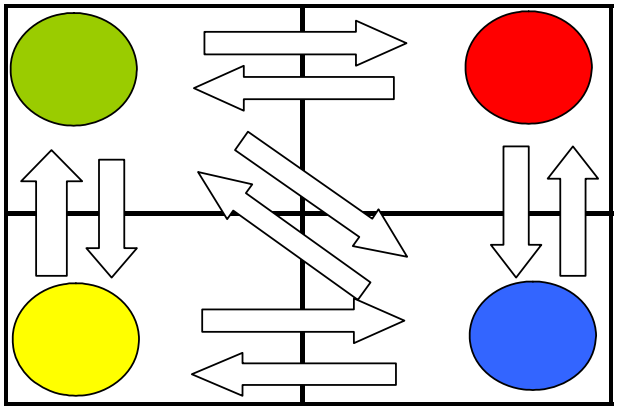
UNIVERSIDAD NACIONAL TRES DE FEBRERO
MAESTRIA EN GENERACIÓN Y ANALISIS DE INFORMACIÓN ESTADÍSTICA

Héctor Maletta

Análisis de panel con variables categóricas



Buenos Aires, 2012

CONTENIDO

1. Introducción al análisis de panel

Se denomina "datos de panel" a las bases de datos sobre una **pluralidad de unidades,** sobre las cuales hay información concerniente al estado de ciertas variables en **varias fechas o períodos a lo largo del tiempo**. El análisis de ese tipo de datos origina interesantes problemas teóricos y metodológicos, y ha motivado el desarrollo de importantes herramientas analíticas. Los estudios de panel forman parte de una familia de métodos de análisis **longitudinal**, en los cuales se cuenta con información **diacrónica** o **intertemporal**, referida a diferentes momentos o períodos a lo largo del tiempo, en oposición a los métodos **transversales** (*cross-section*) en los cuales la información es **sincrónica** o **cotemporal** pues se refiere a un mismo instante o período.

1.1. El desarrollo histórico del análisis de panel

Las estructuras de datos de carácter longitudinal incluyen muchas variantes (como se verá en la sección 1.3). En particular, una distinción inicial permite separar los *paneles de corto plazo* y los *estudios de seguimiento de largo plazo*.

Paneles de corto plazo. La aplicación tradicional del análisis de paneles de corta duración fueron las encuestas pre-electorales, como las realizadas por Lazarsfeld, Berelson y otros en las décadas del cuarenta y cincuenta, publicadas luego bajo los títulos **The people's choice** (Lazarsfeld y otros, 1948) y **Voting** (Berelson y otros, 1954). Todos ellos cubren una duración relativamente breve (unos pocos meses).

Otra aplicación frecuente de los datos de panel son los estudios experimentales o cuasi-experimentales del tipo "antes-después", muy frecuentes en los análisis del impacto de proyectos, efectos de la publicidad o resultados de los tratamientos médicos. Uno de los primeros y clásicos ejemplos del método de panel para análisis cuasi-experimentales del tipo "antes-después" fue el estudio del impacto de la proyección de películas motivadoras y de propaganda sobre la moral de los soldados americanos en la Segunda Guerra Mundial, en el marco de un estudio más amplio de las tropas norteamericanas dirigido por Samuel Stouffer, publicado luego del fin del conflicto con el titulo **The American soldier** (Stouffer y otros, 1949; Merton y Lazarsfeld 1950). En esa tradición de análisis primariamente sociológico y psicosocial las técnicas analíticas eran muy elementales, expresándose sobre todo en la presentación de **tabulaciones cruzadas** de la misma variable (usualmente dicotómica) observada en dos momentos del tiempo, es decir en las llamadas **tablas de rotación** (*turnover tables*) y usando como principal instrumento la **comparación de porcentajes**.

Posteriormente, y sobre todo a partir de las contribuciones de Paul F. Lazarsfeld y James S. Coleman, surgieron herramientas más sofisticadas, que permitieron la aplicación y validación de **modelos** acerca de los cambios en las unidades de análisis a lo largo del tiempo. Estos modelos (como el de las **cadenas de Markov**) permiten formular hipótesis sobre los cambios que se espera que ocurran a los sujetos en el tiempo, y ponerlas a prueba con datos de panel. Uno de los primeros y más importantes esfuerzos en ese sentido fue la aplicación de los conceptos de **variables latentes** al caso de las observaciones de panel, como por ejemplo en Lazarsfeld (1961 y 1965). Un impulso muy fuerte para estos enfoques analíticos fue el desarrollado por James S. Coleman en su ***Introduction to Mathematical Sociology*** (Coleman 1964b) en la cual una buena parte se refiere a modelos que necesitan datos de panel. Ese mismo año Coleman publicó otro trabajo muy importante, ***Models of change and response uncertainty*** (Coleman 1964a), que suministra herramientas para separar, en un análisis de panel, el cambio de las variables subyacentes por un lado, y las meras variaciones aleatorias de las respuestas por el otro. Asimismo Coleman (1968) discutió cuestiones referentes al estudio del cambio no sólo en variables categóricas sino también en variables cuantitativas. Una versión más desarrollada y reciente de los aportes de este autor puede hallarse en su texto ***Longitudinal Data Analysis*** (Coleman 1991). Dentro del ámbito de las variables de tipo categórico otra corriente de análisis utiliza modelos log-lineales, como por ejemplo Hagenaars (1990 y 1994) y Vermunt (1997). El análisis de procesos en Boudon (1967, cap. VII-IX) parte de los modelos basados en cadenas de Markov y en los aportes de Coleman pero además aplica a los datos de panel los "coeficientes de dependencia", en inglés *path coefficients*, y que provienen del análisis de regresión.

Estudios longitudinales de largo plazo. Si bien los paneles pre-electorales abarcan sólo unas pocas "rondas", hay paneles de muy larga duración. En las últimas décadas, desde mediados de los años sesenta en adelante, se iniciaron diversos estudios de seguimiento de muy largo plazo, en los cuales una misma muestra de sujetos es seguida a lo largo de muchos años, algunos de los cuales todavía continúan. Entre ellos están por ejemplo el Estudio Longitudinal sobre la Experiencia de la Juventud en el Mercado Laboral (NLSY) conducido en Estados Unidos desde 1968; la Encuesta Nacional Longitudinal emprendida desde 1972 por el Departamento de Educación del mismo país, o el Estudio Comparativo sobre Dinámica de Ingresos, que se viene llevando a cabo en Estados Unidos, Alemania y Gran Bretaña (Becketti y otros, 1988), entre muchos otros. La principal aplicación de esta metodología han sido los estudios médicos, destinados a observar el comportamiento de variables de salud, dieta y estilo de vida en el largo plazo. Hay un importante estudio sobre la enfermedad de Alzheimer que se realiza sobre una población de monjas, otro estudio sobre proclividad a ciertas enfermedades que hace el seguimiento de una amplia muestra de enfermeras británicas, otro importante estudio que hace el seguimiento de parejas de gemelos (el llamado "Proyecto Minnesota") y muchos otros análogos, con diferentes propósitos. Muchos de ellos realizan el seguimiento de pacientes de ciertas enfermedades, a partir de ser diagnosticados o desde que se sometieron a determinada cirugía o tratamiento (como el seguimiento de personas con trasplante de órganos).

Otra tradición importante de estudios longitudinales es la que se desarrolló en el marco de los estudios sobre **crecimiento y desarrollo infantil**, tanto físico como psicológico, y sobre el envejecimiento. En estos casos el período de seguimiento no es necesariamente tan prolongado, sobre todo en el caso infantil. Excelentes textos sobre los enfoques metodológicos propios de esta tradición son los de Plewis (1985), Magnusson y otros (1994), Hand y Crowder (1996), y Collins y Sayer (2001). Dentro de esta tradición ha habido también diversos intentos de aplicar modelos de variables latentes, tanto a través del análisis factorial como a través de modelos de clases latentes (véase von Eye & Clogg, 1994; Berkane, 1997, especialmente el articulo de Arminger 1997; Nesselroade 1997; y Hagenaars y McCutcheon, 2002). Las curvas de crecimiento normal del peso y talla de los niños de hasta cinco años, que sirve de base para los indicadores antropométricos de desnutrición, se basan en estudios de este tipo, como el Estudio Multicéntrico sobre Patrones de Crecimiento Infantil, realizado con niños de distintos países y continentes desde fines de los años noventa hasta principios del presente siglo, y que ha dado origen a las curvas actuales de la Organización Mundial de la Salud, adoptadas en 2005 (WHO 2006).

Paralelamente se desarrolló una tradición analítica diferente en el marco de la Econometría, referida a situaciones en que se dispone de varias series económicas (usualmente con variables de intervalo) con valores a lo largo del tiempo para diferentes países o para diferentes unidades económicas de cualquier tipo. Esta situación es muy diferente a la anterior, que trata principalmente de individuos y con variables cualitativas, generamente con pocas rondas (dos o tres) mientras en econometría usualmente se dispone de series "largas", con muchos puntos a lo largo del tiempo y con variables cuantitativas. Aquí las unidades son frecuentemente países o empresas, y las variables generalmente son medidas en escalas de intervalo, aunque también se han aplicado las mismas técnicas econométricas a encuestas de hogares, e incluso se han adaptado los procedimientos para incluir variables categóricas en los tratamientos estadísticos derivados de este enfoque, los que suelen ser **modelos de regresión** de varios tipos. A esto contribuye el hecho de que esta clase de estudios suelen disponer de series temporales con muchos puntos sucesivos, lo que es imprescindible para poder aplicar análisis de regresión. Excelentes resúmenes de los aportes de esta tradición puede hallarse en Nerlove (2000) y en el importante texto de Mátyás y Sevestre, 1996. Otros aportes significativos con esta orientación son los de Hsiao (1986), Heckman y Singer (1982, 1985), y Baltagi (1995). Un trabajo que aplica principalmente enfoques de regresión al análisis de panel con aplicaciones al análisis sociológico, aunque se concentra en variables cuantitativas, es el breve texto de Finkel (1995) sobre **inferencias causales a partir de datos de panel**. También desde la tradición econométrica se han hecho avances en el tratamiento de variables categóricas: véase por ejemplo Heckman (1981), así como Hammerle y Ronning (1995). Entre otros muchos ejemplos, pueden encontrarse aplicaciones de modelos econométricos a datos de panel con variables laborales principalmente cualitativas,

basados en encuestas de hogares de América Latina, en el estudio de Pradhan y Van Soest (1997) sobre Bolivia, y el de Gong y Van Soest (2001) sobre México.

Los desarrollos en la formulación de **modelos lineales generalizados**, que incluyen la regresión o el análisis de varianza como subcategorías, incluyen algunos modelos que se han aplicado al análisis estadístico de datos longitudinales, en particular los modelos llamados "mixtos" y "jerárquicos": véase por ejemplo los textos de McCulloch y otros (2000), y de Verbeke y otros (2000). En el texto de Bryk y Raudenbusch (1992) sobre modelos lineales jerárquicos hay también importantes referencias al estudio del cambio y a los diseños de tipo longitudinal. En los modelos jerárquicos, con datos de diferentes niveles anidados unos dentro de otros, hat una variante de panel donde para cada unidad de análisis (por ejemplo una familia) existen varias observaciones a lo largo del tiempo. Estas varias observaciones pueden ser consideradas como meras instancias (no ordenadas intrínsecamente) de una variable subyacente *estática*, donde las observaciones solo difieren entre sí por razones aleatorias, o bien como una *secuencia ordenada* de observaciones a través de las cuales se pueden detectar *patrones temporales de cambio*. Este último caso es el que interesa aquí.

Otra importante contribución desde el territorio de la econometría son los modelos relacionados con los conceptos de **cointegración** y de **raíces unitarias** (véase por ejemplo Rao, 1994). Este enfoque trata de afrontar el problema que presentan las series temporales **no estacionarias**, en las cuales no se cumplen algunos supuestos básicos de la regresión, de modo que la aplicación del método tradicional de regresión conduce a estimaciones sesgadas.

La interrelación de las variables en el tiempo ha sido objeto de análisis no sólo a través de relaciones funcionales que corresponden a **procesos causales** o de interdependencia, como es común en los enfoques econométricos, sino también para identificar **factores subyacentes** que explicarían la correlación o covariación de las variables en el tiempo; en este aspecto se ha desarrollado por ejemplo una serie de métodos de **análisis factorial dinámico** (Tysak y Meredith, 1990; Meredith y Horn, 2001) que han extendido a la dimensión longitudinal los conceptos del análisis factorial clásico. Estos enfoques estiman factores o variables subyacentes inobservables, correlacionadas con las variables manifiestas, y capaces de explicar la covariación de estas últimas en el tiempo.

En la presente introducción metodológica no se cubren todos los aspectos del vasto campo de los estudios longitudinales. Se dedica preferente atención a una subcategoría: los **estudios de panel** donde predominan las **variables de tipo categórico**, y con un **número muy limitado de fechas de observación a lo largo del tiempo**. El ejemplo clásico son las encuestas repetidas sobre la misma muestra de respondentes, realizadas a determinados intervalos, como ocurre con muchas encuestas regulares de hogares. La principal fuente de datos que tuvimos presente para preparar este texto fueron las encuestas de hogares con paneles rotativos que se usan en muchos países. En un panel rotativo de hogares, cada hogar permanece en la muestra durante $\boldsymbol{k}$ rondas del panel, y en cada ronda se reemplaza un contingente de hogares que ya han cumplido el número de rondas previsto. Sin embargo, se incluyen también en este texto algunas consideraciones sobre el tratamiento de variables cuantitativas y sobre los estudios longitudinales más prolongados en los que se generan series temporales con mayor número de observaciones a lo largo del tiempo.

1.2. El prisma de datos

La mayor parte de los datos analizados por las ciencias sociales se pueden representar en una **matriz de datos** (Galtung 1964; véase también una visión más amplia del concepto de matriz de datos en Samaja 1995). Una matriz es un un **arreglo rectangular de datos** con varias filas y varias columnas (véase al final del texto una Nota Técnica sobre matrices). Cada fila representa un **caso**, es decir, una **unidad de análisis**, cada columna una **variable**, y cada celdilla $\boldsymbol{a_{ij}}$ en el cruce de la fila "i" y la columna "j" contiene el **valor** de una variable para una determinada unidad de análisis. El índice $\boldsymbol{i}$ identifica los casos (del caso 1 al caso $\boldsymbol{n}$) y el índice $\boldsymbol{j}$ las variables (de la 1 a la $\boldsymbol{m}$).

Matriz de datos transversal con *n* casos y *m* variables

		Variables		
		1 …….	***j*** **…….. *m***	
	1		:	
	…		:	
Casos	***i***	**……. .**	a_{ij}	**……**
	…		:	
	n		:	

No es necesario que el estudio sea sincrónico, o de corte transversal para generar una estructura de matriz de datos como la que antecede. Aparte de las matrices de datos **transversales** (muchos casos, muchas variables, un solo periodo) puede haber también matrices de datos **longitudinales** pero siempre bidimensionales, donde se estudie un solo caso con muchas variables durante varios períodos, como por ejemplo los datos de un estudio econométrico de un solo país, con varias variables observadas en múltiples ocasiones (las variables serían varias series temporales de datos económicos de un mismo país: PBI, inflación, gasto público, etc., observados en sucesivos años o trimestres); este tipo de matrices de datos son siempre "planas": se refieren a **un solo caso** y están formadas por **t** períodos y **m** variables. Los "casos" aquí no son diferentes **objetos** (países, personas, familias) sino diferentes **períodos**. Este tipo de estudio, aunque es longitudinal, tiene solo dos dimensiones: las variables (de 1 a ***m)*** y las ocasiones (de 1 a ***t***).

Matriz de datos longitudinal de un solo caso con *m* variables

		Variables		
		1 …….	***j*** **…….. *m***	
	1		:	
	…		:	
Periodos	***i***	**……. .**	a_{ij}	**……**
	…		:	
	t		:	

Las matrices de datos tienen, pues, dos dimensiones, que pueden representan muchos casos y muchas variables en una sola ocasión, o muchas ocasiones y muchas variables para un solo caso.

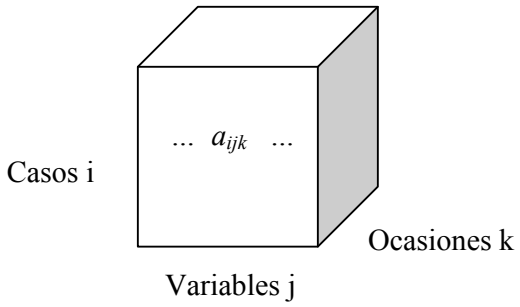
En los estudios de panel, en cambio, el conjunto de datos tiene **tres** dimensiones y no dos:

- las **unidades de análisis** o "casos" sobre las cuales versan los datos;
- las **variables** medidas u observadas en dichas unidades de análisis; y
- los **períodos**, **ocasiones** o **momentos** del tiempo a los que se refieren los datos.

De este modo la matriz de datos se convierte en un **prisma de datos**, cuyas tres dimensiones son los **casos**, los **períodos, fechas** u **ocasiones**, y las **variables**. En una típica situación de panel hay ***m*** variables registradas para ***n*** casos en ***t*** ocasiones.[1] Cada dato a_{ijk} pertenece a un caso ***i***, una variable ***j*** y una ocasión ***k***. En vez de una matriz de datos "plana" se tiene una matriz de datos con "profundidad", es decir un "cubo" o más genéricamente un "prisma" con tres dimensiones.

[1] Los usuarios de programas de cálculo como Excel pueden visualizar esto como la diferencia entre una **hoja de cálculo** con filas y columnas y un **libro de cálculo** con **varias hojas de similar estructura.**

El prisma de datos



1.3. Clasificación de los estudios longitudinales y de panel

Los datos longitudinales se diferencian de los transversales porque contienen información referida a diferentes momentos o períodos a lo largo del tiempo. Los estudios de panel son sólo una de las varias clases de datos longitudinales. Las principales clases son las siguientes:

Principales clases de datos longitudinales	
Mediciones repetidas	Las mismas variables se miden varias veces con el fin de obtener un promedio de varias observaciones como medida más confiable.
Encuestas repetidas	Las mismas preguntas formuladas en varias rondas, a diferentes muestras de la misma población, para analizar cambios agregados
Encuesta retrospectiva	Datos recogidos en una sola oportunidad acerca de diversos momentos del pasado
Panel	Seguimiento de los mismos casos y variables en dos o más rondas
Panel rotativo	En cada ronda, una cierta proporción de los casos se reemplaza por otros casos nuevos, hasta reemplazarlos gradualmente a todos.
Registro continuo	Datos sobre los mismos casos, registrados de manera continua.
Panel simulado	Datos tomados en un cierto período sobre sujetos de diferente edad o antigüedad, aplicados a una "cohorte teórica" de sujetos supuestamente observados a lo largo de su vida.

1.3.1. Mediciones repetidas

Las reiteradas observaciones de los mismos sujetos pueden perseguir diferentes objetivos, y ser utilizadas en diferentes formas. Una primera finalidad es la de obtener **mediciones repetidas de un mismo fenómeno.** Este fenómeno, por lo general, es medido varias veces no tanto para estudiar sus cambios a lo largo del tiempo, sino para obtener una **medición promedio**, más confiable que una medición individual pues las mediciones pueden oscilar aleatoriamente a lo largo del tiempo. Por ejemplo, un médico puede tomar la tensión arterial al paciente varias veces seguidas (o instalarle un aparato que registra la tensión arterial en forma continua durante 24 horas), para tener una medición más confiable de la tensión arterial promedio de ese individuo, haciendo caso omiso de las fluctuaciones normales que se registren alrededor de ese promedio (por supuesto, si alguna de esas observaciones sucesivas revelase un valor totalmente anormal, el médico puede aconsejar nuevos exámenes más profundos o prolongados). En estos casos se trata de observaciones repetidas del mismo fenómeno con el fin de obtener una medición más confiable. El indicador de resumen de todas esas mediciones puede ser, obviamente, la media, pero también se puede resumir la información indicando su grado de variabilidad mediante la desviación estándar, o el coeficiente de variación (que es el cociente de la desviación estándar sobre la media).

Estas **mediciones** repetidas no son en realidad paneles, sino solo una muestra compuesta por varias observaciones acerca de un fenómeno subyacente **que se supone constante**. Las diferencias de una medición a otra se atribuyen totalmente al azar o a errores de medición. En el fondo, estos estudios no son estudios longitudinales: se observa un solo fenómeno, pero se lo mide varias veces para

tener mayor certidumbre. De hecho, en algunas ocasiones las varias mediciones se efectúan simultáneamente, pues lo que se busca es mayor precisión o confiabilidad, y no se pretende estudiar la evolución en el tiempo. En vez de un prisma de datos, lo que se obtiene finalmente es una matriz de datos bidimensional.

Aparte de estos casos de medición repetida con finalidad meramente confirmatoria, que no atribuyen significado sustantivo a las diferencias entre una y otra observación, la mayoría de los estudios longitudinales tratan de capturar **cambios efectivos a lo largo del tiempo.** En el resto de este trabajo la expresión "estudios de panel" se usa casi exclusivamente en este sentido, con la única excepción de los paneles simulados referidos a continuación.

1.3.2. Encuestas repetidas

La clase más elemental de estudio longitudinal son las "series de encuestas repetidas", también llamados "datos de tendencias" (*trend data*). Estos datos contienen información recogida en diferentes momentos y períodos acerca de una determinada **población de referencia**, pero **no sobre la misma muestra de sujetos o unidades de análisis** dentro de esa población. Su forma más habitual son las "encuestas repetidas" (*repeated surveys*).[2] Por ejemplo, varias sucesivas encuestas de opinión a lo largo de una campaña electoral, cada una con una muestra diferente **de la misma población**, pueden indicar una tendencia agregada en las preferencias electorales de la población, pero no permiten hacer el seguimiento de los cambios de opinión de ningún individuo concreto. Miden **cambios poblacionales o agregados**, es decir **cambios a nivel macro**, pero no pueden captar directamente los **cambios a nivel micro** (en este ejemplo, cambios en cada uno de los individuos). Dada esa circunstancia, y considerando que las distintas muestras pueden diferir entre sí aunque todas ellas provengan de la misma población, se deduce que las diferencias entre dos encuestas sucesivas pueden deberse a dos clases de factores: auténticos cambios en el estado de la población, o simples diferencias aleatorias entre las distintas muestras (o una combinación de ambos factores).

Encuestas repetidas emparejadas. Las encuestas repetidas no son paneles en sentido estricto. Hay sin embargo algunas propuestas metodológicas en las cuales se tiende a construir "seudo paneles" con encuestas repetidas. En algunas de dichas propuestas las muestras sucesivas se pueden "emparejar" en cuanto a su composición interna en ciertas variables clave. Algunas referencias sobre esto son Deaton 1985, Verbeek & Nijman 1992, y Verbeek 1996. Así, se comparan muestras diferentes de individuos, pero estos individuos son agrupados según varias variables importantes (p.ej. sexo, edad, educación, ocupación, nivel socioeconómico, zona de residencia, etc.), de modo que se comparan **grupos de individuos similares** en diferentes momentos del tiempo.

1.3.3. Reconstrucciones retrospectivas

Un paso más adelante en cuanto a estudios longitudinales está constituido por las "reconstrucciones retrospectivas". En una sola ocasión, por ejemplo en una encuesta transversal, se recoge información retrospectiva sobre eventos del pasado. Un ejemplo muy conocido es el llamado "calendario de historia de vida" (Freedman y otros, 1988), en el cual se registran retrospectivamente las fechas de una serie de eventos importantes en la vida de cada persona (nacimiento, ingreso a la escuela, egreso de la escuela, iniciación sexual, primer empleo, primer embarazo, matrimonio, divorcio, etc.). También son de este tipo las "historias ocupacionales", las "historias migratorias", y otros estudios de tipo retrospectivo que reconstruyen una serie de eventos situados en el tiempo sin necesidad de extender la recolección de datos a más de un solo momento en el tiempo. El principal inconveniente de muchas encuestas retrospectivas es la dudosa confiabilidad de las respuestas, si estas descansan únicamente en la memoria de los entrevistados. Las que se refieren a eventos muy memorables (nacimiento, matrimonio, etc.) son más confiables que las referidas a eventos cuya ocurrencia y fecha pueden ser más fácilmente olvidables (candidato que votó, episodios de enfermedad, etc.). En algunos casos los datos que se requieren pueden ser respaldados con algún

[2] Acerca de los métodos adecuados para analizar encuestas repetidas (con las mismas variables pero diferentes muestras) véase Firebaugh, 1997.

respaldo objetivo (por ejemplo, si cada entrevistado presenta boletines escolares, historias clínicas, certificados de vacunación, etc.), pero ese diseño no es habitual en encuestas con muestras grandes.

1.3.4. Paneles de registro periódico

En un nivel superior al de las encuestas repetidas y al de las encuestas retrospectivas se encuentran los **estudios de panel** propiamente dichos, que realizan **varias rondas** de recolección de información **sobre las mismas unidades de análisis**. Las rondas se realizan en momentos o períodos específicos, separados entre sí por algún intervalo. Por ejemplo, cada caso puede ser entrevistado con intervalos de un trimestre. Por eso nos referimos a **paneles de datos periódicos.**

Estos diseños registran simultáneamente **macrocambios** y **microcambios**, pues obtienen información de **los mismos sujetos o unidades de análisis en varios momentos del tiempo**, lo cual permite observar cambios a nivel individual así como cambios agregados. En cada ocasión se recoge información sobre **ese momento** (por ejemplo la actual opinión de las personas sobre un cierto tema) o sobre momentos recientes, ocurridos después de la ronda anterior (por ejemplo, el tiempo que ha pasado en situación de desocupado durante ese intervalo).

La observación de los mismos casos en varias ocasiones puede tener dos propósitos: la obtención de mediciones repetidas de un fenómeno que se presume constante, o la obtención de información sobre un fenómeno que cambia a través del tiempo.

En las **mediciones repetidas de un fenómeno constante**, el fenómeno, por lo general, es medido varias veces no tanto para estudiar sus cambios a lo largo del tiempo, sino para obtener una **medición promedio**, más confiable que una medición individual, pues las mediciones pueden oscilar aleatoriamente a lo largo del tiempo sin que ello implique cambios en la realidad subyacente. Por ejemplo, un médico puede tomar la tensión arterial al paciente varias veces en el día, para tener una medición más confiable de la tensión arterial **media** de ese individuo, o de la **varianza** de ese indicador, haciendo caso omiso de las **secuencias temporales** de las fluctuaciones aleatorias que se registren alrededor de ese promedio (por supuesto, si alguna de esas observaciones sucesivas revelase un valor muy anormal, o si se observara una marcada tendencia ascendente o descendente, el médico puede aconsejar nuevos exámenes más profundos o prolongados). El médico no está interesado en principio en la tendencia de la tensión arterial: solo quiere medirla.

La repetición puede hacerse a veces solo para no arriesgar todo en una sola observación. En un caso célebre, cuando en 1919 se trató de comprobar uno de los efectos esperados de la teoría de la relatividad general de Einstein (el efecto gravitatorio sobre la luz) mediante mediciones de la desviación de la luz de las estrellas en las cercanías del sol, obtenidas durante un eclipse total de Sol en una remota isla cerca de Africa, los científicos intervinientes tomaron **varias** fotografías del cielo eclipsado, a fin de no depender de una sola medición. Estas mediciones repetidas no eran en realidad un panel, sino solo una **muestra de varias observaciones** acerca de un fenómeno subyacente que se suponía sustancialmente constante aunque pasible de variaciones aleatorias por imperfecciones de la medición. En el fondo, estos estudios no son estudios longitudinales: se observa un solo fenómeno, pero se lo mide varias veces para tener mayor certidumbre. Los científicos de 1919 podrían haber tenido varias cámaras que obtuviesen una foto cada una, todas simultáneas, en lugar de varias fotos sucesivas con la misma cámara. En todo caso, en vez de un **prisma** de datos, lo que se obtiene es una **matriz** de datos bidimensional; cada elemento es **una** medición de **una** variable.

