La formación de los comandos necesarios 271
2) Escritura del *software* del cálculo 272
3) Archivos que crea el ejecutable 273
4) Comparaciones ulteriores 275

IV. Los mensajes de error 276
1) Los mensajes que provienen de CONTROL.RAT 276
2) Los mensajes que provienen de ORGA.RAT 277
3) Los mensajes que provienen de ROOT.RAT 277
4) Algunas modificaciones posibles 277
V. Bibliografía 277

I. ANÁLISIS NO PARAMÉTRICO: EL MODELO BIVARIADO

Para utilizar las técnicas no paramétricas no se requiere ninguna hipótesis previa, mientras que los métodos de análisis estadístico necesitan que se especifique la distribución funcional de los eventos estudiados en la ausencia de salida de la observación de ciertos miembros de la población sometida a riesgo.

Más allá de un estudio semiparamétrico o paramétrico, las técnicas no parámetricas son herramientas que esclarecen ampliamente el problema estudiado.

El término en inglés *Life tables*, que se emplea comúnmente para designar la estimación no paramétrica de las funciones de supervivencia calculadas a partir de los datos de mortalidad, tiene su equivalente en español con la expresión "Tabla de supervivencia" (en francés *Table de survie*). Nosotros preferimos el término más general de "Tablas de permanencia en un estado" (*Tables de séjour dans un état*), pues el marco del análisis debe permitir el estudio de estados que no concluyen forzosamente con un deceso (soltería, vida de estudiante...). Además se tratará de confrontar las duraciones de permanencia diferenciales según el itinerario de los individuos.

En efecto, tradicionalmente se considera un solo nivel de referencia para situar a los individuos: éstos son solteros y luego se casan; viven con sus padres o dejan el domicilio familiar; ocupan su primer empleo, su segundo empleo, etc., y de esa manera se toman en cuenta características individuales.

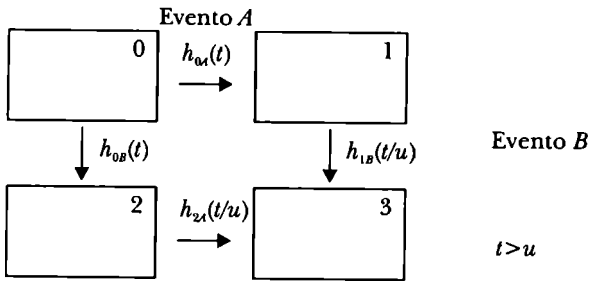
Desde nuestra perspectiva, se busca aprehender la realidad con mayor complejidad, considerando más de un nivel de referencia: la intersección de dos niveles en el caso del modelo bivariado que presentamos aquí.

Un individuo será un soltero que vive con sus padres, un soltero que vive fuera del domicilio familiar, luego un casado que vive con sus padres o un casado que vive fuera del domicilio de éstos, todo lo cual permite tomar en consideración más de un dominio a la vez. Al no aislar un evento en su nivel propio nos acercamos un poco más a la comprensión del campo de las fuerzas de presión y de atracción de las que resultan las elecciones de los individuos.

1) Presentación del modelo

Partimos de la idea simple (y simplificadora) de que es posible asimilar los diferentes estados del ciclo de vida de un individuo a los diversos estados de una cadena de Markov, y las tasas de paso entre los estados a los cocientes instantáneos* de ocurrencia. Los estados pueden ser absorbentes (el fallecimiento de un individuo lo hace pasar hacia un estado indudablemente absorbente) o recurrentes (los estados "activo"/"inactivo").

Modelizamos entonces la situación bivariada mediante el siguiente esquema:



En el instante t encontramos en 0 a los que comienzan el recorrido antes de haber vivido los eventos estudiados A o B. En el caso 1 se encuentran aquellos que ya han experimentado A; en el 2 están quienes ya han experimentado B. En el momento t los cocientes instantáneos de ocurrencia de uno de los eventos, de A por ejemplo, se miden considerando si el individuo ha experimentado o no el otro evento, B, por ejemplo, con antelación.

Entonces se prueban las dos igualdades siguientes:

$$h_{0A}(t) = h_{2A}(t/u) \quad \text{y} \quad h_{0B}(t) = h_{1B}(t/u)$$

De esa manera se tiene la posibilidad de comparar la influencia del evento B sobre la aparición del evento A, y recíprocamente la segunda prueba permite sacar una conclusión acerca del papel del evento A en la aparición del evento B. Probamos, pues, un efecto debido a la dependencia "local" que cada uno ejerce sobre el otro.

*Si se utiliza el término cociente instantáneo de ocurrencia es porque lo que se estima no es la densidad de probabilidad del evento experimentado, sino la densidad condicional de supervivencia de los individuos en el estado donde se encontraban antes del evento.

2) Estimación de los cocientes y estadísticos de prueba

Para estimar esos cocientes, vamos a suponer que son constantes sobre el intervalo de tiempo escogido. Para que la estimación sea posible discretizamos, al no haber disponible todavía ningún método puramente no paramétrico (Cox y Oates, 1984). Esta discretización implica la aparición de casos de ocurrencia simultánea: eventos de los dos tipos considerados que ocurren en el mismo intervalo de tiempo. Tales casos plantean problemas específicos, y en este manual (ref.) sólo presentamos sus soluciones prácticas. Por otra parte, vamos a plantear la hipótesis de la distribución uniforme de los eventos en el curso del intervalo de tiempo escogido (en los ejemplos que siguen el intervalo es de un año).

Sean entonces:

$N_i(t)$ $i = 0, 1, 2$ la población en el estado i al inicio del año t
$n_{ij}(t)$ $j = A, B$ el número de eventos de tipo j ocurridos en la población del estado i durante el año t

los estimadores más simples se calculan según:

$$h'_{0j}(.) = n_{0j}(.) / \left[N_0(.) - 1/2\left(n_{0A}(.) + n_{0B}(.)\right)\right]$$

$$h'_{kj}(.) = n_{kj}(.) / \left[N_k(.) - 1/2\left(n_{kj}(.) - n_{0k}(.)\right)\right]$$

$$j = 2, 1 \quad \text{y} \quad k = 1, 2$$

Al ser los estimadores asintóticamente independientes, normalmente distribuidos y sin sesgos, de varianza estimada $h' .. (t)/N .. (t)$, se forma la diferencia normalizada:

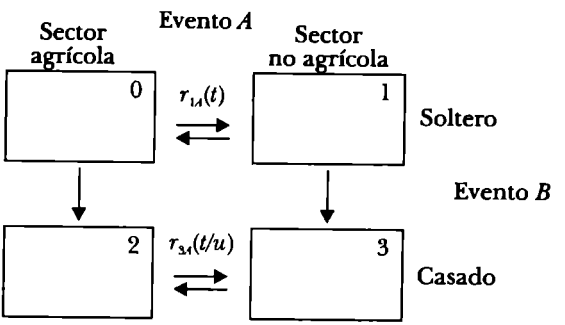
U_t = diferencia de los cocientes / raíz de la suma de sus varianzas
y $U = U_t$/raíz (número de años que se toman en cuenta)

Shou y Vaeth (1980) plantean que la aproximación normal de la distribución de cada $h'..(.)^{1/3}$ es mejor que la de $h'..(.)$; nosotros damos los dos tipos de estadísticos de prueba bajo su forma anual y acumulada.

Este esquema simple no permite describir situaciones variadas a menos que tratemos de adaptarlo a los tipos de eventos y de datos que se ponen en juego en el análisis, lo cual se lleva a cabo con la adición de otros pasos posibles entre los diversos estados.

3) Ejemplo 1

Si se analiza la nupcialidad de los hombres que iniciaron su vida profesional en la agricultura en relación con la modificación de su actividad profesional (en particular su abandono de la agricultura), todos los individuos comienzan en el mismo estado 0 y el siguiente esquema parece ser el que mejor se adapta:

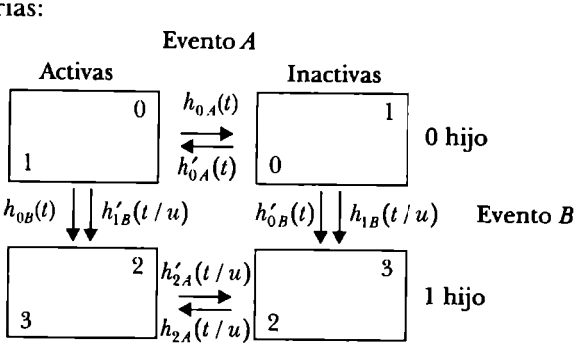


pudiendo cada individuo seguir una trayectoria de idas y venidas en lo que respecta a sus cambios de empleo (Courgeau y Lelièvre, 1986).

4) Ejemplo 2

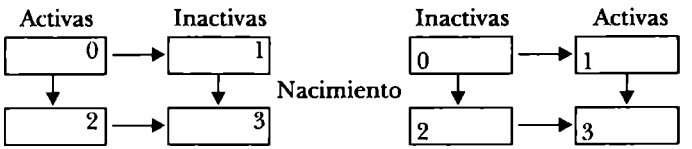
En el caso de la relación de la actividad profesional de las mujeres casadas y la llegada de su primer hijo, conviene esquematizarla como sigue, dejándole a cada individuo de la muestra la posibilidad de comenzar la trayectoria estando activo (cifra en la parte de arriba de los cuadros) o inactivo (cifra en la parte de abajo de los cuadros) y así terminar en estados diferentes, y tomando verdaderamente en cuenta los regresos como una modalidad del evento A.

Sean h y h' los cocientes instantáneos correspondientes a los dos tipos de trayectorias:



este análisis da entonces lugar a pruebas complementarias, puesto que un mismo tipo de paso se puede experimentar en dos contextos diferentes (lo que aquí se simboliza por la presencia de dos flechas verticales) según se haya comenzado estando activo o inactivo.

El esquema puede, de hecho, descomponerse en dos esquemas simples:



superpuestos en la versión precedente (Lelièvre, 1987).

Si los esquemas de los ejemplos 1 y 2 presentan trayectorias de "ida y vuelta" éstas pueden ser de naturaleza diferente si son vividas por el mismo individuo o por dos grupos de individuos distintos; en el ejemplo 1 tomábamos en cuenta las idas y regresos profesionales de un mismo individuo, mientras que en el ejemplo 2 se trataba, por una parte, del cese de la actividad profesional de las mujeres que comenzaron siendo activas, y por otra parte, del inicio de la actividad de aquellas que originalmente fueron inactivas. En este caso, tales diferencias resultan de la manera de abordar el problema: la reversibilidad de un evento se puede considerar sólo como un incidente del recorrido de un individuo o como un evento propiamente.

El caso presentado en el ejemplo 1 ha sido objeto de un estudio en profundidad (Courgeau y Lelièvre, 1986). Ahí se advirtieron algunas particularidades (todos los individuos comienzan el recorrido en el estado 0, los regresos siguen siendo poco numerosos...) que fueron tratadas por otros *softwares*, los cuales toman en cuenta los regresos. Como su utilización se adapta a casos muy particulares, no está disponible ninguna versión que se pueda difundir.

El hecho de que la población que regresa a un estado precedente sea tomada en cuenta en la misma muestra sometida a riesgo puede resultar poco riguroso, pues se introducen sesgos difíciles de medir que falsean tramposamente el análisis.

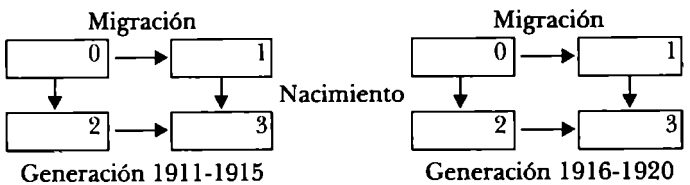
En los cálculos siguientes no se tomarán en cuenta sino los casos en que los eventos se consideran irreversibles, disponiéndose de la posibilidad de probar en ellos los cocientes instantáneos surgidos de tratamientos diferentes. (Por ejemplo, se estudiará la nupcialidad de las mujeres originarias de zonas de fuerte fecundidad, y la de mujeres que provienen de zonas de fecundidad débil, y en un segundo tiempo será posible comparar los cocientes de matrimonio obtenidos en cada uno de los dos análisis).

5) Ejemplo 3

Cuando se analiza la salida del domicilio familiar y la nupcialidad, se observa que son muy numerosos los casos de ocurrencia simultánea de los eventos. Para tomar en cuenta esas simultaneidades hay dos procedimientos posibles: por un lado, discretizar con más finura (siempre se puede decidir que un evento es anterior a otro si las fechas se miden en días, horas, minutos... pero ¿tiene eso sentido en las ciencias sociales?; por otro lado, considerarlas como tales: particulares, ya que se producen en el mismo periodo (ciertamente arbitrario). Por lo tanto, en el *software* previmos la posibilidad de incorporar los casos de ocurrencia simultánea al análisis o sustraerlos de él, haciéndolo de maneras muy diversas: ya existen ocho posibilidades diferentes que se detallan más adelante (Courgeau, Lelièvre y Wagner, 1986).

6) Ejemplo 4

Si en el seno de la muestra disponemos de varias subpoblaciones, quizás queramos realizar comparaciones simples de los cocientes dos a dos. Por ejemplo, comparar la evolución de tendencias por generaciones:



Por lo tanto, hemos previsto la posibilidad de comparar posteriormente la intensidad de los mismos eventos para dos subpoblaciones analizadas por separado. A diferencia del ejemplo 3, donde se buscaba una comparación cruzada para uno de los dos eventos (nacimiento del primer hijo de las activas (respecto de las inactivas) según se estuviera activa (respecto de las inactivas) al casarse o no, aquí pretendemos efectuar comparaciones simples, que además se puedan aplicar sobre los dos eventos (nacimiento anterior a, respecto de, o después de la migración según la cohorte, y migración anterior a, respecto de, o después del nacimiento según la cohorte).

II. PUESTA EN PRÁCTICA DEL ANÁLISIS

Este análisis permite poner en evidencia las interacciones entre dos eventos considerados como definitivos en la biografía de los individuos: migraciones sucesivas, matrimonio, nacimientos de los hijos, movilidad profesional... y *para llevarlo a la práctica se necesita sólo la fecha de aparición de cada uno de los eventos tomados en consideración*, o la edad del individuo en el momento en que ocurren los dos eventos, pero no excluye la introducción en la muestra de parámetros que determinan los subgrupos, que sin embargo siguen siendo facultativos.

En el *software* las fechas de los dos eventos llevan los nombres característicos T1edad y T2edad, que emplearemos de ahora en adelante.

El procedimiento se ejecuta en dos etapas si el archivo admite más informes que las dos fechas (o edades) necesarias:

1) *la selección sobre los datos,* que consiste en escoger una submuestra particular en el seno de los datos y el intervalo de tiempo en el que se quiere realizar el análisis.

(Por ejemplo, seleccionamos a las mujeres y el análisis lo realizamos en el grupo de 15 a 45 años).

2) *la elección de los tratamientos,* que comprende tres modalidades. Por una parte, la selección de las fechas que se tomarán en cuenta si los datos admiten varias fechas para uno, otro, o los dos eventos. Por otra parte, la selección de la forma de considerar los eventos simultáneos. Y, por último, la decisión de conservar o no los resultados para las comparaciones ulteriores.