Estos casos de medición repetida con finalidad meramente confirmatoria no atribuyen significado sustantivo a la **secuencia** de una a otra observación. El orden temporal de las observaciones carece de importancia. Pero la mayoría de los estudios longitudinales tratan de capturar **cambios efectivos a lo largo del tiempo**, y en ese caso el orden temporal es esencial.

Paneles simples y rotativos. Los paneles de registro periódico pueden ser simples o rotativos en cuanto a la composición de la muestra de casos. Los **paneles simples**, o **paneles de seguimiento**, observan **el mismo grupo de personas** a lo largo de toda la duración del estudio (salvo por desgranamiento espontáneo debido a muertes, abandonos, etc.). Ejemplos de este tipo de estudios son por ejemplo los estudios médicos de largo plazo sobre salud y estilo de vida, en que la misma muestra

es seguida durante muchos años, y también en plazo más corto el seguimiento de una misma muestra de votantes a lo largo de una campaña electoral.

Los **paneles rotativos** se utilizan cuando interesan procesos concentrarse en procesos de plazo más breve, aunque el estudio se realiza de manera permanente a lo largo de mucho tiempo. En estos casos, los miembros de la muestra permanecen en el estudio durante un cierto número de rondas, y luego son reemplazados. Los paneles rotativos se revisan con más detalle en la sección 1.4, y el desgranamiento espontáneo en la sección 1.5.

1.3.5. Paneles simulados y cohortes teóricas

En algunas circunstancias, es imposible o poco práctico hacer un seguimiento efectivo de los mismos casos o sujetos a lo largo del tiempo. En tales casos, sin embargo, el análisis de panel puede aplicarse a un "panel teórico", simulado a partir de datos ***captados en una sola fecha***. Unos datos esencialmente sincrónicos se usan para "construir" una secuencia simulada a lo largo del tiempo.

En las encuestas de reconstrucción retrospectiva (sección 1.3.3) se da una situación parecida: en una misma ocasión se recogen datos de cada sujeto referidos a diferentes fechas del pasado. Pero todavía en ese caso retrospectivo, si bien la recolección de datos se hace en una sola fecha, los datos en sí siguen siendo históricos: se recoge información de cada sujeto en diferentes fechas. En el caso presente, en cambio, la información se recoge en una sola ocasión y se refiere a un solo período, pero sirve para construir una reconstrucción histórica simulada, con datos de *diferentes* sujetos.

Un ejemplo clásico son las **tablas de vida**. Esas tablas describen la forma en que una población nacida en un cierto momento se va extinguiendo gradualmente mediante defunciones ocurridas a diferentes edades. El promedio de años vividos por los miembros de esa población es lo que se denomina "expectativa de vida al nacer".

Una forma de construir esa tabla sería tomar una cohorte (por ejemplo los nacidos en 1920) y seguirlos a lo largo de los años subsiguientes, a fin de registrar la edad en que cada uno de ellos muere; de la población nacida en 1920 quedan muy pocos sobrevivientes en 2012; esa cohorte histórica se extinguirá totalmente en pocos años más. Pero un estudio de ese tipo es muy engorroso y prácticamente no se ha llevado a cabo nunca. Además de su costo y dificultad, sería necesario ir siguiendo todas las cohortes (por ejemplo las nacidas en 1921, 1922, y así sucesivamente) a fin de actualizar permanentemente los riesgos de mortalidad.

Ese tipo de estudio, en teoría, podría realizarse a partir de las actas de nacimiento y defunción en los registros civiles, y de hecho hay algunos estudios de ese tipo en países donde el Registro Civil ha sido completo y fidedigno, y donde hay pocos emigrantes (por ejemplo Islandia). Pero normalmente las tablas de vida no se construyen de ese modo.

Las tablas de vida, en la práctica, no se construyen mediante el seguimiento de cada cohorte de individuos desde el nacimiento hasta la muerte. En realidad, se toman las **tasas de mortalidad por edad observadas en un cierto período**, por ejemplo en un cierto año, en que ocurrieron defunciones de **personas de diferentes edades**, es decir pertenecientes a diferentes cohortes. Las probabilidades de morirse a diferentes edades van cambiando con el tiempo (por ejemplo en virtud de la introducción de adelantos médicos, o la ocurrencia de epidemias). Las personas de 70 años fallecidas en 2010 habían nacido (aproximadamente) en 1940, y por ende pertenecían a una cohorte completamente diferente de la formada por personas de 50 años fallecidas en 2010, las cuales habían nacido en 1960. Aquellas personas nacidas en 1940, cuando tenían 50 años (en 1990) estaban expuestas a una tasa de mortalidad diferente a la que rigió en 2010 para personas que (en 2010) tenían 50 años.

Del mismo modo, los niños de 10 años existentes en el período de referencia (2010), algunos de los cuales murieron en ese período, tenían una tasa de mortalidad (experimentada en 2010) que no es la misma tasa que experimentaron los adultos o ancianos de ese período **cuando ellos tenían 10 años** (muchos años atrás). La tasa de mortalidad que sufrirán en el futuro los niños de 2010 cuando lleguen a tener 70 años no será igual a la tasa de mortalidad que sufrieron en 2010 las personas de 70 años. En realidad no existe ninguna cohorte real de personas que haya estado sometida a las tasas

de mortalidad por edad observables hoy: estas tasas son las que afectan hoy a las diferentes cohortes existentes hoy, nacidas en diferentes años, cada una de las cuales tiene hoy diferentes edades. Cada una de esas cohortes reales estuvo sujeta, a lo largo de su vida, a una secuencia específica de riesgos de mortalidad, que no coinciden con los riesgos de otras cohortes a las mismas edades.

Una tabla de vida de 2010 no se refiere a ninguna de las cohortes realmente existentes en 2010. Se refiere a una **población teórica** que nunca existió realmente. Sobre la base de las tasas de mortalidad por edad observadas en 2010, correspondientes a personas de diferentes cohortes históricas, se construye una **cohorte teórica**, supuestamente formada por personas nacidas todas ellas al mismo tiempo, y a esa población teórica se le van aplicando las tasas de mortalidad observadas (para cada edad) en el período de referencia. Esto hace que el tamaño inicial de la cohorte teórica (por ejemplo cien mil personas) vaya disminuyendo a medida que sus miembros mueren a diferentes edades.

Cuando se define la "expectativa de vida al nacer" sobre la base de una tabla de vida, se calcula cuánto viviría en promedio un miembro de la cohorte teórica, pero ello no es directamente aplicable a los niños nacidos hoy: esos niños experimentarán, dentro de 50 o 70 años, las tasas de mortalidad que regirán **en esa época futura** para personas de 50 o de 70 años, las cuales serán seguramente distintas de las que rigen ahora para los actuales cincuentones o setentones. La tabla de vida corresponde, pues, a una cohorte teórica.

El mismo principio de la cohorte teórica utilizado en las tablas de vida puede ser aplicado en otros campos. Por ejemplo, las tasas de abandono escolar observadas en un cierto año (sobre estudiantes que cursaban en ese año diferentes niveles de educación) se pueden aplicar a una cohorte teórica de estudiantes, sometida supuestamente a esas mismas tasas de abandono escolar a lo largo de su proceso educativo. En forma similar, los datos sobre divorcios ocurridos en un cierto año a matrimonios que se formaron en diferentes momentos del pasado, permite calcular una "expectativa de vida de la pareja", referida a una cohorte teórica de parejas formadas todas ellas en un mismo año y sometidas a lo largo del tiempo a las tasas de divorcio observadas en el año de referencia (que afectaron a matrimonios recientes o antiguos, celebrados en diferentes momentos del pasado).

Si se genera una cohorte teórica, es posible que se pueda saber sobre ella no solo el hecho fundamental (por ejemplo la edad de muerte) sino también otros hechos (su educación, su estado civil a diferentes edades, el número de hijos que tuvieron, su condición socioeconómica. Por ejemplo, es posible que los registros de mortalidad tengan información sobre área de residencia, estado civil y otros datos; también la tabla de vida podría basarse sobre una encuesta demográfica en la cual se averigüen las defunciones ocurridas en los hogares entrevistados, y en tal caso se contará con información sobre el fallecido y sobre su hogar. Esa base de datos, referida a una población teórica, podría permitir un análisis de panel simulado. Por ejemplo, se podría comparar la expectativa de vida de personas residentes en diferentes zonas, o con diferente nivel socioeconómico, entre otros muchos análisis posibles. El riesgo de mortalidad o la expectativa de vida podría así calcularse separadamente para muchas subpoblaciones, o calificarse según diversos condicionantes (no solo sexo y edad, como es habitual, sino status socioeconómico, zona urbana o rural, región de residencia, etc.).

1.3.6. Paneles de registro continuo

Los paneles como el de la EPH son de **registro periódico**: se recogen datos de cada caso en fechas o períodos determinados y discontinuos (aun cuando los casos de la muestra estén distribuidos en el tiempo). Es decir que aun en la EPH "continua" cada hogar es entrevistado en ocasiones separadas por intervalos de uno o más trimestres, y no en forma continua.

En cambio los estudios de **registro continuo** son habitualmente los que se basan en datos de registro, como por ejemplo los datos extraídos de los legajos de personal de las organizaciones, donde se anotan todos los eventos concernientes a cada trabajador: hora de llegada y salida, licencias, ausentismo, accidentes, promociones o ascensos, vacaciones, etc., con la fecha exacta de cada evento. Lo mismo pasa con datos sobre movimientos de cuenta y transacciones de cada cliente en bancos u otras empresas, o datos de ventas o inventario de mercadería en supermercados, o registros de signos vitales de pacientes internados en salas de terapia intensiva.

En algunos casos se trata de datos fechados de manera continua, pero que son **registrados a intervalos**, de modo que los eventos intermedios están sujetos a un **registro retrospectivo**. Aun cuando los datos son continuos, la recolección de los datos no lo es. Por ejemplo, un médico puede anotar mensualmente en la historia clínica del paciente los acontecimientos relevantes de todo el mes, con sus respectivas fechas. En un intervalo más breve, la enfermera puede anotar diariamente en la historia clínica los eventos relevantes de un paciente ocurridos a diferentes horas del día. De todas maneras, si los registros son de alta frecuencia, y la longitud del intervalo es breve, los datos se pueden considerar como registrados en forma continua.

1.4. Paneles rotativos

Hay estudios longitudinales de largo plazo que intentan reconstruir toda la historia de vida de cada sujeto. Hay así estudios médicos que siguen a una misma muestra de personas durante 40 o 50 años para analizar la evolución de su salud y su estilo de vida. Otros estudios de este tipo pueden ser de corto plazo, como el seguimiento de una muestra de votantes a lo largo de una campaña electoral.

En esos estudios no hay un plan de rotación o reemplazo de los casos. Los sujetos en principio deben permanecer en la muestra durante todo el estudio; pueden dejar el estudio por eventos exógenos no programados (muerte, emigración, abandono voluntario del estudio, etc.) pero de cualquier modo **no son reemplazados**. Lo que se intenta es el seguimiento **del mismo grupo** (o de quienes permanezcan en el grupo) a lo largo de todo el proceso bajo estudio (por ejemplo durante toda su vida, o durante toda una campaña electoral). Si se desea llegar al final con un cierto número de casos que permita inferencias estadísticas válidas, se comienza con un grupo más numeroso en previsión del inevitable desgranamiento.

Pero aquí nos interesa otra clase de estudios, los paneles **rotativos.** En esos estudios de panel el estudio se puede prolongar por muchos años, pero interesan primordialmente procesos de duración más breve, y donde por lo tanto no es necesario seguir a los mismos sujetos durante tanto tiempo. Por ejemplo en una encuesta de empleo puede interesar el seguimiento de cada sujeto durante un año o dos, a fin de observar cambios en su situación ocupacional, pero no se pretende seguirlo durante muchos años.

En ese caso, los sujetos de la muestra pueden ser reemplazados por otros **en forma programada**, y para ello el mecanismo más usual son los **paneles rotativos.** De hecho, la rotación también resuelve otros problemas, que también afligen a los estudios de largo plazo: es difícil mantener a los mismos sujetos durante muchas rondas, pues los sujetos evidencian fatiga, abandonan el panel voluntaria o involuntariamente, o bien reducen la calidad y veracidad de sus respuestas. Por otro lado, los hogares o personas que eran representativos de la población al inicio del estudio podrían ya no ser representativos al cabo de uno, dos o más años. Los estudios de largo plazo deben lidiar con estos problemas, pero los paneles rotativos se libran de ellos mediante la renovación gradual de los casos.

En los **paneles rotativos**, cada sujeto permanece en la muestra durante un número limitado de rondas, tras de lo cual es reemplazado por un caso "fresco". Este proceso es gradual y programado. Normalmente en cada ronda hay un cierto número de casos que están siendo entrevistados por primera vez, otros que están en su segunda aparición, otros en la tercera, y así sucesivamente, hasta aquellos que están respondiendo a la encuesta por última vez. Los esquemas de rotación pueden ser de distinto tipo. Los siguientes ejemplos muestran dos esquemas de rotación muestral.

Ejemplo 1: La EPH puntual (1974-2002) en la Argentina. En la Encuesta Permanente de Hogares de la Argentina tal como estaba organizada entre 1974 y 2002, con dos rondas por año, los hogares permanecían en la muestra durante cuatro rondas (por ejemplo mayo, octubre, mayo, octubre) en dos años sucesivos. En cada ronda hay un 25% de la muestra que es entrevistado por primera vez, 25% por segunda vez, 25% por tercera vez, y 25% por cuarta y última vez.

En ese tipo de panel rotativo, donde en cada ronda se reemplazan **M** casos, que representan por ejemplo **1/4** del total, se tiene en todos los períodos una muestra de **4M** casos entrevistados, y al mismo tiempo se cuenta con un 75% de los casos (**3M**) para comparar el período corriente con el anterior, un 50% de los casos (**2M**) para compararlo con los dos anteriores, y un 25% de los casos

(es decir **M** casos) para comparaciones entre el periodo corriente y los tres anteriores. Se pueden así captar trayectorias de hasta cuatro períodos, con datos sobre cada uno de esos períodos.

Ejemplo 2: La EPH Continua de la Argentina (2003-). En el nuevo esquema de esa Encuesta Permanente de Hogares de la Argentina, que comenzó a aplicarse en 2003, la encuesta se realiza de manera continua, con rotaciones trimestrales. La muestra total se divide en cuatro partes iguales (A, B, C, D) con 1/4 de los casos totales cada uno de ellos; el grupo A ingresa en el primer trimestre, el grupo B en el segundo, el C en el tercero y el D en el cuarto. Cada uno de estos grupos es encuestado durante dos trimestres sucesivos, luego "descansa" dos trimestres, y vuelve a ser entrevistado por otros dos trimestres (esquema 2-2-2). Esto permite observar, para cada hogar, los cambios ocurridos entre dos trimestres sucesivos en el primer año, entre otros dos trimestres sucesivos en el segundo año, y también los cambios inter-anuales (entre un trimestre del primer año y el mismo trimestre del año siguiente).[3] El esquema general de rotación es el siguiente:

Año t				Año t+1				Año t+2				Año t+3			
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A	**A**			**A**	**A**			**A**	**A**			**A**	**A**		
	B	**B**			**B**	**B**			**B**	**B**			**B**	**B**	
		C	**C**			**C**	**C**			**C**	**C**			**C**	**C**
D			**D**	**D**			**D**	**D**			**D**	**D**			**D**

Pero los hogares de cada grupo deben ser **renovados.** En el caso de la EPH continua, los hogares de cada grupo de rotación son renovados (sustituidos por hogares nuevos) a razón de 50% por semestre. En otras palabras, solo el 50% de la muestra **de cada grupo de rotación** sobrevive hasta el año siguiente. En cada grupo (por ejemplo A) hay dos mitades: una mitad que va a ser dada de baja al final de los próximos dos trimestres, y otra mitad que sobrevivirá hasta volver a entrar en escena un semestre después. Si denotamos estas sucesivas mitades de hogares con índices 1, 2, 3, etc., por ejemplo A1, A2, etc., el gráfico precedente debería ser expandido para mostrar claramente estos **contingentes de renovación**. En cada trimestre, cada grupo (A, B, C o D) está representado por una mitad "vieja" o "saliente", que no sobrevivirá más allá del semestre, y una mitad "nueva" o "entrante", destinada a ser entrevistada de nuevo más adelante, como se muestra en la tabla siguiente.

En el primer trimestre del año t la muestra trimestral (que en el caso de la EPH es de unos 25000 hogares) está compuesta por hogares del grupo A (encuestado en el primer y segundo trimestre) y por hogares del grupo D (que son entrevistados en el primer y cuarto trimestre), subdivididos en cuatro contingentes: **A1** (que es un grupo saliente), **A2** (entrante), **D1** (saliente) y **D2** (que ingresó en el cuarto trimestre del año t-1). Cada uno de esos contingentes tiene aproximadamente 6250 casos. En el segundo trimestre solo reaparece la mitad de los casos del trimestre anterior (**A1** y **A2**). Los grupos **D1** y **D2** no reaparecen porque a ellos les toca aparecer solo en los trimestres 1 y 4 de cada año. En el primer trimestre del año t+1, se repite la mitad de la muestra de un año antes, es decir a los contingentes A2 y D2, mientras que A1 y D1 han sido dados de baja. En otras palabras, la comparación de un trimestre con el trimestre precedente, o con el mismo trimestre del año anterior, se puede hacer con el 50% de los casos encuestados en cada trimestre. Dado que la muestra trimestral (en la EPH argentina) es de 25000 hogares, la mitad comparable son aproximadamente 12500 hogares.

[3] Aparte de este esquema de rotación trimestral de grupos, la muestra de hogares de cada trimestre es distribuida a lo largo de dicho trimestre, de modo que los resultados son representativos de todo el trimestre, aun cuando las preguntas específicas pueden referirse a la fecha de la entrevista, como en el caso de la edad, o a la semana anterior, como en el caso de la situación de empleo. Cada hogar es asignado a una cierta semana dentro del trimestre, y la mantiene en los sucesivos trimestres en los cuales es entrevistado. Un hogar del grupo A, asignado por ejemplo a la segunda semana del primer trimestre, se mantiene en esa semana en los sucesivos trimestres en que sea entrevistado (INDEC 2003).

Año t				Año t+1				Año t+2				Año t+3			
1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
A1	**A1**														
A2	**A2**			**A2**	**A2**										
				A3	**A3**			**A3**	**A3**						
								A4	**A4**			**A4**	**A4**		
												A5	**A5**		
	B1	**B1**													
	B2	**B2**			**B2**	**B2**									
					B3	**B3**			**B3**	**B3**					
									B4	**B4**			**B4**	**B4**	
													B5	**B5**	
		C1	**C1**												
		C2	**C2**			**C2**	**C2**								
						C3	**C3**			**C3**	**C3**				
										C4	**C4**			**C4**	**C4**
														C5	**C5**
D1															
D2			**D2**	**D2**											
			D3	**D3**			**D3**	**D3**							
							D4	**D4**			**D4**	**D4**			
											D5	**D5**			**D5**
															D6

La renovación de los grupos por mitades procede como se indica en la tabla. En los dos primeros trimestres la muestra del grupo A se compone de un contingente saliente A1, que está en su segunda aparición semestral, y un contingente entrante de casos frescos (A2) que acaba de entrar en la muestra. Ese contingente fresco A2 reaparecerá en los dos primeros trimestres del año siguiente, pero A1 va a ser sustituido por un contingente nuevo, A3. Este último es a su vez encuestado nuevamente en los dos primeros trimestres del año t+2, pero no acompañado por A2 sino por un contingente nuevo (A4); y así sucesivamente. Lo mismo pasaría con los grupos de rotación B, C y D. La renovación hace que un hogar no permanezca indefinidamente en la muestra; ello evita la "fatiga del respondente" y permite reflejar mejor los cambios que vayan ocurriendo en la población. De este modo, la muestra de un trimestre comparte el 50% de los casos con la muestra del trimestre sucesivo, y también 50% con la de igual trimestre del año siguiente. En general, y dejando de lado la denotación de los contingentes de renovación, la muestra de cada trimestre estará compuesta por dos grupos de igual tamaño, con la alternancia siguiente:

Trimestre			
1	2	3	4
A	**A**	**C**	**C**
D	**B**	**B**	**D**

En el primer trimestre del año la muestra proviene de los grupos A y D (con un 50% de casos del grupo A y otro 50% del grupo D). En el segundo trimestre se encuesta a los grupos A y B; en el tercero a los grupos B y C, y en el cuarto a los grupos C y D. Al año siguiente se repite el esquema.

Como se vio antes, en su tercera aparición, después de "descansar" por dos trimestres, la muestra de cada grupo es renovada en un 50%. Así, por ejemplo, el grupo D que ingresa en el trimestre 4 contiene la mitad de los casos que integraban el mismo grupo D en el trimestre 1, y otra mitad de casos "frescos". Este reemplazo por contingentes "frescos" tiene por objeto evitar la fatiga o acostumbramiento de los encuestados, y reflejar también cualquier cambio que hubiese en la población de base. En el caso de la EPH continua en la Argentina, cada uno de los cuatro grupos (A, B, C, D) se compone de aproximadamente 12500 hogares; en cada trimestre se entrevista a 25000 hogares,

pertenecientes por partes iguales a dos de esos grupos. En cada uno de esos grupos, a su vez, hay contingentes "nuevos" y "viejos" de aproximadamente 6250 casos cada uno.

En este esquema, un mismo hogar es entrevistado cuatro veces a lo largo de un año y medio, en los dos primeros y los dos últimos trimestres de ese período. Es así posible estudiar trayectorias de hasta cuatro períodos, como en la antigue EPH puntual, pero los períodos ya no son equidistantes: hay dos períodos intermedios sin entrevista. Habría un trimestre de diferencia entre las ocasiones 1 y 2 y también un trimestre entre las ocasiones 3 y 4, pero en cambio habría nueve meses entre las ocasiones 2 y 3. Esto es importante tenerlo en cuenta si se quieren calcular "tasas de transición", pues estas deben referirse a algún período estándar, por ejemplo un trimestre. Los cambios ocurridos en un intervalo de nueve meses deberían ser expresados en "cambios por trimestre" para poder ser comparables. Es la misma operación que realizamos cuando el combustible utilizado para diferentes viajes en auto son estándarizados en forma de "litros por cada 100 kilómetros", independientemente de la distancia cubierta en cada uno de los viajes. Las tasas de transición son examinadas en los capítulos 3 y 4.

1.5. Desgranamiento y reemplazo

La rotación planificada de las muestras se sobrepone a un proceso no planificado de **desgranamiento** (*attrition*). De una onda a otra hay siempre algunos sujetos que "desaparecen" de la muestra por distintas razones: muerte, emigración o cambio de domicilio, negativa a seguir en el estudio, etc. Algunas de estas desapariciones pueden ser por causas "sustantivas", que de por sí constituyen un dato, por ejemplo la muerte o la emigración. En otros casos se trata de una "desaparición" cuya causa se desconoce, o de un rechazo a la nueva entrevista. En el caso de las encuestas de hogares puede haber **hogares completos** que "desaparecen" y también **individuos determinados** que dejan de aparecer dentro de algunos hogares aun cuando el hogar como tal siga incluido en la muestra.

Según el modelo muestral que se utilice estos sujetos "desaparecidos" pueden o no ser reemplazados. Naturalmente, si el hogar continúa en la muestra, entonces no se reemplaza a los miembros "desaparecidos" (más bien se puede investigar la razón de su ausencia). Si se reemplazan los hogares desgranados, el hogar que ingresa como reemplazante representa una "entrada tardía", que permanecerá en la muestra solo el tiempo necesario para completar el período programado para el hogar al cual está reemplazando. Por ejemplo, si en el esquema de la EPH continua hubiese un hogar reemplazante del grupo A1, que entrase en la segunda ocasión (trimestre 2 del primer año), solo será entrevistado una vez, porque todo el contingente A1 es dado de baja al final de ese trimestre. En cambio, si se reemplaza en ese mismo período un hogar del contingente A2, ese hogar será entrevistado en total tres veces: en el segundo trimestre del primer año, y en los dos primeros trimestres del segundo año. Tanto los hogares "desaparecidos" como los de "entrada tardía" no llegar a permanecer en la muestra durante todo el ciclo programado de cuatro observaciones. Esto puede introducir problemas en el análisis, así como en el cálculo estadístico de los errores de estimación.[4] Por ejemplo, para analizar transiciones entre dos trimestres solo podrán utilizarse los casos que efectivamente hayan sido entrevistados en ambos trimestres considerados, dejando de lado aquellos que se desgranaron (que figuran solo en el primero de esos trimestres) y dejando también de lado a los reemplazantes. (que solo figuran en el segundo).

1.6. Estados y eventos

Aparte de considerar la composición y dinámica de **los casos**, es importante analizar las características de los distintos **tipos de información** que se recoge sobre ellos.

En este texto nos concentramos en paneles con variables categóricas, las cuales incluyen varias categorías en las cuales puede estar ubicado cada sujeto en un momento dado. Para los fines del

[4] Sobre el fenómeno del desgranamiento o desgaste (*attrition*) y sus implicaciones estadísticas véase Alderman y otros (2001), van der Berg y Lindeboom (1998), Lillard y Panis (1998), Fitzgerald y otros (1998), Zabel (1998), Ziliak y Kniesner (1998) y el capítulo 5 de Kish (1987). El problema del muestreo en paneles, excelentemente tratado por Kish, es también abordado en Kyriazidou (1997 y 1999).

análisis secuencial propio de los paneles, hay que distinguir entre la información referida al **estado** del sujeto (respecto de una variable) en un **momento** determinado, y por otro lado la información referida a los **eventos** (cambios de estado) experimentados por una variable a lo largo de un **período** determinado en el cual se reconocen diferentes momentos.

Tipos de información de una variable para cada sujeto o caso		
Concepto	**Definición**	**Referencia temporal**
Estado	Valor de la variable	Ocasión determinada
Evento	Cambio de valor de la variable	Tiempo transcurrido entre dos ocasiones

La "ocasión" o "momento" en que se mide el estado de cada sujeto rara vez se refiere a un "instante". Usualmente ese momento es a su vez un **período** considerado en conjunto, es decir como un período indivisible dentro del cual no se reconocen subdivisiones, fases o subperíodos. Por ejemplo la edad del entrevistado a la fecha de la entrevista se mide en años cumplidos por el individuo hasta el día de la entrevista inclusive; en este caso el "momento" es un día entero (sin distinguir horas o minutos). Si el individuo cumple años el mismo día de la entrevista, no importa si nació por la mañana o por la noche: solo se registra que nació (y por lo tanto cumple años) **en algún momento** del día. Si las entrevistas de una ronda se distribuyen a lo largo de un mes, o de un trimestre, el mismo individuo podría resultar con diferentes edades según que sea entrevistado al principio o al final de ese período. En la práctica, esas pequeñas diferencias se ignoran, y se considera que esa persona tenía esa edad "en el período en que se recogieron los datos", es decir en un cierto mes o trimestre, sin mayor precisión. Para otros conceptos puede tratarse de un período más largo o más impreciso, también considerado en conjunto y sin reconocer fases o subperíodos dentro del mismo. Por ejemplo: un sujeto es clasificado como económicamente activo si trabajó o buscó trabajo **en algún momento** de la semana anterior.

De todas maneras, los "estados" caracterizan a ese sujeto **en el período de referencia**, sea este un día, una semana o cualquier otro. En cambio los "eventos" son **cambios de estado**, y por lo tanto involucran la comparación de dos momentos u ocasiones, y requieren tener en cuentas el tiempo transcurrido entre esos dos momentos.

En una variable con varios valores categóricos (por ejemplo el estado civil) cada persona en cada ocasión se encuentra en alguno de los varios **estados** posibles (soltero, casado, viudo, divorciado, etc.). En el período entre dos observaciones pueden haber ocurrido uno o más **eventos** respecto de esa variable (algunos sujetos pueden haberse casado, otros se divorciaron o enviudaron, etc.). En algunas variables, y si el período de observación es suficientemente largo, puede haber **más de un evento por período** (por ejemplo, una persona puede haberse casado y luego enviudado dentro del mismo período; puede haberse quedado sin empleo y luego conseguir trabajo nuevamente). Estos casos donde ocurren **eventos intermedios** son examinados más tarde con mayor detalle.