(Por ejemplo: para cada una de las mujeres tenemos su edad en la primera migración después del matrimonio, después del primer hijo y en los nacimientos sucesivos. En esta etapa, por ejemplo, escogemos emprender dos tratamientos: interacciones entre la primera migración después del matrimonio y el primer nacimiento, con exclusión de los casos de simultaneidad; y luego, interacciones entre la primera migración después del primer nacimiento y el segundo nacimiento con inclusión de los casos de simultaneidad).

A) Estructura de los datos

El archivo de los datos debe estar compuesto por *enteros*; así pues, hay que convertir las fechas o las edades a meses para que no incluyan una parte

decimal, y en el formato que desee el usuario.** Las diferentes características que podrán ser utilizadas por el programa también deberán ser enteras.

- El archivo mínimo. No incluye más que dos fechas, que son las de la ocurrencia de cada uno de los eventos.

(Por ejemplo: T1edad = duración transcurrida desde el matrimonio hasta el primer nacimiento; T2edad = duración transcurrida desde el matrimonio hasta la primera migración).

- Ciertos individuos se salen de la observación: truncamiento a la derecha, el archivo contiene dos fechas y dos índices de truncamiento.

(Por ejemplo: T1edad = edad al primer empleo y obs1 = 1 (donde T1edad = edad a la salida de la observación y obs1 = 0 si el primer empleo no es observado), y T2edad = edad al matrimonio y obs2 = 1 (donde T2edad = edad a la salida de la observación y obs2 = 0 si el matrimonio no es observado)).

- Además, se poseen criterios de selección.

(Por ejemplo: las datos reúnen a los hombres, las mujeres, varias generaciones, varias localidades de origen. Cada una de estas características se marca mediante un índice crit(i)).

- Se tienen más de dos fechas por individuo.

(Por ejemplo: T1edad(n) = edad a los n nacimientos sucesivos y T2edad = edad al producirse la primera migración después del matrimonio; entonces se puede hacer el análisis de las interacciones entre esta migración y el nacimiento del rango escogido).

- Para cada individuo se tienen fechas de naturaleza diferente.

(Por ejemplo: T1edad = edad al primer hijo y T2edad (j) = edad a la primera migración después del matrimonio e ind2 = j; ahora bien, para cada individuo se ha hecho la distinción de la migración según su destino; de ahí que para cada uno se formará una serie de fechas de las cuales sólo una es distinta de cero, y se indicará el rango de esta fecha, lo que dará la naturaleza de ésta: ind2 = 1 si es una migración hacia la metrópoli, ind2 = 2 si es una migración en el interior de la metrópoli).

Las órdenes necesarias para efectuar la selección de los datos y la elección de los tratamientos están reagrupadas en un archivo de comandos que se crea (y nombra) al ejecutar el programa CONTROL.RAT, el cual le exige sucesivamente al usuario las informaciones que damos a continuación.

** NOTA: los formatos previstos para la presentación de los resultados están diseñados sólo para poblaciones de menos de 10 000 personas, y las características únicamente pueden probar respecto de los valores enteros de cuatro cifras como máximo.

B) Selecciones sobre los datos

El *software* ofrece, pues, la posibilidad de hacer elecciones por medio de criterios de selección que determina el usuario, lo que significa seleccionar una submuestra en el archivo de datos.

La selección de datos sigue siendo la misma para todos los tratamientos de una misma sesión, y corresponde a la elección de los criterios, cada uno acompañado de una prueba, que describimos a continuación.

1) Los criterios

El *software* puede tomar en cuenta nueve criterios de selección (sexo, cohorte, categoría socioprofesional...) y el usuario dispone de 11 pruebas posibles para cada criterio (una prueba por tratamiento para cada uno de los criterios; por ejemplo, sexo = 1, cohorte = 3, categoría socioprofesional comprendida entre 10 y 20...).

2) Las pruebas posibles y su código

Se dispone de tantas selecciones posibles entre los datos como de criterios por individuo, ya que sólo se prevé una prueba por criterio, siendo nueve el máximo. La selección se realizará empleando valores de referencia y una prueba escogida por el usuario.

Las pruebas disponibles presentadas a continuación van a seleccionar a los individuos según el resultado de la confrontación del valor leído del criterio y del valor (o los valores) de referencia: de acuerdo con el resultado, el individuo será incorporado o no al análisis.

El criterio se lee para cada individuo y las referencias se dan en los comandos.

Enunciado	Código
Cuatro pruebas simples:	
criterio = referencia	10
criterio = referencia	20
criterio = referencia	30
criterio = referencia	40
Siete pruebas dobles:	
criterio = ref1 o = ref2	11
criterio = ref1 o = ref2	13
criterio = ref1 o = ref2	14

criterio = ref1 y = ref2	22
criterio = ref1 y = ref2	23
criterio = ref1 y = ref2	24
criterio = ref1 y = ref2	34

Si se escoge una prueba simple, se indicará un solo valor a continuación del código de la prueba (valor entero de cuatro caracteres como máximo); el código de una prueba doble será el seguido de dos valores de referencia (misma condición de tamaño y de tipo de variable).

Si no se requiere ninguna prueba, a ésta se le asigna el valor cero.

3) Las edades mínimas y máximas para el análisis

La última selección sobre los datos es la que concierne a las edades (las fechas), mínimas y máximas, entre las que se quiere hacer el análisis.

En efecto, el *software* va a calcular unidad de tiempo por unidad de tiempo (para cada edad) las tasas instantáneas de transición, comenzando por el mínimo indicado y terminando por la edad (la fecha) máxima especificada.

Atención: a ninguna edad debe asignarse el valor 0.

C) La selección de los tratamientos

Una vez efectuada la selección de los datos, quedan por determinar los tratamientos sucesivos escogidos que van acompañados por el título del análisis al que corresponden. Su número no está limitado, es sólo el tiempo CPU acordado por el sistema del usuario lo que impone un límite.

En los tratamientos se procede a la selección de la edad (o la fecha) que determina la naturaleza de cada uno de los eventos. En efecto, los datos pueden incluir varias fechas por individuo y se trata de escoger entre ellas; esas series de edades pueden ser de dos tipos diferentes, que detallamos a continuación.

1) Las series de edad, sean cronológicas o exclusivas

Cualquiera que sea su tipo, esas series se limitan a ocho edades diferentes por evento y son:

- series cronológicas de edades (por ejemplo: edades al producirse los nacimientos, migraciones sucesivas);

- series en las que sólo uno de los valores no es igual a cero para cada individuo (por ejemplo: edad al producirse el cambio profesional o al migrar en caso que se tome en cuenta el origen de la migración; los individuos que migran hacia la ciudad parten de lugares diferentes, y se puede entonces imaginar una serie de edades que forma un vector donde el rango de las coordenadas corresponde a la región de origen del individuo).

En el primer caso se indicará la selección sobre el rango deseado para los dos eventos y el conjunto de los datos se tomará en cuenta para el análisis de los eventos indicados (por ejemplo: segundo nacimiento, sexta migración...).

En el segundo caso, el usuario *debe* proporcionar al *software* el indicador (ind1/ind2) del rango del valor distinto de cero para cada individuo (es decir, prever esta variable en el archivo de los datos); su omisión acarrea una sobreevaluación de la población sometida al riesgo: "habiendo experimentado el evento cuya fecha se leería como igual a cero y no habiendo experimentado aún el segundo evento".

Esta vez, cuando uno indica su selección se efectúa en la misma ocasión una selección de datos que eventualmente se pueden suprimir, para hacer el análisis de tipo "todos los orígenes confundidos" (por ejemplo: se puede analizar la interacción entre el matrimonio y la promoción profesional de los empleados, y luego rehacer el análisis cualquiera que sea el origen profesional de los individuos).

Así, en el primer caso basta con que el usuario indique su selección en el archivo de comando, y en el segundo caso, si se indica igualmente su selección para el análisis, *debe* disponerse para cada individuo de un índice que indique su "origen", lo que corresponderá al rango de la coordenada distinta de cero en la serie de las edades retenidas. Ese índice se leerá en el fichero de los datos para cada individuo bajo el nombre ind1/ind2 según corresponda a la serie de edad T1edad o T2edad.

Por último, por razones de pesadez de la programación, las dos series de edad, T1edad y T2edad, no desempeñan exactamente el mismo papel: si ellas pueden ser de uno u otro de los dos tipos descritos antes, efectivamente no se podrá tratar de la manera "todos los orígenes confundidos" a los dos eventos a la vez; sólo aquel cuya fecha se lee bajo el nombre de "T1edad" podrá ser objeto de un reagrupamiento como ése. Sin embargo, es posible intentarlo para cada combinación entre una verdadera serie cronológica y las series de edades indexadas.

2) Tratamiento de los eventos concurrentes o en competencia

Hay ocho opciones disponibles para el tratamiento de los eventos concurrentes (el caso en el que dos eventos se producen en la misma fecha —a la misma edad— para un individuo dado). Esas opciones corresponden a la posibilidad de sustraer o incorporar al análisis los casos en que dos eventos se producen en la misma fecha, o también de calcular aparte los cocientes instantáneos que tienen relación con ellos.

La hipótesis que se retiene en todas las opciones, con excepción de "apar 2", es la de sometimiento al riesgo durante 1/2 intervalo de tiempo de aquellos que experimentan un evento en ese intervalo de tiempo.

1) Eliminación total de los eventos concurrentes [sin 1]
   Se suprime por completo a los individuos y se les excluye de la muestra; así se obtienen cocientes instantáneos de transición que conciernen a una población que ha espaciado sus decisiones o que ha experimentado, con un año o menos de intervalo, los dos eventos considerados. En el caso en el que se advierte una partición neta de la población observada según ese criterio de simultaneidad, este tipo de aproximación puede proporcionar un amplio esclarecimiento sobre los dos comportamientos, cuando se completa mediante el cálculo de los cocientes instantáneos de transición "simultánea".
2) Cálculo aparte del quinto cociente instantáneo [apar 2]
   Cuando la población está sometida al riesgo durante todo el intervalo de tiempo.
   Aquí se calculan los cinco cocientes instantáneos de transición.
3) Cálculo aparte del quinto cociente instantáneo [apar 1]
   Se calculan los cinco cocientes instantáneos de transición. El quinto cociente representa la transición directa de la subpoblación de partida hacia el estado final que se considera en el análisis.
4) Distribución uniforme de los casos simultáneos entre los dos grupos de eventos [sin 2]
   Los individuos en cuestión, el año de su paso simultáneo a dos estados estarán sometidos a riesgo 1/4 de año en la subpoblación 0, luego 1/4 de año en las subpoblaciones 1 y 2, y los eventos se van a repartir por cuartos para entrar en el cálculo de los cuatro cocientes instantáneos de transición.
5) Descomposición de la trayectoria de los casos simultáneos, con indicación del último evento experimentado [last T1/last T2]
   Los individuos serán sometidos a riesgo 1/4 de año en la subpoblación de partida, luego 1/4 de año en la subpoblación correspondiente al último evento experimentado, y los eventos se dividirán por la mitad entre los dos cocientes de la trayectoria elegida.

6) Consideración de las simultaneidades [con T1 / con T2]
Según la hipótesis elegida para el análisis, hemos preservado la posibilidad de enlazar los eventos en competencia con los eventos de tipo *A* o *B*: entonces, el usuario puede decidir, *a priori*, si en los casos de simultaneidad conviene contar a los individuos que la experimentan dentro de la subpoblación de aquellos que ya han experimentado uno de los eventos y, por extensión, contar esas simultaneidades como eventos del tipo T1edad o T2edad. La población sometida al riesgo en 0, el año t, no comprende entonces a ningún individuo que haya experimentado los dos eventos en ese año.

(Por ejemplo: para una parte del análisis, los casos de matrimonio registrados el mismo año del abandono de la agricultura han sido enlazados a los matrimonios de los individuos que ya han salido del sector agrícola, pues la hipótesis de trabajo estipulaba en ese caso que dichas personas habían contraído matrimonio cuando ya su futuro estaba considerado fuera del mundo agrícola).

3) Ejemplos de resultados de los diversos procedimientos de tratamiento de las simultaneidades

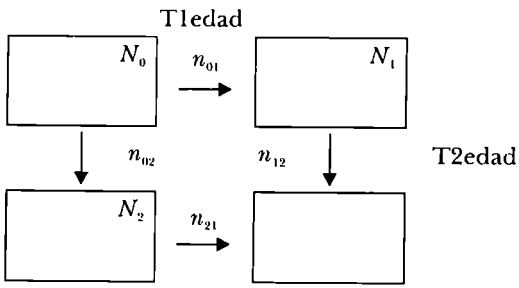
A continuación presentamos los resultados obtenidos para esos ocho tratamientos si se toma una muestra tal que:

$N_0(3)$: la población sometida a riesgo en el estado 0 al inicio de la fecha $t = 3$ sea igual a 20 individuos.
$N_1(3)$: la población sometida a riesgo en el estado 1 al inicio de la fecha $t = 3$ sea igual a 10 individuos.
$N_2(3)$: la población sometida a riesgo en el estado 2 al inicio de la fecha $t = 3$ sea igual a 10 individuos.
$n_{01}(3)$: el número de eventos de tipo T1edad acaecidos en la población del estado 0 durante el periodo 3, sea de 5.
$n_{02}(3)$: el número de eventos de tipo T1edad acaecidos en la población del estado 0 durante el periodo 3, sea de 2.
$s(3)$: el número de eventos simultáneos registrados durante el periodo 3, sea de 3.
$n_{12}(3)$: el número de eventos de tipo T2edad acaecidos en la población del estado 1 durante el periodo 3, sea de 0.
$n_{21}(3)$: el número de eventos de tipo T1edad acaecidos en la población del estado 2 durante el periodo 3, sea de 0.

Dándose por entendido que para cada uno de los eventos considerados se comparan los cocientes instantáneos siguientes:

Para evt = T1edad Pob. antes = antes T2edad (estim. de N_0)
Pob. después = después T2edad (estim. de N_2).
Para evt = T2edad Pob. antes = antes T1edad (estim. de N_0)
Pob. después = después T1edad (estim. de N_1).

según el esquema siguiente:



Observaciones

En los resultados que se dan a continuación se observarán las diferencias entre los tratamientos que son consecuencia de la inclusión diferencial de casos de simultaneidad.

Los otros eventos, por su parte, acarrean un descuento en la población sometida al riesgo correspondiente, de una media unidad de tiempo según la hipótesis de ocurrencia uniforme de los eventos en el periodo. Sólo el tratamiento de la supresión total de las simultaneidades ("apar 2") considera la población sometida a riesgo durante toda la unidad de tiempo.