1.7. Tipos de análisis longitudinal de variables categóricas

Según el tipo de datos que se obtenga sobre estados y eventos, los principales tipos de análisis de panel (o longitudinal) cuando se trata de variables **categóricas** son los siguientes.

Tipos de análisis longitudinal con variables categóricas	
Serie de estados	Se mide el **estado** de los sujetos en cada fecha, y se compara cada estado con el anterior. No se registran eventos intermedios
Conteo de eventos	Se registra la **cantidad de eventos intermedios**, pero no su orden secuencial ni la fecha en que ocurrieron
Secuencia de eventos	Se registra la **cantidad y orden secuencial de los eventos intermedios** pero no la fecha en que ocurrieron
Historia de eventos	Se registran los **eventos intermedios** y las **fechas** en que ocurrieron

Los estudios de panel consisten básicamente en una **serie de cortes transversales** de la misma muestra, que proveen información sobre el **estado** de las unidades de análisis en el momento de ca-

da observación, y sobre algunos flujos o cambios ocurridos en el período intermedio, pero usualmente **no cubren el flujo de cambios** ocurridos a lo largo del tiempo. Por ejemplo, una encuesta de empleo puede incluir una pregunta sobre la situación laboral "actual" del sujeto (ocupado, desocupado, inactivo). Esta información puede efectivamente ser instantánea, o bien puede referirse a un breve período inmediatamente anterior a la entrevista (por ejemplo durante la última semana), pero de todas maneras apunta a registrar la situación de la población en ese momento, y no intenta reconstruir la evolución de esa situación desde la anterior entrevista realizada varios meses antes.
En esos casos, los estudios de panel permiten **comparar estados**, y por lo tanto dan información sobre los **cambios netos** ocurridos entre una y otra ronda, pero no sobre la **secuencia de cambios** que puede haber ocurrido en el interín. Por ejemplo, si un sujeto estaba "ocupado" en la ronda 1 de una encuesta, y "desocupado" en la ronda 2 realizada varios meses después, se sabe que al menos ha habido un cambio (el sujeto en algún momento quedó desocupado), pero no se sabe si ése fue el único cambio que hubo en la situación laboral de ese individuo: el sujeto puede haber perdido y encontrado empleo varias veces durante el período intermedio, o puede haber pasado transitoriamente a la inactividad económica, o quizá se ausentó del país durante un tiempo: el panel sólo registró su **estado inicial** y su **estado final**, y no los estados intermedios. Aun si el individuo no presentase variación alguna en su situación (por ejemplo, si estaba ocupado en ambas rondas), no se podría aseverar que no haya sufrido cambios en su situación laboral: pudo haber estado desocupado en algún momento intermedio sin que la pregunta formulada lo registre pues sólo se refiere a la situación inmediatamente previa a la entrevista (a menos que la encuesta incluya una pregunta específica de tipo retrospectivo). Tampoco se registra el tiempo que permaneció en cada estado: el sujeto pudo haber quedado desocupado inmediatamente después de la primera ronda, permaneciendo así hasta la segunda, o pudo estar ocupado casi todo el intervalo entre las rondas quedando desocupado poco antes de la segunda entrevista. En muchos de estos casos la fecha exacta del evento no es registrada en estudios de panel que se limiten a medir el estado "actual" del sujeto en el momento de la entrevista o en un breve período precedente. **La fecha exacta de la transición** (la fecha, o mejor dicho la última fecha, en que el sujeto quedó desocupado), puede haber sido en cualquier momento dentro del intervalo entre las dos rondas.

En muchos estudios de panel se incluyen, de todas maneras, preguntas referidas al período intermedio (por ejemplo ¿cuántas veces perdió su empleo en los últimos seis meses? O bien ¿cuánto tiempo lleva buscando trabajo?), lo que permite remediar en parte ese problema. Pero debe tenerse en cuenta que el intervalo entre las rondas es usualmente arbitrario. El resultado es siempre referido al momento de la encuesta, y resultaría diferente si se eligiese otro momento para realizarla: si la encuesta se realiza en octubre, los sucesos ocurridos "en los últimos seis meses" no son los mismos que se registrarían si la encuesta se realizase en julio o en otro momento del año. En cualquier caso, ese período (los últimos seis meses) puede ser comparado al mismo período registrado en la encuesta anterior, de modo que aun en ese caso el análisis compara dos estados instantáneos (los **cambios acumulados al día de la encuesta** durante los seis meses precedentes), y generalmente no permite un análisis detallado del período intermedio como tal. Registrar retrospectivamente una historia detallada de un período de esa longitud, incluyendo fechas, es un recurso posible para superar esta limitación, pero la retrospección frecuentemente no arroja resultados confiables, sobre todo si se pretenden respuestas detalladas y fechas precisas.

Para concluir, los modelos de panel más simples incluyen sólo "variables de estado" medidas en el momento de cada ronda de la encuesta, y no tienen información sobre el período intermedio, sino sólo sobre el **estado del sujeto al momento de cada una de las rondas**. En el presente análisis esos son los datos de panel que primariamente se tendrán presentes, a menos que se especifique lo contrario expresamente.

Cuando se registran eventos intermedios de manera retrospectiva se pueden realizar algunos análisis especiales, en particular los denominados **conteo de eventos** (*event count*) y **secuencia de eventos** (*event sequence*). Los datos de conteo de eventos indican **cuántos** eventos de cierto tipo ocurrieron en el período (por ejemplo, cuantas veces estuvo desocupado, o cuántas veces fue a comer a un restaurante). Los datos de secuencia de eventos registran no sólo la cantidad de veces sino también

la secuencia o sucesión de eventos **en el orden en que ocurrieron**. Por ejemplo, en un estudio laboral con conteo de eventos podría registrarse para cada trabajador la cantidad de varios posibles eventos (nombramientos, despidos, ascensos, licencias, vacaciones, huelgas), y en un estudio de secuencia de eventos se registrarían en el orden temporal en que ocurrieron, aunque no necesariamente su fecha exacta. Este tipo de estudios pueden basarse en una **observación continua** o equivalentemente en un **registro continuo** (como el que se lleva en los legajos individuales del personal de las organizaciones o empresas), o bien pueden basarse en las **preguntas retrospectivas** incluidas en las encuestas de panel. Aun cuando existan datos sobre fechas (por ejemplo en un registro de personal) el análisis de los conteos o secuencias de eventos sólo está interesado en la presencia o ausencia de los eventos, o en su orden de sucesión temporal. Las fechas no juegan ningún papel esencial en esas clases de análisis.

Otro tipo de análisis longitudinal son las llamadas **historias de eventos** (*event histories)* que se parecen a los de secuencia de eventos pero además incluyen como elemento central del análisis **la fecha exacta en que cada evento ocurre**. Esto puede lograrse mediante encuestas retrospectivas o mediante un registro continuo. Por ejemplo, los datos de mortalidad registran la fecha (y por lo tanto la edad exacta) en que ocurre la muerte, y los legajos de personal incluyen la fecha en que ocurrió cada ascenso de categoría, cada aumento de sueldo, cada vez que el trabajador llegó tarde o no concurrió a trabajar. Estos datos pueden ser frecuentemente datos de registro, como en el caso de la mortalidad o del registro de personal, o también pueden obtenerse retrospectivamente en encuestas transversales o de panel, siempre que el intervalo no sea muy largo.

Un ejemplo de observación continua (o casi continua) son los estudios longitudinales de largo plazo, como es frecuente en la investigación médica, en los cuales un grupo de sujetos es observado a lo largo de un período prolongado, con entrevistas frecuentes, registrando la fecha en que ocurren diferentes eventos o el cambio de ciertas variables. Si bien la observación no es estrictamente continua, las entrevistas son frecuentes y la precisión de las fechas y eventos es suficiente como para considerar que la información es continua. Del mismo tipo son los estudios del desarrollo de carreras personales a través de los legajos individuales del personal de una o varias organizaciones. Otro ejemplo de paneles con observación continua son los estudios cuyas unidades de análisis son países, de los cuales se conocen diferentes estadísticas o eventos a lo largo del tiempo. La observación, en sentido estricto, no es continua sino intermitente, pero dado que se cubren períodos muy largos con muchas observaciones a lo largo del mismo, y los procesos que se investigan son en general de larga maduración, la serie resultante puede ser considerada prácticamente como una serie continua.

Muchos estudios longitudinales se refieren al registro de **eventos** o de **cambios de estado** en relación a variables de tipo cualitativo con dos o más posibles categorías o estados. Por lo tanto, las **variables de observación** más usuales son de tipo **categórico** o **cualitativo** (muerte o sobrevivencia, estado civil, condición de salud o enfermedad, condición laboral, etc.). Pero estos estudios pueden también observar **variables cuantitativas o de intervalo**, como los ingresos mensuales, la estatura, el peso, el valor del patrimonio familiar, el nivel de colesterol en la sangre, o cualquier otra variable de tipo continuo. En este caso, lo que se observa no son "estados" y "cambios de estado" sino "valores" y "variaciones de valor", que no ocurren por saltos entre estados discretos sino por una variación gradual a lo largo de un continuo.

1.8. Paneles sincrónicos y asincrónicos

En una base de datos de panel los casos son observados en varias ocasiones, por ejemplo en varios trimestres sucesivos, o en determinados meses del año. Este diseño general, en su forma más simple, corresponde a casos que cumplen las siguientes condiciones:

(a) Ingresan al estudio en la misma fecha histórica.
(b) Son observados todos al mismo tiempo en las sucesivas observaciones.
(c) Los intervalos entre observaciones son todos iguales entre sí.

Por ejemplo, se puede tomar una muestra de hogares que sean entrevistados por primera vez en el mes de enero del año ***t***, y que vuelven a ser entrevistados a intervalos de tres meses durante ese año:

en abril, en julio, y en octubre. Todos ingresan al mismo tiempo, todas las observaciones son simultáneas, y todas las sucesivas observaciones están equiespaciadas a intervalos constantes.

En una base de datos con estas características se tiene un panel totalmente *sincrónico*: coinciden las fechas históricas (todos comenzaron en el mismo mes del mismo año), coinciden las fechas de las sucesivas observaciones (todos fueron observados por ejemplo en ciertos meses de cierto año), y los intervalos entre sucesivas observaciones son a su vez todos iguales entre sí.

Pero puede haber casos de *asincronicidad*. La asincronicidad tiene tres posibles formas, no excluyentes entre sí: los sujetos pueden ingresar en diferentes fechas, los intervalos pueden ser de diferente duración, y las observaciones de los diferentes sujetos pueden tener distintas periodicidades.

1.8.1. Ingreso en fecha variable

Aun cuando los *intervalos* entre las observaciones sean los mismos para todos los casos, la *fecha* de ingreso en el panel a veces puede variar. Dos típicos ejemplos de ello son los distintos contingentes que ingresan en las muestras rotativas de las encuestas de hogares, y los estudios de seguimiento de pacientes en la investigación médica.

En las bases de datos de panel obtenidas en encuestas con muestras rotativas es habitual que cada sujeto permanezca en la muestra por un número limitado de ocasiones, por ejemplo cuatro rondas. Esos sujetos son reemplazados gradualmente, de modo que en cada ronda se reemplaza una parte de la muestra. Por ello en una ronda determinada, solo una parte de los sujetos ha permanecido ya durante cuatro rondas. Otros han permanecido solo tres o dos rondas, y hay algunos que son observados por primera vez. En una determinada ronda los casos que son entrevistados por cuarta vez pueden ser demasiado pocos para los fines del estudio. Para tener un número más grande de sujetos observados cuatro veces, habría que tomar sujetos que ingresan y salen de la muestra *en diferentes rondas*. En este caso, los sujetos habrían sido observados en diferentes fechas. Esto solo es razonable si el fenómeno que se quiere estudiar **no depende de la fecha**.

Hay variables relacionadas con la fecha, por ejemplo el status ocupacional que depende del ciclo económico, o la asistencia escolar que depende de los meses del año en que no hay vacaciones escolares. Hay otras variables en que la fecha tiene menos importancia, como por ejemplo la evolución del estado de salud de las personas. Este es el caso más frecuente en la investigación médica de **seguimiento de pacientes**, por ejemplo un estudio de personas que han sido diagnosticadas con una enfermedad, que han sido sometidas a una operación quirúrgica o que han iniciado un tratamiento. En este tipo de estudios se tienen datos sobre diferentes pacientes que (por lo general) comenzaron a ser observados *en diferentes fechas*, cuando se produjo el hecho desencadenante de su inclusión (el infarto, la cirujía, el diagnóstico, el inicio del tratamiento). Esto es permisible siempre que el fenómeno estudiado no sea afectado por la fecha de la observación (usualmente no lo es en el caso médico), y por otro lado ello es inevitable en muchos casos: no se puede programar de antemano, por ejemplo, que todos los infartos ocurran al mismo tiempo.

Cuando el fenómeno que se quiere analizar es considerado como independiente de la fecha, un panel con ingresos asincrónicos no debería causar ningún problema. Por ejemplo, las condiciones que operan sobre eventos de salud como los accidentes cardiovasculares no se espera que varíen significativamente entre una y otra fecha, excepto tal vez en plazos muy largos en que se registren cambios importantes en las condiciones de salud. La fisiología humana no será diferente en los pacientes de este año comparados con los del año pasado. Si la fecha no tiene relevancia, se puede formar una base de datos de panel con ingreso asincrónico, conformada por participantes, cada uno de ellos con varias observaciones, que ingresaron al estudio en diferentes fechas.

Cuando el fenómeno, en cambio, depende de la fecha (por ejemplo la condición de actividad económica, que depende del ciclo económico), la agregación de casos de diferentes períodos (personas que entraron en la encuesta en diferentes rondas) se vuelve problemática, aunque no imposible. Las informaciones obtenidas de un sujeto en su segunda o tercera entrevista deberán valorarse de manera diferente en caso que la tercera entrevista haya ocurrido, digamos, en el año (o mes) $\boldsymbol{t}$ o en el año (o mes) $\boldsymbol{t}$**+1**. Las distintas variables observadas en la encuesta, como la condición de actividad y

otras, no solo deberán ser caracterizadas por el *orden secuencial* de la observación (desde la primera hasta la última ronda considerada), sino también por la *fecha* en que el sujeto inició su participación. Por ejemplo, la probabilidad de quedar desocupado no es la misma para trabajadores que ingresaron al estudio al final de un período de auge económico, o en el momento más grave de una crisis, o en un período de recuperación y crecimiento de la economía. Si además de conocer el número ordinal de la observación (primera, segunda, etc.) se conoce también la fecha de inicio de la participación, esa variable (la fecha) debe ser considerada en el análisis de los fenómenos bajo estudio, como por ejemplo en un estudio sobre las causas por las cuales la gente pierde el empleo, o sobre la probabilidad de perder el empleo.

1.8.2. Intervalos variables entre observaciones

Aparte de la fecha en que cada sujeto ingresó al estudio, puede haber variabilidad en los intervalos de observación. La variabilidad puede ser **entre observaciones** y/o **entre sujetos.** En otras palabras, la variabilidad en la periodicidad puede ser la misma para todos los sujetos, o puede a su vez variar entre un sujeto y otro. En esta sección se revisa el caso en que todos los sujetos son entrevistados en la misma secuencia de intervalos, pero esa secuencia se compone de intervalos diferentes entre sí.

Por ejemplo, una encuesta de hogares podría entrevistar simultáneamente a todos los sujetos en marzo, abril, agosto y noviembre. Quizá todos iniciaron en la misma ocasión, y todos son entrevistados simultáneamente, pero el intervalo entre dos observaciones sucesivas no es siempre el mismo: entre marzo y abril hay un mes, entre abril y agosto hay cuatro meses, y entre agosto y noviembre hay tres meses.

Algo parecido, aunque más regular, ocurre con la Encuesta Permanente de Hogares de la Argentina, que desde 2003 tiene un diseño 2-2-2: cada hogar que ingresa en la muestra es entrevistado en dos trimestres sucesivos, descansa durante los dos trimestres siguientes, y finalmente es entrevistado en los dos trimestres subsiguientes que ya corresponden al año siguiente. Entre las dos primeras observaciones de cada hogar hay tres meses de intervalo, y lo mismo entre las dos últimas, pero entre la segunda y la cuarta hay nueve meses. Los intervalos programados entre sucesivas observaciones son, por lo tanto, variables.

1.8.3. Intervalos variables entre sujetos

Hay casos en que los intervalos varían no solo entre un intervalo y el siguiente, sino también entre un sujeto y otro. Esto puede ocurrir, por ejemplo, para bases de datos que han sido reconstruidas a partir de *registros imperfectos del pasado*, o basadas en *decisiones espontáneas* de los casos.

Puede ocurrir, por ejemplo, que las historias clínicas de los pacientes no sean actualizadas a intervalos regulares, sino solo *cuando ellos visitan al médico*, y estas visitas podrían ocurrir con diferente periodicidad: si tenemos una muestra de mujeres embarazadas, entre las sucesivas visitas de la embarazada ***i*** pueden haber transcurrido 15 días, un mes y dos meses, pero estos intervalos no tienen por qué ser los mismos para la embarazada ***j.*** Cada sujeto tendrá su propio calendario de observaciones, con diferentes intervalos: estos intervalos pueden variar entre observaciones sucesivas del mismo sujeto, y además pueden variar entre un sujeto y otro.

Si varía la longitud de cada intervalo de un sujeto a otro, también varía como consecuencia la **longitud promedio** de cada intervalo. De hecho, ya no existe una duración única de cada intervalo, pues cada caso individual (cada embarazada en el ejemplo anterior) tiene intervalos individuales diferentes: solo se podría calcular la duración promedio de cada intervalo, si es que se considera que esa medida es útil. Entre la primera y segunda visita no solo hay diferentes intervalos para la embarazada ***i*** que para la embarazada ***j***, sino que además la longitud promedio del primer intervalo (entre las visitas 1 y 2) va a ser diferente a la duración promedio del segundo intervalo (entre las visitas 2 y 3). Por supuesto, en estos casos tampoco suele haber simultaneidad en el ingreso: cada embarazada se presentó por primera vez en diferente fecha, por ejemplo cuando descubrió o sospechó que podría estar en esa condición. Cada una, además, llegó con diferente grado de avance de su embarazo.

Al igual que en los casos anteriores, esta clase de paneles radicalmente asincrónicos requiere caracterizar cada observación no solo por la **fecha de ingreso** en el estudio sino también por el **número**

de orden de la observación (primera, segunda, etc.) y asimismo por la **longitud del intervalo** entre esa observación y la precedente.

Los paneles realizados con *intervalos variables*, ya sean que varíen entre individuos o a lo largo del tiempo (o ambas cosas a la vez), tienen una complicación adicional respecto de los paneles con periodicidad uniforme. Por ejemplo, las tasas porcentuales de variación entre una ronda y la siguiente corresponderán a intervalos de distinta duración, lo cual los tornaría **no comparables** entre sí. La probabilidad de perder el empleo a lo largo de nueve meses puede ser diferente de la probabilidad de perderlo en tres meses, aunque la "velocidad" con que se pierden empleos sea siempre la misma. En tales casos, las tasas de variación deberían ser estandarizadas, reduciéndolas a un común denominador, por ejemplo a una tasa *mensual* o *trimestral* de variación, aunque solo se cuente con observaciones cada dos, tres, seis o nueve meses.

1.9. La dimensión temporal de las variables

Un panel observa la misma muestra en varias "ondas" (*waves*) o "rondas" (*rounds*), también llamadas "cortes temporales" o "cortes transversales". Los datos obtenidos en cada ronda o corte temporal pueden referirse a un **intervalo** (una "feta de tiempo" o *time slice* delimitada entre dos fechas, por 15 días precedentes a la entrevista) o bien a una **fecha** determinada, por ejemplo al 30 de junio.[5] Los intervalos a su vez pueden ser "absolutos" (entre dos fechas del calendario, como el 1 y el 30 de junio) o bien "relativos" (como los "15 días precedentes a la entrevista", cuando las entrevistas no se hacen todas en el mismo día sino a lo largo de un cierto período de trabajo de campo.

Período de recolección de datos y período de referencia. Es importante en este punto distinguir el período (o fecha) de **recolección** de los datos, y el período (o fecha) de **referencia** de los datos. Por ejemplo, una encuesta semestral de hogares puede tomarse durante un período de recolección de datos que puede ser un determinado día, o distribuirse a lo largo de una semana, o de todo un mes. Ese es el período o fecha **de recolección** de los datos. En esa misma encuesta, por otro lado, la información recogida sobre ciertas variables puede referirse a un cierto día, o a un cierto período (una semana, un mes, un semestre), que sería entonces la fecha o período **de referencia**. Una persona puede ser entrevistada el 5 de agosto (fecha de recolección) acerca de su situación laboral en la última semana de julio (período de referencia). A veces el período de referencia se estipula explícitamente (existencia de ganado al 30 de junio de este año), y otras veces en una forma más genérica que puede permitir cierta ambigüedad. Por ejemplo, si las entrevistas de una encuesta se distribuyen a lo largo de todo un mes, y una de las preguntas se refiere a "los últimos siete días", es evidente que esos siete días no serán los mismos para todas las personas encuestadas.

Stocks y flujos. Es importante distinguir, como se hace habitualmente en Economía, entre variables de estado o stock, y variables de proceso o flujo. Las variables de stock son variables que reflejan el **estado de las unidades en un momento determinado** (por ejemplo: edad, número de miembros del hogar, número de hijos vivos, valor del patrimonio hogareño, número de personas ocupadas). Estas variables se expresan usualmente en sus propias unidades de medida (en personas, unidades monetarias, años de edad, etc.), **sin dimensión temporal,** aunque sí con una **referencia temporal** (corresponden a un período determinado).

Las variables de flujo se refieren a **sucesos ocurridos a lo largo de un período** (por ejemplo: ingresos obtenidos durante el último mes, hijos nacidos vivos en los últimos cinco años, libros leídos durante el último año, etc.) y se expresan en cantidades **por período** (ingresos mensuales, libros leídos por año, etc.). Estas variables no sólo tienen una **referencia** temporal (ingresos mensuales **en el último mes**) sino también una **dimensionalidad** temporal que se manifiesta en su unidad de medida (ingresos **mensuales**, es decir medidos en cantidad de unidades monetarias **por mes**). La distancia recorrida en un viaje es una variable de stock (medida en kilómetros); la velocidad del viaje es una variable de flujo (medida en kilómetros **por hora**).

[5] En el vocabulario habitual de los estudios de panel, el momento en que se realiza cada ronda se denomina a menudo un *período,* en sentido genérico, entendiéndose que el "período" puede ser de duración instantánea.

En general las variables de estado o stock se refieren a un **instante** o **fecha** determinados, pero algunas variables de estado se refieren a un **período**, aunque ello a veces genera cierta ambigüedad; por ejemplo la pregunta "¿Estuvo Ud ocupado durante los últimos siete días?" podría significar "¿Estuvo Ud ocupado [en algún momento] durante los últimos siete días?" lo cual es completamente distinto a la pregunta "¿Durante cuántos días [u horas] estuvo Ud ocupado durante los últimos siete días?" En el primer caso la respuesta se mide en unidades, sin dimensión temporal (en este caso, se mide en personas ocupadas). En cambio si la pregunta hubiese sido "¿[Por cuántas horas] estuvo Ud ocupado durante los últimos siete días?" entonces podría dar lugar a dos variables en el archivo de datos. Una primera variable de estado, como en el caso anterior, clasificaría a la persona como ocupada o no ocupada durante los últimos siete días, de acuerdo al tiempo que haya estado ocupado. Por ejemplo, se la podría considerar ocupada aunque haya tenido sólo una hora de trabajo, o podrían exigirse como mínimo 15 horas, etc.; esta variable reflejaría un "estado" y se mediría en "personas ocupadas", sin dimensión temporal. Otra variable derivada de la misma pregunta sería "horas semanales de trabajo", que se mide en "horas por semana", y que es una variable de flujo por unidad de tiempo, igual que "ingreso por mes" o "días de ausencia por año".

Dimensionalidad. El prisma de datos de panel contiene ambos tipos de variable, capturadas en las varias rondas del panel y referidas a momentos o períodos muy variados, y reflejando tanto stocks o estados en momentos determinados, como flujos por unidad de tiempo a lo largo de un período determinado. Las variables de estado o stock y las variables de flujo se diferencian por su **dimensionalidad**, o en términos más sencillos, por tener o no tener al tiempo incorporado en su unidad de medida. En las variables de flujo, el tiempo figura en la unidad de medida (horas por semana, kilómetros por hora, producción por año). Las variables de stock se miden en cantidades de su propia unidad de medida (personas, dólares, kilómetros). Por ejemplo, el peso de un bebé en cada visita a su pediatra es una variable de stock o de estado, y lo mismo ocurre con la cantidad de artículos en un inventario, o el saldo de una cuenta corriente. Tiene una referencia temporal, porque se refiere a un momento dado, pero se mide en sus propias unidades de medida **sin dimensión temporal alguna**: se mide en kilogramos, en dólares, en metros, en número de unidades existentes. En cambio las variables de flujo se miden en **unidades de medida por período** (kilómetros recorridos por hora, dólares ingresados por año, unidades vendidas por mes, etc.). En estos casos **el tiempo forma parte de la unidad de medida**. Si **M** es una unidad de medida genérica y **T** es el tiempo, se dice que las variables de estado o de stock tienen dimensionalidad **M**, y las variables de flujo más sencillas tienen dimensionalidad **M/T**. De hecho en muchos casos las variables de flujo pueden definirse como el cociente entre una cantidad **M** y el tiempo durante el cual esa cantidad se ha estado acumulando (velocidad=distancia recorrida/tiempo; es decir distancia recorrida entre dos puntos (**M**) dividida por el tiempo (**T**) que se tardó en llegar desde uno de esos puntos hasta el otro).

Estas son las variables de flujo **más sencillas**. Puede haber variables de dimensionalidad más complicada como las que reflejan **cambios en el flujo**. Una de ellas es la **aceleración** (aumento de la velocidad por unidad de tiempo): no se mide en distancia por unidad de tiempo, **M/T**, por ejemplo metros por segundo, sino en "metros por segundo *por segundo*", los "metros por segundo" *adicionales* recorridos por cada segundo adicional, **(M/T)/T**, o en forma abreviada equivalente: $\mathbf{M/T^2}$. La unidad de medida en ese caso no sería m/s, sino m/s^2. En el campo de las variables socioeconómicas hay muchas de este tipo, que se presentan sobre todo en economía. Por ejemplo, el producto bruto interno es un flujo (dimensionalidad **M/T**) pues representa la cantidad de bienes y servicios producidos **durante un año**. La tasa anual de crecimiento del producto bruto tiene dimensionalidad $\mathbf{M/T^2}$, pues mide el incremento *por año* de los bienes producidos *anualmente*, o sea la producción anual adicional por año. En general, el aumento o variación de una variable de stock es una variable de flujo. Por ejemplo, la población en un momento dado es una variable de stock; el **aumento** de esa población **entre dos momentos** es una variable de flujo. El aumento o variación de una variable de flujo, a su vez, es una variable de flujo de dimensión inmediatamente superior (así el incremento del producto bruto, que tiene dimensión **M/T**, es una variable de flujo con dimensión $\mathbf{M/T^2}$).

Una variable de flujo sencilla, con dimensión **M/T**, se suele denominar "variable de flujo de primer orden", mientras que una variable con dimensión $\mathbf{M/T^2}$ es una "variable de flujo de segundo orden",

y las puede haber, en principio, de orden superior. Las variables de stock, según esta nomenclatura, podrían ser consideradas como "variables de flujo de orden cero". Su dimensionalidad sería $\mathbf{M/T^0}$: cualquier número elevado al exponente cero es igual a 1, de modo que $\mathbf{M/T^0}$ es equivalente a **M.** En la siguiente tabla se suministran algunos ejemplos sencillos de variables de stock y de flujo, y sus respectivas unidades de medida.