DISTRIBUCIÓN UNIFORME

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
15.75	5.75	0.3651	11.75	0.75	0.0638	-1.7810	-1.7810	-1.7205	-1.7205
Evt=T2edad									
15.75	2.75	0.1746	13.25	0.75	0.0566	-0.9522	-0.9522	-0.9425	-0.9425

Cocientes acumulados T1edad y T2edad

cqA(3)=0.3651 cqB(3)= 0.0638 cqA(3) = 0.1746 cqB(3) = 0.0566

DESCOMPOSICIÓN LAST T1

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
15.75	5.00	0.3175	11.75	1.50	0.1277	-1.0776	-1.0776	-1.0469	-1.0469
Evt=T2edad									
15.75	3.50	0.2222	12.50	0.00	0.0000	-1.8708	-1.8708	0.0000	0.0000

Cocientes acumulados T1edad y T2edad

cqA(3)=0.3175 cqB(3)= 0.1277 cqA(3) = 0.2222 cqB(3) = 0.0000

DESCOMPOSICIÓN LAST T2

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
15.75	6.50	0.4127	11.00	0.00	0.0000	-2.5495	-2.5495	0.0000	0.000
Evt=T2edad									
15.75	2.00	0.1270	13.25	1.50	0.1132	-0.1069	-0.1069	-0.1066	-0.106

Cocientes acumulados T1edad y T2edad

cqA(3)=0.4127 cqB(3)= 0.000 cqA(3) = 0.1270 cqB(3) = 0.1132

TOMAR EN CUENTA CON T1

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
13.50	5.00	0.3704	12.50	3.00	0.2400	-0.6037	-0.6037	-0.6024	-0.6024
Evt=T2edad									
13.50	2.00	0.1481	12.50	0.00	0.0000	-1.4142	-1.4142	-0.0000	-0.0000

Cocientes acumulados T1edad y T2edad

cqA(3)=0.3704 cqB(3)= 0.2400 cqA(3) = 0.1481 cqB(3) = 0.0000

TOMAR EN CUENTA CON T2

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
13.50	5.00	0.3704	11.00	0.00	0.0000	-2.2361	-2.2361	-0.0000	-0.0000
Evt=T2edad									
13.50	2.00	0.1481	14.00	3.00	0.2143	0.4080	0.4080	0.4081	0.4081

Cocientes acumulados T1edad y T2edad

cqA(3)=0.3704 cqB(3)= 0.0000 cqA(3) = 0.1481 cqB(3) = 0.2143

ELIMINACIÓN TOTAL

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
13.50	5.00	0.3704	12.50	3.00	0.2400	-0.6037	-0.6037	-0.6024	-0.6024
Evt=T2edad									
13.50	2.00	0.1481	12.50	0.00	0.0000	-1.4142	-1.4142	-0.0000	-0.0000

Cocientes acumulados T1edad y T2edad

cqA(3)=0.3704 cqB(3)= 0.0000 cqA(3) = 0.1481 cqB(3) = 0.0000

CÁLCULO APARTE RUDIMENTARIO

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
20.00	5.00	0.2500	10.00	0.00	0.0000	-2.2361	-2.2361	-0.0000	-0.0000
Evt=T2edad									
20.00	2.00	0.1000	10.00	0.00	0.0000	-1.4142	-1.4142	-0.0000	-0.0000

Cocientes acumulados T1edad y T2edad

cqA(3)=0.2500 cqB(3)= 0.0000 cqA(3) = 0.1000 cqB(3) = 0.0000

CÁLCULO APARTE NORMAL

Edad: 3 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=T1edad									
15.00	5.00	0.3333	11.00	0.00	0.0000	-2.2361	-2.2361	-0.0000	-0.0000
Evt=T2edad									
15.00	2.00	0.1333	12.50	0.00	0.0000	-1.4142	-1.4142	-0.0000	-0.0000

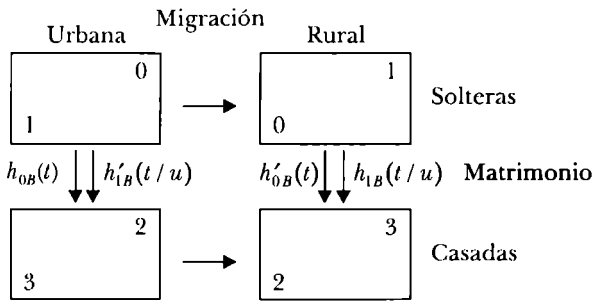
Cocientes acumulados T1edad y T2edad

cqA(3)=0.3333 cqB(3)= 0.0000 cqA(3) = 0.1333 cqB(3) = 0.0000

4) Caso de reversibilidad de los eventos

Se trata de eventos para los que existe una posibilidad de ocurrencia en sentido inverso: si se migra de una zona rural hacia una metrópoli, el hecho de migrar de una zona urbana hacia una zona rural se puede considerar como el movimiento inverso. Si se estudia, por ejemplo, la nupcialidad y la migración en interacción, se va a poder comparar la nupcialidad en las metrópolis de las mujeres originarias de la zona urbana con la de las mujeres que han inmigrado allí. Asimismo, se comparará la nupcialidad de las mujeres de las zonas rurales según hayan vivido siempre allí o hayan inmigrado hacia allí.

Sean h y h' los cocientes instantáneos correspondientes a los dos tipos de trayectoria:



El *software* aquí presentado toma en cuenta ese tipo de caso descomponiendo el esquema inicial de las trayectorias posibles en dos esquemas distintos (*cf.* ejemplo 2, capítulo I) y reservándole al usuario la posibilidad de probar la igualdad de los cocientes instantáneos h y h' obtenidos con dos tratamientos consecutivos. Esto corresponde, una vez hechos los primeros análisis, a comparar las flechas dobles entre ellas.

Durante esos dos tratamientos basta con indicar el hecho de que se desea prolongar el análisis para otras comparaciones. El *software* conserva entonces los cocientes de los dos tratamientos que serán objeto de comparaciones ulteriores en el mismo archivo: TAUX.DAT

Ese archivo está bajo una forma inmediata legible por el programa COMPARE.RAT de comparaciones finales.

Los datos

Los datos deben ser necesariamente del tipo de serie de edades (o de fechas) exclusiva que corresponde al evento reversible. Para cada individuo, la fecha del evento vivido se almacena en un vector cuyas coordenadas son iguales a cero, con excepción de esta fecha colocada en un rango que especifica su naturaleza. Ese vector se llamará necesariamente T1edad. El otro evento —aquel para el que se quieren hacer comparaciones según el origen— será llamado T2edad.

Ejemplo:

Si se analizan las migraciones entre zonas urbanas y zonas rurales en relación con la nupcialidad, los datos estarán formados al menos por un vector de las edades de tipo exclusivo (*cf. Las series de edad*): el de las edades de migración (que se llaman entonces T1edad).

Un individuo originario de una zona rural será identificado por un indicador que corresponde al rango del vector donde se encuentra la edad de la migración, y ese vector (aquí de dimensión 2, pues sólo se toman en cuenta dos orígenes, el rural y el urbano) del cual uno de los componentes es igual a cero.

Si se le asigna el rango 1 a las migraciones hacia las zonas urbanas y el rango 2 a las migraciones hacia las zonas rurales, y siendo T1edad la edad al producirse el matrimonio:

- los datos que corresponden a una parisina que a los 38 años se fue a vivir a la región de Limousin (zona rural) comprenderán el vector T1edad = (0.38) y el indicador ind1 = 2.
- los datos concernientes a un campesino de la Ardéche llegado a Lyon (zona urbana) a los 16 años comprenderán el vector T1edad = (16.0) y el indicador ind1 = 1.

Resultados proporcionados por el software

Una vez que se han hecho los dos análisis y que el archivo TAUX.RAT está completado, una ejecución de COMPARE.RAT suministra un archivo COMPARE.DAT que admite los resultados de la última comparación.

5) Comparación de eventos en secuencias de desenlace variable

Por secuencia de desenlace variable se designa a los periodos de la biografía que pueden terminar de diversas maneras: una unión puede concluir con un matrimonio o romperse por una separación; un periodo de actividad se puede interrumpir por un abandono de la actividad o terminar por un cambio profesional que marca entonces el inicio de otro periodo de actividad.

En ese caso, quizás se quiera estudiar la llegada de un evento en el seno de la población de origen, según se le haya confrontado con uno u otro tipo de desenlace. Entonces se emprenderán sucesivamente dos análisis independientes de interacción tomando el evento para el cual se quiere efectuar la comparación en T2edad, e indicando que se quiere prolongar el análisis mediante comparaciones.

Las tasas se almacenarán en archivos COMPT2.DAT que se deberán fusionar antes de lanzar el programa COMPARE.RAT.

Éste es un análisis delicado, y hay que ser extremadamente prudente al establecer la hipótesis de partida y al seleccionar la muestra. En efecto, todos los individuos al inicio del periodo están sometidos al riesgo de todos los desenlaces, razón por la cual, a diferencia de las comparaciones en el marco de eventos reversibles, no se puede considerar una serie de edad exclusiva que selecciona las poblaciones sometidas a riesgo.

III. LOS *SOFTWARES*. PUESTA EN PRÁCTICA DE LOS PROCEDIMIENTOS INFORMÁTICOS

En este capítulo presentaremos dos ejemplos de utilización del *software*, con una intención pedagógica.

El paquete se compone de:

CONTROL.RAT	programa de formación del archivo de los comandos necesarios para ejecutar los cálculos.
ROOT.RAT	programa fuente del cálculo.
ORGA.RAT	construye un ejecutable de ROOT a partir del archivo de comando y de la fuente ROOT.RAT.
COMPARE.RAT	programa de comparaciones posteriores de los resultados obtenidos.

Los archivos creados por el ejecutable son:

(nombres dados cuando se crea el archivo de comandos)	archivos de los comandos archivos de los resultados
COMPTE.DAT	si se han pedido conteos.
APART.DAT	para los tratamientos "apar" de las simultaneidades.
TAUX.DAT	cuando se quiere hacer comparaciones cruzadas de los cocientes instantáneos de T2edad, caso de reversibilidad de los eventos considerados.
COMPT1.DAT	cuando se quiere hacer comparaciones
COMPT2.DAT	simples de los cocientes de T1edad, T2edad.

A) Actividad femenina y fecundidad

Para este ejemplo se dispone de un archivo que reúne a la población de un país dado, de cuatro cohortes diferentes, consigna para cada individuo su actividad profesional y el nivel de ingreso de la pareja al casarse; los dos eventos cuyas interacciones van a estudiarse son el primer nacimiento y el abandono de la actividad y, respectivamente, el reinicio de la actividad o el primer ingreso a la actividad de las mujeres.

Se tienen así cuatro criterios de selección (*cf.* II.B.1), una serie de edades exclusiva: la que comprende ya sea la edad cuando se abandona la actividad, sea la edad cuando se la retoma o cuando ingresa por primera vez a ella (*cf.* II.C.1), y una edad cuando se produce el primer nacimiento. A esto hay que

añadirle, para la serie de las edades, un indicador (*cf.* II.C.1), que especifica el rango del valor distinto a cero de la serie, sea un indicador que precisa si la mujer está activa o inactiva al casarse (este indicador puede no existir para los hombres de la muestra, pues éstos no entran jamás en el análisis).

Extracto del archivo:

1	2	2100	23	2	22	0	23
2	2	1800	18	4	0	26	99
1	2	2100	24	2	22	0	37
1	2	2400	32	1	26	0	32
0	1	2000	21	3	21	33	
0	1	2100	46	4	22	38	
2	2	2700	59	3	0	28	59
1	2	2100	29	2	24	0	99
0	1	2200	34	2	23	42	
0	1	2100	21	3	22	99	
0	1	1800	19	1	20	22	

Observaciones

El archivo formado por enteros contiene las informaciones de un individuo por línea: indicador de actividad al casarse para las mujeres (activa = 1/ inactiva = 2), el sexo, el nivel de ingresos de la pareja al casarse, ocupación profesional del individuo, cohorte de nacimiento (código de 1 a 4), edad de salida de la actividad, edad cuando se retoma la actividad y edad cuando se produce el primer nacimiento para las mujeres, y para los hombres su edad al casarse y cuando se produce el primer nacimiento.

Dado que el análisis sólo se interesa por la población femenina, la no homogeneidad de los datos entre hombres y mujeres no plantea problemas. Basta solamente con que el criterio que sirve para separarlos se encuentre en un mismo lugar para los dos sexos.

El indicador de actividad es igual al rango distinto de cero de la serie de edades de entrada y salida de actividad.

1) La formación de los comandos necesarios

Con tal propósito se ejecuta el programa CONTROL.RAT:

Ese *software* permite crear el archivo de comandos en la forma de "conversación". En nuestro ejemplo:

nombre del archivo de comandos:	PUBLI.DAT
¿nombre del archivo de datos?	ACTENF.DAT
¿nombre del archivo de resultados?	RESU.RES

| ¿dimensión del vector de las edades T1? (1-8) 2
| ¿dimensión del vector de las edades T2? (1-8) 1
| ¿núm. de criterios de selección? (ninguno = REGRESO, 1-9) 4
| dé las variables tal como serán leídas
| ellas deben comprender T1edad(1), T2edad(1) y eventualmente
| pueden comprender ind1, ind2, crit(1 a 9)
| T1edad (2 a 8), T2edad (2 a 8), Obs(1 a 8)
| La lista de esas variables excede los 50 caracteres
| escriba (sí/1, no/0) 1
|
| La lista se escribirá en dos líneas
| dar la primera línea terminada por una , (coma)
| Señal 1 2 3 4 5
| *ind1, crit(1), crit(3), crit(4),T1edad(1)*
| dar la última línea, puede tener 80 caracteres
| Señal 1 2 3 4 5 6
|
| *T1edad(2), T2edad(1)*
|
| formato FORTRAN de los datos por leer 2i2,1x,i4,5i3
|
| nombre de 1 evento T1 entrada/salida
|
| nombre de 1 evento T2 1er nac.
|
| prueba escogida para el primer criterio 10
|
| usted escogió una prueba simple
| ¿valor de referencia de la prueba? 2
|
| prueba escogida para el segundo criterio 40
|
| usted escogió una prueba simple
| ¿valor de referencia de la prueba? 1500
|
| prueba escogida para el tercer criterio 0
|
| usted no realiza ninguna prueba
| prueba escogida para el cuarto criterio 11
|
| usted ha escogido una prueba doble
| ¿primer valor de referencia de la prueba? 3

¿segundo valor de referencia de la prueba?	4
edad mínima de 1 análisis	23
edad máxima de 1 análisis	28
núm. de tratamientos demandados	2
Primer tratamiento, título: ACTIVAS AL CASARSE, SALIDA DE ACTIVIDAD Y PRIMER NACIMIENTO	
rango de 1 edad elegida para el primer evento	2
tratamiento de simultaneidades ¿con/sin/apar/last?	APAR
¿qué método? (estimación de las partidas y llegadas = 1 /rudimentario = 2)	1
¿quiere obtener los conteos? (sí/1, no/0)	0
¿quiere prolongar 1 análisis mediante una comparación cruzada? (sí/1, no/0)	1
¿quiere prolongar 1 análisis mediante una comparación simple de los cocientes de T1edad? (sí/1, no/0)	0
¿quiere prolongar 1 análisis para una comparación simple de los cocientes de T2edad? (sí/1, no/0)	0
Segundo tratamiento, título: INACTIVAS AL CASARSE, ENTRADA EN ACTIVIDAD Y PRIMER NACIMIENTO	
rango de la edad escogida para el primer evento	1
tratamiento de las simultaneidades ¿con/sin/apar/last ?	APAR
¿qué método? (estimación de las partidas y llegadas = 1 / rudimentario = 2)	1
¿quiere obtener los conteos? (sí/1, no/0)	0

```
|
| ¿quiere prolongar el análisis mediante una
|  comparación cruzada? (sí/1, no/0)                          1
|
| ¿quiere prolongar el análisis mediante una
| comparación simple de los cocientes T1edad?
| (sí/1,no/0)                                                 0
|
| ¿quiere prolongar el análisis mediante
| una comparación simple de los cocientes T2edad? (sí/1,no/0)  0
```

Este tipo de "conversación" difiere de un caso a otro según el número de criterios, el tamaño de los vectores de edad, etcétera.