Variables de estado o de stock (referidas a un momento dado)	**Unidad de medida (dimensión M)**
Distancia entre dos puntos geográficos	Kilómetros (km)
Población	Personas
Capital fijo existente	Unidades monetarias
Existencias de una mercadería en el almacén	Unidades físicas
Hijos nacidos vivos en toda la vida	Hijos nacidos vivos
Valor del patrimonio hogareño	Unidades monetarias
Existencias ganaderas	Cabezas de ganado
Variables de flujo de primer orden (referidas a un período determinado)	**Unidad de medida (dimensión M/T)**
Velocidad	Kilómetros por hora (km/h)
Nacimientos	Niños nacidos vivos por año
Defunciones	Personas fallecidas por año
Aumento de la población	Personas adicionales por año
Producto bruto	Unidades monetarias por año
Ventas de mercadería	Unidades vendidas por mes (u otro período)
Hijos nacidos vivos en los últimos doce meses	Hijos por año
Gastos mensuales en alimentación	Unidades monetarias por mes
Comidas en restaurante en los últimos 30 días	Comidas por mes
Extracción ganadera	Cabezas (faenadas o vendidas) por año
Variables de flujo de segundo orden (referidas al cambio ocurrido en un flujo en un período determinado, respecto al mismo flujo en un período anterior)	**Unidad de medida (dimensión $\mathbf{M/T^2}$)**
Aceleración	Kilómetros por hora por hora (km/h^2)
Incremento del producto bruto	Dólares anuales adicionales por año
Aumento de la natalidad	Nacimientos anuales adicionales por año
Aumento de la mortalidad	Defunciones anuales adicionales por año
Incremento de las ventas anuales	Ventas anuales adicionales por año

Las variables, tanto de stock como de flujo, pueden ser expresadas en forma **absoluta** o **relativa**. Las variables relativas más usuales son los **ratios** y las **tasas de variación**. Los ratios son cocientes entre dos magnitudes, como por ejemplo la población y la superficie de un país (densidad demográfica) o el PBI per cápita (producto bruto dividido por la población). El ratio de dos variables de estado **X** e **Y** tiene dimensionalidad $\mathbf{M_X/M_Y}$ (por ejemplo habitantes por kilómetro cuadrado). Un ratio entre una variable de flujo y una de estado tiene dimensionalidad $\mathbf{(M_X/M_Y)/T = M_X/M_YT}$ (por ejemplo, producto bruto del año *t* dividido por la **población promedio** del mismo año). Este último tipo de variable es también una variable de flujo que se mide en cantidad de unidades monetarias por unidad de tiempo (valor del PBI por habitante **durante un determinado año**).

Los ratios, aunque son relativos, son siempre expresados en valores absolutos con una cierta unidad de medida (por ejemplo kilómetros por hora, o PBI por habitante). Las **tasas de variación** no son valores absolutos sino **proporciones de cambio** de los valores entre dos períodos. Pueden referirse a variables de stock o de flujo. Las tasas de variación no tienen dimensión física **M** pero siguen teniendo la dimensión temporal, y su dimensionalidad es **1/T**: sus unidades físicas desaparecen cuando el flujo de expresa en forma de proporción o porcentaje. La **tasa de crecimiento** de la población (proporción de aumento anual de la población) elimina las unidades naturales **M** (número de personas) pero mantiene la dimensionalidad temporal ("por año"), por lo cual su dimensionalidad es **1/T**

(porcentaje de crecimiento **anual** en el número de habitantes).[6] En el caso de las variables de flujo de segundo orden la situación es similar. Si la variable original de por sí ya tenía dimensión **M/T**, como por ejemplo el producto bruto (producción **por año**), su variación anual tendrá dimensión **M/T^2** (aumento **por año** del producto **por año**, o más elegantemente, aumento anual del producto anual; su **tasa de crecimiento** tendrá dimensión **$1/T^2$ (porcentaje** de aumento **anual** del producto **anual).** La tasa de crecimiento del PBI tiene una dimensión temporal (es la tasa **anual**) pero ha perdido la conexión con las unidades intrínsecas respectivas. Ya no se refiere a las unidades monetarias en que se mide el producto bruto: es un puro porcentaje (aunque es un porcentaje **por año**).

En el caso de las variables categóricas en los estudios de panel, generalmente los **estados** se expresan en variables de stock, y miden la ubicación de los sujetos en las distintas categorías de las variables **en un momento dado**. En cambio los **eventos** (cambios de estado) se suelen medir como variables de flujo (eventos por unidad de tiempo). Por ejemplo, la condición ocupacional de las personas económicamente activas es una variable categórica con dos estados posibles (ocupado o desocupado) que constituye una variable de stock o de estado cuya información se refiere a un momento determinado; en cambio, una variable referida al **cambio** en la condición ocupacional durante el intervalo entre dos ondas del panel registra **eventos** (personas ocupadas que pasaron a estar desocupadas, o desocupados que pasaron a estar ocupadas, **durante ese intervalo**) y constituye por lo tanto una variable de flujo. Se puede expresar en una simple variable dicotómica (personas que cambiaron de estado o que no cambiaron), o puede incluir datos sobre el **número de eventos** ocurridos (cantidad de veces que la persona se quedó sin empleo, o que encontró un nuevo empleo, a lo largo del año).

Aparte del tipo de variables que se pretenda medir, los estudios longitudinales pueden diferir también en la **forma en que tratan la variable "tiempo"**. Ella puede ser considerada como una variable continua o como una variable discreta. En los paneles constituidos por encuestas periódicas se registran observaciones en momentos discretos, que luego son comparados entre sí; en general el número de períodos es pequeño, y la separación entre las rondas es considerable, de modo que el analista **trata el tiempo como una variable discreta** (si bien su concepto interpretativo subyacente acerca del tiempo puede concebirlo como una variable continua). Así, por ejemplo, puede comparar la situación ocupacional en el momento *t* con la situación ocupacional del momento *t+1*, y carece de datos sobre el período intermedio, pero conceptualmente el analista es conciente que los cambios de status ocupacional pueden ocurrir en cualquier momento, y que durante el período intermedio pudieron ocurrir uno o más cambios, que pueden haber ocurrido en cualquier momento intermedio dentro de ese período.

Cuando se trata, en cambio, de una gran cantidad de momentos muy cercanos entre sí, como ocurre por ejemplo en muchos datos de registro, el tiempo podría ser considerado como una variable continua. De hecho, el concepto matemático de una variable continua es usualmente explicado como una sucesión de datos discretos en que la longitud del intervalo tiende a cero. Implícitamente, el analista supone que entre una observación y la siguiente se ha producido una variación continua en las variables, a lo largo de un tiempo continuo; más aún, si el intervalo es suficientemente breve se descarta la posibilidad de que haya habido saltos o variaciones no observadas entre una observación y otra: se supone que entre una observación y otra ha habido un progreso suave o monótono.

Por ejemplo, si en una onda la persona tiene un gasto mensual de $1000 y en la siguiente tiene un gasto de $1100, la concepción continua supone que el gasto ha ido aumentando gradualmente entre esas dos fechas, de modo que en cualquier fecha intermedia se habría observado un gasto intermedio (por ejemplo $1040 o $1050). Por supuesto, ello no necesariamente ha sido así: tal vez entre una y otra fecha esa persona puede haber estado gastando mucho más (por ejemplo $5000), regresando a su nivel habitual antes de la segunda fecha. Quizá un poco después de la primera fecha

[6] La dimensión 1/T puede también expresarse como T^{-1}, y la dimensión $1/T^2$ como T^{-2}. Una discusión de la dimensionalidad de las variables económicas, que puede aplicarse a variables sociológicas sin mayores dificultades, puede hallarse en muchos textos sobre medición y sobre econometría, por ejemplo en Lange (1964).

esa persona recibió una pequeña herencia o un premio de lotería, que decidió gastar íntegramente, de modo que por un tiempo su nivel de gasto aumentó hasta cerca de $5000 por mes, regresando a alrededor de $1000 cuando se le acabó ese dinero llegado "del cielo". Pero si las fechas son más o menos cercanas entre sí la probabilidad de que ocurran esas cosas es muy pequeña, y por lo general esa posibilidad se descarta. Esto será importante cuando tratemos de insertar los datos de panel en un **modelo** de flujos continuos: normalmente supondremos un proceso intermedio sin altibajos u oscilaciones, por ejemplo sin premios de lotería, salvo que tengamos información específica.

Del mismo modo, la diferencia de estado que se observa entre dos fechas a menudo es interpretada como una **transición directa**, cuando en realidad podría haber habido una serie de **cambios intermedios no observados**. Así, por ejemplo, si la condición laboral es registrada una vez por año o por semestre siempre existe la posibilidad de que el sujeto haya sufrido cambios no registrados durante el intervalo intermedio, pero si las observaciones fuesen mensuales esa posibilidad es mucho menor (poca gente queda desempleada más de una vez en el curso de un mismo mes), y si la frecuencia de las observaciones fuese semanal la posibilidad de cambios de condición laboral no registrados prácticamente desaparece. Si un sujeto estaba ocupado en la primera fecha y también ocupado en la segunda, es **posible** que en el período intermedio haya estado temporariamente desocupado, pero ello no es muy **probable**: si el período es breve esa posibilidad es muy remota y podría ser descartada.

Estas reflexiones indican que hay que distinguir entre cuatro aspectos. En primer lugar el carácter discreto o continuo **de las observaciones** (que por lo general son discretas pero si son muy frecuentes podrían considerarse como continuas); en segundo lugar el carácter discreto o continuo **de las variables** observadas (cuyo nivel de medición puede ser nominal u ordinal, es decir discreto; o de intervalo, es decir continuo); en tercer lugar la **operacionalización del tiempo** como variable discreta o continua; y en cuarto lugar la **conceptualización del proceso de cambio** como un proceso que ocurre en forma continua o a través de saltos discretos.

Si se tiene una gran cantidad de observaciones sucesivas tomadas con intervalos muy breves, esa sucesión de observaciones podría considerarse como una buena aproximación a una medición continua (bajo la hipótesis de que el intervalo es tan breve que no puede haber ningún cambio intermedio no captado por la serie de observaciones efectuadas). Los cambios de estado entre dos observaciones observados en las variables categóricas pueden modelizarse alternativamente como un proceso continuo, donde diferentes individuos cambian de estado, quizá repetidas veces, en diferentes momentos a lo largo de ese intervalo, o bien como un cambio simultáneo de todos ellos entre dos momentos discontinuos del tiempo.

Cuando se tienen pocas observaciones tomadas a intervalos considerables, los instrumentos de análisis se suelen basar en el análisis de variaciones discretas, aunque las variables de observación sean intrínsecamente variables de intervalo; en ese caso se supone que los sujetos "saltan" de un valor inicial a un valor final, sin considerar su posible paso por valores intermedios, ni tomar en cuenta el estado o valor en que se encontrarían en el período intermedio. En cambio, cuando hay muchas observaciones con intervalos breves entre sí, resulta posible aplicar instrumentos analíticos que suponen considerar los procesos de cambio (y el tiempo mismo) como variables continuas. Pero también es posible suponer un proceso continuo (no observado) aun cuando se disponga solamente de algunas observaciones discretas.

Estas situaciones no son necesariamente incompatibles entre sí, aunque cada análisis concreto debe caer en alguna de ellas. Por ejemplo, en el mismo conjunto de datos puede haber variables categóricas con estados discretos y al mismo tiempo variables continuas, y una sucesión de muchos paneles semestrales podría considerarse como una observación "continua", al menos en aquellas variables que varían sólo gradualmente y en las que no se esperan fluctuaciones muy grandes durante cada período. Por ejemplo, un econometrista puede considerar una serie de datos trimestrales (de producción, de oferta monetaria, de precios, etc.) como variables continuas, y aplicar en consecuencia métodos de análisis que así lo suponen, como la regresión; pero si sólo tiene esas variables medidas en dos o tres períodos difícilmente pueda aplicar esos enfoques (entre

otras cosas porque tendría muy pocas observaciones y los resultados de la regresión no serían estadísticamente confiables). Además hay variables que registran fluctuaciones dentro de cada intervalo, por lo cual su estado en el momento de la observación podría no ser suficiente para reconstruir el movimiento continuo de la variable. Por ejemplo, la población total de un país generalmente cambia de manera bastante regular, de modo que dos censos espaciados diez años permiten interpolar la población de los años intermedios con poco margen de error; en cambio, la desocupación pueden variar bastante de un mes al otro, de modo que si el dato existe en encuestas separadas por intervalos semestrales o anuales, no se puede aseverar que no haya habido cambios de situación ocupacional no registrados, ocurridos durante el intervalo entre las rondas.

1.10. Estados y eventos en paneles de datos categóricos

La situación más simple para el análisis de panel es aquella en que se tienen **variables de tipo categórico**, que se observan en **momentos discretos del tiempo**. En esta situación, las variables representan **estados** que corresponden por lo general a un momento en el tiempo que usualmente es el mismo momento de la observación, aunque a veces corresponden a **eventos** ocurridos en algún momento del período precedente, o incluso a **procesos** desarrollados a lo largo de dicho período. La comparación de los estados de los individuos en dos o más momentos en el tiempo permite observar los **cambios** ocurridos en el estado de cada individuo, es decir, el pasaje de los individuos de un estado inicial a un estado final (que puede ser el mismo estado del principio).

Es conveniente definir aquí con mayor precisión los conceptos de **estados, eventos**, y **procesos**, que hemos venido usando de manera intuitiva. Los estados son categorías de una variable cualitativa (o valores de una variable continua) en que puede resultar clasificada cada unidad de análisis **en un momento determinado**. Ese estado puede ser observable o inobservable. Por el momento suponemos que el estado "interno", inobservable o latente del sujeto está unívocamente asociado a una determinada categoría manifiesta de la variable, es decir, a una determinada respuesta observable, aunque más adelante introduciremos el concepto de **incertidumbre de respuesta** con el cual se admite que una misma respuesta manifiesta puede corresponder a varios estados latentes, y viceversa, un mismo estado latente puede dar origen a diferentes respuestas manifiestas.

Los **eventos** no son otra cosa que los **cambios de estado** (manifiestos o latentes) de los sujetos. Por ejemplo, si los estados son dos, "ocupado" y "desocupado", un evento sería el paso de una situación a otra (encontrar empleo, o quedar sin trabajo). Si el estado es "tener 30 años de edad" los eventos relevantes podría ser "cumplir 30 años" y "cumplir 31 años". El primero de estos eventos constituye la "entrada" en la situación de tener 30 años, y el segundo corresponde a la "salida" o "egreso" (otro posible "egreso" sería morir antes de cumplir 31).

Los **procesos**, también llamados **trayectorias**, son **secuencias de eventos** a lo largo de un período de tiempo. Cuando se registra el estado del sujeto en dos rondas del panel, la **secuencia manifiesta o aparente** (o cambio neto) es simplemente el registro de un pasaje desde su estado en la primera ronda a su estado en la segunda, pero si el proceso subyacente es un proceso que opera en plazos más breves, o es un proceso continuo, podría haber una **secuencia no registrada de eventos intermedios**. Por ejemplo, si el sujeto estaba ocupado en ambas rondas, podría haber tenido de todas maneras algún período no registrado de desocupación en el lapso intermedio.

El estado de un sujeto en un momento determinado puede referirse a su situación "actual" (tener empleo o no tenerlo en el momento de la entrevista), o bien puede incluir información referida al pasado inmediato. Por ejemplo, un sujeto puede encontrarse hoy en el estado de "haber tenido empleo continuadamente durante los últimos siete días", y esa información no se refiere solamente a hoy sino a los siete días anteriores. Sin embargo, aun ese caso la "fecha de referencia" sigue siendo hoy, ya que se estipula un período (últimos siete días) que corresponde a "hoy"; si la encuesta se realizara "mañana", la respuesta se referiría a otro conjunto de siete días. A veces la variable se cuantifica en función de un período anterior no necesariamente conectado con el "hoy", como por ejemplo "Número de horas semanales trabajadas durante la última semana en que estuvo empleado", o bien "Número de horas semanales trabajadas durante la semana del 20 al 27 de julio".

Variables tan inocentes como la edad o la antigüedad en el empleo pueden ocultar complicaciones de este tipo. La edad (en años) significa: "Cantidad de años completos en que el sujeto ha estado vivo desde su nacimiento hasta el presente". La antigüedad en el empleo significa "tiempo en que el sujeto ha estado empleado en este mismo empleo". Estas variables no reflejan un estado actual en sentido estricto, sino que reflejan un proceso continuado (consistente en estar vivo, o en estar empleado en determinado puesto de trabajo) durante cierto tiempo. Sin embargo, aun en esos casos la fecha de referencia es "hoy", ya que mañana o ayer el individuo podría haber tenido otra edad (si es que su cumpleaños ocurre en los días adyacentes a la fecha de la encuesta), y del mismo modo podría cambiar el número de años (o meses) de antigüedad en el empleo.

Las variables pueden tener una **fecha (o período) de referencia** no necesariamente igual a la **fecha de observación** o **fecha de registro**. Por ejemplo en un Censo Agropecuario realizado en agosto o septiembre se puede preguntar sobre la cantidad de cabezas de ganado que cada unidad productiva poseía al 30 de junio último, fecha que no coincidirá con la fecha en que el agricultor responde las preguntas del Censo. Además, en muchas encuestas o censos el período de observación no es una fecha fija, sino que las entrevistas se extienden a lo largo de un período de relevamiento, que por ejemplo puede abarcar todo un mes, de modo que las preguntas referidas a "hoy" o a "los últimos siete días" pueden en realidad corresponder a fechas o períodos diferentes para cada sujeto.

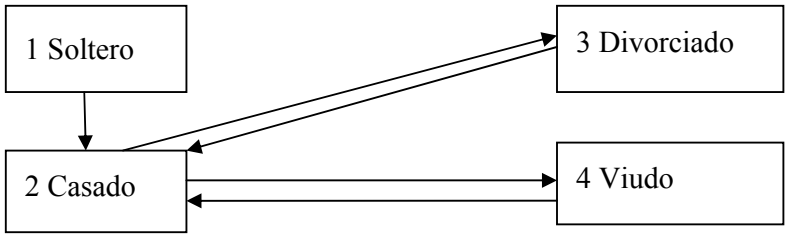
La distinción entre estados y eventos puede ilustrarse mediante la variable Estado Civil, una tipica variable categórica que describe el **estado actual** de los sujetos en un momento dado. Si nos atenemos al estado civil **legal**, se consideran en principio los siguientes **estados**:

1. Soltero
2. Casado
3. Divorciado
4. Viudo

Estos son todos los estados civiles, al menos los reconocidos legalmente; sin embargo, aun prescindiendo de los estados conyugales "informales" (conviviente o unido), esas categorías no son todos los estados posibles. Si se parte en el momento inicial con sujetos situados en alguno de estos cuatro estados civiles, en el segundo momento algunos podrían estar, por ejemplo, **muertos**, o "**No vivos**". Eso añadiría un quinto "estado", que consiste en "no estar vivo". También podría haber sujetos que en la primera ronda aún no habían nacido, pero que en la segunda ronda aparecen, probablemente como "solteros". Habrían salido del estado "**No vivo**" ingresando al estado "**Soltero**".[7] Esta complicación no es significativa para los presentes propósitos, por lo cual se deja de lado por el momento en el ejemplo de los estados civiles. También se dejan de lado por el momento los individuos "perdidos", "desgranados" o "desertores", que no reaparecen la segunda vez por haber cambiado de domicilio, por rechazar la segunda entrevista o por cualquier otra razón, así como los de "aparición tardía", que no fueron encuestados la primera vez pero aparecen con información en la segunda ronda. Para el ejemplo se presume que los individuos fueron entrevistados en las dos ocasiones. En caso contrario habría otro posible estado: "Ignorado (por falta de datos)". Si se dejan de lado los nacimientos, muertes y desgranamientos, los **eventos básicos posibles** serían los siguientes:

Casamiento de un soltero	**Pasaje del estado 1 al estado 2**
Casamiento de un divorciado	**Pasaje del estado 3 al estado 2**
Casamiento de un viudo	**Pasaje del estado 4 al estado 2**
Disolución del vínculo por divorcio	**Pasaje del estado 2 al estado 3**
Disolución del vínculo por muerte del cónyuge	**Pasaje del estado 2 al estado 4**

[7] Para ser estrictos, los niños aun no nacidos en la primera ronda podrían corresponder a embarazos ya comenzados, de modo que en rigor ya estaban vivos; pero aquí se usa "vivo" en el sentido de "nacido vivo", o "vivo como un individuo autónomo", excluyendo la condición fetal.



Si se incluyera el estado "No vivo" y el estado "Fuera del panel", habría otros estados posibles: "Nacer", "Morir", "Aparecer" y "Desaparecer", que no son contemplados aquí para no complicar el esquema, pero la lógica de su eventual introducción no es difícil de entender.

Es fácil advertir, de todos modos, que la lista no es "completa". No hay flechas entre todos los recuadros, por ejemplo no hay ninguna entre 1 y 3, o entre 4 y 3, o entre 3 y 1. Ningún divorciado o soltero puede pasar a ser viudo, ningún viudo o casado puede pasar a ser soltero.

Estas exclusiones obedecen a dos clases de razones. Algunos cambios de estado son directamente **imposibles** (por ejemplo nadie vuelve a ser soltero); mientras que otros son posibles pero sólo si se admite la existencia de **eventos intermedios** no registrados. Es posible que una persona sea soltera la primera vez y viuda en la segunda, pero ello implica que **en el período intermedio** se casó y luego enviudó. Nadie puede pasar de soltero a viudo, a menos que en el interin se haya casado.

Los cinco eventos mencionados precedentemente son en realidad los únicos **eventos básicos** que pueden ocurrir con la variable Estado Civil, pero entre dos rondas del panel puede haber algunos cambios de estado **intermedios**, no observados o no registrados. La secuencia observada puede incluir **uno o más eventos básicos** no observados, ocurridos en el período intermedio. Alguien puede pasar de estar soltero en una ronda a estar viudo en la ronda siguiente, siempre que se haya casado en el período intermedio y haya enviudado poco después. La secuencia observada soltero-viudo implica en realidad dos cambios intermedios: de soltero a casado, y de casado a viudo. De este modo, algunos eventos aparentes son el **resultado neto** de dos o más eventos intermedios.

En primer lugar, entonces, hay algunos cambios intrínsecamente imposibles. En este ejemplo, la categoría "soltero" se refiere al estado civil **inicial** de las personas, que se adquiere al nacer y se pierde **para siempre** al casarse.[8] Nadie puede pasar de casado a soltero, o de viudo a soltero, o de divorciado a soltero, ya que "soltero" es un estado civil del cual se puede salir pero en el cual no se puede ingresar a partir de otros estados. Todos los flujos desde otros estados civiles al estado de soltería son entonces, por definición, inexistentes e imposibles.

Otros cambios de estado civil entre dos diferentes momentos del tiempo son imposibles como cambio directo, y sólo pueden existir si ha habido un evento intermedio (o más de un evento intermedio). Por ejemplo, para pasar de ser soltero en la fecha **t** a estar divorciado en la fecha ***t+k*** se requiere que a lo largo de ese intervalo el sujeto primero haya pasado de soltero a casado, y luego de casado a divorciado. Pasar de soltero a divorciado no es un evento **básico** sino un **cambio neto** observable al cabo de un tiempo como resultado de **dos o más eventos básicos ocurridos en el período intermedio**.

Eventos conceptualmente imposibles	Pasaje de casado a soltero, de viudo a soltero, o de divorciado a soltero.
Eventos imposibles en forma directa, pero posibles con eventos intermedios	Pasaje de soltero a viudo, o de soltero a divorciado, o de divorciado a viudo

[8] Aquí se usa el concepto de soltero como equivalente a "nunca casado", como es usual en castellano, equivalente al concepto inglés *bachelor* (que cada vez se usa menos en inglés contemporáneo). En inglés la palabra *single* significa "sin vínculo actual de pareja" (y se dice *single* o bien *unmarried*), y designa tanto a solteros como a divorciados o viudos, lo cual no es equivalente a "nunca casado" (*bachelor* o *never married*). No hay una palabra castellana equivalente a *single*, excepto expresiones más complejas como "actualmente sin pareja".

Si la observación es muy frecuente o se hace mediante registro continuo no existe posibilidad práctica de observar cambios de soltero a divorciado, o de divorciado a viudo, entre dos observaciones sucesivas, pero si se trata de intervalos más largos esos casos pueden aparecer. Habría en esos casos uno o más **eventos intermedios no registrados**. El proceso completo, que podría ser capturado por una secuencia de observaciones más frecuentes, incluiría ese evento intermedio (soltero→**casado**→divorciado, o bien divorciado→**casado**→viudo). Si el período es suficientemente largo podría haber incluso más de uno o dos eventos intermedios: el cambio neto de soltero a divorciado podría reflejar una secuencia no observada más compleja, como la siguiente: soltero→**casado**→**viudo**→**casado**→divorciado.

Es importante distinguir, entonces, entre cambios intrínsecamente imposibles y cambios netos **aparentemente** imposibles, pero que se pueden explicar por la existencia de **fases intermedias no observadas**. Del mismo modo hay que distinguir entre la **secuencia neta** registrada por el panel (estado en ***t*** y estado en ***t+h***) de la **secuencia completa** o **trayectoria** que incluiría todos los eventos intermedios entre esas dos fechas. Si se carece de datos sobre el período intermedio, no hay manera de saber cuál fue la trayectoria seguida por un determinado individuo, pero mediante la aplicación de modelos que postulan un **proceso continuo subyacente** que genera los datos observados, a veces es posible incluso **estimar la frecuencia** de determinadas secuencias completas o trayectorias a partir solamente de las secuencias netas, como se verá más adelante.

Nótese que una persona observada en el mismo estado en dos ocasiones sucesivas podría ser una persona que persistió en el mismo estado inicial (por ejemplo casado→casado) o bien que haya pasado por una etapa intermedia (casado→viudo→casado, o bien casado→divorciado→casado). Si solo se conoce el estado existente en cada observación, sin conocer las transiciones intermedias, estas dos trayectorias no podrían ser distinguidas una de la otra.

2. Análisis descriptivo de panel

Es conveniente distinguir entre el simple uso de tabulaciones cruzadas de las variables registradas en diferentes períodos, destinado a **describir** los cambios experiementados por los sujetos, y la aplicación de modelos teóricos como por ejemplo los modelos de Markov, que se introducen para **explicar** los datos de panel. La tabulación es sólo un **instrumento descriptivo** que permite observar cómo han cambiado los sujetos (y la población en su conjunto) entre un período y otro. Los modelos, en cambio, postulan un cierto tipo de **proceso subyacente** (no observable) que generaría o explicaría los datos observados. En esta sección se examinan métodos para el análisis descriptivo de los datos de panel sin imponer todavía ningún modelo explicativo, y se introducen algunos términos técnicos de uso frecuente en este contexto.