Se creó así el archivo de comandos PUBLI.DAT siguiente:

```
publi.dat
actenf.dat
resu.res
2
1
4
ind2, crit(1), crit(2), crit(3), crit(4), T1edad(1), T1edad(2),T2edad(1)

2i2, 1x, i4, 5i3
Evt=entrada/salida
Evt=1er nac.
10      2      0
40   1500      0
 0      0      0
11      3      4
23     28
ACTIVAS AL CASARSE, SALIDA DE ACTIVIDAD Y PRIMER NACIMIENTO
2  1 apar 1 0 1 0 0
INACTIVAS AL CASARSE, ENTRADA EN ACTIVIDAD Y PRIMER NACIMIENTO
1  1 apar 1 0 1 0 0
```

2) Escritura del software *de cálculo*

Se hace mediante el *software* ORGA.RAT a partir de la fuente ROOT.RAT y del archivo de comandos (aquí PUBLI.DAT).

Este *software* solicita dos informaciones en la forma de "conversación",

"$ nombre del programa ejecutable":

el nombre del ejecutable que se crea debe tener un .RAT como extensión (la fuente fue escrita en Ratfor)

"$ nombre del archivo de comandos":

el nombre del archivo de comandos.

3) Los archivos creados por el ejecutable

APAR.DAT y RESU.RES archivos de los resultados

ACTIVAS AL CASARSE, SALIDA DE ACTIVIDAD Y PRIMER NACIMIENTO
apar 1
COCIENTES INSTANTÁNEOS DE SIMULTANEIDAD
APAR.DAT

Edad selección:	Pob. 1	Evt.	Coc.	Acumulado
23	807.50	22.00	0.0272	0.0272
24	699.00	18.00	0.0258	0.0530
25	597.50	13.00	0.0218	0.0748
26	507.00	15.00	0.0296	0.1043
27	440.00	7.00	0.0159	0.1202
28	382.50	11.00	0.0288	0.1490

RESU.RES (SÓLO EL ÚLTIMO TRATAMIENTO SE MUESTRA A CONTINUACIÓN)
INACTIVAS AL CASARSE, ENTRADA EN ACTIVIDAD Y PRIMER NACIMIENTO
apar 1
Edad: 23 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=entrada/salida									
596.00	5.00	0.0084	410.00	12.00	0.0293	2.2595	2.2595	2.4766	2.4766
Evt=primer nacimiento									
596.00	96.00	0.1611	9.00	1.00	0.0000	0.0000	0.0000	0.0000	0.0000

Edad: 24 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=entrada/salida									
490.50	4.00	0.0082	498.50	11.00	0.0221	1.7827	2.8576	1.8075	3.0293
Evt=primer nacimiento									
490.50	104.00	0.2120	13.00	0.00	0.0000	0.0000	0.0000	0.0000	0.0000

Edad: 25 ----------

Pob.antes	Evt	Coc.	Pob.después	Evt	Coc.	H(t)anual	H(t)	S(t)anual	S(t)
Evt=entrada/salida									
392.50	2.00	0.0051	581.50	10.00	0.0172	1.8551	3.4042	1.7616	3.4905
Evt=primer nacimiento									
392.50	63.00	0.2115	13.00	6.00	0.4615	1.3172	1.3172	1.6483	1.6483

Cocientes acumulados T1edad y T2edad

cqA(23)=0.0084 cqB(23)= 0.0293 cqA(23) = 0.1611 cqB(23) = 0.0000

cqA(24)=0.0165 cqB(24)= 0.0513 cqA(24) = 0.3731 cqB(24) = 0.0000

cqA(25)=0.0216 cqB(25)= 0.0685 cqA(25) = 0.5846 cqB(25) = 0.4615

TAUX.DAT constituido por dos tratamientos consecutivos

ACTIVAS AL CASARSE, SALIDA DE ACTIVIDAD Y PRIMER NACIMIENTO

Evt = 1er nacimiento

Edad	Pob. antes	Evento	Cociente	Pob. después	Evento	Cociente
23	807.50	82.00	0.1015	10.50	4.00	0.3810
24	699.00	74.00	0.1059	16.00	6.00	0.3750
25	597.50	77.00	0.1289	20.50	6.00	0.2927
26	507.00	63.00	0.1243	23.50	1.00	0.0426
27	440.00	40.00	0.0909	24.00	7.00	0.2917
28	382.50	45.00	0.1176	24.50	4.00	0.1633

INACTIVAS AL CASARSE, ENTRADA EN ACTIVIDAD Y PRIMER NACIMIENTO

Evt = 1er nacimiento

Edad	Pob. antes	Evento	Cociente	Pob. después	Evento	Cociente
23	596.00	96.00	0.1611	9.00	1.00	0.0000
24	490.50	104.00	0.2120	13.00	0.00	0.0000
25	392.50	83.00	0.2115	13.00	6.00	0.4615
26	310.50	75.00	0.2415	11.50	1.00	0.0870
27	241.50	55.00	0.2277	13.50	1.00	0.0741
28	191.50	34.00	0.1775	17.00	0.00	0.0000

4) Las comparaciones ulteriores

Cuando se ha solicitado extender el análisis mediante una comparación cruzada, el *software* produce el archivo TAUX.DAT que presentamos antes. Se efectuarán entonces las comparaciones ejecutando el programa COMPARE.RAT. Este *software* interroga de manera interactiva sobre el tipo de comparación deseado:

	respuesta
¿quiere una comparación cruzada? (sí=1/no=0)	1
nombre del archivo de cocientes a comparar: si se hace una comparación cruzada, TAUX.DAT	TAUX.DAT
nombre del archivo de resultados	COMPARE.DAT
edad de inicio de la comparación	23
edad máxima de la comparación	27

Si las edades leídas en los datos (TAUX.DAT) no corresponden a las que se dan, el *software* lo indica y se interrumpe.

Éste es el archivo COMPARE.DAT que se obtiene:

COMPARE.DAT
1er análisis = ACTIVAS AL CASARSE, SALIDA DE ACTIVIDAD Evt = 1er nac.
2o análisis = INACTIVAS AL CASARSE, ENTRADA EN ACTIVIDAD Evt = 1er nac.
COMPARACIÓN : Tx antes del primer análisis
Tx después del segundo análisis

Edad	Coc.	Coc.	H(t) anual	H(t)	S(t)anual	S(t)
23	0.1015	0.0000	0.0000	0.0000	0.0000	0.0000
24	0.1059	0.0000	0.0000	0.0000	0.0000	0.0000
25	0.1289	0.4615	1.7599	1.7599	2.5036	2.5036
26	0.1243	0.0870	-0.4221	0.9460	-0.3751	1.5051
27	0.0909	0.0741	-0.2226	0.6439	-0.2085	1.1085
28	0.1176	0.0000	0.0000	0.6439	0.0000	1.1085

COMPARACIÓN : Tx antes del segundo análisis
Tx después del primer análisis

Edad	Coc.	Coc.	H(t) anual	H(t)	S(t)anual	S(t)
23	0.1611	0.3810	1.1501	1.1505	1.4793	1.4793
24	0.2120	0.3750	1.0550	1.5593	1.2479	1.9284
25	0.2115	0.2927	0.6671	1.6583	0.7333	1.9979
26	0.2415	0.0426	-3.9079	-0.5178	-2.3010	0.5798
27	0.2277	0.2917	0.5592	-0.2131	0.5976	0.7858
28	0.1775	0.1633	-0.1630	-0.2610	-0.1595	0.6522

B) PRIMERA MIGRACIÓN DESPUÉS DEL MATRIMONIO Y PRIMER NACIMIENTO

Para este ejemplo disponemos de un archivo que nos da, para cada individuo, su edad al casarse y luego las duraciones transcurridas en meses desde el matrimonio hasta la primera migración y el primer nacimiento. Esta vez puede ser que los eventos no hayan sido observados en la fecha de la encuesta y se dispone entonces de un índice de truncamiento que vale 1 si el evento es observado, y 0 en el caso contrario. La duración transcurrida desde el matrimonio hasta la migración se clasifica en un vector, con un rango que corresponde a la región que constituye el destino del desplazamiento. También se presentan algunos criterios de selección.
Extracto del archivo:

1	2	21	23	2	0	5	0	0
1	2	18	18	4	0	0	0	21
1	2	11	24	2	0	6	0	0
1	2	14	15	1	17	0	0	0
0	1	20	21	3	0	0	48	0
0	1	1	16	4	0	0	0	48
1	2	27	19	3	0	0	13	0
1	2	6	15	2	0	16	0	0
0	1	2	14	2	0	48	0	0
0	1	21	21	3	0	0	48	0
0	1	10	19	1	48	0	0	0
1	2	16	18	4	0	0	0	1
1	2	11	24	2	0	6	0	0

Observaciones

Por individuo (por línea) tenemos entonces el índice de truncamiento, el sexo, la duración transcurrida desde el matrimonio, la edad al casarse, el indicador del rango de la duración transcurrida hasta la migración, y por último la duración transcurrida hasta la migración.

1) La formación de los comandos necesarios

Mediante CONTROL.RAT se crea un archivo de comando llamado PUBCOM.DAT, que selecciona a los individuos de ISLAND.DAT casados antes de los 21 años y pide que se efectúen los cinco tratamientos para duraciones a partir del matrimonio de seis a 20 meses.

Los cuatro primeros análisis atenderán al destino de los migrantes (rango 1 del vector de las migraciones = destino el Caribe, rango 2 = destino la

Metrópoli, rango 3 = destino el Continente Americano, rango 4 = todos los demás destinos). Para esos tratamientos, los casos de simultaneidad se incorporan uniformemente entre los dos eventos (*cf.* II.C.2.(4) y, para el primer análisis se recuperan los conteos.

El último análisis se hace de la forma "todos los destinos confundidos" con las simultaneidades contadas como migraciones, y se recuperan los conteos (*cf.* II.C.2.(5)).

```
Archivo PUBCOM.DAT
pubcom.dat
island.dat
test.dat
4
1
2
obs1(1), crit(1), T2edad(1), crit(2), ind1, T1edad(1),
T1edad(2), T1edad(3), T1edad(4)
i1, 5x, i2, 2i4, 5i3
evt: migración
Evt: primer nacimiento
0       0       0
30      20      0
6       20
PRIMER NACIMIENTO Y MIGRACIÓN AL CARIBE
1   1       sin     2       1       00      0
PRIMER NACIMIENTO Y MIGRACIÓN DEL CARIBE HACIA LA METRÓPOLI
2   1       sin     2       0       01      0
PRIMER NACIMIENTO Y MIGRACIÓN DEL CARIBE HACIA EU
3   1       sin     2       0       01      0
PRIMER NACIMIENTO Y MIGRACIÓN DEL CARIBE HACIA OTROS DESTINOS
4   1       sin     2       0       00      0
PRIMER NACIMIENTO Y MIGRACIÓN FUERA DEL CARIBE
9   1       last    1       1       00      0
```

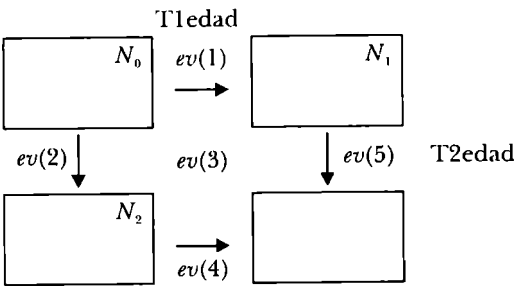
2) Escritura del software *del cálculo*

Se crea un ejecutable diferente del anterior con ORGA.RAT, que explota entonces los datos del archivo de comando que sigue.

3) Archivos que crea el ejecutable

TEST.DAT archivo de resultados para este análisis.
COMPTE.DAT conteos efectuados por el primero y el último tratamientos.
COMPT1.DAT archivo creado con los tratamientos 2 y 3 para comparaciones simples.

Los conteos se dan antes de hacerse cualquier estimación; volvemos a encontrar así la situación observada al inicio del periodo. Se tendrán las cuentas:



dispuestas como sigue:

Pob. sometida al riesgo N_0, N_1, N_2
eventos ev(1), ev(2), ev(3), ev(4), ev(5)

COMPTE.DAT
PRIMER NACIMIENTO Y MIGRACIÓN AL CARIBE
Edad: 6, rango escogido para T1edad: 1, rango escogido para T2edad: 1

Pob. en riesgo 647. 368. 7.
evt 6. 96. 1. 12. 1.

Edad: 7, rango escogido para T1edad: 1, rango escogido para T2edad: 1

pob. en riesgo 545. 452. 11
evt 4. 104. 1. 11. 0.

Edad: 8, rango escogido para T1edad: 1, rango escogido para T2edad: 1

pob. en riesgo 436. 545. 15.
evt 2. 83. 2. 10. 6.

Edad: 9, rango escogido para T1edad: 1, rango escogido para T2edad: 1

pob. en riesgo 349. 618. 11.
evt 2. 75. 0. 12. 1.

COMPTE.DAT al final del listado, el último tratamiento proporciona las cuentas para el análisis de la forma "todos los destinos confundidos" en mayor detalle.

Podemos además destacar que la selección 1 es el conteo para el rango 1 del vector T1edad y que se encuentra en las cuentas de la página anterior.

Edad: 6, rango escogido para T1edad: 9, rango escogido para T2edad: 1

selección: 1
pob. en riesgo 647. 368. 7.
evt 5. 96. 1. 12. 1.

selección: 2
pob. en riesgo 701. 1. 127.
evt 5. 96. 1. 12. 1.

selección: 3
pob. en riesgo 13. 76. 24.
evt 4. 9. 5. 20. 1.

selección: 4
pob. en riesgo 216. 122. 58.
evt 57. 48. 67. 0. 0.

selección: 5
pob. en riesgo 0. 0. 0.
evt 0. 0. 0. 0. 0.

selección 6
pob. en riesgo 0. 0. 0.
evt 0. 0. 0. 0. 0.

selección 8
pob. en riesgo 0. 0. 0.
evt 0. 0. 0. 0. 0.

Edad: 7 rango escogido para T1edad: 9, rango escogido para T2edad: 1

COMPT1.DAT El archivo para la comparación se presenta de manera idéntica a TAUX.DAT

PRIMER NACIMIENTO Y MIGRACIÓN DEL CARIBE HACIA LA METRÓPOLI
evt= migración

Edad	Pob.antes	Evt	Coc.	Pob. después	Evt	Coc.
23	807.50	52.00	0.1015	10.50	4.00	0.3810
24	699.00	74.00	0.1059	16.00	6.00	0.3750
25	597.50	77.00	0.1289	20.50	6.00	0.2927
26	507.00	63.00	0.1243	23.50	1.00	0.0426
27	440.00	40.00	0.0909	24.00	7.00	0.2917

PRIMER NACIMIENTO Y MIGRACIÓN DEL CARIBE HACIA EU
evt= migración

Edad	Pob.antes	Evt	Coc.	Pob. después	Evt	Coc.
23	596.00	5.00	0.0084	410.00	12.00	0.0293
24	490.50	4.00	0.0082	498.50	11.00	0.0221
25	392.50	2.00	0.0051	591.50	10.00	0.0172
26	310.50	2.00	0.0064	649.50	12.00	0.0185
27	241.50	4.00	0.0166	700.50	16.00	0.0228

4) Comparaciones ulteriores

Esta vez es diferente el diálogo cuando se ejecuta COMPARE.DAT:

	respuesta
¿hace una comparación cruzada (sí= 1/no=0)?	0
nombre del archivo de los cocientes a comparar: si se hace una comparación cruzada, TAUX.DAT	COMPT1.DAT
nombre del archivo de los resultados	DIFMIGR.DAT
edad de inicio de la comparación	23
edad máxima de la comparación	27

El resultado se encuentra a partir de ahora en el archivo DIFMIGR.DAT siguiente:

DIFMIGR.DAT
1er análisis = PRIMER NACIMIENTO Y MIGRACIÓN DESTINO EL CARIBE Evt = migración
2o análisis = PRIMER NACIMIENTO Y MIGRACIÓN DESTINO EL CARIBE Evt = migración
Comparación de los Tx antes de la perturbación

Edad	Coc.	Coc.	H(t) anual	H(t)	S(t)anual	S(t)
23	0.1015	0.3810	1.4647	1.4647	2.1180	2.1180
24	0.1059	0.3750	1.7521	2.2747	2.4843	3.2544
25	0.1289	0.2927	1.3606	2.6428	1.7193	3.6498
26	0.1243	0.0426	-1.8010	1.3882	-1.2665	2.5276
27	0.0909	0.2917	1.8061	2.0494	2.4591	3.3605

Comparación de los Tx después de la perturbación

Edad	Coc.	Coc.	H(t) anual	H(t)	S(t)anual	S(t)
23	0.0084	0.0293	2.2595	2.2595	2.4762	2.4762
24	0.0082	0.0221	1.7790	2.8557	1.8000	3.0238
25	0.0051	0.0172	1.8655	3.4087	1.7605	3.4853
26	0.0064	0.0185	1.7269	3.8154	1.5569	3.7968
27	0.0166	0.0228	0.6161	3.6881	0.5852	3.6577

IV. LOS MENSAJES DE ERROR

1) Los mensajes que provienen de CONTROL.RAT

El *software* controla:

- el tamaño máximo de los vectores T1edad y T2edad, que no puede ser mayor que 8;
- el número de criterios de selección, 9 como máximo;
- la codificación de las pruebas, pero ese filtro no es impermeable y deja pasar valores no previstos que no serán detectados como falsos hasta hacerse los cálculos;
- el rango escogido del vector de las edades respecto del tamaño de éste;
- cuando se escoge el tratamiento de las simultaneidades, el mensaje "error de ortografía" advierte sobre una mala selección sin interrumpir el desarrollo de CONTROL.RAT. Habrá entonces que recomenzar o modificar el valor erróneo entre (con/sin/apar/last) en el archivo creado.

No se ha establecido ningún control sobre:

- la denominación de las variables o de su formato. El usuario debe desconfiar, aunque los errores en esta etapa inevitablemente se indican en el cálculo.

2) Los mensajes que provienen de ORGA.RAT

No se ha previsto ningún mensaje específico; únicamente archivos no encontrados o mal construidos, lo que implica errores de lectura o escritura.

3) Los mensajes que provienen de ROOT.RAT

Inicialmente el *software* lee ciertos valores del archivo de comando y realiza un nuevo control sobre el rango escogido para las edades, respecto del valor máximo del vector y sobre las especificaciones de tratamiento de las simultaneidades (su nombre o su versión).

En el momento de la ejecución se efectúan pruebas para cada individuo sobre los valores ind1 e ind2 (índices facultativos de rango leídos para cada individuo) que se comparan con el tamaño del vector.

Finalmente, si un código de prueba de selección de un criterio es falso, el *software* lo indica.

Esos errores señalados no siempre interrumpen los cálculos, sin embargo los resultados obtenidos luego de una sesión donde se presentan errores son siempre falsos.

4) Algunas modificaciones posibles

Sin tener que reescribir el ejecutable, hay ciertos valores del archivo de los comandos que pueden ser modificados:

Las pruebas sobre los criterios;

las edades mínima y máxima del análisis, y

la especificación de los tratamientos: el título y las selecciones que le siguen.

V. BIBLIOGRAFÍA

A continuación presentamos algunas referencias sucintas a los trabajos metodológicos que se relacionan con este *software*, así como artículos sobre la aplicación de los métodos.

Aalen, O. (1978), "Nonparametric Interference for a Family of Counting Processes", *The Annals of Statistics*, vol. 6, núm. 4, pp. 701-726.

Aalen, O., O. Borgan, N. Keiding y J. Thorman (1980), "Interaction Between Life History Events: Nonparametric Analysis for Prospective and Retrospective Data in Presence of Censoring", *Scandinavian Journal of Statistics*, núm. 7, pp. 161-171.

Courgeau, D. (1984), "Relations entre cycle de vie et migrations", *Population*, núm. 3, pp. 483-514.

Courgeau, D. (1987), "Constitution de la famille et urbanisation", *Population*, núm. 1, pp. 57-82.

Courgeau, D. y E. Lelièvre (1985), "Estimations of transition rates in dynamic household models", en N. Keilman (comp.), *Modelling Household Formation and Dissolution*, Oxford University (de próxima publicación).

Courgeau, D. y E. Lelièvre (1986), "Nuptialité et agriculture", *Population*, núm. 2, pp. 303-326.

Courgeau, D., E. Lelièvre y M. Wagner (1986), "Leaving home and marriage in France and Germany" (de próxima publicación).

Hoem, J. y U. Funck Jensen (1982), "Multistate life table methodology: a probabilistic critique", en Land and Rogers (comps.) *Multidimensional Mahtematical Demography*, Nueva York, Academic Press, pp. 155-264.

Lelièvre E. (1986), "The analysis of interactions between phenomena: data –a french survey–, tools, first results", septiembre, Sopron (Hungría), Conferencia IIASA.

Lelèvrie E. (1987), "Activité professionnelle et fécondité: les choix et les déterminations des femmes françaises entre 1930 y 1960", Finlandia, Coloquio Internacional de Demografía IUSSP.

Schou, G. y M. Vaeth (1980), "A small sample study of ocurrence/exposure rates for rare events", *Scandinavian Actuarial Journal*, núm. 4, pp. 209-225.

ANEXO 3

MANUAL DE UTILIZACIÓN DE EVACOV. FOR

Febrero de 1986

Éva Lelièvre
INED

INTRODUCCIÓN

EVACOV.FOR es un *software* Fortran puesto en funcionamiento por el Instituto Nacional de Estudios Demográficos (INED), dentro del marco de una tesis sobre "Los métodos matemáticos y estadísticos de análisis de historias de vida individuales".

Este *software* de evaluación del papel de las covariables en el análisis de las interferencias entre dos fenómenos, se ha inspirado ampliamente en el que proponen Kalbfleisch y Prentice (1980); sin embargo, éste permite analizar archivos de mayor envergadura, hasta de 2 000 individuos cada uno, con un vector que puede incluir 30 variables explicativas.

Empero, las rutinas de maximización y el espíritu de este análisis —capaz de obtener resultados significativos incluso cuando se trata de efectivos reducidos— consumen un tiempo CPU muy importante y por ello no se justifica una extensión en el futuro de las capacidades máximas de este programa.

Esta versión, aún poco perfeccionada, sólo debe considerarse como una fase de la puesta en funcionamiento de un *software* más flexible y potente; por esta razón desearía recibir sugerencias del usuario que ayudaran a mejorar este programa, y le pido que sea indulgente con las imperfecciones de la versión actual.

Análisis semiparamétrico

Este análisis permite estudiar las interacciones entre dos eventos (matrimonio y abandono de la agricultura; migración y fecundidad, Courgeau y Lelièvre, s.f.; 1986), señalados por T_1 y T_2, su fecha de ocurrencia, que comúnmente es la edad del individuo cuando estos eventos se producen. Este análisis, que formaliza la noción intuitiva de que un proceso estocástico, puede influir localmente en el desarrollo de otro en la fecha t y permite poner en evidencia el efecto de las diversas características del individuo sobre la ocurrencia de uno u otro de los eventos seleccionados. Sucesivamente T_1 (respecto de T_2) será el evento final estudiado cuya ocurrencia resulta perturbada, o no, por T_2 (respecto de T_1) y después el evento perturbador cuando ocurre T_2 (respecto de T_1).

Aquí se supone que las características individuales actúan de manera multiplicativa sobre los cocientes instantáneos (o intensidades) calculados. Sin embargo, el efecto de una característica puede ser modificado según el individuo haya experimentado o no el evento perturbador.

En el estudio de los efectos que ocasiona el abandono de la agricultura sobre la nupcialidad, el hecho de provenir de una familia campesina multiplica en cada edad el cociente instantáneo de nupcialidad de un individuo

por el mismo factor que indica su origen; pero como se admite que esta influencia puede cambiar si el individuo sale del mundo agrícola, ese factor varía también en función de su actividad profesional.

Formalización matemática

Sean T_1, T_2 dos variables aleatorias que corresponden respectivamente a la edad (o la fecha) de ocurrencia de dos eventos.

Sea Z un vector de n variables explicativas

$$Z = \left(z_1, \underbrace{\ldots\ldots, z_r}_{r}, z_{r+1}, \underbrace{\ldots\ldots, z_n}_{s} \right)$$

compuesto de:

- r variables explicativas independientes de la perturbación;
- s variables explicativas asociadas a la perturbación que son nulas antes de la fecha de la perturbación.

Tomando la hipótesis de que sólo interviene la edad en el momento del evento final y no la duración transcurrida desde la perturbación (de no ser así se requeriría una estimación para cada valor de $T_2 - T_1$), se consideran las densidades siguientes:

$$\lambda_{0i}(t; z) = \lambda(t) \exp(z.\beta_1)$$

$$\lambda_{ij}(t / u; z) = \lambda(t) \exp(z.\beta_2) \qquad u < t$$

donde β_1 y β_2 son los vectores parámetros de la regresión que mide, respectivamente, el efecto de Z sobre la densidad marginal y condicional. Es importante señalar que así se introducen en la densidad condicional nuevas variables explicativas: las que están ligadas a la perturbación.

Los parámetros se estiman mediante los métodos de verosimilitud parcial propuestos por Kalbfleisch y Prentice (1980); siguiendo a Crowley y Hu (1977) utilizamos una hipótesis simplificadora de proporcionalidad de los cocientes, es decir, que la densidad puede expresarse de manera más condensada por:

$$\lambda(t) \exp[Z\beta_1 + H(t-u)(\beta_0 + Z\beta_2)]$$

con

$$H(x) = \begin{array}{l} 0 \text{ si } x <= 0 \\ 1 \text{ si } x > 0 \end{array}$$

β_0, β_1 y β_2 también se estiman mediante el método de verosimilitud parcial.

En ese modelo:

$Z\beta_1$ mide los efectos principales entre las variables de Z y el evento final;

$\beta_0 + Z\beta_2$ una vez experimentada la perturbación, mide el efecto de la perturbación y la interacción entre las variables y la ocurrencia del evento final.

En los cálculos, el vector de las características evoluciona de la manera siguiente:

Disponemos de r características independientes de la perturbación y de s características adquiridas con la perturbación,

antes de la fecha de ocurrencia del evento perturbador

Z = serie de r valores seguidos de $(n+1)$ valores iguales a cero

a partir de esa fecha

r valores de las características independientes

$Z = 1$

r valores (repetidos)

s valores de las características adquiridas.

Esto permite mediar la evolución del papel de una característica antes y después de la perturbación.

Observaciones sobre la interpretación de los parámetros

La interpretación absoluta de β_0 (parámetro que mide el efecto de la perturbación misma) es difícil, pues su significación se ve afectada por la modificación de Z.

Si una variable independiente se incluye (para el análisis) en la parte "dependiente" del vector después de la perturbación y es excluida antes, la interpretación del resultado es poco clara, pues el valor del parámetro estará influido a la vez por la dependencia de la variable y de la edad en el estado prefinal, así como por la dependencia de la variable y de la perturbación.

El programa también permite analizar los datos clasificados en diferentes niveles (cuatro como máximo); tal partición sólo será legítima en caso de que un análisis anterior (por ejemplo de tipo no paramétrico) haya permitido confirmar el carácter proporcional del comportamiento interno en los grupos diferenciados por la partición, y un efecto no proporcional de las variables consideradas entre los diversos grupos.

CONFORMACIÓN DE LOS DATOS; PUESTA EN MARCHA DE EVACOV

Datos de análisis

Los datos necesarios para el análisis deben presentarse bajo la forma siguiente:

• una línea para cada individuo en el siguiente formato:	|_|_ _|_ _|_ _ _|_ _ _|_ ...
• salto de un carácter	1x
• nb de NIVELES	I2 (entero 2 caracteres)
• índice de CENSURA	I2 (entero 2 caracteres)
• OCURRENCIA (del evento final)	
• PERTURB	F3.0 (relativos sin parte decimal)
• nvb1	variables independientes
• nvb2	asociadas.

El índice de censura es relativo al evento final, OCURRENCIA, si éste es observado el índice vale 1, si la información es censurada ese índice vale 0.

OCURRENCIA y PERTURB son la edad o la fecha de ocurrencia de los dos eventos.

Esos datos se deben clasificar por orden creciente de OCURRENCIA, y en cada grupo de edad formado de esa manera los individuos no censurados (censura = 1) deben preceder a los que se salen de la observación (censura = 0).

Órdenes necesarias para poner en funcionamiento el software

Hay un pequeño *software*, ORDRES.FOR, que está a disposición del usuario para formar este conjunto de valores de comando en la forma de "conversación". Este *software* pregunta sucesivamente sobre:

- tamaño del archivo de datos (< = 2000 ind)
- núm. de niveles (comprendidos entre 1 y 4)
- núm. de características independientes de la perturbación (< =8)

- núm. de características asociadas a la perturbación (< =13)
- núm. de tratamientos demandados (< =30)
  además,
- el nombre de los diferentes niveles
- el nombre de las características independientes y sus utilizaciones sucesivas [1 incluida en el tratamiento, –1 excluida del análisis]
- el nombre de las características asociadas y sus utilizaciones sucesivas [*idem*].

Nombre de los archivos utilizados por EVACOV.FOR

unit = 7: archivo de comandos ORDRES.DAT
unit = 8: archivo de datos EVACOV.DAT
unit = 9: archivo de resultados RESUL.DAT

El usuario tiene la libertad de utilizar nombres elegidos por él al inicio del programa en los "open(..."

EVACOV.FOR

Este *software* Fortran (FORTRAN VAX 77) se presenta con subrutinas de base incorporadas, pero hay que ligarlo con una subrutina de matriz inversa a la que debemos transmitirle:

ENN, la matriz cuya inversa, VAR, es la matriz de varianza/covarianza de los estimadores β. Esas dos matrices se presentan con dimensión 30 en el programa principal (núm. máximo de variables del análisis), pero son de tamaño mc1 × mc1 (núm. de variables del análisis en curso, mc1 = 1 en 30).

ENN y VAR son los nombres que utiliza el programa, una subrutina de la biblioteca IMSL:

LINVIF (ENN, mc1, 30, VAR, idec, wpet, ier) se puede utilizar cuando la biblioteca está disponible.

Ejemplo: Las reproducciones del listado correspondientes se encuentran en el anexo.

Archivo de los comandos creados por ORDRES.FOR

En la primera línea se indican el número de los individuos del archivo de datos, el número de las variables independientes de la perturbación, el número total de variables del vector suministrado en los datos, el número

total de las variables que se toman en cuenta en el análisis, y el número de niveles.

(¡Atención! El vector de las variables explicativas de este ejemplo no comprende más que nueve coordenadas en f3.0 y el vector de trabajo comprende $17 = 2 \times 7 + 1 + 2$).

En la segunda línea se lee el nombre del nivel

Las líneas siguientes llevan el nombre de las variables independientes (antes de la perturbación y después), y luego el de las dos variables asociadas a la perturbación precedidas por el "estatus modificado", variable que toma el valor 1 desde la fecha de la perturbación.