2.1. La tabla de rotación

2.1.1. Características generales

El instrumento fundamental del análisis de panel con datos discretos es la **tabla de rotación** (*turnover table*)**,** que también puede ser denominada **tabla intertemporal univariada**, en la cual se tabulan las frecuencias cruzadas de **la misma variable observada en dos períodos o fechas diferentes**. Cada una de estas observación se suele llamar una "onda" o bien una "ronda". En este texto usaremos indistintamente ambos vocablos, o bien otros equivalentes como "observación". Supongamos por ejemplo una variable **X** con sólo dos valores posibles. Los subíndices indican estados, los superíndices denotan fechas o períodos.[9]

---

[9] Algunos autores usan otras convenciones en su notación. Por ejemplo en el caso de variables dicotómicas se pueden usar subíndices para designar los sucesivos momentos de observación (t=1 y t=2), y usar una barra horizontal encima de esos valores para denotar los estados de la variable. Así, los contingentes en los diferentes flujos serían N_{12}, $N_{1\bar{2}}$, $N_{\bar{2}1}$ y $N_{\bar{1}\bar{2}}$. El flujo $N_{1\bar{2}}$, por ejemplo, representaría la cantidad de personas que tenían un cierto valor de la variable X en el primer período y el otro valor en el segundo período. Nótese que en nuestra notación los subíndices se refieren a los **valores de la variable** (X=1 y X=2, o bien X=0 y X=1), mientras que aquí los mismos subíndices se refieren a los

Tabla de rotación referida a la variable X en dos rondas del panel			
Primera ronda	**Segunda ronda**		**Total**
	X=1	**X=2**	
X=1	N_{11}	N_{12}	N_1^t
X=2	N_{21}	N_{22}	N_2^t
Total	N_1^{t+1}	N_2^{t+1}	N

Según la nomenclatura ejemplificada en esta tabla, *N* designa una cantidad de individuos o casos. En las celdillas interiores de la tabla, donde las cantidades son del tipo N_{ij}, el primer subíndice (***i***) se refiere al valor de la variable en la primera observación (en este caso los valores pueden ser 1 o 2), y el segundo subíndice (***j***) al valor de la misma variable en la segunda observación. De ese modo, N_{12} es la cantidad de individuos con valor 1 en la primera observación y valor 2 en la segunda. En las **frecuencias marginales** (la fila y columna de totales) aparece la cantidad de personas en cada uno de los estados, en cada una de las rondas. Se usan los superíndices ***t*** y ***t+1*** para indicar la primera o segunda ronda. Así, N_1^t representa todos los sujetos que tuvieron valor 1 en la primera oportunidad, independientemente del valor que hayan registrado en la segunda, mientras N_1^{t+1} es el número de sujetos con valor 1 en la segunda ronda, independientemente de su estado en la primera. La cifra *N* en la celdilla inferior derecha es el número total de participantes del panel.

Por simplicidad, salvo que sea necesario, se omite la referencia temporal en las cantidades que indican flujos, es decir en las celdillas interiores de la tabla. Si se desea incluir esa referencia, la cantidad N_{ij} se debería indicar como $N_{ij}^{t,t+1}$. Por ejemplo si se trata del flujo desde el estado 2 al estado 1, en el período que va de t=1 a t=2, el flujo N_{21} se denotaría como N_{21}^{12}. Es obvio que las celdillas interiores, que representan **flujos de transición** entre el primer y segundo momento de observación, son **cantidades por período** (flujos) mientras las celdillas marginales son **cantidades instantáneas** (stocks). En otras palabras, las celdillas marginales representan cantidades existentes **en un momento dado**, mientras las celdillas interiores representan sujetos que se movieron de un estado inicial a un estado final **durante un determinado período de tiempo**. Dado que no se registran eventos intermedios, esos flujos son **cambios netos** de estado sufridos por los individuos entre la primera y la segunda observación.

2.1.2. Estabilidad e inestabilidad en la tabla de rotación

Como resultado de los flujos de transición ocurridos entre las dos observaciones, la distribución marginal final podría ser diferente a la distribución marginal inicial; asimismo, diversos individuos pueden acabar en un estado diferente al que ocupaban al inicio. En este punto puede resultar conveniente distinguir entre la estabilidad o cambio **de los individuos** por un lado, y **de la población en su conjunto** por el otro. Si ningún individuo cambia de estado, obviamente tampoco cambia la distribución agregada. Por ejemplo:

Estabilidad individual y agregada			
Primera ronda	**Segunda ronda**		**Total**
	X=1	**X=2**	
X=1	100		100
X=2		300	300
Total	100	300	400

períodos (t=1 y t=2) y en cambio los valores de la variable son denotados por la **presencia o ausencia de la barra horizontal**. Esta notación con barras fue introducida hace muchos años por Paul F. Lazarsfeld en varias de sus obras sobre atributos dicotómicos, como por ejemplo Lazarsfeld 1961, 1965, y 1968. Podría verse también Maletta 1970 para una introducción general al enfoque de Lazarsfeld. Esta notación con barras, a diferencia de la usada en este trabajo, no es extensible en forma directa al caso de variables con más de dos categorías, aunque con algunas modificaciones ello también puede lograrse.

Los 100 individuos que estaban en el estado 1 siguieron en ese estado, y del mismo modo los otros 300 sujetos permanecieron en el estado 2. Nadie cambió de estado, y por lo tanto la distribución marginal de sujetos también siguió siendo la misma: 25% en el estado 1, y 75% en el estado 2. Esa situación de **estabilidad individual** (que implica también la **estabilidad agregada**) en la práctica no es muy común. Lo más factible es que algunos individuos cambien de estado entre una observación y otra. Estos movimientos podrían implicar o no un cambio en la distribución agregada de la variable. Por ejemplo en la siguiente tabla existen cambios individuales pero ellos **se compensan mutuamente**, de modo que se mantiene la estabilidad agregada.

Estabilidad agregada con cambios de estado a nivel individual			
Primera ronda	**Segunda ronda**		**Total**
	X=1	**X=2**	
X=1	80	20	100
X=2	20	280	300
Total	100	300	400

Hay veinte sujetos que cambian del estado 1 al estado 2, y otros veinte que sufren el cambio opuesto. Como resultado, la distribución agregada no cambia, a pesar de que hay 40 sujetos (un 10% del total) que han cambiado de estado. El **flujo de entrada** en el estado 1 (veinte sujetos) es igual al **flujo de salida** del mismo estado 1 (otros veinte sujetos), y lo mismo ocurre con el estado 2. Esta situación requiere **que para cada estado el flujo de entrada sea igual al flujo de salida**, pero no depende de la magnitud misma de esos flujos. En este caso los sujetos "móviles" representan el 10% del total de sujetos, pero podrían representar cualquier otra proporción. De hecho pueden darse situaciones en que la mayor parte o incluso la totalidad de los sujetos cambie de estado sin que por ello se altere la distribución agregada. Considérese por ejemplo el siguiente ejemplo:

Estabilidad agregada aunque todos los individuos cambian de estado			
	Segunda ronda		**Total**
Primera ronda	**X=1**	**X=2**	
X=1		200	200
X=2	200		200
Total	200	200	400

Aquí **todos** los individuos han cambiado de estado sin que la distribución agregada se haya modificado en lo más mínimo. Esta situación no es totalmente imaginaria: piénsese por ejemplo en una empresa con 400 trabajadores que los distribuye en contingentes iguales para trabajar respectivamente en horario diurno y nocturno. Cada trabajador durante una semana trabaja en horario diurno (estado 1) y a la semana siguiente en horario nocturno (estado 2). Siempre hay 200 trabajadores en cada turno, pero los trabajadores diurnos de la primera semana son los que trabajan de noche en la segunda semana, y viceversa. Si los 400 trabajadores declararan en qué turno trabajan en dos semanas sucesivas la tabla sería como la que antecede. Estabilidad agregada, pero cambio total a nivel individual.

Esta situación de **cambios individuales compensados**, en que la distribución agregada se mantiene invariable, no siempre ocurre en la práctica. La distribución de la variable puede cambiar de un período a otro, y de hecho esa es la situación más frecuente. En esos casos, los cambios ocurridos a los individuos no están compensados entre sí. En consecuencia, la distribución agregada termina siendo diferente. Por ejemplo:

Cambios individuales no compensados, resultando en inestabilidad agregada			
Primera ronda	**Segunda ronda**		**Total**
	X=1	**X=2**	
X=1	50	50	100
X=2	20	280	300
Total	70	330	400

En este caso, 50 sujetos pasan del estado 1 al 2, pero sólo 20 pasan del estado 2 al estado 1, de modo que la distribución agregada se modifica: de una distribución inicial de 100 individuos en el estado 1 y 300 en el estado 2 se llega a una distribución final con sólo 70 sujetos en el estado 1 y 330 en el estado 2.

Como se vio antes, para que exista **estabilidad agregada** se requiere que los flujos de **entrada** y de **salida** de cada estado, en este caso los flujos N_{12} y N_{21}, sean **de igual magnitud**. Esos flujos son N_{12}=0 y N_{21}=0 en la primera de estas tablas, N_{12}=20 y N_{21}=20 en la segunda tabla, y N_{12}=200 y N_{21}=200 en la tercera, involucrando respectivamente un total de cero, cuarenta y cuatrocientos sujetos moviéndose en ambas direcciones. En la cuarta tabla los flujos son desiguales: 50 en una dirección y 20 en la dirección opuesta. En otras palabras, la condición necesaria y suficiente para que haya estabilidad agregada en el caso de una **variable dicotómica** (con sólo dos estados posibles) es:

$$N_{12} = N_{21}$$

Esta condición se complica un poco cuando hay más de dos estados. Para que se mantenga la distribución marginal no es necesario que **cada uno** de los flujos en un sentido se compense con un flujo igual de sentido contrario. Sólo es necesario que la **suma** de los **flujos de salida** de los varios estados se compense con la **suma** de los **flujos de entrada** en los distintos estados. En símbolos, habrá estabilidad agregada si se cumple la siguiente condición:

$$\sum_j N_{ij} = \sum_j N_{ji} \quad (i \neq j) \text{ para todo estado } i$$

Por ejemplo, si hay tres estados que corresponden a las situaciones laborales de la fuerza laboral (inactivos, ocupados y desocupados), para que haya estabilidad agregada se requiere que el número de personas que deja de ser inactivo (para pasar a ser ocupado o desocupado) iguale al número de activos (ocupados y desocupados) que pasan a ser inactivos, y que ocurra lo mismo para los otros estados (el número de personas que dejan de ser desocupados debe igualar al número de personas que entran en la desocupación, y el número de personas que encuentra trabajo debe igualar al número de personas que deja de tener trabajo). No es necesario que cada flujo esté compensado. Por ejemplo no es necesario que el flujo, digamos, de ocupado a inactivo se compense específicamente con el flujo de inactivo a ocupado, ya que el número de inactivos y de ocupados también está afectado por lo que suceda con los desocupados, y su magnitud puede conservarse constante por la combinación de flujos de entrada y salida, aunque cada flujo separadamente no esté exactamente compensado por uno de sentido opuesto. Basta con que el total de los flujos de salida y el total de los flujos de entrada estén compensados en todos los estados.

2.1.3. Variables exhaustivas y estabilidad agregada

Para que haya estabilidad agregada se requiere que los estados sean **exhaustivos**, es decir, que cubran todas las posibilidades. Piénsese por ejemplo en un panel con la variable "estado civil". En cada período han cambios de estado civil de distinto tipo, pero hay una restricción: nadie puede "convertirse en soltero". Ahora bien, si en cada período una cierta cantidad de solteros se convierte en casado, ¿cómo se reponen los solteros? En el caso del estado civil tal como ha sido reflejado en la tabla anterior es imposible que haya estabilidad agregada porque uno de los estados ("soltero") es un estado "originario" sin ninguna "reposición"; tiene flujos de salida pero no tiene flujos de entrada: por definición ningún casado, divorciado o viudo puede convertirse en soltero.

Para que hubiese estabilidad habría que convertir la variable en exhaustiva introduciendo **dos estados adicionales**. Si el estado civil califica la población de todas las edades, esos estados adicionales serían "estar vivo" y "no estar vivo", y además de los eventos ya mencionados se añadirían otros dos eventos relevantes: "nacer" (pasar a integrar la población compuesta por los que "están vivos") y "morir" (abandonar esa población pasando a "no estar vivo"). En ese caso, el número de solteros puede mantenerse estable si la cantidad de personas que nacen (todas ellas necesariamente solteras) es igual al número de solteros que se casan o mueren en el período. Si la tabla considerara solamente la población a partir de cierta edad, por ejemplo desde los 15 años, el número

de "solteros de 15 y más años" no se incrementaría mediante nacimientos, sino cuando los sujetos de 14 años cumplan 15 y accedan así a la población considerada. La cantidad de solteros de 15 años o más permanecería constante si el número de solteros que se casa o fallece equivaliese al número de personas que cumplieron 15 años de edad en el período intermedio y permanecieron solteras hasta la segunda observación.

Estas consideraciones muestran que para un análisis realmente sistemático de los cambios de estado es menester definir la variable de manera **exhaustiva,** de tal modo que se cubran todos los estados posibles en ambos momentos de observación. No estar vivo o no ser encuestado podrían convertirse así en "estados" legítimos, aparte de los estados ya reconocidos como soltero, casado, divorciado o viudo (que obviamente se aplican solamente a las personas vivas que hayan respondido a la encuesta en una observación determinada). Del mismo modo, además de eventos como casarse, divorciarse o enviudar deben reconocerse eventos adicionales, por ejemplo: Nacer como soltero, Morir como soltero, Morir como casado, Morir como divorciado, Morir como viudo. Aparte de no estar vivo, otro posible estado sería "No entrevistado", que se comporta más o menos en forma similar. La siguiente tabla representa la evolución del estado civil considerando nacimientos, defunciones, y también no-respuestas.

	Segunda ronda						**Total**
Primera ronda	**Solt.**	**Casado**	**Div.**	**Viudo**	**No vivo**	**Ausente, ignorado**	
Soltero	N_{ss}	N_{sc}	N_{sd}	N_{sv}	N_{sz}	N_{sa}	N_s^t
Casado	--	N_{cc}	N_{cd}	N_{cv}	N_{cz}	N_{ca}	N_c^t
Divorciado	--	N_{dc}	N_{dd}	N_{dv}	N_{dz}	N_{da}	N_d^t
Viudo	--	N_{vc}	N_{dv}	N_{vv}	N_{vz}	N_{va}	N_v^t
No vivo	N_{zs}	--	--	--	--	--	N_z^t
Ausente, ignorado	N_{as}	N_{ac}	N_{ad}	N_{av}	--	--	N_a^t
Total	N_s^{t+1}	N_c^{t+1}	N_d^{t+1}	N_v^{t+1}	N_z^{t+1}	N_a^{t+1}	N

El estado N_s^t corresponde a "No vivo en el momento de la primera ronda", y el estado N_a^t corresponde a "Ausente o con estado civil ignorado en el momento de la primera ronda". La celdilla N_{zs} en la fila "No vivo" incluye personas solteras en la segunda ronda que nacieron durante el período intermedio (por definición, las personas nacen solteras, y se supone aquí por simplicidad que no se pueden casar entre su nacimiento y la próxima ronda de la encuesta). Las celdillas N_{sz}, N_{cz}, N_{dz} y N_{vz} en la columna "No vivo" corresponden a personas de todos los estados civiles que han fallecido antes de la segunda ronda. Las celdillas N_{sa}, N_{ca}, N_{da} y N_{va} contienen personas de los cuatro estados civiles que sobreviven hasta la segunda ronda pero no son encuestadas por segunda vez por razones cualesquiera (ausencia, o simple negativa a responder) o cuyo estado civil en la segunda ronda ha quedado sin registrar.

Se han indicado con guiones las celdillas imposibles o necesariamente vacías: nadie puede nacer en otro estado civil sino soltero, nadie puede pasar a ser soltero a partir de otro estado civil, y nadie puede figurar en la base de datos si no ha sido encuestado al menos una vez.[10] Esta tabla permite que haya "**desertores**", es decir, personas con su estado civil registrado en la primera ronda, pero ausentes o con estado civil ignorado en la segunda oportunidad, como en las celdillas N_{sa}, N_{ca}, N_{da} y N_{va}; y también "**entradas tardías**", o sea personas que no fueron encuestadas o no declararon su estado civil en la primera ronda, pero sí lo hicieron en la segunda, como en las celdillas N_{as}, N_{ac}, N_{ad} y

[10] Se podría incluir también el flujo de personas que en la primera ronda aún no habían nacido y que en el momento de la segunda ronda ya habían fallecido. Esto corresponde a niños nacidos en el período intermedio que hayan fallecido antes de la segunda ronda. Esas personas nunca llegaron a ser encuestadas. Si el panel, como suponemos por simplicidad, sólo registra el estado inicial y el estado final, estos casos serán en efecto omitidos, pero puede haber encuestas que registren retrospectivamente los nacimientos y fallecimientos ocurridos entre las dos rondas, como se suele hacer en las encuestas demográficas, y en ese caso el flujo de "No vivo" a "No vivo" incluiría niños que nacieron y murieron entre las dos rondas.

N_{av}. También aparecen separadamente los "**fallecidos**" y los "**nacidos**". Esta tabla es ciertamente "exhaustiva", y por lo tanto es capaz de producir situaciones de estabilidad agregada, si la suma de los flujos de entrada en cada estado está compensada por la suma de los respectivos flujos de salida.

Sin embargo, la inclusión de estos estados adicionales (que incluyen personas "fuera de la población") plantea algunos problemas. El principal de ellos es que la población incluida en la tabla es necesariamente superior a la población inicial y también superior a la población final. Por ejemplo, supóngase que en la primera oportunidad fueron encuestadas 1000 personas, y en la segunda se volvió a encuestar a 950 de esas 1000 personas (el resto falleció o no fue encuestado), pero entretanto han nacido 50 bebés. La tabla dará un total de 1050 "casos" en ambas rondas, ya que se están registrando eventos que afectan a las 1000 personas entrevistadas en la primera ronda (algunas de las cuales mueren antes de la segunda o no son re-encuestadas por alguna otra razón) y además se incluyen los 50 niños nacidos entre la primera y la segunda ronda. En esta clase de situaciones la población total considerada no coincide con la cantidad total de personas vivas y presentes en ninguna de las dos rondas. En ambas rondas hubo mil personas encuestadas, pero las distribuciones marginales se referirán a 1050 personas: en la primera ronda se contabilizan 1000 entrevistados y 50 niños aún no nacidos; en la segunda ronda se incluyen 950 entrevistados por segunda vez, 50 no entrevistados en la segunda ronda por muerte, ausencia o no respuesta; y 50 recién nacidos entrevistados por primera vez en la segunda ronda.

Lo mismo sucedería con las personas que no fueron encuestadas en la primera ronda por encontrarse ausentes o por cualquier otro motivo, pero que sí fueron encuestadas la segunda vez, en la cual aparecieron en cualquiera de los estados civiles. La población incluida en la tabla podría definirse como la suma de las personas respondentes en la primera ronda, más las personas nacidas entre la primera y segunda ronda que fueron contabilizadas en la segunda, más las personas omitidas en la primera ronda que fueron incluidas en la segunda.

Asimismo, por razones de coherencia y de representatividad, si se incluyen los desertores y las entradas tardías, habría que incluir también aquellas personas que debieron figurar en el panel pero estuvieron ausentes o con datos ignorados en ambas rondas, o estuvieron ausentes o ignorados en la primera y muertos en la segunda. Sin embargo, usualmente esas personas no generan ningún registro y son directamente ignoradas.

La existencia de desertores, ingresos tardíos, y casos directamente omitidos en ambas rondas plantea problemas respecto a la representatividad del panel. Si los desertores fuesen una muestra al azar de la población general de sujetos que ingresó en la primera ronda y debió permanecer hasta la segunda, y si los tardíos fuesen asimismo representativos del total de personas que debió entrar en la primera ronda (incluyendo los que nunca entraron ni en la primera ni en la segunda), entonces el problema no tendría mayor importancia. Pero existe la posibilidad de que entre todas las personas que debieron entrar pero no entraron en la primera ronda, aquellos que ingresan tardíamente sean estadísticamente diferentes de los demás, en cuyo caso los resultados del panel podrían tener una distorsión. Lo mismo puede pasar con los desertores, quienes tal vez pertenezcan preferentemente a ciertos tipos específicos (por ejemplo, que sean principalmente personas de menor nivel educativo). En otras palabras, las deserciones, las entradas tardías y las exclusiones pueden no ser aleatorias sino sesgadas en alguna dirección, lo cual distorsionaría los resultados.

Sobre este tema las soluciones sólo pueden ser parciales. Es mejor incluir los tardíos y los desertores, al menos para verificar si tienen características similares al resto o si forman una población especial. En cuanto a los que nunca fueron encuestados, no hay manera de tomarlos en cuenta, excepto tal vez comparando los resultados de la encuesta con el perfil de la respectiva población total, al menos en aquellas variables para las cuales se dispone de información censal reciente o actualizada. Ese tipo de comparaciones podría servir para evaluar si la muestra del panel, con los inevitables abandonos y entradas tardías, no adolece de alguna distorsión respecto a la población total.

2.1.4. Transiciones indirectas

Como ya se ha comentado, la tabla de rotación, en sí misma, no registra un **proceso de cambio**, sino que relaciona dos **situaciones estáticas**: el estado del sujeto en la ronda 1 y el estado del sujeto

en la ronda 2. Esa comparación no permite saber, en principio, si el sujeto ha pasado del estado inicial al estado final en forma directa, o si ha pasado por algún estado intermedio (o más de uno). Por ejemplo, la celdilla N_{sv} incluiría personas que eran solteras en la primera ronda y viudas en la segunda, lo cual supone que pasaron por un estado intermedio (casadas) en algún momento durante el período intermedio. Si las dos fechas no están muy alejadas en el tiempo, y tratándose de estados civiles, esos casos son sumamente improbables, porque la mayoría de los cambios sucesivos de estado civil no ocurren en forma rápida dentro de plazos breves. Sin embargo esos casos son en principio posibles. Para que esta situación se presente se requiere que el período transcurrido haya sido suficientemente largo para que hayan ocurrido el casamiento del sujeto y el posterior fallecimiento del cónyuge. Si las rondas están muy próximas en el tiempo (por ejemplo si se realizan con frecuencia mensual o trimestral) esos casos serán extremadamente raros, ya que el porcentaje de solteros que se casa por trimestre o semestre es de por sí muy bajo, y a su vez una bajísima proporción de recién casados enviuda a los pocos meses de su casamiento. Pero en otras variables los cambios son más veloces, y por lo tanto puede haber cambios múltiples aun durante períodos sumamente breves (por ejemplo cambios en las intenciones de voto en un panel de encuestas pre-electorales), que no siempre son capturables por medio del panel.

El método general para captar cambios intermedios son las preguntas retrospectivas, pero éstas no siempre arrojan respuestas confiables. Con el uso de técnicas únicamente descriptivas es prácticamente imposible distinguir las transiciones directas y las indirectas. Si una persona aparece como ocupado en la primera ronda y como desocupado en la segunda, es prácticamente imposible saber (excepto a través de preguntas retrospectivas) si pasó en forma directa de su anterior empleo a su condición final de desocupado, o si pasó por alguna condición intermedia (por ejemplo, puede haber quedado desocupado, conseguir otro empleo, y volver a quedar desocupado antes de la segunda ronda, o puede haber dejado su anterior empleo para pasar a la inactividad, y varios meses después empezó a buscar trabajo de modo que la segunda ronda lo clasificó como desocupado). Estas trayectorias son invisibles para el panel, salvo que se incluyan preguntas retrospectivas muy específicas. Sin embargo, mediante algunos modelos teóricos y matemáticos que se analizan más tarde es posible estimar la incidencia porcentual de esas trayectorias, aunque no es posible identificar las personas específicas que las han recorrido.

2.2. Porcentajes y proporciones en tablas de rotación

Las tablas de rotación, como las presentadas anteriormente, pueden ser objeto de un tratamiento estadístico un poco más elaborado, además de usarse para describir las cantidades de sujetos que ocupan cada celdilla. El más simple de esos tratamientos consiste en el uso de **porcentajes, proporciones** o **frecuencias relativas** en lugar de las frecuencias absolutas. En principio hay tres maneras de obtener esos porcentajes: respecto al total de cada fila, de cada columna, o de la totalidad de los sujetos.[11] Se supone en los siguientes parágrafos que hay una fila para cada **estado inicial** y una columna para cada **estado final**. El análisis de distintos tipos de porcentaje permite responde a varias preguntas sobre los distintos individuos: ¿a dónde van los que estaban en un cierto estado inicial?. ¿de dónde vienen los que están en un cierto estado final?, ¿cuáles son las transiciones más comunes? [12]

2.2.1. Porcentajes de fila: ¿A dónde van?

Los porcentajes de fila son los más importantes, ya que indican las **proporciones de transición** (o **probabilidades de transición**) entre diferentes estados a lo largo de un determinado período. Indican el porcentaje o proporción de los sujetos que estaban en un cierto estado ***i*** en la primera observación y aparecen en un estado ***j*** en la segunda observación.

[11] Se usan aquí intercambiablemente proporciones o porcentajes. Muchas veces en la presentación de las tablas las fracciones aparecen como porcentajes, pero en los cálculos matemáticos se las considera como proporciones.

[12] Sobre el uso de porcentajes en el análisis de relaciones entre atributos véase Zeisel, 1966, y Hyman, 1965.

$$r_{ij}^{t,t+1} = \frac{N_{ij}^{t,t+1}}{N_i^t}$$

En la tabla de rotación de una variable dicotómica estas probabilidades aparecen como proporciones o porcentajes de fila, como se ve en la siguiente tabla.

Primera ronda (t)	Segunda ronda (t+1)		Total
	X=1	X=2	
X=1	0.50	0.50	1.00
X=2	0.10	0.90	1.00

Estas probabilidades son el instrumento fundamental de uno de los modelos teóricos más usuales para este tipo de datos, los modelos de Markov. Es importante destacar que estas probabilidades están definidas **en función de un intervalo de tiempo de una cierta duración**. Un período más corto o más largo alteraría indudablemente la probabilidad. Por ejemplo, supongamos que los estados que se consideran sean ***i***="soltero" y ***j***="casado"; la probabilidad de que un soltero pase a estar casado será mayor cuanto más tiempo transcurra entre la primera y segunda observación. Si la segunda ronda se realiza un mes después de la primera posiblemente haya muy pocos cambios de estado en ese lapso; si se realiza luego de dos o tres años seguramente habrá más. Esta advertencia anticipa que si se comparan los cambios ocurridos en períodos de diferente longitud, los porcentajes o probabilidades de cambio no serán comparables, a menos que todas ellas sean normalizadas, es decir reducidas a un común denominador temporal (convirtiéndolas por ejemplo en tasas anuales, bajo el supuesto de que el período empírico puede ser extrapolado o interpolado para estimar la probabilidad anual). El mismo recurso utilizamos al medir la velocidad en "kilómetros por hora", aunque las distancias y los tiempos varíen entre un caso y otro. Más adelante veremos que en efecto los principales modelos teóricos utilizan este enfoque, introduciendo unidades de estandarización (como la "hora" en los "kilómetros por hora") o nociones más abstractas como las **tasas instantáneas de transición**.

Cuando en dos rondas sucesivas se observa un cambio de estado, no se sabe en principio cuánto tiempo ha pasado el sujeto en cada estado. El cambio puede haber ocurrido en cualquier momento dentro del periodo intermedio transcurrido entre ambas rondas. En algunas ocasiones esto puede ser importante, porque el "tiempo de exposición" puede ser un aspecto significativo del nuevo estado. Por ejemplo, supongamos que se analizan transiciones laborales, como encontrar y perder empleos, y para ello se averigua la condición de ocupación en dos rondas sucesivas. Habrá un grupo que pasó, por ejemplo, de ocupado a desocupado. Obviamente, este grupo incluye personas que quedaron sin empleo en diferentes momentos dentro del intervalo intermedio. En general, no se sabe en qué momento ocurrió el paso de ocupado a desocupado, salvo que haya alguna pregunta específica para aclarar este punto. Todo lo que se puede hacer es imputar a esos sujetos una "fecha presunta" que usualmente es el punto medio del intervalo. Si el intervalo va desde ***t*** hasta ***t+h***, la fecha estimada promedio de los eventos ocurridos será ***t + (h/2)***. Mediante la aplicación de modelos que reflejan procesos subyacentes es posible hacer estimaciones más precisas, pero en el plano del análisis descriptivo no se puede llegar más lejos que esto.

2.2.2. Porcentajes de columna: ¿De dónde vienen?

Los porcentajes de columna se calculan sobre la cantidad de sujetos en cada estado **final**. Se los puede usar para reflejar la **distribución de origen** de aquellos sujetos que se encuentran en un determinado estado en la segunda observación. Por ejemplo, permiten contestar preguntas como éstas: ¿dónde residían hace cinco años quienes ahora viven en esta ciudad? ¿En qué situación laboral estaban el año pasado los actuales desocupados? Esta clase de estructuras porcentuales tiene sobre todo un sentido descriptivo. Para usos explicativos, que implican relaciones de causa y efecto, una regla universal indica que se deben usar más bien los porcentajes basados en las posibles causas (es decir, en las variables antecedentes en el tiempo) y no los basados en los efectos (es decir en las variables consecuentes o posteriores en el tiempo).

Primera ronda (t)	Segunda ronda (t+1)	
	X=1	X=2
X=1	0.40	0.25
X=2	0.60	0.75
Total	1.00	1.00

2.2.3. Porcentajes sobre el total de la tabla

Los porcentajes sobre el total de la tabla también son bastante frecuentes. Ellos incluyen, en primer lugar, los **porcentajes marginales** (en la fila o columna de totales), que representan la **distribución porcentual de la variable** en una fecha determinada. Estos porcentajes indican qué proporción de la población se encontraba en cada estado en el momento de cada ronda del panel. Por otro lado se pueden calcular también las **proporciones de las celdillas interiores** sobre el total de sujetos involucrados en la tabla de rotación. Estas celdillas representan cantidades de personas que se encontraban en el estado ***i*** en la fecha ***t*** y en el estado ***j*** en la fecha ***t+h***. Las proporciones resultantes se suelen llamar **proporciones de flujo:**

$$p_{ij}^{t,t+h} = \frac{N_{ij}^{t,t+h}}{N}$$

Primera ronda (t)	Segunda ronda (t+1)		
	X=1	X=2	Total
X=1	0.20	0.25	0.45
X=2	0.10	0.45	0.55
Total	0.30	0.70	1.00

Estas proporciones, que representan el porcentaje de un cierto **flujo** de sujetos entre diferentes estados, respecto al total de unidades consideradas, son utilizadas en algunos modelos matemáticos que se analizan más tarde. Desde el punto de vista puramente descriptivo que en este momento nos interesa, la utilidad primordial de estos porcentajes es **tipológica**, ya que clasifican los sujetos en función de la permanencia o cambio de sus características a lo largo del tiempo. Supongamos una tabla sobre la pobreza de los hogares en dos fechas en el tiempo, y que en cada fecha se clasifica a los hogares como "pobres" o "no pobres". Estos datos podrían usarse para determinar cuatro tipos de hogares, que sin mucha exigencia de precisión conceptual podrían denominarse pobres crónicos, no pobres, empobrecidos, y emergentes.

Primera ronda	Segunda ronda	
	Pobres	No pobres
Pobres	Pobres crónicos	Emergentes
No pobres	Empobrecidos	No pobres

Esta clase de tipología puede ser muy útil para analizar la naturaleza, la dinámica y la evolución de un fenómeno, en este caso la pobreza, y suministra las bases para estudiar las conductas de los distintos tipos de hogares, ya que muchas otras variables pueden ser cruzadas con la tipología surgida de esas dos mediciones sucesivas de la pobreza. Obviamente, ser pobre en dos rondas sucesivas quizá no es suficiente (conceptualmente) para ser declarado "crónico", y haber mejorado entre dos rondas tal vez no es suficiente para ser "emergente", pero los rótulos son solo ilustrativos, y no pretenden reflejar conceptos válidos sobre la pobreza. Si entre las rondas mediara un período más largo, tal vez los rótulos tendrían más solidez conceptual.

2.2.4. Distribuciones marginales

En la columna de totales a la derecha de las tablas de rotación aparece la distribución de la variable en el momento ***t***, mientras que en la fila inferior aparece la distribución de la variable en el momento ***t+1***, o más genéricamente ***t+h***. Estas son las llamadas **distribuciones marginales** de la variable. Estas distribuciones pueden modificarse o permanecer estables a lo largo del tiempo (independientemente de lo que hagan los individuos). Por ejemplo, si lamisma cantidad de individuos pasa de ocupado a desocupado, y viceversa, entonces la distribución marginal por

condición de ocupación podría permanecer constante, aun cuando los individuos específicos (o muchos de ellos) pueden haber cambiado de estado. En cambio, si prevalece un flujo sobre el otro, la distribución marginal cambiará. Este tema será tratado con mayor extensión más adelante (sección 3.6).

2.2.5. Convenciones de notación

En general, las proporciones marginales que corresponden a la distribución de los sujetos en el momento inicial, se denotan como p_i^t. Las referidas a la distribución en el momento final se indican del mismo modo aunque modificando el superíndice temporal: p_i^{t+h}. La proporción de una celdilla interior respecto al total de fila, que representa el porcentaje de sujetos originalmente en el estado ***i*** que son encontrados luego en el estado ***j***, y que se usan para estimar las **probabilidades de transición,** se denota con los dos subíndices ***ij*** de la celdilla, y con el superíndice temporal de los dos momentos involucrados, como en $r_{ij}^{t,t+h}$. Lo mismo ocurre con las proporciones de flujo, que expresan cada celdilla interior como porcentaje del total de la tabla con la notación $p_{ij}^{t,t+h}$. Por ejemplo, la proporción (respecto al total de casos) representada por los sujetos que en el primer momento (t=0) estaban en el estado 1 y en el segundo momento (t=1) estaban en el estado 3 se denotaría como p_{13}^{01}, y la proporción de sujetos que hicieron ese mismo cambio de estado expresados como proporción de aquellos que originalmente estaban en e estado 1, se expresa como r_{13}^{01}. En todos estos casos los superíndices de tiempo, que indican períodos o rondas, pueden ser omitidos cuando ello no genera confusión.

2.3. Tablas de rotación multivariadas

Una extensión natural de las tablas de rotación univariadas consiste en introducir una segunda variable. El propósito de esta extensión puede ser, entre otros, analizar el efecto de un posible factor explicativo, o comprobar si las dos variables varían asociadamente en el tiempo. Supóngase que la variable de interés es la intención de voto por determinado candidato **A** en las próximas elecciones. La segunda variable puede ser observada solamente la primera vez, o solamente la segunda, o en ambas oportunidades, y puede ser una condición estable de los sujetos (como por ejemplo el sexo) o una condición variable (por ejemplo el conocimiento de las propuestas del candidato A por parte de los votantes). Un ejemplo podría ser el siguiente, donde se analiza la transición entre diferentes intenciones de voto según sexo.

Población electoral según sexo e intención de voto en dos rondas de una encuesta de opinión						
	Segunda observación					
Primera observación	**Varones**			**Mujeres**		
	Partido A	**Otra opinión**	**Total**	**Partido A**	**Otra opinión**	**Total**
Partido A	N_{AAV}	N_{AOV}	N_{AV}^1	N_{AAM}	N_{AOM}	N_{AM}^1
Otra opinión	N_{OAV}	N_{OOV}	N_{OV}^1	N_{OAM}	N_{OOM}	N_{OM}^1
Total	N_{AV}^2	N_{OV}^2	N_V	N_{AM}^2	N_{OM}^2	N_M

Nótese que la tabla tiene dos subtablas indepedientes en el sentido horizontal (una para varones y otra para mujeres) pero sólo un grupo de filas, sin distinción de sexo, porque el sexo no varía entre una y otra observación. Equivalentemente, la variable sexo podría haber estado en la dimensión vertical solamente, de modo que la subtabla de Varones y la de Mujeres estarían una debajo de la otra, en lugar de estar una al lado de la otra. Esto es sólo una cuestión de disposición tipográfica y no afecta sustantivamente el análisis.

Si la variable "Sexo" se reemplazara con una que pueda variar entre las dos rondas, como por ejemplo "Conocimiento de las propuestas del candidato A", la tabla tendría más filas, como se

muestra a continuación. Habría que crear al menos cuatro subtablas, ya que ambas variables pueden variar en las dos rondas del panel.

Población electoral según intención de voto y conocimiento de las propuestas del candidato A en dos rondas de una encuesta de opinión					
	Segunda observación				
	Recuerda propuestas de A		**No recuerda propuestas de A**		
Primera observación	**Votará A**	**Otra intención**	**Votará A**	**Otra intención**	**Total**
Recuerda propuestas de A					
Votará A	N_{RARA}	N_{RARO}	N_{RANA}	N_{RANO}	N^1_{RA}
Otra intención	N_{RORA}	N_{RORO}	N_{RONA}	N_{RONO}	N^1_{RO}
No recuerda propuestas de A					
Votará A	N_{NARA}	N_{NARO}	N_{NANA}	N_{NANO}	N^1_{NA}
Otra intención	N_{NORA}	N_{NORO}	N_{NONA}	N_{NONO}	N^1_{NO}
Total	N^2_{RA}	N^2_{RO}	N^2_{NA}	N^2_{NO}	N
R=Recuerda las propuestas del candidato A; **N**=No las recuerda **A**=Votará por el candidato A; **O**=Otra intención de voto					

Este tipo de tabla puede usarse para fines explicativos de varias maneras. La variable central (intención de voto en este caso) es la **variable dependiente**, mientras la otra variable es el posible factor explicativo, o **variable independiente**. La variable independiente puede tener varias ubicaciones en el tiempo. Puede ser una **variable antecedente** (como el sexo), que tiene su valor establecido desde antes de la primera ronda, y que además es **fija**, es decir que **no varía en el tiempo que transcurre entre ambas observaciones.** Otro caso podría ser el de una **variable interviniente**, que para poder serlo tiene que poder cambiar de valor **entre las dos observaciones**, como por ejemplo el conocimiento de las propuestas del candidato sobre el cual versa la encuesta.[13] Suele así distinguirse entre factores "fijos" (como el sexo) que no cambian a lo largo del tiempo, y factores "variables" (como el status ocupacional o el conocimiento de las propuestas de un candidato) que pueden cambiar entre una ronda y otra.

Este tipo de tabla permitiría investigar diversas hipótesis sobre la relación que existe entre conocer al candidato y tener la intención de votarlo, considerando el conocimiento anterior y el conocimiento "sobreviniente", y la influencia del conocimiento sobreviniente sobre los cambios en la intención de voto. Estas hipótesis deben derivar de un modelo teórico sobre el proceso subyacente, que en este caso debe ser formulado de acuerdo con algunos de los esquemas metodológicos disponibles (como por ejemplo los modelos de Markov) que serán tratados más adelante.

2.4. Trayectorias

Cuando se tienen dos momentos solamente, las posibles transiciones denotan diferentes *trayectorias* de los individuos. Si hay dos estados (A y B), las trayectorias posibles son: A→A, A→B, B→A y B→B. Siempre considerando dos ocasiones, en caso de haber tres estados (A, B y C) el número de trayectorias aumenta a nueve posibilidades (AA, AB, AC, BA, BB, BC, CA, CB, CC). En general, cuando hay *k* estados el número de las trayectorias posibles entre dos períodos es k^2. Por eso para dos estados (k=2) hay 2^2=4 trayectorias, para tres estados hay 3^2=9 trayectorias, y así sucesivamente.

El número de trayectorias posibles aumenta notablemente cuando se consideran tres o más rondas. Por ejemplo, con solo dos estados A y B, pero con tres rondas, las trayectorias posibles serían A→A→A, A→A→B, A→B→A, A→B→B para el estado inicial A, y otras cuatro similares para

[13] Por lo general el conocimiento de las propuestas de un candidato no puede esfumarse: las personas que declararon conocerlas en la primera ronda muy probablemente seguirán declarando lo mismo en la ronda siguiente; el único cambio posible sería de la ignorancia al conocimiento del candidato; pero en la práctica pueden presentarse casos de "olvido", donde alguien que había declarado conocer las propuestas posteriormente declara no conocerlas o no recordarlas. Por eso la tabla incluye casillas que reflejan ese posible "olvido".

quienes arrancan en el otro estado B: B→A→A, B→A→B, B→B→A, B→B→B. En total habrá k^3 trayectorias, en este caso 2^3=8. Si además de haber más rondas hay también más estados posibles, el número de trayectorias rápidamente se torna muy elevado. Con tres estados y tres rondas habría 3^3=27 trayectorias posibles. El número total aumenta muy velozmente: tres estados con cinco rondas generan 243 trayectorias, y los mismos tres estados con diez rondas implican 59049 trayectorias posibles.

Ante ese panorama, el análisis de trayectorias requiere algún mecanismo para *limitar* el número de trayectorias relevantes. En primer lugar, la propia realidad impone límites: usualmente las trayectorias *efectivamente observadas* son mucho menos numerosas que las trayectorias *posibles*. Por lo menos, se requiere una muestra al menos tan grande como el número de trayectorias observadas: no hay posibilidad alguna de observar 59049 trayectorias si la muestra es de solo 3000 casos. En segundo lugar, aun cuando dispusiéramos de 60000 o 100000 casos, posiblemente muchas de las 59049 trayectorias posibles nunca se presentarían: hay algunas trayectorias más probables que otras, y hay algunas que nunca aparecen en la práctica. En tercer lugar, aun cuando se observaran todas las trayectorias posibles, rara vez se dispone de una *teoría* capaz de prever y dar significado a un gran número de trayectorias diferentes: la teoría quizá es "de grano grueso", y solo distingue grandes tipos de trayectorias. En cuarto lugar, muchos paneles son breves, y no generan trayectorias muy prolongadas: si cada sujeto permanece en la muestra solo cuatro periodos, no se podrán observar trayectorias que impliquen cinco o diez estados sucesivos.

Por tales razones, en el análisis de trayectorias se suelen adoptar alguno de los varios caminos posibles para reducir el problema a dimensiones manejables. Entre esas alternativas están las siguientes (no excluyentes entre sí):

- Limitar las trayectorias a pocas rondas (no más de tres o cuatro).
- Agrupar *cualitativamente* las trayectorias, construyendo *tipologías de trayectoras*, en lugar de considerar separadamente cada una de ellas. Por ejemplo, si los tres estados son: ocupado, desocupado e inactivo, se puede crear una tipología con tres *tipos de trayectorias:* un primer grupo de trayectorias *estables* (donde el estado no cambia a lo largo de la trayectoria, como en AAAA o CCCC): otro grupo de trayectorias *oscilantes* (donde se produce un cambio pero luego se retorna al estado original, ya sea retornando al primero (ABCA o bien ABBA) o manteniendo una oscilación recurrente (ABAB); y otro grupo de trayectorias *divergentes* donde el estado original conduce a un estado final distinto (como en ABBC, o bien AACB).
- Agrupar las trayectorias según que incluyan o no algún estado determinado (por ejemplo, trayectorias que incluyan algún período de desocupación, y las que no lo incluyan). En este enfoque, en realidad, no se analiza la trayectoria como tal: se considera el *conjunto de estados sucesivos* (sin considerar su orden temporal) y se los clasifica según un criterio *atemporal* (que en algún período el sujeto haya estado desocupado, sin importar en qué periodo ello ocurrió).
- Mantener individualizadas las trayectorias más frecuentes, y agrupar en categorías las trayectorias minoritarias, de modo que muchas trayectorias cuyas diferencias son relativamente irrelevantes y cuya frecuencia es muy baja terminan agrupadas, en lugar de ser analizadas en forma separada. Por lo general, cuando hay un número elevado de trayectorias posibles, lo más probable es que sólo una fracción pequeña sean realmente observadas, y que muchas de ellas aparezcan solo en un caso, o en unos pocos casos.

Estas posibilidades de análisis, de todos modos, implican un **análisis descriptivo** de las trayectorias, que no les imponga a priori ninguna restricción. Otra estrategia posible consiste en ver si las trayectorias observadas se ajustan a ciertos patrones. Un enfoque de ese tipo son los procesos de Markov, analizados a continuación.

3. Procesos de Markov

La clasificación y tabulación de los datos de panel, aparte de servir con fines descriptivos, puede usarse como base a fin de aplicar hipótesis explicativas. La "explicación", en este caso, significa

postular algún **mecanismo subyacente**, generalmente a nivel de los individuos o unidades de análisis, que sea capaz de generar los datos observados. Uno de los esquemas metodológicos más usuales, y que suministra una buena introducción al vocabulario técnico aplicable a este tema, son las llamadas "cadenas de Markov" o "modelos de Markov".

3.1. Características generales de los procesos de Markov

Si los sujetos de estudio pueden estar en dos o más estados en un momento dado, y pueden cambiar de estado con el correr del tiempo, el proceso de Markov analiza esas transiciones de los individuos entre un estado y otro. Un individuo que en una primera observación estaba en el estado 1 puede seguir en el estado 1 o haber pasado al estado 2 cuando se lo observe nuevamente un tiempo más adelante. El análisis de Markov se centra en la **probabilidad** de que ocurra una u otra de esas transiciones. En su forma más sencilla, el modelo markoviano hipotetiza un proceso a nivel micro donde **cada individuo** en un estado ***i*** en el momento ***t*** está afectado por determinadas **probabilidades de transición**, que se denotan como $r_{ij}^{t,t+1}$, y que miden sus probabilidades de estar en el estado **j** en el momento ***t+1*** si es que estuvo en el estado ***i*** en el momento ***t*** (incluyendo la probabilidad $r_{ii}^{t,t+1}$ de **permanecer** en el mismo estado ***i***). En los modelos de Markov se supone que estas probabilidades son **iguales para todos los individuos**, **constantes en el tiempo**, e **independientes de la trayectoria anterior o posterior de cada sujeto**. Por supuesto, las observaciones pueden confirmar o refutar la hipótesis de que el proceso responde a los supuestos de los modelos de Markov.

Los modelos de Markov consideran al tiempo como una variable **discreta**. Se comparan momentos ***t*** y ***t+1*** pero no se hace ningún supuesto sobre el período intermedio entre ambas fechas. Esas probabilidades de transición markovianas son **condicionamientos internos (no observables) de cada individuo**, y no deben confundirse conceptualmente con la **proporción observada de cambios de estado** entre dos rondas sucesivas de la encuesta. Esas proporciones observadas (los porcentajes de fila considerados anteriormente) son propiedades de la población en su conjunto y no propiedades de cada individuo, si bien a menudo las proporciones observadas se usan como mediciones estimativas de aquellas probabilidades no observables.

Si se supone que todos los individuos están afectados por las mismas probabilidades de transición, y se supone asimismo que cada individuo actúa solamente en función de sus probabilidades subjetivas de transición, la proporción de sujetos que cambia de estado (magnitud observable en un panel) coincidirá con las probabilidades individuales de transición postuladas por el modelo de Markov. Pero esta coincidencia es una **implicancia lógica del modelo teórico**, y no un dato de la realidad. Por lo tanto no puede considerarse a las proporciones observadas como una **definición** de las probabilidades de transición. Si los supuestos del modelo no se cumplen, las proporciones observadas no podrían equipararse a las probabilidades subjetivas de los individuos. Por ejemplo, si hubiese probabilidades subjetivas **heterogéneas** entre los individuos, la proporción agregada representaría sólo una media estadística y no coincidiría con las probabilidades de transición de ningún individuo (excepto las de alguno cuyas probabilidades coincidan por casualidad con el promedio).

Las características generales de las probabilidades de transición en un proceso de Markov son las siguientes:

Homogeneidad	Las probabilidades de pasar a otro estado, o de permanecer en el mismo estado, son las mismas para todos los sujetos
Constancia	Las probabilidades de transición son constantes en el tiempo
Amnesia	Las probabilidades de transición de un individuo dependen sólo de su estado actual, y no de su trayectoria anterior. En forma equivalente: todos los individuos que están en un cierto estado tienen las mismas probabilidades de transición hacia otros estados.

La **homogeneidad** de las probabilidades individuales de transición significa que todos los sujetos o unidades en un cierto estado tienen las mismas probabilidades de estar en otro estado o de volver a aparecer en el mismo estado en la siguiente fecha de observación. Versiones más complicadas del modelo permiten diferenciar entre sujetos, mediante la introducción de variables que los tipifican

(por ejemplo, las mujeres podrían tener una mayor o menor probabilidad de transición que los varones, siendo el sexo la variable que tipifica los casos y permite asignarles una probabilidad específica) o mediante el supuesto de que las probabilidades de los individuos varían aleatoriamente en torno a la probabilidad promedio. Por el momento nos referimos a los modelos de Markov en su forma más simple, sin variables de tipificación que rompan el supuesto de homogeneidad, y asumiendo que todos los individuos están afectados por las mismas probabilidades de pasar a otros estados.

Hay sin embargo un enfoque que se aparta de este supuesto: es el que divide a los sujetos en dos grandes clases: los móviles y los inmóviles (denominados respectivamente *movers* y *stayers* en la jerga habitual de esos modelos). Los móviles tienen tasas de transición diferentes de cero, las mismas para todos ellos, mientras para los inmóviles no hay cambio alguno a lo largo del tiempo: permanecen siempre en el mismo estado inicial, con tasas de transición iguales a cero para todo $i \neq j$, e iguales a 1 para $i=j$. Esta simplificación puede ser útil en algunas situaciones, pero es sin duda una simplificación extrema. Versiones más sofisticadas de este enfoque suponen una gradación de sujetos con mayor o menor volatilidad, sin necesidad de suponer que los menos volátiles sean totamente incapaces de cambiar de estado.

La **constancia** significa que la probabilidad de pasar de un estado ***i*** a un estado ***j*** es la misma en todos los períodos. También aquí, por supuesto, el supuesto puede ser flexibilizado haciendo que la probabilidad sea función del tiempo. Esto cubre dos casos diferentes. Por una parte, la probabilidad de transición puede cambiar entre sucesivas rondas; por ejemplo, para una mujer la probabilidad de pasar de "económicamente inactiva" a "económicamente activa" puede tener tendencia creciente si en el país se registra una tendencia general al aumento de la participación laboral de la mujer durante el período cubierto por el panel. Por otra parte, la probabilidad puede cambiar según el tiempo transcurrido desde el evento anterior; por ejemplo, la probabilidad de que una mujer en edad fértil tenga un nuevo hijo se incrementará al pasar más tiempo desde el nacimiento de su hijo anterior, para luego disminuir una vez que el tiempo transcurrido haya llegado a cierto límite. Pero en cualquier caso estas variaciones inter-temporales en las probabilidades de transición representarían una violación de uno de los supuestos que caracterizan a los procesos de Markov.[14]

La constancia de las probabilidades individuales no significa que haya proporciones constantes a nivel agregado. Un modelo con probabilidades de transición constantes a nivel de los individuos puede generar un proceso en el cual se registra una creciente cantidad de individuos en determinados estados; podría así suceder que un modelo, en el cual la probabilidad de entrar en actividad sea constante para cada individuo, explique sin embargo un proceso en el cual hay cada vez más mujeres activas. Ello podría ser así, por ejemplo, si al mismo tiempo hay un alargamiento de la vida o una disminución de la fecundidad que hace permanecer más años a las mujeres en la esfera laboral.

La **amnesia**, o más técnicamente la **independencia de la trayectoria anterior** significa que los sujetos **"no tienen memoria"**: la probabilidad de transición de ***i*** a ***j*** entre los momentos ***t*** y ***t+1*** es igual para todos los sujetos que están en el estado ***i*** en el momento ***t***, independientemente del estado en que se encontraban en el momento ***t-1*** o en períodos anteriores. Una natural extensión de este supuesto de ausencia de memoria es el supuesto de **ausencia de previsión**, según el cual las probabilidades son también **independientes de la trayectoria futura**, es decir, independientes del estado en que se encontrarán los sujetos en el momento ***t+2*** o posteriormente.

El supuesto de amnesia es bastante fuerte. Por ejemplo, significa que la probabilidad de pasar de "Ocupado" a "Desocupado" entre dos fechas es independiente de la historia ocupacional anterior del individuo. Obviamente es posible que la probabilidad de caer en el desempleo en ese lapso dependa de que el individuo haya estado siempre ocupado o haya estado desocupado en algún momento antes de la primera fecha considerada.

El supuesto de independencia respecto al futuro no es tan restrictivo en la práctica, pero a veces cobra relevancia, pues el evento futuro de algún modo puede ser indicador de algún suceso no obser-

[14] Hay algunos modelos markovianos más complejos, o más heterodoxos, con **probabilidades variables en el tiempo**. Sin embargo en el presente caso estamos circunscribiéndonos al caso de **probabilidades constantes.**

vado del pasado. Por ejemplo, en un panel sobre fumadores crónicos, pasar de fumador a no fumador puede no ser independiente del hecho (futuro) de que el sujeto muera de cáncer de pulmón, pues este hecho futuro puede ser un indicador de que al momento de dejar el cigarrillo esa persona ya tenía indicios de que podría estar padeciendo el cáncer o estaría próximo a desarrollarlo, y por eso precisamente dejó de fumar. El supuesto de "independencia respecto al futuro" dice que las probabilidades de dejar de fumar son las mismas para sujetos fumadores que terminarán teniendo cáncer de pulmón como para aquellos fumadores que se librarán eventualmente de contraer esa enfermedad.

La ausencia total de "memoria" o de "previsión" configura un caso extremo, que corresponde a un modelo de Markov **de primer orden**: sólo el estado del sujeto en el momento inicial determina las probabilidades de transición. En un modelo **de segundo orden** las probabilidades que afectan al individuo en el momento ***t*** están determinadas por el estado del sujeto en el momento ***t*** y en el momento ***t-1***, pero no por su estado en el momento ***t-2*** o en otros anteriores. El sujeto tiene buena memoria en el corto plazo pero sufre de amnesia en cuanto al largo plazo. En general, en un modelo **de orden *h*** las probabilidades de transición en el momento **t** están determinadas por los estados en que estuvo el sujeto en los momentos ***t, t-1, t-2, ..., t-h***.

Sin embargo, esto no cambia la esencia del modelo, ya que las trayectorias completas durante ***h*** períodos pueden ser consideradas como "estados". Por ejemplo, en el momento inicial algunos sujetos están en el estado "ocupado desde hace más de seis meses" mientras otros están como "ocupados ahora que seis meses antes estaban desocupados". Los estados finales pueden definirse de igual modo, con lo cual los sujetos pasarían de ostentar una cierta trayectoria acumulada hasta el momento ***t*** a tener otra trayectoria acumulada hasta el momento ***t+1***.

Algunas transiciones entre trayectorias serían imposibles, porque gran parte de la trayectoria inicial debe reaparecer intacta en la trayectoria final; por ejemplo, los que estaban ocupados en el momento ***t,*** sea cual fuere su trayectoria anterior, no podrían pasar al estado final "ocupados que en la ronda anterior estaban desocupados". Esto involucraría una contradicción.

En esta transformación del modelo, en la cual en lugar de observar transiciones entre estados se observen transiciones **entre trayectorias**, el número de "estados" se multiplica, haciendo que se necesiten muestras más grandes para llegar a resultados significativos. Esta consideración práctica hace menos atractivo el uso de modelos de Markov de orden superior. Parece más lógica la estrategia de probar primero un modelo de Markov simple o con pocas complicaciones, tratando así de predecir las probabilidades de transición de la manera más simple posible. En general, cuando ello no es suficiente se puede modificar el modelo, pero no necesariamente introducir procesos de Markov de orden superior, por las dificultades que ellos implican para el análisis. Más adelante se retorna sobre este punto.

3.2. Probabilidades de transición

El instrumento fundamental del análisis markoviano es una **matriz de probabilidades de transición**. A cada individuo le corresponde un conjunto de probabilidades de estar en cada estado ***j*** en el momento ***t+1***, dependiendo únicamente del estado ***i*** en que se encontraba en el momento ***t***. Para un sujeto ***k*** la probabilidad $r_{ijk}^{t,t+1}$, que a veces se denota incluyendo sólo el punto inicial del intervalo, r_{ijk}^{t}, indica la probabilidad de que ese sujeto ***k*** pase de estar en el estado ***i*** en el momento ***t*** a estar en el estado ***j*** en el momento ***t+1***. Los supuestos de **homogeneidad** y de **amnesia** exigen que esta probabilidad sea la misma para todos los sujetos, y el supuesto de **constancia** asume que esta probabilidad es la misma en todos los períodos. Por ello los modelos de Markov postulan lo siguiente:

$$r_{ijk}^{t,t+1} = r_{ijk}^{t} = r_{ij} \qquad \textit{para todo } \mathbf{k} \textit{ y para todo } \mathbf{t}$$

Por tal razón en la notación habitual, y salvo situaciones especiales que aconsejen lo contrario, se omiten tanto el subíndice ***k*** que identifica a cada sujeto como el superíndice ***t*** o ***t+1*** que identifica el período o fecha de la observación.