Por último, las series de 1 y −1 que corresponden a los ocho diferentes tratamientos solicitados (observar que el "estatus modificado" que indica el hecho de haber experimentado la perturbación o no, siempre se toma en cuenta en el análisis).

Archivo de los resultados: RESUL.DAT

El que se presenta como anexo corresponde al primer tratamiento solicitado por el archivo anteriormente comentado.

La variable analizada indica el papel de los diplomas; el parámetro es de valor opuesto antes y después de la perturbación, pero su papel sólo es significativo antes de la perturbación donde el estadístico de prueba vale 0.02 (x_2 con 928 grados de libertad).

En seguida están disponibles los cocientes instantáneos de ocurrencia en cada edad de los individuos.

Los mensajes de error

Los errores que se presentan suelen deberse a la imposibilidad de invertir la matriz de varianza-covarianza y con frecuencia son producto de una particularidad de los datos, o de un error de especificación de las dimensiones de los diferentes vectores cuando se crea el archivo de comandos (en las versiones futuras del *software* se incluirá una descripción completa de los comandos, el valor mínimo y máximo de cada uno, la desviación estándar); sin embargo, pueden deberse a una mala presentación del archivo de datos.

También puede suceder que debido al papel simétrico de dos o más variables, la matriz estimada sea la matriz de identidad, en cuyo caso las estimaciones siguientes no tendrían una verdadera significación.

Por último, si el archivo que se procesa es de gran tamaño se puede alcanzar el tiempo CPU límite, lo que da un mensaje de error. Entonces no conviene enviar más que un número restringido de datos a la vez.

ARCHIVO DE DATOS

```
1   1   7.  40.  0.  1.  9.  0.  0.  0.  1.  0.  0.
1   1   7.  41.  0.  0.  4.  0.  0.  0.  1.  0.  0.
1   1   7.  36.  1.  0.  7.  0.  0.  0.  1.  0.  0.
1   1   7.  35.  3.  0.  7.  0.  1.  1.  2.  0.  0.
1   1   7.  37.  0.  0.  5.  0.  0.  1.  2.  0.  0.
1   1   8.  39.  2.  0.  1.  0.  0.  1.  4.  0.  0.
1   1   8.   7.  1.  0.  7.  1.  0.  1.  2.  1.  2.
1   1   8.  39.  0.  1.  1.  0.  0.  0.  2.  0.  0.
1   1   8.  41.  0.  0.  5.  0.  0.  1.  2.  0.  0.
1   1   8.  38.  1.  0.  1.  0.  0.  1.  2.  0.  0.
1   1   8.   7.  3.  1.  0.  0.  0.  1.  4.  1.  0.
1   1   8.  11.  1.  0.  9.  0.  1.  0.  2.  1.  3.
```

ARCHIVO DE COMANDOS

```
 929  7  9          17  1
umag 1 - 3 f
diploma
hijo mayor
núm. fs
res par
ne etr
cspr
csc1
diploma
hijo mayor
núm. fs
res par
ne etr
cspr
csc1
estatus modifiado
loc sor
csmig
```

```
 1 -1 -1 -1 -1 -1 -1  1 -1 -1 -1 -1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1  1 -1 -1 -1 -1 -1 -1  1 -1 -1 -1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1  1 -1 -1 -1 -1 -1 -1  1 -1 -1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1  1 -1 -1 -1 -1 -1 -1  1 -1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1  1 -1 -1 -1 -1 -1 -1  1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1  1 -1 -1 -1 -1 -1 -1  1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1  1 -1 -1 -1 -1 -1 -1  1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
-1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1      1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1
```

ARCHIVO DE RESULTADOS

Valores iniciales y datos

El número total de individuos es 929

la variable núm.	1	diploma	se itera	valor inicial del parámetro = 0.0000
la variable núm.	2	diploma	se itera	valor inicial del parámetro = 0.0000
la variable núm.	10	estatus modificado	se itera	valor inicial del parámetro = 0.0000

El número total de nivel es 1

Núm. del nivel	nombre del nivel	núm. de individuos en el nivel
1	umag 1-3 f	929

Estimadores del máximo de verosimilitud

el coeficiente de la variable	1	diploma	es	0.078190
el coeficiente de la variable	8	diploma	es	–0.078910
el coeficiente de la variable	15	estatus modificado	es	0.178681

cv se alcanza en tres iteraciones

el máximo de 1 log-verosimilitud es –5157.28759766

La matriz de covarianza estimada es

0.001662	–0.001618	0.001381
–0.001668	0.018335	–0.022256
0.001381	–0.022256	0.042216

Resumen

Coeficiente núm	N. D.	p valor (1 - sided)
1	1.917964	0.0276
8	–0.582765	0.2800
15	0.869645	0.1922

Rango	Edad en que ocurre el evento	Cociente
1	4.00	0.0060170
2	5.00	0.0110996
3	6.00	0.0219373
4	7.00	0.0533056
5	8.00	0.0620040
6	9.00	0.1105908
7	10.00	0.1255040

8	11.00	0.1670425
9	12.00	0.1404964
10	13.00	0.1229043
11	14.00	0.1237812
12	15.00	0.1139734
13	16.00	0.0953012
14	17.00	0.0910126
15	18.00	0.1177944
16	19.00	0.1301398
17	20.00	0.0671189
18	21.00	0.0845341
19	22.00	0.0731555
20	23.00	0.0433056
21	24.00	0.0227014
22	25.00	0.0853655
23	26.00	0.0170327
24	27.00	0.0173662
25	28.00	0.0353456
26	29.00	0.0091628
27	30.00	0.0135078
28	31.00	0.0094399
29	33.00	0.0190834
30	34.00	0.0097453
31	35.00	0.0098495
32	37.00	0.0099475
33	39.00	0.0100515
34	43.00	0.0203171
35	48.00	0.0217622

Bibliografía

Courgeau, D. y E. Lelièvre (s. f.), "Estimation of transition rates in dynamic household models", *Modelling Household Formation and Dissolution* (será publicado en Oxford University Press).

Courgeau, D. y E. Lelièvre (1986), "Nuptialité et Agriculture", *Population*, núm. 2.

Crowley, J. y M. Hu (1977), "Covariance analysis of heart transplant data", *Journal of the American Statistical Association (JASA)*, 72, pp. 27-36.

Hoem, J. y U. Funck Jensen (1982), "Multistate life table methodology: a probabilist critique", en Land y Rogers (comps.), *Multimensional Mathematical Demography*, Nueva York, Academic Press, pp. 155-264.

Kalbfleisch, J. y L. Prentice (1980), *The Statistical Analysis of Failure Time Data*, Nueva York, Wiley and Sons.

Schou, G. y M. Vaeth (1980), "A small sample study of occurrence/exposure rates for rare events", *Scandinavian Actuarial Journal*, núm. 4, pp. 209-225.

BIBLIOGRAFÍA POR TEMAS

MANUALES Y LIBROS

Allison, P.D. (1980), *Event History Analysis: Regression for Longitudinal Event Data*, Sage University Paper.
(Presentación para investigadores en ciencias humanas de los modelos paramétricos y semiparamétricos, con ejemplos de aplicación y una revisión de los *software* existentes.)

Anales de Vaucresson (1987), *Histoires de vie, Histoires de famille, Trajectoires sociales*, núm. 26, Trabajos del taller "Constitución de las trayectorias sociales", CRIV, 21, 22 y 23 de mayo de 1986.

Back, K.W. (comp.) (1980), *Life Course: Integrative Theories and Exemplary Populations*, American Association for the Advancement of Science, Selected Symposium 41.
(Una compilación de trabajos multidisciplinarios (antropología, sociología, demografía, historia...) sobre el tema de las trayectorias de vida, en la que todos los autores son estadunidenses: una visión estadunidense del tema.)

Blossfeld, H.P., A. Hamerle y K.U. Mayer (1986), *Ereignisanalyse. Statistische Theorie und anwendung in den Wirtschafts und Sozialwissenschaften*, Campus Verlag, Francfurt.

Blumen, I., K. Marvin y P.J. McCarthy (1955), *The Industrial Mobility of Labour as a Probability Process*, Nueva York, Cornell Studies of Industrial Labour Relations, núm. 4, vol. 6.

Bocquier, Philippe (1996), *L'analyse des enquêtes biographiques à l'aide du logiciel STATA*, Documents et Manuels du CEPED, París.

Coleman, J.S. (1981), *Longitudinal Data Analysis*, Nueva York, Basic Books, Inc. Publishers.

Collomb, Ph. (1987), "La mort de l'orme séculaire. Crise agricole et migration dans l'Ouest audois des années cinquante", *Cahier de l'INED*, núms. 105 y 106, PUF.

Courgeau, D. (1980), *Analyse Quantitative des Migrations Humaines*, París, Masson.

Courgeau, D. (1988), *Méthodes de mesure de la mobilité spatiale. Migrations internes, mobilité temporaire, navettes*, París, Ediciones del INED.

Cox, D.R.y D.V. Hinkley (1974), *Theoretical Statistics*, Nueva York y Londres, Chapman and Hall.

Cox, D. y D. Oakes (1984), *Analysis of Survival Data*, Londres, Chapman and Hall.

Chiang, C. (1968), *Introduction to Stochastic Processes in Biostatistics*, Nueva York, Wiley and Sons.

Deroo, M. y A.M. Dussaix (1980), *Pratiques et Analyse des Enquêtes par sondage*, París, PUF.

Eland-Johnson, R.C. y N.L. Johnson (1980), *Survival Models and Data Analysis*, Nueva York, Wiley and Sons.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Nueva York, Wiley and Sons.

Henry, L. (1972), *Démographie: analyse at modèles*, París, Ediciones del INED, (1984).

Johnson, N.L. y S. Kotz (1970), *Continuous Univariate Distributions*, Nueva York, Wiley and Sons, vols. I y II.

Kalbfleisch, J. y R. Prentice (1980), *The Statistical Analysis of Failure Time Data*, Nueva York, Wiley and Sons.

Keilman, N., A. Kuijsten y A. Vossen (1988), *Modelling Household Formation and Dissolution*, Oxford, Clarendon Press.

Keyfitz, N. (1985), *An Introduction to the Mathematics of Demography*, Reading Mass., Addison-Wesley.

Land, K.C. y A. Rogers (comps.) (1982), *Multidimensional Mathematical Demography*, Academic Press.
(Compilación de los trabajos presentados en la conferencia del mismo nombre. Allí se pueden encontrar numerosos artículos de base para la modelización.)

Lelièvre, E. (1988), *Méthodes mathématiques et statistiques pour l'analyse d'histories de vie*, tesis de doctorado del EHSS, marzo.

Lelièvre, Éva y Armand Bringi (1998), *Practical guide to event history analysis using SAS, TDA and STATA*, Méthodes et savoir, núm. 2, INED-PUF, París.

McCullagh, P. y J.A. Nelder (1983), *Generalized Linear Models*, Londres, Chapman and Hall.

Pons, O. y E. de Turckeim (1983),* "Modèles de régression de Cox périodique et étude d'un comportement alimentaire", Versailles, *Cahiers de Biométrie*, INRA.

Pourcher, G. (1964), "Le peuplement de Paris, origine régionale, composition sociale, attitudes et motivations", *Cahier de l'INED*, núm. 43, PUF.

* N. del T. Falta en el original.

Pressat, R. (1961), *L'analyse démographique*, París, PUF.
Pressat, R. (1966), *Principes d'analyse. Cours d'analyse démographique de l'IDUP*, París, Ediciones del INED.
Prigogine, I. y I. Stengers (1988), *Entre le temps et l'eternité*, Fayard.
Sorensen, A.B., F.E. Weinert y L.R. Serrod (comps.) (1986), *Human Development and the Life Course: Multidisciplinary Perspectives*, Hillsdale, Nueva Jersey, Laurence Erlbaum Associates.
Tapinos, G. (1985), *Eléments de Démographie*, Armand Collin, Colección U.
Trusell, J. (comp.) (1989), "Demographic Applications of Event History Analysis", Seminario IUSSP, *Studies in Demography*, Oxford University Press.
Tuma, N. (comp.) (1985), *Sociological Methodology*, San Francisco, Jossey-Bass Publishers.
Tuma, N. y M. Hannan (1984), *Social Dynamics. Models and Methods*, Colección Quantitative Studies in Social Relations, Academic Press.
Wendel, B. (1953), *A Migration Schema, Theory and Observation*, Lund Studies in Geography, Serie B, *Human Geography*, núm. 9.
Wunsch, G. (1988), *Causal Theory and Causal Modelling*, Lovaina, Leuven University Press.

ARTÍCULOS

Aalen, O. (1977), "Weak convergence of stochastic integrals related to counting processes", *Zeitschirft für Wahrscheinlichkeitstheorie*, núm. 38, pp. 261-277.
(Explora los comportamientos asintóticos de las integrales estocásticas ligadas a los procesos de conteo.)
Aalen, O. (1978), "Nonparametric Interference for a Family of Counting Processes", *The Annals of Statistics*, vol. 6, núm. 4, pp. 701-726.
(Demuestra la aplicación de la teoría de las martingalas a los procesos de conteo multivariados y a las modelizaciones de los procesos markovianos no homogéneos. Presenta la aproximación no paramétrica. Justifica la utilización de las curvas de los cocientes acumulados, esclarece las relaciones con los estimadores de Kaplan-Meier y propone pruebas de comparación.)
Aalen, O. (1982), "Pratical Applications of the Nonparametric Statistical Theory for Counting Processes", *Statistical Research Report*, núm. 2, Instituto de Matemáticas, Universidad de Oslo.
(Ejemplos de aplicaciones médicas de las demostraciones de 1978.)
Aalen, O., O. Borgan, N. Keiding y J. Thorman (1980), "Interaction Between Life History Events: Nonparametric Analysis for Prospective and Retrospective Data in Presence of Censoring", *Scandinavian Journal of Statistics*, núm. 7, pp. 161-171.

(Una aplicación práctica del modelo no paramétrico a las interacciones entre la menopausia y las enfermedades de la piel.)

Aalen, O. y J. Hoem (1978), "Random time changes for multivariate counting processes", *Scandinavian Journal of Statistics*, núm. 2, pp. 81-101.
(Relación de los procesos de conteo multivariados con los procesos de Poisson. Posibilidad de aplicación a todos los datos de supervivencia.)

Aalen, O. y S. Johansen (1978), "An Empirical Transition Matrix for Non Homogeneous Markov Chains Based on Censored Information", *Scandinavian Journal of Statistics*, núm. 5, pp. 141-150.
(Presenta un estimador de las probabilidades de transición de una cadena de Markov no homogénea en un espacio finito de estados, construido en términos de producto integral. Las propiedades de este estimador se estudian en el cuadro de la teoría de las martingalas integrables.)

Allison, P. (1982), "Discret-time methods for the analysis of event histories", en S. Leinhardt (comp.), *Sociological Methodology*, San Francisco, Jossey Bass, pp. 61-98.

Andersen, P. (1980), "Testing Goodness-of-Fit of Cox's Regression and Life Model", Informe de investigación 80/1, Unidad de Investigación Estadística Danesa.
(Presenta técnicas gráficas y pruebas para asegurarse de la hipótesis de proporcionalidad en los modelos semiparamétricos de tipo Cox.)

Andersen, P. (1981), "Comparing Survivals via Hazard Ratio Estimates", Informe de investigación 81/7, Unidad de Investigación Estadística Danesa.
(Compara las varianzas asintóticas de los estimadores de los cocientes instantáneos; demuestra la preferencia que se le da a los estimadores surgidos de los modelos de riesgos proporcionales planteados por Cox.)

Andersen, P. (1981), "Measuring and Evaluating Prognosis using the Proportional Hazards Model", Informe de investigación 81/8, Unidad de Investigación Estadística Danesa.
(Aplicaciones a los datos longitudinales para la cirrosis hepática. Se encaran dos problemas: la selección de una medida de previsión —estimación de la función de supervivencia, del tiempo promedio, media de supervivencia con las desviaciones estándar— y evaluación de las previsiones cuando hay pocos datos disponibles [pruebas].)

Andersen, P. (1981), "On the Application of the Theory of Counting Processes in the Statistical Analysis of Censored Survival Data", Informe de investigación 81/10, Unidad de Investigación Estadística Danesa.
(Retoma las demostraciones de Aalen (1978) en el marco teórico general que los procesos de conteo multivariados le suministran a los procesos de Markov no homogéneos. Aplica esos resultados a la comparación de más de dos distribuciones y al caso de los modelos de regresión para datos longitudinales.)

Andersen, P. y R. Gill (1982), "Cox's Regression Model for Counting Processes: A Large Sample Study", *The Annals of Statistics*, 10, pp. 1100-1120.
(El modelo de Cox le confiere a las variables explicativas un efecto proporcional sobre la densidad condicional: los cocientes instantáneos. Los autores muestran que es posible transformarlo en un modelo donde las variables tienen un efecto proporcional sobre las intensidades —probabilidades de transición—.)

Andersen, P., O. Borgan, R. Gill y N. Kiding (1982), "Linear nonparametric tests for comparison of counting processes, with applications to censored survival data", *International Statistical Review.*
(Hace una presentación general de las pruebas no paramétricas existentes para muestras simples y múltiples, y de la teoría estocástica subyacente.)

Andersen, P. y O. Borgan (1985), "Counting Process for Life History Data: A Review", *Scandinavian Journal of Statistics,* vol. 12, pp. 97-158.

Andress, H.J. (1983), "The first 10 years of a working career: An ilustration of event history analysis with West German mobility data", *Computational Statistics and Data Analysis,* núm. 1, pp. 111-135.
(Presentación del marco de aplicación de los procesos estocásticos a los datos longitudinales de las ciencias sociales. Aplicación a la movilidad profesional. Revisión de los programas disponibles para el análisis paramétrico de los datos biográficos.)

Approches longitudinales (1987), *Rapport du groupe de travail interdisciplinaire de réflexion dans le cadre du colloque "Bilan Sociologique"*, relatores: Eva Lelièvre y Daniel Courgeau. Trabajo presentado en el Coloquio de Sociología III en Estrasburgo, 14, 15 y 16 de mayo de 1987.

Arjas, E. y P. Kangas (1988), "Discrete-time method for longitudinal analysis in demography: a comparative study of the data on third births in Sweden", trabajo presentado en el "Seminar on event history analysis", IUSSP, realizado en París del 14 al 17 de marzo de 1988.

Arjas, E. y D. Venzon (1988), "A Test for Discriminating Between Additive and Multiplicative Relative Risk in Survival Analysis", *Applied Statistics,* núm. 1, pp. 1-11.

Bertaux, D. (1980), "L'approche biographique: sa validité méthodologique, ses potentialités", *Cahiers Internationaux de Sociologie,* vol. LXIX, pp. 197-224.

Bocquier, Ph. (1987), "Retours dans le pays d'origine des immigrants en Suède", notas y documentos, *Population,* núm. 3, pp. 544-548.
(Utilización del GLIM.)

Bourdieu, P. (1986), "L'illusion biographique", *Actes de la Recherche en Sciences Sociales,* núms. 62-63, pp. 69-72.
(Artículo virulento que pone en cuestión el nuevo interés por lo biográfico, tan generalizado en las ciencias humanas.)

Breslow, N.E., J.H. Lubin, P. Marek y B. Langholz (1983), "Multiplicative models and cohort analysis", *Journal of the American Statistical Association,* vol. 78, núm. 381, pp. 1-12.

Brillinger, D. (1986), "The Natural Variability of Vital Rates and Associated Statistics (A biometrics invited paper with discussion)", *Biometrics*, núm. 42, pp. 693-734.
(Artículo muy interesante sobre el desarrollo de las aproximaciones de las distribuciones de las tasas de mortalidad brutas, por edad, estandarizadas. Utiliza una doble distribución de Poisson seguida por el número de decesos y la población sometida al riesgo. Plantea seguidamente una modelización para tomar en cuenta variables explicativas.)

Bretagnole, J. y C. Huber-Carol (1985), "Effet de l'omission de covariables dans le modèle de Cox", *Statistiques des processus en milieu médical,* Seminario 85, Huber, Lelouch, Prieur, París V, pp. 1-20.

Bretagnole, J. y C. Huber-Carol (1988), "Effects of Omitting Covariates in Cox's Model for Survival Data", *Scandinavian Journal of Statistics,* núm. 15, pp. 125-138.

Brouard, N. (1980), "Espérance de vie active, reprise d'activité féminime: un modèle", *Revue Economique,* vol. 31, núm. 6, pp. 1260-1287.
(A partir de un modelo no paramétrico explora las proyecciones de las tasas de actividad.)

Buckley, J.D. (1984), "Additive and multiplicative models for relative survival rates", *Biometrics,* 40, pp. 51-62.

Cambois, M.A. (1987), "Rapport sur la mobilité professionnelle", datos de la encuesta 3B, INED, p. 52.

Chang, N. y G.L. Yang (1987), "Strong Consistency of a nonparametric estimation of the survival function with doubly censored data", *The Annals of Statistics,* vol. 15, núm. 4, pp. 1536-1547.

Coale, A. y D. McNiel (1972), "The distribution by age of the frequency of first marriage in a female cohort", *JASA*, núm. 67, pp. 27-52.

Courgeau, D. (1973), "Migrants e migrations", *Population,* vol. 28, núm. 1, pp. 92-129.

Courgeau, D. (1984), "Relations entre cycle de vie et migrations", *Population,* núm. 3, pp. 483-514.

Courgeau, D. (1984), "Analysis of the French Migration, Family and Occupation History Survey", *Materialien Bevölkerungswissenschaft,* BIB, núm. 38, pp. 86-102.

Courgeau, D. (1985a), "Effect de déclarations erronées sur une analyse de données migratoires", *Chaire Quetelet: Migrations internes,* Lovaina.

Courgeau, D. (1985b), "Interaction between spacial mobility, family and career life-cycle: A French survey", *European Sociological Review,* vol. 1, núm. 2, septiembre, pp. 139-162.

Courgeau, D. (1985c), "Bases théoriques et modèles pour une enquête sur la biographie familiale, professionnelle et migratoire", *Espace, Population, Société,* núm. 1, pp. 240-247.

Courgeau, D. (1985d), "Changements de logement, changements de département et cycle de vie", *L'Espace Géographique,* núm. 4, pp. 289-306.

Courgeau, D. (1987), "Pour une approche statistique des histoires de vie", *Annales de Vaucresson,* núm. 26, pp. 25-36.

Courgeau, D. (1987), "L'analyse des enquêtes rétrospectives", *Chaire Quetelet: L'explication en Sciences Sociales, la recherche des causes en Démographie,* Lovaina.

Courgeau, D. (1987), "Constitution de la famille et urbanisation", *Population,* núm. 1, pp. 57-82.

Courgeau, D. (1988), "Migration, Family and Career: A Life Course Approach", vol. 10, *Life-Span Dvelopment and Behavior,* Hillsdale, Nueva York, Baltes, Featherman y Lerner (comps.), Lawrence Erlbaum Ass. (en prensa).

Courgeau, D. y E. Lelièvre (1985), "Estimation of transition rates in dynamic household models", en N. Keilman, A. Kuijsten y A. Vossen (comps.), *Modelling Household Formation and Dissolution,* Clarendon Press Oxford (1988), pp. 160-176.

Courgeau, D. y E. Lelièvre (1986), "Nuptialité et Agriculture", *Population,* núm. 2, pp. 303-326.

Courgeau, D. y E. Lelièvre (1988), "Interrelation between first home ownership, constitution of the family and professional occupation", trabajo presentado en el Seminar on Event History Analysis, IUSSP, París, del 14 al 17 de marzo de 1988.

Courgeau, D. y E. Lelièvre y M. Wagner (1986), "Leaving home and marriage in France and Germany", manuscrito inédito.

Cox, D.R. (1972), "Regression Models and Life Tables (with discussion)", *Journal of Royal Statistical Society,* B34, pp. 187-220.
(Presentación inicial de los modelos semiparamétricos.)

Cox, D.R. (1975), "Partial Likelihood", *Biometrika,* núm. 62. 2, pp. 269-276.
(Definición y estimación.)

Crowley, J. (1974), "Asymptotic Normality of a New Nonparametric Statistic for Use in Organ Transplant Studies", *Journal of the American Statistical Association,* vol. 69, núm. 348, pp. 1006-1011.
(Demuestra la normalidad asintótica de los estadísticos de prueba no paramétricos, propuestos recientemente en el análisis del trasplante de corazón.)

Crowley, J. y M. Hu (1977), "Covariance analysis of heart transplant data", *Journal of the American Statistical Association (JASA),* vol. 72, pp. 27-36.
(Aplicación del modelo semiparamétrico.)

Davies, R.B. (1984), "A generalised beta-logistic model for longitudinal data

with an application to residential mobility", *Environment and Planning A*, vol. 16, pp. 1375-1386.

Davies, R.B. y R. Crouchley (1984), "Calibrating longitudinal models of residential mobility and migration", *Regional Science and Urban Economics*, núm. 14, pp. 231-247.

Davies, R.B. y R. Crouchley (1985), "Longitudinal versus cross-sectional methods for behavioural research: a first-round knockout", *Environment and Planning A*, vol. 17, pp. 1315-1329.

Davies, R.B. y R. Crouchley (1985), "A panel study of life-cycle effects in residential mobility", *Geographical Analysis*, vol. 17, núm. 3, pp. 199-216.

Diamond, I., J. McDonald y I. Shah (1986), "Proportional Hazards Models for Current Status Data: Application to the Study of Differentials in Age et Weaning in Pakistan", *Demography*, vol. 23, núm. 4, pp. 607-620.
(Estimación semiparamétrica no clásica con la ayuda de GLIM.)

Duchene, J. (1985), "Un test de fiabilité des enquêtes rétrospectives Biographie Familiale Professionnelle et Migratoire", *Chaire Quetelet: Migrations internes*, Lovaina.

Elder, G. (1978), "Approaches to social changes and the family turning points, Historical and Sociological essays on the family", *American Journal of Sociology*, vol. 84, número especial, pp. 1-39.

Finkelstein, D.M. (1986), "A proportional hazard model for interval-censored failure time data", *Biometrics*, 42, pp. 845-854.

Foner, A. y D. Kertzer (1978), "Transition over the Life Course: lessons from age-set societies", *American Journal of Sociology*, vol. 83, núm. 5, pp. 1081-1105.

Fougere, D. y G. Tahar (1987), "Participation au marché du travail et nuptialité: Etude des interdépendances au sein d'une cohorte", U A 921 del CNRS. Nota número 60 (87-09).
(Medida de las interacciones en un marco markoviano, además de la evaluación de la influencia de variables explicativas en diversos modelos.)

Gill, R.D. (1985), "On estimating transition intensities of a markov process with aggregate data of a certain type", *Scandinavian Journal of Statistics*, núm. 4.

Gill, R.F. y M. Schumaker (1987), "A simple test of the proportional hazards assumption", *Biometrika*, vol. 74, núm. 2, pp. 289-300.

Ginsberg, R.B. (1971), "Semi-Markov Processes and Mobility", *Journal of Mathematical Sociology*, vol. 1, pp. 233-262.
(Presentación de los procesos semimarkovianos y aplicación a la movilidad social, artículo de base.)

Heckman, J. y B. Singer (1982), "Population Heterogeneity in Demographic Models", en Kenneth y Rogers (comps.), *Multidimentional Mathematical Demography*, Academic Press, pp. 567-604.

Heckman, J. y B. Singer (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data", *Econometrica,* vol. 52, núm. 2, pp. 271-320.

Henry, L. (1959), "D'un problème fondamental de l'analyse démographique", *Population,* núm. 1, pp. 9-32.

Henry, L. (1966), "Analyse et mesure des phénomènes démographiques par cohortes", *Population,* núm. 3, pp. 465-482.

Hobcraft, J. y M. Murphy (1986), "Demographic event History Analysis: a selective review", *Population Index,* vol. 52, núm. 1, pp. 3-27.
(Muestra la obtención de resultados contradictorios según la distribución *a priori* dada a la heterogeneidad subyacente y diferente. Esta no estabilidad revela claramente la precariedad de tal aproximación esencialmente cuantitativa de la heterogeneidad no observada. Se trata de una revisión muy completa que da una buena ojeada sobre los métodos, los resultados y los problemas que se encuentran en este campo.)

Hoem, J. (1976), "The Statistical Theory of Demographic Rates", *Scandinavian Journal of Statistics,* núm. 3, pp. 169-185.
(Revisión de la aproximación estadística de las tasas de ocurrencia sobre riesgo.)

Hoem, J. (1985), "Weighting, Misclassification and other Issues in the Analysis of Survey Samples of Life Histories", en J. Heckman y B. Singer (comps.), *Longitudinal Analysis of Labour Market Data,* Cambridge University Press.
(Sesgo causado por los diferentes diseños de muestreo y soluciones propuestas.)

Hoem, J. (1987), "The Issue of Weights in Panel Surveys of Individual Behavior", Informe de investigación, Departamento de Estadística de Estocolmo, núm. 39.

Hoem, J. (1987), "Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review", *International Statistical Review,* vol. 55, núm. 2, pp. 119-152.

Hoem, J. y U. Funck Jensen (1982), "Multistate life table methodology: a probabilist critique", en Land y Rogers (comps.), *Multidimensional Mathematical Demography,* Nueva York, Academic Press, pp. 155-264.

Hogan, D.P. (1978), "Order of Events in the Life Course", *American Sociological Review,* vol. 43, núm. 4, pp. 573-586.

Jayet, H. y A. Moreau (1988), "Proportional Hazard Model: Estimation and Specification Test Using Asymtotic Least Squares", documento de trabajo núm. 8804 del INSEE.

Johansen, S. (1983), "An Extension of Cox's Regression Model", *International Statistical Review,* vol. 51, pp. 165-174.
(Muestra la manera de modelizar un proceso de saltos dependientes de

una medida de intensidad arbitraria, con la propiedad de que si dicha medida es absolutamente continua, el modelo se resume en una regresión de tipo Cox.)

Kalbfleisch, J. y R. Prentice (1973), "Marginal Likelihood based on Cox's regression and life model", *Biometrika,* núm. 60, pp. 267-278.

Kaplan, E. y P. Meier (1958), "Nonparametric Estimation from Incomplete Observations", *Journal of the American Statistical Association* (JASA), vol. 53, pp. 457-278.
(Artículo seminal.)

Kay, R. (1986), "A Markov Model for Analysing Cancer Markers and Disease States in Survival Studies", *Biometrics,* núm. 42, pp. 855-865.
(Una aplicación médica de un modelo markoviano.)

Kimball, T. y M. Pearsall (1954), "The Nature of Human Groups", *The Talladega Story,* University of Alabama Press.