Otro aspecto digno de destacar es que los modelos de Markov no presuponen, de por sí, un **proceso continuo**. Operan como si el tiempo estuviese constituido por una sucesión de "instantes" discretos, y sólo comparan el estado de los sujetos en dos o más de estos "instantes". El tiempo intermedio no forma parte del modelo de Markov como tal.

En un caso simple con tres estados posibles, la matriz de probabilidades de cada individuo (y por consiguiente del conjunto de la población) sería así:

Una matriz de probabilidades de transición entre tres estados				
	Estado final			
Estado inicial	**1**	**2**	**3**	**Total**
1	r_{11}	r_{12}	r_{13}	1
2	r_{21}	r_{22}	r_{23}	1
3	r_{31}	r_{32}	r_{33}	1

Como se indica en la tabla, las probabilidades de Markov son exhaustivas para cada estado inicial: la suma de todos los estados finales para un cierto estado inicial es igual a la unidad:

$$\sum_j r_{ij} = 1 \quad \text{para todo estado inicial } \boldsymbol{i} \text{ (incluyendo } \boldsymbol{i}=\boldsymbol{j})$$

Aparte de las probabilidades de transición el modelo markoviano utiliza la **distribución marginal** de los sujetos entre los diferentes estados, también llamadas "**probabilidades de estado**" aunque este nombre no es muy correcto. Habrá una distribución **anterior (**también llamada **"inicial")** y una distribución **posterior** (o **"final"**) de los sujetos entre los varios estados posibles. Siguiendo la notación establecida para las cifras absolutas *N*, la definición general de estas proporciones de estado para cualquier estado ***i*** en un momento ***t*** es:

$$p_i^t = \frac{N_i^t}{N}$$

También para algunos propósitos los flujos de un estado a otro en el intervalo ***t,t+h***, es decir $N_{ij}^{t,t+h}$, pueden expresarse como proporción del número total de casos:

$$p_{ij}^{t,t+h} = \frac{N_{ij}^{t,t+h}}{N}$$

Estas proporciones sobre el total de la tabla se pueden denominar **proporciones de flujo**, y no deben ser confundidas con las **probabilidades de transición**, en las cuales el flujo no es dividido por el total de casos, sino por los casos que se encontraban en el estado inicial ***i***, o sea los que estaban **en riesgo** de efectuar el pasaje de ***i*** a ***j***:

$$r_{ij}^h = \frac{N_{ij}^{t,t+h}}{N_i^t}$$

Las proporciones de flujo suman 1 en el total de casos, es decir en el total de la tabla de rotación. Las probabilidades de transición, en cambio, suman 1 para cada **fila** de la tabla:

$$\sum_i \sum_j p_{ij}^{t,t+h} = 1 \qquad \sum_j r_{ij}^h = 1$$

En una población de 10000 personas económicamente activas, donde 8000 están ocupados y 2000 desocupados en el momento ***t***, y 400 personas desocupadas encuentran empleo entre el período ***t*** y el período ***t+h***, la probabilidad de transición (de ocupado a desocupado) será de 400/20000=0.20, mientras que la probabilidad de ese flujo será 400/10000=0.04. Cuatro de cada cien personas por período realizan ese cambio de estado, lo que corresponde a veinte de cada cien desocupados.

En un modelo de Markov, las probabilidades de transición entre dos períodos en general se suponen constantes, por lo cual la referencia temporal a los instantes ***t*** y ***t+h*** puede ser omitida cuando ello no da lugar a confusión. Por el mismo motivo, y si bien esas probabilidades de transición dependen de la longitud (***h***) del intervalo considerado, esa referencia también puede generalmente ser omitida, para así denotar esas probabilidades con la notación r_{ij}. Sin embargo, es menester recordar que esas

probabilidades de transición **se refieren siempre a un período determinado** y cambiarían sus valores si, por ejemplo, los intervalos entre las rondas fuesen más breves o más prolongados.

Si se estiman las probabilidades de transición en un modelo de Markov, r_{ij}, a partir de una tabla de rotación con datos en panel, y se supone que esas probabilidades son homogéneas y constantes, ellas podrían usarse para **predecir** la distribución de sujetos en una fecha ulterior, es decir, se puede estimar la distribución **esperada** de los sujetos al cabo de un período adicional. Es evidente que la **proporción esperada** de sujetos en un determinado estado ***j*** en el momento ***t+h*** es igual a la suma de los productos entre las proporciones iniciales de sujetos en los varios estados multiplicadas por sus respectivas probabilidades de transición:

$$p_{ej}^{t+h} = \sum_i p_i^t r_{ij}^h \qquad \text{(Ec. 1)}$$

Por ejemplo, en una encuesta que se realiza en mayo y octubre, supóngase que las probabilidades de transición entre mayo y octubre se han estimado sobre la base de las encuestas realizadas en esos meses. Si esas probabilidades fuesen constantes, el porcentaje esperado de desocupados en octubre **del año siguiente** será la suma de:

(a) el porcentaje de desocupados en mayo de ese año, multiplicado por su probabilidad de que permanezcan desocupados hasta octubre, más:
(b) el porcentaje de ocupados en mayo, multiplicado por su probabilidad de estar desocupados en el siguiente mes de octubre, más:
(c) el porcentaje de inactivos en mayo, multiplicado por su probabilidad de entrar en actividad en el período intermedio y estar desocupados en octubre.

De modo similar se puede estimar la **cantidad esperada** de sujetos en cada estado. Si la cantidad inicial de sujetos en un estado ***j*** en la ronda ***t*** es igual a N_j^t, y las probabilidades de transición son r_{ij}, la cantidad esperada en la ronda siguiente (***t+1***) será:

$$N_{ej}^{t+1} = \sum_i N_i^t r_{ij} \qquad \text{(Ec. 2)}$$

En este ejemplo se ha supuesto que permanecen constantes las probabilidades de transición entre el estado observado en mayo y el estado observado en octubre. No se ha dicho nada sobre la transición desde octubre de un cierto año hasta mayo del año siguiente. El modelo puede prever que esas probabilidades también sean iguales a las probabilidades mayo-octubre, o puede suponer que existe estacionalidad, de modo que las probabilidades sean constantes para iguales períodos de años sucesivos (las de mayo-octubre de este año permiten prever las de mayo-octubre del año próximo, pero no las de octubre-mayo). En un modelo no estacional se supone directamente que las probabilidades son constantes en todo momento. En realidad, el modelo "estacional" no es un solo modelo, sino dos modelos de Markov que operan alternadamente.

Si se supone que el proceso de Markov opera en forma constante a lo largo del tiempo, se podría también predecir la probabilidad de llegar de un estado a otro **a lo largo de varios períodos.** Dado que la probabilidad de pasar de uno a otro estado solo depende del estado inicial, la probabilidad de pasar del estado inicial *i* al estado final *j* pasando por una determinada trayectoria con *k* etapas intermedias será igual al ***producto*** de las sucesivas probabilidades de transición que describen esa trayectoria. Por ejemplo, una de las trayectorias que unen A con D en tres etapas es la trayectoria ABCD; otras trayectorias posibles podrían ser AADD, ACBD, ADBD u otras (algunas serán imposibles, por ejemplo si implican pasar de "casado" a "soltero"). Si hay varias trayectorias que lleguen del estado *i* al estado *j* en un proceso de *k* etapas, la probabilidad total de pasar de *i* a *j* en *k* etapas será la suma de las probabilidades de lograrlo a través de todas las trayectorias posibles que unen el estado *i* con el estado *j* en *k* etapas. Por ejemplo, si conocemos las probabilidades de transición entre todos los estados, es decir la probabilidad de pasar de un estado *i* a un estado *j* en un período, la probabilidad combinada de transición desde el estado A hasta el estado D en *tres* períodos mediante la trayectoria ABCD será:

$$r_{ABCD} = r_{AB} r_{BC} r_{CD}$$

Esta no es la única manera de llegar desde A hasta D en tres períodos. Cada trayectoria posible origina una de estas probabilidades combinadas, y la suma de todas ellas equivale a la probabilidad conjunta de llegar desde A hasta D en tres períodos *por cualquier trayectoria*:

$$r_{AijD} = \sum_{ij} r_{Ai} r_{ij} r_{jD}$$

Tomando la proporción inicial de casos en el estado A, es decir p_A, se puede estimar la proporción esperada (del total de casos) que se espera que lleguen al estado D por cualquier trayectoria de tres pasos, habiendo partido de A. Esa proporción sería $p_{AD}^{t+3} = p_A^t r_{AijD}$, es decir la proporción que había en el estado A en el momento inicial *t*, multiplicada por la tasa combinada o acumulada de transición hasta llegar al estado D en el momento t+3 por cualquier trayectoria, calculada de acuerdo a la ecuación precedente. En esa ecuación, se quiere estimar la probabilidad de ese evento **en un período futuro** a partir del estado actual, y de las probabilidades de transición **estimadas a partir de períodos anteriores.**

Ahora bien, esa predicción (aplicada al futuro) **quizá no se cumpla**. La expectativa calculada se basa en el supuesto de que las transiciones futuras sean iguales a las probabilidades estimadas en el pasado. La realidad no tiene por qué obedecer un proceso de Markov al pie de la letra. El proceso de Markov, una vez cuantificado, permite estimar las proporciones o cantidades ***esperadas*** de casos que estarían en un cierto estado después de uno o más períodos, *siempre que se cumplan los supuestos del proceso de Markov*, y siempre que además los valores de las probabilidades de transición *se mantengan en el tiempo* al menos durante el número de períodos considerado. Hoy quizá las probabilidades son r_{AB}, r_{BC} y r_{CD}, pero podría ocurrir que en la práctica ellas no se mantengan iguales a lo largo del tiempo. Tal vez dentro de dos o tres períodos el valor de r_{CD} ya no sea igual al que se ha estimado hoy. De este modo, la proporción efectivamente observada en el estado "final" (D) en el momento ***t+1 puede no coincidir*** con la proporción esperada. Diferirá de ella por algún margen de error (ε):

$$p_j^{t+1} = p_{ej}^{t+1} + \varepsilon$$

Ese error puede estar asociado a algunas variables explicativas. Tal vez depende de la edad del sujeto, de su sexo, de su antigüedad en el empleo, o de la fase del ciclo económico en cada uno de los períodos. La proporción efectiva final puede entonces ser considerada como la suma de la proporción esperada $\boldsymbol{r_{ej}}$, más el efecto de diversas variables explicativas, más un error aleatorio:

$$p_j^{t+1} = p_{ej}^{t+1} + f(X, W, Z) + \varepsilon \qquad \text{(Ec.3)}$$

En esta formulación, la función explicativa *f(X, W, Z)* apunta a explicar no la probabilidad final como tal, sino la desviación entre esa probabilidad final y la que podía preverse en función de observaciones efectuadas en períodos anteriores. Esto puede reducirse a una función donde la variable dependiente no sea la probabilidad final sino una variable ***y*** equivalente a la **diferencia** o **desviación** entre la probabilidad observada y la esperada:

$$y = p_j^{t+1} - p_{ej}^{t+1} = f(X, W, Z) + \varepsilon \qquad \text{(Ec.4)}$$

Estas mismas relaciones pueden formularse no solo para las proporciones de sujetos (*p*) sino también respecto a la cantidad de sujetos (*N*). Esto no cambia mayormente las cosas, excepto que las diferencias absolutas dependerán del tamaño absoluto de la muestra, lo cual no parece muy interesante como opción metodológica.

3.3. Modelos de Markov con memoria de orden superior

En los ejemplos anteriores con trayectorias del tipo ABCD se suponía que las transiciones de cada período (de A a B, de B a C, etc.) estaban gobernadas por tasas de transición que expresaban un proceso de Markov simple, con sus tres atributos de homogeneidad, constancia y amnesia. En esta sección se discute la posibilidad de que la amnesia no sea total. Quizá la probabilidad de que un sujeto en el estado A pase a B pueda variar según la proveniencia de ese sujeto: ¿dónde estaba en el período anterior? En el momento anterior (t-1) puede haber estado en el mismo estado A, o en

cualquier otro estado. Tal vez sus probabilidades de transición desde t hasta t+1 dependan por lo menos de su *último* estado anterior, como si los sujetos tuvieran una "memoria" que alcance a "recordar" solo lo que pasó ayer, pero no lo que pasó anteayer o antes de anteayer. O bien podría ser que la "memoria" alcance a ayer y anteayer (dos períodos) pero no más. La amnesia total define un proceso de Markov "de primer". Recordar solo el día de ayer sería una amnesia "de segundo orden"; recordar también anteayer equivale a una amnesia "de tercer orden", y así sucesivamente. Esto permitiría definir movelos markovianos de orden superior, con algún grado de "memoria".

En forma más general: Las probabilidades de transición entre ***t*** y ***t+1*** pueden ser uniformes para todos los sujetos que están en un determinado estado en la ronda ***t,*** o pueden variar según los **estados anteriores** que hayan atravesado los sujetos (por ejemplo su condición de ocupación en una ronda anterior), con lo cual el modelo markoviano simple de primer orden se transformaría en un modelo markoviano con memoria, o de orden superior. Pero en ese caso, algunas de las probabilidades de transición serían por definición iguales a cero, pues corresponden a situaciones imposibles. El siguiente ejemplo ilustra las probabilidades de transición entre los momentos o fechas ***t*** y ***t+1*** en un modelo de segundo orden con dos estados posibles (denominados 1 y 2).

Estado inicial		**Estado final**			
Estado en t-1	**Estado en t**	**t=1, t+1=1**	**t=1, t+1=2**	**t=2, t+1=1**	**t=2, t+1=2**
1	**1**	p_{1111}	p_{1112}	$p_{1121}=0$	$p_{1122}=0$
1	**2**	$P_{1211}=0$	$p_{1212}=0$	p_{1221}	p_{1222}
2	**1**	P_{2111}	p_{2112}	$p_{2121}=0$	$p_{2122}=0$
2	**2**	$p_{2211}=0$	$p_{2212}=0$	p_{2221}	p_{2222}

Las celdillas sombreadas representan situaciones imposibles, cuya probabilidad es cero. Por ejemplo, en la primera fila la probabilidasd p_{1121} es cero porque incluye como estado inicial sujetos que en el período **t** estaban en el estado 1, y como estado final sujetos que en el mismo período **t** estaban en el estado 2, lo cual es imposible: los sujetos involucrados tendrían que estar en dos estados distintos en el mismo período **t**. Sólo las celdillas no sombreadas pueden tener probabilidades diferentes de cero. En estas probabilidades "posibles", los dos subíndices internos del cuarteto de subíndices (es decir, el estado final de la trayectoria inicial, y el estado inicial de la trayectoria final) deben coincidir, como por ejemplo en p_{1221}.

Las transiciones factibles (no necesariamente iguales a cero) podrían también ser expresadas en forma de **probabilidades de trayectorias** con tres puntos. Por ejemplo p_{1221} es en realidad la probabilidad p_{121}, es decir la probabilidad de una trayectoria que empieza en el estado 1 en el tiempo **t-1**, pasa al estado 2 en **t**, y regresa al estado 1 en **t+1**. Todos los modelos markovianos de orden superior se reducen a un modelo de trayectorias por múltiples puntos. En general, trayectorias que tocan **k** puntos en el tiempo requieren modelos de orden **k-1**.

La representación precedente es sumamente engorrosa y poco intuitiva, aparte de tener una cantidad de celdillas "censuradas" o "imposibles". En una representación alternativa de la matriz de transición para modelos de orden superior se pueden usar matrices **rectangulares**, definiendo los estados iniciales a partir de dos o más rondas, y los estados finales sólo por el estado en la próxima ronda.. Si el estado inicial se define según los últimos dos estados por los que pasó el sujeto, la matriz de transición sería la siguiente:

Estado inicial de 2º orden		**Estado en t+1**	
Estado en t-1	**Estado en t**	**1**	**2**
1	**1**	p_{111}	p_{112}
1	**2**	p_{121}	p_{122}
2	**1**	p_{211}	p_{212}
2	**2**	p_{221}	p_{222}

En esta representación no hay celdillas "imposibles". Aparecen solamente las celdillas "posibles" de la matriz anterior. Esta representación puede ser extendida fácilmente a modelos de orden 3 o superior, añadiendo mayor número de "antecedentes", con lo cual naturalmete el número de filas del

cuadro iría aumentando: si se tomaran también los estados en ***t-2*** habría, en el caso del esquema precedente, no ya cuatro sino ocho estados iniciales, que representarían ocho "trayectorias" posibles a lo largo de las tres rondas anteriores (***t-2***, ***t-1*** y ***t***).

Si bien es posible que las probabilidades de transición estén de algún modo afectadas por la "historia anterior" de cada individuo, la experiencia ha mostrado que rara vez este problema se encara exitosamente mediante modelos de Markov de orden superior. Una de las razones para ello es el hecho de que los períodos entre rondas son generalmente arbitrarios, por lo cual no resulta plausible que el registro de la variable en la fecha de las rondas anteriores tenga tanta importancia. En cambio, en la mayor parte de los casos estas situaciones son enfrentadas mediante modelos que introducen a hipótesis de un **proceso continuo** en el tiempo, con probabilidades de transición constantes o variables, usando las observaciones de panel como una "muestra de instantes" en los cuales fue registrado el estado de los sujetos, y tratando de inferir los parámetros del proceso continuo subyacente a partir de esos datos. Más adelante se tratan algunos de estos enfoques.

3.4. Procesos de Markov multivariados

En los ejemplos precedentes se trata de una sola variable categórica. Sin embargo, es posible concebir un proceso de Markov que involucre dos o más variables categóricas. Con dos variables dicotómicas **X** y **Z**, los sujetos pueden estar, en cada instante, en **cuatro** estados definidos por la **combinación** de valores de las dos variables, y entre dos instantes pueden pasar de la combinación inicial a cualquier otra combinación final. Una notación que se puede usar para este caso es r_{ijkm}, donde los dos primeros subíndices (***i, j***) indican el valor inicial de **X** y **Z** y los dos últimos subíndices (***k, m***) se refieren a los valores finales de ambas variables. Así r_{0110} será la probabilidad de transición desde un estado inicial con **X**=0 y **Z**=1 hasta un estado final con **X**=1 y **Z**=0. La siguiente sería la tabla completa de rotación.

		Período t+1			
		Valor de X			
		0		1	
Período t		Valor de Z		Valor de Z	
Valor de X	Valor de Z	0	1	0	1
0	0	r_{0000}	r_{0001}	r_{0010}	r_{0011}
	1	r_{0100}	r_{0101}	r_{0110}	r_{0111}
1	0	r_{1000}	r_{1001}	r_{1010}	r_{1011}
	1	r_{1100}	r_{1101}	r_{1110}	r_{1111}

Es fácil ver que este caso es formalmente igual al de un proceso de Markov con una sola variable de cuatro categorías. Basta para ello con postular una variable **W** cuyos estados 0, 1, 2 y 3 corresponden a las combinaciones 00, 01, 10 y 11 de los valores de **X** y **Z**. En otras palabras, el análisis de los procesos de Markov multivariados puede ser fácilmente reducido al de un proceso de Markov univariado, por lo cual no es necesario el desarrollo de métodos especiales para su tratamiento.

Sin embargo, si el proceso involucra una relación causal determinada entre las variables, por ejemplo si se supone que **X** es la causa de los cambios que ocurren en **Z**, podría ser conveniente mantener ambas variables separadas. Esto no altera esencialmente el proceso pero de este modo se pueden tornar más claras las relaciones entre las variables.

3.5. Aplicaciones prospectivas de procesos de Markov

Las ecuaciones anteriores permiten obtener las probabilidades esperadas de estado del período **siguiente**. Como ya se mencionó, si se llega a la conclusión de que ciertas probabilidades de transición van a seguir rigiendo en el futuro, una de las posibles aplicaciones del modelo consiste en predecir la distribución en un momento futuro, separado del actual no ya por uno sino por dos o más períodos. Si se repite la aplicación de las probabilidades de transición sobre las probabilidades de estado esperadas resultantes de la primera proyección, se obtienen probabilidades esperadas para dentro de dos períodos, y así sucesivamente para períodos ulteriores.

$$p_{ej}^{t+2} = \sum_i p_{ei}^{t+1} r_{ij}$$

$$p_{ej}^{t+3} = \sum_i p_{ei}^{t+2} r_{ij}$$

$$p_{ej}^{t+u} = \sum_i p_{ei}^{t+u-1} r_{ij}$$

Si se representa como **R** la matriz de probabilidades de transición, y se define el vector-fila $\boldsymbol{p_i^t}$ de probabilidades de estado iniciales y el vector-fila $\boldsymbol{p_i^{t+1}}$ de las probabilidades finales, la ecuación 1 se puede expresar en notación matricial como sigue.[15]

$$p_{ej}^{t+1} = p_i^t R \qquad \text{(Ec. 5)}$$

Para predecir los valores del momento ***t+2*** se requeriría multiplicar por la matriz **R** a las probabilidades esperadas del momento ***t+1***:

$$p_{ej}^{t+2} = p_{ei}^{t+1} R \qquad \text{(Ec. 6)}$$

Reemplazando en la ecuación 6 las probabilidades de estado de ***t+1*** por su equivalente según la ecuación 5, la ecuación 6 queda en la forma siguiente:

$$p_{ej}^{t+2} = p_i^t R^2 \qquad \text{(Ec. 7)}$$

La notación R^2 corresponde al cuadrado de la matriz *R*, es decir al producto matricial *RR*. En general, a partir de la distribución observada en el período ***t***, la distribución esperada en el período ***t+u*** (donde ***u*** es un número de intervalos, cada uno de ellos de longitud ***h***) podría ser predicha mediante la ecuación matricial:

$$p_{ej}^{t+u} = p_i^t R^u \qquad \text{(Ec. 8)}$$

Muchos procesos socioeconómicos tienden a ser inestables o cíclicos, y en algunos de ellos las probabilidades de transición difícilmente sean constantes. Pero se debe notar que es posible que las proporciones marginales tengan un comportamiento tendencial o cíclico aunque las probabilidades de transición sean constantes; y por otra parte, hay procesos más constantes o regulares que pueden ser predichos eficientemente por este enfoque. Por ejemplo, supongamos que se trate de estadísticas escolares, y que los estados básicos consistan en estar cursando diferentes grados, o convertirse en desertor, o ser un graduado. Las transiciones desde cada grado pueden consistir en repetir el grado, pasar al grado siguiente, graduarse, o desertar, y probablemente las probabilidades de transición no varíen mucho a lo largo del tiempo. A partir de una distribución inicial de sujetos y del número esperado de ingresantes en primer grado en sucesivos años podría aplicarse la ecuación 8 un número suficiente de veces para prever la población escolar en cada grado, así como el número de egresados del último grado y de desertores de cada grado, en los sucesivos períodos futuros.

Si las probabilidades de transición son constantes, la aplicación de la ecuación 8 permite estimar la distribución de los sujetos de estudio después de cualquier número de períodos. Un ejemplo sencillo permite ilustrarlo. Supóngase que se desea estimar de antemano el estado de salud de una población de 10.000 niños en sucesivos períodos futuros (por ejemplo, en los sucesivos meses). Los niños son revisados una vez por mes y clasificados como sanos o enfermos. El interés fundamental es saber cuántos niños estarán enfermos en cada mes, para poder planificar la prestación de servicios de salud. Para ello se cuenta con una estimación de la probabilidad de que un niño que estaba sano en un cierto mes esté enfermo en el mes siguiente, y de que un niño enfermo en un cierto mes haya sanado para el mes siguiente. Las probabilidades mensuales de transición que se han estimado son las que figuran en la siguiente **matriz de probabilidades de transición**:

[15] Los lectores no familiarizados con la notación matricial y las operaciones con matrices podrian consultar la nota técnica al final.

		Mes t+1	
		Sano	**Enfermo**
Mes t	**Sano**	0.95	0.05
	Enfermo	0.80	0.20

Supongamos que en el mes inicial (t=1) todos los niños están sanos. Las cantidades esperadas al mes siguiente (t=2) serán de 9500 niños sanos y 500 enfermos, ya que hay una probabilidad de 0.05 de pasar del estado sano al estado enfermo en el curso de un mes. La operación matricialmente indicada sería la siguiente:

$$[1000 \quad 0]\times\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}=\{[1000\times 0.95+0\times 0.80] \quad [1000\times 0.05+0\times 0.20]\}=[9500 \quad 500]$$

Repitiendo la operación para el siguiente período, el vector inicial [10000 0] se reemplaza por [9500 500], que se vuelve a multiplicar por la matriz de probabilidades de transición. Así para el momento t=3, los 9500 niños sanos se transforman en 9025 sanos y 475 enfermos, pero al mismo tiempo un 80% de los 500 niños enfermos, es decir 400, se sana. De modo que en t=3 se espera que haya 9025+400=9425 niños sanos y 475+100=575 enfermos. La operación equivale a volver a multiplicar el vector inicial por la matriz de probabilidades de transición, o en otros términos, multiplicar el vector inicial por el **cuadrado** de la matriz:

$$\begin{bmatrix}9500 & 500\end{bmatrix}\times\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}=\begin{bmatrix}10000 & 0\end{bmatrix}\times\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}\times\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}=\begin{bmatrix}10000 & 0\end{bmatrix}\times\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}^2$$

Esta operación por lo tanto puede ser expresada en dos formas. La primera, que se indica en el miembro izquierdo de la ecuación precedente, consiste en multiplicar el vector [9500; 500] resultante de la primera operación por la matriz de probabilidades de transición.

$$\begin{bmatrix}9500 & 500\end{bmatrix}\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}=\begin{bmatrix}9500\times 0.95+500\times 0.80\\ 9500\times 0.05+500\times 0.20\end{bmatrix}=\begin{bmatrix}9425 & 575\end{bmatrix}$$

Equivalentemente, se puede elevar al cuadrado la matriz de probabilidades de transición, y luego multiplicar el vector inicial [10000; 0] por esa matriz elevada al cuadrado, como se indica en los miembros del centro y de la derecha de la penúltima ecuación. La matriz cuadrática es la siguiente:

$$\begin{bmatrix}0.95 & 0.05\\ 0.80 & 0.20\end{bmatrix}^2=\begin{bmatrix}0.95\times 0.95+0.05\times 0.80 & 0.95\times 0.05+0.05\times 0.20\\ 0.80\times 0.95+0.20\times 0.80 & 0.80\times 0.05+0.20\times 0.20\end{bmatrix}=\begin{bmatrix}0.9425 & 0.0575\\ 0.92 & 0.08\end{bmatrix}$$

El producto en este caso es:

$$\begin{bmatrix}10000 & 0\end{bmatrix}\times\begin{bmatrix}0.9425 & 0.0575\\ 0.92 & 0.08\end{bmatrix}=\begin{bmatrix}9425 & 575\end{bmatrix}$$

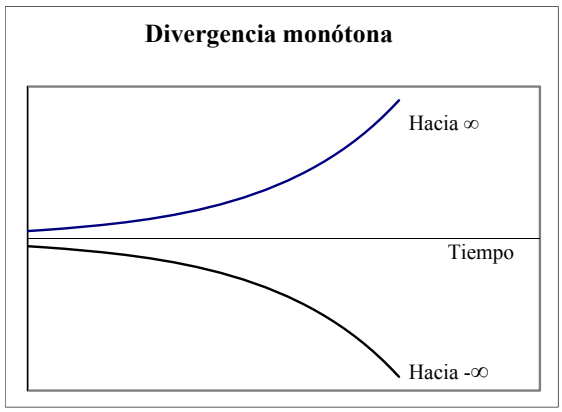
En el mes siguiente el proceso continuaría en la misma forma. Un 95% de los 9425 niños sanos (es decir 8954 niños) se espera que siga sano, y además un 80% de los 575 enfermos se sanará (añadiendo así otros 460 sanos), de modo que en t=4 habrá 8954+460=9414 niños sanos y consiguientemente 586 enfermos. El proceso podría continuar indefinidamente.