Kirwan, F. y F. Harrigan (1986), "Swedish-Finnish Return Migration, Extent, Timing, and Information Flow", *Demography,* vol. 23, núm. 3, agosto de 1986, pp. 313-327.

Kluzing, E., J. Siegers, N. Keilman y L. Groot (1988), "Static versus Dynamic Analysis of the Interaction between female labour force participation and fertility", *Europen Journal of Population,* vol. 4, núm. 2.
(Determinación del mejor intervalo de tiempo para el análisis y comparación del punto de vista transversal *versus* el longitudinal.)

Lagakos, S., L. Barraj y V. de Gruttola (1988), "Nonparametric analysis of truncated survival data, with application to AIDS", *Biometrika,* núm. 75, pp. 515-523.

Langevin, A. (1986), "Rythmes sociaux et réinterprétation individuelle des critères d'âge dans le parcours de la vie", *Annales de Vaucresson,* núm. 26, pp. 169-180.

Lelièvre, E. (1986), "The analysis of interactions between phenomena: data —a french survey—, tools, first results", septiembre, Sopron (Hungría), Conferencia IIASA.

Lelièvre, E. (1987a), "Activité professionnelle et fécondité: les choix et les déterminations des femmes françaises entre 1930 y 1960", *Cahiers Québéquois de Démographie,* vol. 16, núm. 2, pp. 207-236.

Lelièvre, E. (1987b), "Migrations définitives vers la France et constitution de la famille", *Revue Européenne des Migrations Internationales,* vol. 3, núm. 1-2.

Lelièvre, E. (1988), "Interactions entre l'acquisition du premier logement et la naissance du dernier enfant", documento de trabajo para el seminario Stratégies Résidentielles, enero.

Lelièvre, E. (1988), "L'étude des interactions entre phénomènes: dépendance unilatérale et causalité", presentación en el Coloquio Internacional de la

AIDELF Démographie et Différences, Montreal, del 7 al 10 de junio.

Lelièvre, E. (1988), "Constitution de la famille et urbanisation du Mexique", presentación en las jornadas demográficas de la ORSTOM, Migration, Changements sociaux et Développement, del 20 al 22 de septiembre de 1988.

Lindsay, B. A. (1983), "The geometry of mixture likelihoods: a general theory", *The Annals of Statistics,* vol. 11, núm. 1, pp. 86-94.

Lindsay, B. A. (1983), "The geometry of mixture likelihood: the exponential family", *The Annals of Statistics,* vol. 11, núm. 3, pp. 783-792.

Lyberg, I. (1983), "The effect of sampling and nonresponse on estimates of transitions intensities: some empirical results from the 1981 Swedish Fertility Survey", *Stockholm Research Reports in Demography,* núm. 14.

McGinnis, R. (1968), "Stochastic model of social mobility", *American Sociological Review,* núm. 33, pp. 712-721.
(Introducción del concepto de inercia acumulada.)

Mau, J. (1986), "On a Graphical Method for the Detection of Time-Dependent Effects of Covariates in Survival Data", *Applied Statistics,* núm. 3, pp. 245-255.
(Al comparar los cocientes instantáneos de riesgo inicial obtenidos mediante tres modelos: Aalen, Cox y Anderson y Senthilselvan —extensión del modelo de riesgos proporcionales—, el autor plantea que utilizar sólo el modelo de Cox puede conducir a importantes faltas de información. En ese caso es preferible no prescindir de las gráficas que el análisis de Aalen proporciona tan fácilmente.)

Mayer, K.U. y N.B. Tuma (1987), "Applications of Event history Analysis in Life Course Research", *Materialien aus der Bildungsforschung,* núm 30, Instituto Max-Planck, Berlín.
(Compilación de los trabajos de un coloquio.)

Modell, J., F. Furstenberg y D. Strong (1978), "The timing of marriage in the transition to adulthood: continuity and change, 1860-1975, Turning points", Historical and sociological essays on the family, *American Journal of Sociology,* vol. 84, número especial, pp. 120-150.

Monnier, A. (1987), "Projets de fécondité effective, une enquête longitudinale: 1974, 1976, 1979", *Population,* núm. 6, pp. 819-842.

Menken, J. y J. Trussell (1981), "Proportional hazards life table models: An illustrative analysis of socio-demographic influences on marriages dissolution in the United States", *Demography,* núm. 18 (2), pp. 181-200.
(Puesta en práctica de los modelos de riesgos proporcionales con la utilización de variables dependientes del tiempo.)

Murphy, M. (1984), "The influence of fertility, early housing career and socio-economic factors on tenure determination in contemporary Bri-

tain", *Environment and Planning A*, vol. 16, núm. 10, pp. 1303-1318.
(Esta investigación pudo realizarse gracias a una encuesta efectuada en Gran Bretaña en 1976 sobre la formación de la familia. El efecto de las características sobre el cambio de estatus de ocupación se estima mediante métodos de análisis de las historias de vida.)

Murphy, M. y O. Sullivan (1985), "Housing tenure and family formation in contemporary Britain", *European Sociological Review*, vol. 1, núm. 3, pp. 230-243.
(Los autores se abocan a demostrar la reciprocidad de los nexos entre los cambios de estatus de ocupación de la vivienda y los cambios de comportamiento de fecundidad de las parejas.)

Nelson, W. (1969), "Hazard plotting for incomplete failure data", *Journal of Qual. Techn.*, núm. 1, pp. 27-52.

Nelson, W. (1972), "Theory and Application of hazard plotting for censored failure data", *Technometrics*, vol. 14, pp. 945-965.
(Desarrolla los esquemas que serán retomados por Aalen para representar las intensidades acumuladas del fenómeno estudiado.)

Neveu, J. (1986), "Arbres et processus de Galton-Watson", *Ann. Inst. Henry Poincaré*, vol. 22, núm. 2, pp. 199-207.

Oakes, D. (1981), "Survival times: aspects of partial likelihoods", *International Statistical Review*, vol. 49, pp. 235-264.

Petersen, T. (1986), "Estimating fully parametric hazard rate models with time-dependent covariates", *Sociological Methods and Research*, vol. 14, núm. 3, pp. 219-246.

Petersen, T. (1986), "Fitting parametric survival models with time-dependent covariates", *Applied Statistics*, vol. 35, núm. 2, pp. 281-288.

Peto, M. y R. Peto (1972), "Asymptotically efficient rank invariant test procedures (with discussion)", *Journal of Royal Statistical Society A*, núm. 135, pp. 185-206.

Peto, R. y M.C. Pike (1973), "Conservatism of the approximation $\Sigma(O-E)^2 / E$ in the Logrank Test for Survival Data", *Biometrics*, pp. 579-584.
(Eficacia de las pruebas de rango en el caso de modelos no paramétricos.)

Pickles, A. y R.B. Davies (1983), "The longitudinal analysis of housing careers", *Journal of Regional Science*, vol. 75, pp. 85-101.

Prentice, R. (1975), "Discrimination among some parametric models", *Biometrika*, núm. 62, pp. 607-614.

Prentice, R. (1973), "Exponential survivals with censoring and explanatory variables", *Biometrika*, 62, núm. 2, pp. 279-288.

Prentice, R. (1982), "Covariate measurement errors end parameter estimation in a failure time regression model", *Biometrika*, 69, núm. 2, pp. 231-242.

Prentice, R. (1986), "On the design of synthetic case-control studies", *Biometrics*, 42, pp. 301-310.

Prentice, R. y N. Breslow (1978), "Retrospective studies and failure time models", *Biometrika,* 65, 1, pp. 153-158.

Riandey, B. (1985), "L'enquête biographie familiale, professonnelle et migratoire (INED 81). Le bilan de la collecte", *Chaire Quetelet: Migrations internes,* Lovaina.

Ricoeur, P. (1983), *Temps et récit,* tomo I, *La configuration du temps dans le récit de fiction,* tomo II, *Le temps raconté,* tomo III, Colección L'ordre philosophique, París, Seuil.

Rogers, A. (1973a), "The Multiregional Life Table", *Journal of Mathematical Sociology,* núm. 3, pp. 127-137.

Rogers, A. (1973b), "The Mathematics of Multiregional Demographic Growth", *Environment and Planning,* núm. 5, pp. 3-29.

Rouy, E. (1986), "Rapport de stage sur les troncatures á droite", INED.

Sandefur, G. (1985), "Variation in interstate migration of men across the early stages of the life cycle", *Demography,* vol. 22, núm. 3, pp. 353-366.
(Demuestra que la importancia de los motivos de la migración depende fuertemente de la etapa del ciclo de vida familiar en la que se encuentra el individuo.)

Sandefur, G. y W. Scott (1981), "A dynamic analysis of migration: an assessment of the effect of age, family and career variables", *Demography,* vol. 18, núm. 3, pp. 355-368.
(Una vez que se toman en cuenta las variables familiares —estado matrimonial, tamaño de la familia— y las económicas, desaparecen las diferencias de movilidad según la edad.)

Schoen, R. (1979), "Calculating increment-decrement life tables by estimating mean durations at transfer from observed rates", *Mathematical Biosciences,* 47, pp. 255-269.

Schou, G. y M. Vaeth (1980), "A small sample study of occurrence/exposure rates for rare events", *Scandinavian Actuarial Journal,* núm. 4, pp. 209-225.
(Demostración empírica de la convergencia asintótica de los estimadores en el caso de muestras pequeñas.)

Schweder, T. (1970), "Composable Markov Processes", *Journal of Applied Probability,* núm. 7, pp. 400-410.
(Introducción al concepto de dependencia local en el caso de dos procesos estocásticos.)

Singer, B. y S. Spillerman (1974), "Social Mobility Models for Heterogeneous Population", en Costner (comp.), *Sociological Methodology,* Jossey Bass, pp. 356-401.

Sorensen, A. (1977), "Estimating Rates from Retrospectives Questions", en D.R. Heise (comp.), *Sociological Methodology,* San Francisco, Jossey Bass, pp. 209-223.
(Interesante análisis matemático de la inadecuación del tiempo medido

por la edad en el análisis de los comportamientos humanos. Problema de los truncamientos de la observación.)

Spillerman, S. (1972), "Extension of the Mover-Stayer Model", *American Journal of Sociology,* vol. 78, núm. 3, pp. 599-626.

Struthers, C. y Kalbfleish (1986), "Misspecified proportional hazard models", *Biometrika,* núm. 2, pp. 363-369.
(Artículo esclarecedor sobre los debates acerca de la legitimidad del empleo de los modelos de riesgos proporcionales.)

Suzuki, K. (1985), "Nonparametric Estimation of Life Time Distributions from a Record of Failures adn Follow-ups", *JASA,* vol. 80, núm. 389, pp. 68-72.

Swain, M. (1987), "Theories of Causation", *Chaire Quetelet: L'explication en Sciences Sociales, la recherche des causes en Démographie,* Lovaina.

Trussell, J. y C. Hammerslough (1983), "A hazard model analysis of the covariates of infant mortality in Sri Lanka", *Demography,* núm. 20 (1), pp. 1-26.
(A partir de la Encuesta Mundial de Fecundidad (WFS), los autores plantean un análisis de la mortalidad infantil. Hay que señalar que J. Trussell, asociado con otros autores, ha escrito varios artículos que giran en torno a la mortalidad o la fecundidad a partir de la WFS en los más diversos países, como Paquistán, Indonesia, Filipinas, publicados además en *Demography.*)

Trusell, J. y T. Richard (1985), "Correcting for unmeasured heterogeneity in hazard models using the Heckman-Singer procedure", en N. Tuma (comp.), *Sociological Methodology,* San Francisco, Jossey Bass, pp. 242-276.

Tsai, W., S. Leurgans y J. Crowley (1986), "Nonparametric Estimation of a Bivariate Survival Function in Presence of Censoring", *Annals of Stadistics,* vol. 14, núm. 4, pp. 1351-1365.
(Se propone una nueva familia de estimadores construida a partir de la descomposición de la función de permanencia con dos variables. Se demuestran sus propiedades.)

Tuma, N. y M. Hannan (1984), "Models for Heterogeneous Populations", en *Social Dynamics,* Academic Press, pp. 155-186.

Tuma, N., M. Hannan y L. Groenevelt (1979), "Dynamic Analysis of Event Histories", *American Journal of Sociology,* vol. 84, núm. 4, pp. 820-854.

Turnbull, B. (1974), "Nonparametric Estimation of a Survivorship Function with Doubly Censored Data", *JASA,* vol. 69, núm. 345, pp. 169-173.
(Aborda el problema de los truncamientos a la izquierda.)

Vallin, J. y A. Nizard (1977), "La mortalité par état matrimonial. Mariage sélection ou mariage protection", *Population,* número especial, pp. 95-125.

Vaupel, J. y A. Yashin (1985), "Heterogeneity ruses: some surprising effects of selection on population dynamics", *The American Statistician* (próxima publicación).

Vincent, P. (1947), "Nomogrammes pour la détermination des différences significatives entre deux taux", *Population,* núm. 2, pp. 313-322. (Presentación de los principios de construcción de nomogramas —impresos por el INED— que permiten resolver inmediatamente y mediante procedimientos elementales algunos de los problemas más frecuentes que plantean las dimensiones limitadas de las poblaciones.)

Waters, H.R. (1984), "An approach to the study of multiple state models", *Journal of the Institute of Actuaries,* 111, parte II, pp. 363-374.

Wing Hung Wong (1986), "Theory of partial likelihood", *The Annals of Statistic,* vol. 14, núm. 1, pp. 88-123.

Yashin, A., K. Manton y J. Vaupel (1985), "Mortality and Aging in Heterogeneous Population: A Stochastic Process Model with Observed and Unobserved Variables", *Theoretical Population Biology,* vol. 27, núm. 2, pp. 154-175.

Análisis demográfico de las biografías
se terminó de imprimir en abril de 2001
en los talleres de Impresores Aldina,
Obrero Mundial 201, col. Del Valle, 03100 México, D.F.
Se tiraron 1 000 ejemplares más sobrantes para reposición.
Composición tipográfica y formación: Solar, Servicios Editoriales, S.A. de C.V.
La edición estuvo al cuidado del Departamento
de Publicaciones de El Colegio de México.

CENTRO DE ESTUDIOS DEMOGRÁFICOS
Y DE DESARROLLO URBANO

Durante los últimos años, las técnicas de análisis demográfico han alcanzado un desarrollo importante, tal es el caso del análisis de historia de eventos. Incluso esta metodología ha sido utilizada ampliamente desde los ochenta en áreas tales como la economía y la bioestadística.

Para la demografía, el análisis de la historia de eventos amplía el alcance del enfoque longitudinal tradicional y provee resultados adicionales que son esenciales para el entendimiento del comportamiento humano en toda su complejidad.

El análisis de historia de eventos tiene su origen en el análisis longitudinal tradicional. Después de la segunda guerra mundial, el análisis de cohortes reemplazó al análisis transversal, el cual se utilizaba desde fines del siglo XVII. El análisis longitudinal estudia la ocurrencia de un evento demográfico (matrimonio, nacimiento, defunción, migración, etc.) dentro de un determinado grupo homogéneo o cohorte.

Por otra parte, el análisis de historia de eventos estudia la aparición de uno o más fenómenos demográficos interrelacionados. Puede involucrar simultáneamente a un gran número de características individuales, algunas de las cuales podrían cambiar con el tiempo. Esto proporciona una adecuada solución a algunos de los problemas que se presentan en el análisis longitudinal tradicional.

ISBN 9-681-20968-0