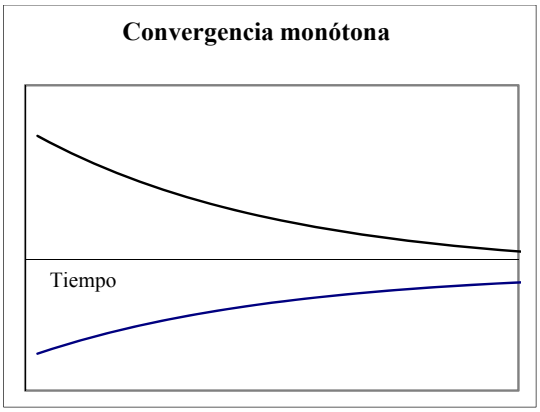
Este ejemplo sencillo demuestra que aun cuando las probabilidades de transición sean constantes, la distribución de la población entre los diferentes estados no resulta necesariamente constante. En este caso, el número de niños sanos va bajando desde los 10000 iniciales a 9500, luego 9425, luego 9414 y así sucesivamente. Nótese que en este caso el número de sanos va disminuyendo poco a poco, y además su disminución es cada vez menor: en el primer intervalo cae en 500 unidades (de 10000 a 9500), en el segundo cae 75 unidades (de 9500 a 9425), en el tercer período cae en solo 11 unidades (de 9425 a 9414). ¿Cómo continuaría esa evolución? ¿Se estabilizará en algún punto la proporción de sanos y enfermos? En la siguiente sección se analiza el patrón que siguen estas extrapolaciones o proyecciones si se las sigue un número suficiente de veces.

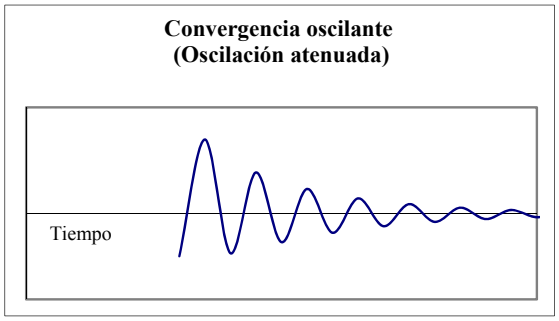
3.6. Convergencia y equilibrio

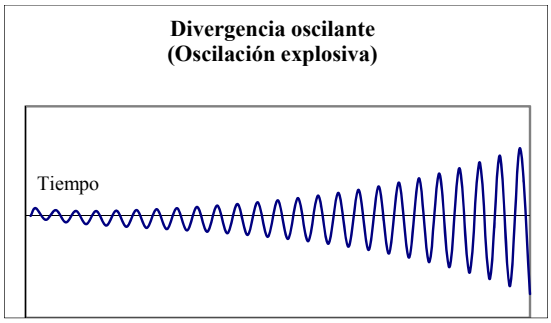
El proceso generado por la aplicación de una matriz de probabilidades de transición puede ser convergente, estable o divergente, y además puede ser monótono u oscilante, dando lugar a seis posibilidades. La siguiente tabla y los subsiguientes gráficos ilustran estas posibilidades (se omite el gráfico del equilibrio estable pues consistiría sólo en una línea recta horizontal).

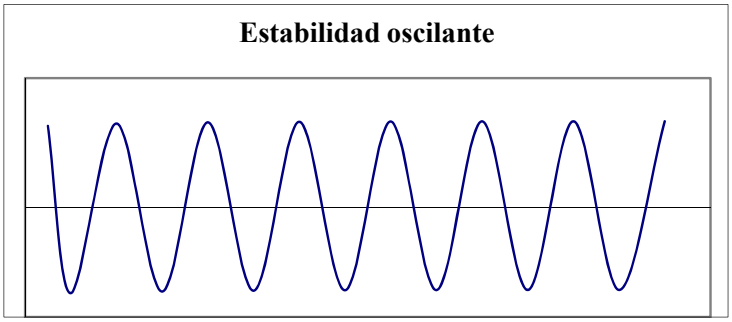
-	**Monótono**	**Oscilante**
Convergente	Convergencia monótona	Oscilación atenuada
Estable	Equilibrio estable	Equilibrio oscilante
Divergente	Divergencia monótona	Oscilación explosiva











Si un sistema se encuentra inicialmente en equilibrio, permanecerá en equilibrio, pues reproduce las magnitudes iniciales en todos los períodos subsiguientes. Si un sistema se encuentra inicialmente apartado de su punto de equilibrio, y el proceso es convergente, las magnitudes irán convergiendo hacia el equilibrio en forma gradual, ya sea en forma monótona o mediante una oscilación cada vez más atenuada. En el ejemplo anterior de los niños sanos y enfermos la convergencia se producía en forma gradual y monótona, sin oscilaciones.

Si la variable es de tal naturaleza que ella puede asumir cualquier valor, incluso valores negativos, un proceso divergente tiende a $+\infty$ o a $-\infty$, mediante una divergencia monótona o mediante una oscilación explosiva. En la práctica, sin embargo, esa divergencia indefinida no ocurre. Si se trata de tablas de rotación empíricas, las cantidades de sujetos en cada estado tienen que estar entre 0 y N, de modo que nunca podrían divergir hacia $\pm\infty$. Si la serie disminuye, termina cuando se llega a cero, y si aumenta la progresión acaba cuando se lega a N.

Cualquier proceso de este tipo, convergente o divergente, tiene un punto o situación de **equilibrio** (representado por la línea recta horizontal en los gráficos precedentes) Si el proceso es convergente, se llega gradualmente a esa situación, en la cual hay una cierta cantidad de casos en cada estado que se mantiene a lo largo del tiempo de allí en adelante. Si en el momento inicial reina ya la situación de equilibrio, las proporciones marginales se mantendrán indefinidamente. Si la situación inicial es otra, diferente a la de equilibrio, las proporciones irán evolucionando gradualmente, hasta alcanzar la situación de equilibrio si el proceso es convergente, o apartándose cada vez más del equilibrio si el proceso es divergente. Si el proceso es de oscilación estable, los valores oscilarán por encima y por debajo del valor de equilibrio, que en ese caso es simplemente el valor promedio de largo plazo de esos valores oscilantes.

Suele hacerse una distinción entre **equilibrio estable e inestable**. En realidad, el concepto de equilibrio como tal requiere estabilidad (un sistema en equilibrio sigue en equilibrio), pero esta distinción se refiere a lo que ocurre en caso de una perturbación externa. En un equilibrio estable, si se produce una perturbación del equilibrio el sistema retorna al equilibrio gradualmente. En un equilibrio inestable, cualquier perturbación conduce a una evolución divergente. Los sistemas convergentes tienen equilibrio estable, los divergentes tienen o pueden tener puntos de equilibrio inestables.

Equilibrio en este caso significa ausencia de cambios a lo largo del tiempo, pero más específicamente significa ausencia de cambios **agregados**: no necesariamente ausencia de cambios **individuales**. Aun en la situación de equilibrio habrá un cierto número de cambios de estado balanceados entre sí. Esto significa que, en equilibrio, las proporciones marginales son constantes, pero puede haber cambios de estado a nivel individual. En el ejemplo del examen médico de los niños, en una situación de equilibrio cada mes habrá una cierta cantidad de niños sanos que se enferman, pero se compensarán con una igual cantidad de niños enfermos que se sanan, de modo que el número de sanos y enfermos no variará de un mes al otro.

La condición que debe regir para que haya equilibrio agregado, pues, es que el número niños sanos que se enferman cada mes sea igual al número de niños enfermos que se sanan en el mismo mes. La cantidad esperada de niños sanos que se enferman es el producto del número de niños sanos por la probabilidad de que los niños sanos se enfermen, y de igual modo con la cantidad esperada de niños enfermos que se sanan. En equilibrio (con dos estados posibles) debe regir la siguiente igualdad:

$$N_s r_{se} = N_e r_{es}. \qquad \text{(Ec. 9)}$$

De esta relación necesaria de equilibrio entre los flujos opuestos de cambio de estado se desprende la proporción que deben tener entre sí las cantidades de sujetos en cada estado cuando se llegue al estado de equilibrio, cantidades que una vez alcanzadas serían constantes en el tiempo y se indican con un asterisco en lugar del superíndice t o $t+1$:

$$\frac{N_s^*}{N_e^*} = \frac{r_{es}}{r_{se}} \qquad \text{(Ec. 10)}$$

Las probabilidades de sanar y de enfermar son, en este ejemplo, 0.80 y 0.05 respectivamente. Transponiendo términos se tiene una expresión que permite relacionar esas probabilidades de transición con las cantidades de niños sanos y enfermos que debe haber en una situación de equilibrio:

$$\frac{N_s^*}{N_e^*} = \frac{r_{es}}{r_{se}} = \frac{0.80}{0.05} = 16$$

La razón de equilibrio entre niños sanos y niños enfermos debe ser igual a la razón entre la probabilidad de sanar y la probabilidad de enfermar. De ello se deduce el *número* de niños sanos y enfermos en la situación de equilibrio. Esto significa que en este ejemplo el proceso en cuestión alcanzaría el equilibrio cuando la cantidad de niños sanos sea dieciséis veces más alta que la cantidad de niños enfermos, porque la probabilidad de sanar (0.80) es dieciséis veces mayor que la probabilidad de enfermar (0.05). Si el número total de niños es diez mil, esto se logra con 9411.77 niños sanos y 588.23 niños enfermos. Como los niños no pueden existir en forma fraccionaria, las cifras se redondean a 9412 niños sanos y 588 enfermos.[16]

Una vez que llegue a esa situación de equilibrio la población **esperada** de niños sanos y enfermos permanecería en torno a 9412 niños sanos y 588 enfermos, con pequeñas fluctuaciones no significativas debidas al redondeo. En sólo tres rondas después de la ronda inicial el número de niños sanos en t=4 ya había descendido desde 10000 hasta 9414, cifra muy cercana a la de equilibrio (9412). Es obvio que en este ejemplo el equilibrio se alcanza muy rápidamente. De hecho se alcanzaría en el período siguiente. En efecto, en t=5 permanecerán sanos un 95% de los 9414 niños sanos de t=4, es decir 8943, y de los 586 enfermos sanará el 80%, es decir 469, dando un total de 9412 sanos y 588 enfermos. De allí en adelante el número de sanos permanecerá en 9412 y el de enfermos en 588, salvo pequeñas fluctuaciones por redondeo. La siguiente secuencia muestra los resultados de las seis primeras observaciones mensuales, de t=1 hasta t=6. En esta secuencia se advierte que inicialmente los individuos que cambian de estado en un sentido no concuerdan con los que cambian de estado en sentido contrario. Pero posteriormente ambos flujos convergen a alrededor de 470 en cada sentido: 470 sanos se enferman, y 470 enfermos se sanan, de modo que la distribución global no cambia. Las celdillas sombreadas indican los flujos de cambio de estado en cada una de las fases mensuales, que se igualan entre sí al llegar a la situación de equilibrio.

	t = 2			**t = 3**			**t = 4**			**t = 5**			**t = 6**	
t=1	**S**	**E**	**t=2**	**S**	**E**	**t=3**	**S**	**E**	**t=4**	**S**	**E**		**S**	**E**
S	9500	500	**S**	9025	475	**S**	8954	471	**S**	8943	471		8942	470
E	-	-	**E**	400	100	**E**	460	115	**E**	469	117		470	118
	9500	500		9425	575		9414	586		9412	588		9412	588

Este ejemplo presuponía que al inicio los diez mil niños estaban sanos. Si en el momento inicial todos los niños hubiesen estado enfermos, o hubiese habido cualquier otra distribución intermedia, el proceso de convergencia hubiese sido similar. Cada mes se curaría el 80% de los enfermos y se enfermaría el 5% de los sanos, convergiendo al mismo estado de equilibrio al cabo de algunas rondas. La cantidad de rondas necesarias para alcanzar el equilibrio con suficiente exactitud dependerá de cuáles sean las condiciones iniciales, y de cuán lejos estén del equilibrio. En el ejemplo se partía de 10000 sanos cuando en equilibrio debían ser 9412, de modo que ya la situación inicial estaba

---

[16] La ecuación 10 sólo suministra el **cociente** entre las cantidades de sujetos que (en equilibrio) deben estar en cada estado. Para encontrar los valores absolutos de N_s y N_e , que en este caso son 9412 y 588 respectivamente, se debe usar también la información sobre el número total de sujetos, N, en este caso 10000. Para ello se formula el problema como un sistema de dos ecuaciones lineales con dos incógnitas:

$$N_s + N_e = N = 10000$$

$$N_s = 16N_e$$

De estas ecuaciones se desprende la expresión que permite obtener la proporción de sanos o de enfermos en la situación de equilibrio (indicada con un asterisco):

$$p_s^* = \frac{N_s}{N_s + N_e} = \frac{N_s}{N} = \frac{r_{es}}{r_{se} + r_{es}}$$

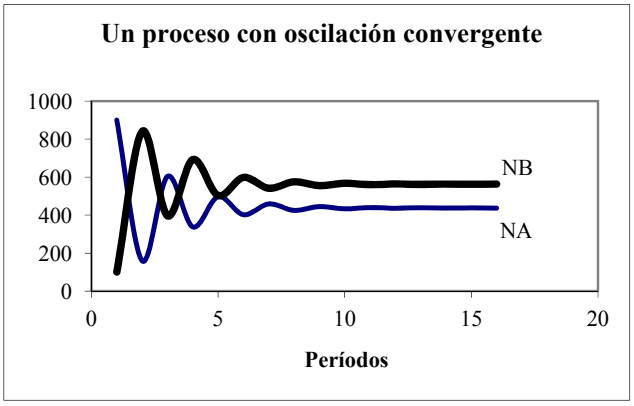
$$p_e^* = \frac{N_e}{N_s + N_e} = \frac{N_e}{N} = \frac{r_{se}}{r_{se} + r_{es}}$$

cerca del equilibrio. Si se hubiese partido de una situación muy alejada del equilibrio, por ejemplo con todos (o casi todos) los niños enfermos, el número de rondas necesario para llegar a la distribución de equilibrio (9412 niños sanos) hubiera sido mayor.

En este ejemplo, la convergencia hacia el estado de equilibrio es monótona, es decir que procede en forma directa y gradual: las cantidades se van aproximando gradualmente a su punto de equilibrio sin ninguna oscilación. Eso es lo que ocurre cuando las probabilidades de transición más elevadas están situadas en la diagonal principal, es decir cuando hay una correlación o asociación directa entre el estado antecedente y el estado consecuente. Si todos los niños empezaran sanos, pero las probabilidades mayores estuviesen concentradas en la diagonal secundaria, es decir, cuando hay una asociación **inversa** entre el estado antecedente y el consecuente, el proceso de convergencia hacia el punto de equilibrio puede ser **oscilante**, aunque la amplitud de las oscilaciones va disminuyendo a medida que se avanza hacia el equilibrio. Por ejemplo supóngase un proceso con dos estados A y B, cuya distribución inicial es N_A=900 y N_B=100, y cuya matriz de probabilidades de transición es:

$$R = \begin{bmatrix} 0.1 & 0.9 \\ 0.7 & 0.3 \end{bmatrix}$$

En este caso, la aplicación reiterada de estas probabilidades al vector inicial arroja la siguiente evolución en cuanto a la cantidad de gente en los dos estados. La tabla y gráfico siguientes ilustran la evolución de las cifras durante 16 períodos, evidenciando que el proceso converge de manera oscilante hacia sus dos valores de equilibrio. Aquí se puede ver cómo las cantidades oscilan con amplitud decreciente hasta converger a la distribución de equilibrio que se sitúa alrededor de un 43.7% en el estado A y un 56.3% en el estado B. Al llegar al período 16 el sistema ya está prácticamente en equilibrio, variando sólo marginalmente con cambios que no se pueden distinguir de los errores de redondeo.



Oscilación convergente		
Período	N_A	N_B
1	900	100
2	160	840
3	604	396
4	338	692
5	497	503
6	402	598
7	459	541
8	425	575
9	445	555
10	433	567
11	440	560
12	436	564
13	439	561
14	437	563
15	438	562
16	437	563

Nótese también que en el ejemplo anterior la situación inicial [10000; 0] está muy cerca de la situación de equilibrio [9412; 578], y por consiguiente el equilibrio se alcanzaba en sólo tres o cuatro pasos. En cambio en el ejemplo último la situación inicial [900; 100] está muy alejada y hasta en relación inversa de las proporciones de equilibrio [437; 563] por lo cual alcanzar el equilibrio requiere muchos más pasos intermedios (en este caso 14 pasos).

3.7. Evaluación empírica del ajuste del modelo de Markov

No todos los datos de panel se ajustan a las previsiones de un modelo de Markov. Si se desea aplicar ese modelo será necesario verificar si se cumplen sus supuestos. El modo de hacerlo es verificar si las cifras esperadas (generadas por el modelo) coinciden con las observadas.

Es frecuente que se cuente con un panel breve, en el que se observen sólo dos o tres períodos. A menudo se usan los dos primeros períodos para estimar las probabilidades de transición, que luego se aplican para generar las cifras esperadas de los períodos subsiguientes. Pero no hay nada especial a priori en los dos primeros períodos, o en cualquier par de períodos que se elija para basar la estimación de las probabilidades. La estimación de las probabilidades de transición podría basarse, en realidad, en cualquier par de períodos arbitrariamente elegidos, o en un promedio de todos ellos. El modelo implica que esas probabilidades son constantes, de modo que **si el modelo de Markov es válido** todas esas variantes deberían arrojar más o menos los mismos resultados. La evaluación suele concentrarse en tres preguntas cruciales: (1) ¿Se cumplen los supuestos de homogeneidad, constancia y "amnesia"? Es decir: ¿Obedece el proceso al modelo markoviano con probabilidades de transición homogéneas y constantes? y (2) ¿Se encuentra el proceso en una situación de equilibrio?

3.7.1. La contrastación empírica de los supuestos de Markov

La hipótesis de la **constancia** puede ser puesta a prueba si se dispone de varios períodos de observación: si esa hipótesis fuese cierta, las diferencias de las proporciones de flujo entre períodos tendrían que ser poco significativas, atribuibles a errores de muestreo. En el caso de las variables de empleo, que están fuertemente influidas por los ciclos económicos, es difícil presuponer esa constancia. Pero algunas otras transiciones (por ejemplo los cambios de estado civil de la población, o las transiciones de los estudiantes entre distintos niveles del sistema educativo) probablemente sean más constantes. La hipótesis de "amnesia" también puede ser verificada comparando las proporciones de flujo para sujetos con diferentes experiencias anteriores (por ejemplo la probabilidad de perder el empleo por parte de personas con o sin experiencias recientes de desocupación).

En general ninguna de las hipótesis (homogeneidad, constancia y amnesia) se cumplirá. Pero ellas no están ahí para cumplirse, sino para determinar un modelo de "línea de base", al cual luego se le incorporan variables para explicar las diferencias entre ese modelo y la realidad. Por ejemplo, una variable del ciclo económico (como la tasa de crecimiento del PBI) puede usarse para como indicadora del contexto macro para separar el componente "constante" y el componente "cíclico" en la transición entre ocupación y desocupación o viceversa. Según esa clase de enfoque, en cada período hay dos clases de transiciones entre empleo y desempleo: algunas de esas transiciones son "normales" y ocurren en todos los períodos; otras son inducidas por la situación económica general, de modo que aumentan las transiciones hacia el desempleo en épocas de recesión, y disminuyen en épocas de expansión económica.

Este enfoque se relaciona estrechamente con el que Coleman llama "método de los residuos" (Coleman, 1964b, capítulo 15). Aquí no se supone que tenga que haber un solo valor normal: puede haber datos que permitan esperar diferentes valores para diferentes individuos. La idea general de ese método consiste en generar un modelo teórico donde los valores esperados reflejen el efecto de variables ya conocidas, restringiendo la investigación a las **desviaciones** o **residuos** respecto a lo que debería observarse en el caso normal. Por ejemplo, un factor que puede influir muchas variables económicas es el nivel de ingreso. Si la conducta de los individuos, por ejemplo sus hábitos de consumo, responde nítidamente a diferencias en sus ingresos, no hay mucho más que decir, pues ya se sabe de antemano que el ingreso tiene ese efecto. Pero si se observan individuos con fuertes desviaciones respecto a esa explicación basada en datos conocidos, será necesario encontrar otras explicaciones. La investigación, según este enfoque, debe descartar primero los factores ya conocidos, para luego concentrarse en los residuos de origen no evidente o no trivial. No se supone que el ingreso sea el único factor en juego, pero se sabe que ejerce un cierto influjo, entonces primero se descuenta ese influjo para poner en evidencia la presencia de otros factores.

De este modo, no importa si el modelo teórico (que solo considera el ingreso) explique acabadamente la realidad o no: es posible que el modelo se use sólo para "descartar" o "descontar el efecto" de una explicación trivial cuyo efecto pueda ser cuantificado y estimado separadamente, dejando los residuos o desviaciones como fenómenos a explicar. Una aplicación posible de este método suele aparecer en los estudios de psicología económica o de sociología económica. Se procura primero explicar la conducta de los sujetos (por ejemplo consumidores) mediante hipótesis basadas en el modelo microeconómico convencional de maximización de utilidades: se aplica un modelo que supone agentes económicos racionales maximizadores de utilidad; si esos supuestos explican los fenómenos observados dejando residuos no significativos, allí acaba la investigación; si en cambio hay residuos significativos, ello indica que no se cumple alguno de los supuestos del modelo microeconómico convencional, y ello origina un interesante programa de investigación.

Si se adopta ese punto de vista, un modelo extremadamente sencillo, como los más simples modelos de Markov, puede servir para generar una línea de base, explicando "trivialmente" la conducta de los individuos por el supuesto de que todos ellos poseen probabilidades homogéneas y constantes de pasar de un estado a otro. Descontando el efecto esperado de ese modelo permite concentrar luego la investigación en las desviaciones observadas respecto a esa explicación trivial. En líneas generales, entonces, el análisis markoviano parte de un "modelo básico" con probabilidades homogéneas, constantes y sin memoria, y luego incorpora factores que puedan explicar desviaciones entre la realidad y ese modelo básico.

La "memoria" de los sujetos respecto a su experiencia anterior también puede afectar su comportamiento tanto como la heterogeneidad de los sujetos en cuanto a determinados factores diferenciadores o el cambio de las probabilidades a través del tiempo. Al incorporar el factor memoria o experiencia anterior pueden originarse diversas líneas interpretativas. Una de ellas es la hipótesis de la **autoselección**: la experiencia de haber estado desocupado no sería la "causa" de que el sujeto pase nuevamente a la desocupación, sino sólo un indicador de su tendencia a quedarse sin empleo. Si el individuo cayó ya alguna vez en el desempleo, "por algo habrá sido", y se presume que ese "algo" (aunque no se conozca su naturaleza) hará que tenga más chance de quedar desocupado más tarde.

Otra es la hipótesis del **aprendizaje**, que puede ser positivo o negativo. El haber tenido una experiencia aumenta o disminuye la probabilidad de volverla a tener. El **aprendizaje negativo** reduce la posibilidad de volver a caer en situaciones desagradables cuando el sujeto ya ha tenido la experiencia, y podría llamarse "hipótesis del gato escaldado": si un individuo una vez perdió su empleo, la próxima vez se cuidará más de no perderlo, y por lo tanto su chance de quedar sin trabajo será menor que de la de sus colegas que aun no han pasado por la experiencia de la desocupación. El **aprendizaje positivo** incrementa la posibilidad de pasar a estados agradables o deseables para los individuos que ya tuvieron esa experiencia positiva en el pasado. Por ejemplo: si un individuo estuvo desocupado en el pasado y luego encontró empleo, la próxima vez que caiga en la desocupación tendrá más probabilidad de encontrar empleo que los que están desempleados por primera vez, por su mayor experiencia anterior en la búsqueda de trabajo. Discriminar cuál de estas posibles explicaciones es la que vale (autoselección, aprendizaje positivo o negativo) es un válido problema de investigación que puede ser enfocado mediante el análisis de panel.

Si se dispone al menos de tres observaciones se podría poner a prueba la hipótesis, más importante, de que las cifras observadas son efectivamente generadas por un proceso subyacente de tipo markoviano, caracterizado por **probabilidades constantes de transición** entre un estado y otro. Los datos de los dos primeros períodos observados se pueden usar para estimar las probabilidades de transición, y con ellas predecir las cifras del tercer período; a la inversa, las probabilidades podrían estimarse a partir del segundo y tercer período para luego estimar retrospectivamente el primero. Si esas cifras esperadas coinciden cercanamente con las observadas, no se podría rechazar la hipótesis de que los datos hayan sido generados por un proceso subyacente del tipo que se ha expuesto. Esta comprobación adquiere más fuerza si se tienen cuatro o más períodos de observación, y las cifras esperadas coinciden cercanamente con las observadas para todos ellos. En caso que esas predicciones se aparten fuertemente de la realidad observada, será necesario otro modelo que contemple otros factores.

Una de las maneras de tratar este tema consiste en introducir **otras variables** que ayuden a predecir mejor los cambios, lo cual de hecho hace que las probabilidades de transición dejen de ser constantes y se vayan modificando con el correr del tiempo. Esto equivale a explorar la posibilidad de que las probabilidades de transición sean **heterogéneas**, es decir, que algunos subgrupos de individuos tengan probabilidades diferentes a otros subgrupos. Por ejemplo, podría ocurrir que las probabilidades de caer en el desempleo o de encontrar empleo varíen en función de la edad, el sexo o el nivel educativo de los trabajadores. En tal caso, la proporción de sujetos que pasa del estado i al estado j no sería una estimación correcta de r_{ij} para todos los sujetos, sino sólo el promedio ponderado de las probabilidades heterogéneas que están operando. Por otra parte, al actuar esas distintas probabilidades la composición de los contingentes en cuanto a esos factores va cambiando, y en consecuencia también cambiarán las proporciones promedio de pasaje de un estado a otro. Si en un panel con tres o más períodos se observa que las proporciones de pasaje van variando, tal vez esas variaciones puedan ser explicadas al subdividir el grupo total de acuerdo a variables relevantes.

Otro enfoque consiste en utilizar el concepto de "cohorte teórica" en lugar de hacer siempre referencia a cohortes empíricas o reales. Este enfoque en realidad no resuelve el problema sino que lo ignora, sin comprobar en momento alguno la adecuación del modelo de Markov, y sin comprobar la existencia o no de una situación de equilibrio. El enfoque de cohorte teórica toma las probabilidades obtenidas a partir de **diferentes cohortes empíricas** en un determinado intervalo, o en el promedio de varios intervalos adyacentes, y las utiliza para estimar la probable evolución de **una cohorte teórica** que a lo largo del tiempo estuviese sometida a las probabilidades estimadas a partir del panel. El uso más frecuente de las cohortes teóricas es la construcción de tablas de vida o tablas de supervivencia. Pueden referirse al evento consistente en morir, o a eventos menos trascendentales como ingresar en la actividad económica o comprar un auto. Para el caso de la mortalidad se miden en un período determinado las tasas de mortalidad por edad y con ellas se estiman las probabilidades de morir en cada edad o tramo de edades (es decir las probabilidades de pasar del estado "Vivo a la edad t" al estado "Muerto antes de la edad $t+k$"), las cuales se aplican luego a una cohorte teórica, como si esas tasas afectaran a esa población teórica a lo largo de su vida. En este

caso el "tiempo" es reemplazado por la "edad" (tiempo transcurrido desde el nacimiento) pero es siempre un caso de análisis longitudinal.

Aplicar las tasas de mortalidad que hoy afectan a personas de diferentes edades, a una cohorte de personas teóricas es un artificio teórico. En realidad, los niños de hoy nunca van a ser afectados por la mortalidad que hoy sufren los ancianos, pues cuando esos niños lleguen a la vejez las tasas de mortalidad en la vejez serán otras (posiblemente inferiores). Igualmente, los ancianos de hoy nunca fueron afectados por la mortalidad infantil de hoy, pues cuando ellos eran niños las tasas de mortalidad infantil eran otras (probablemente muy superiores a las de hoy). La expectativa de vida que surge de una tabla de mortalidad equivale a la cantidad de años vividos por una población teórica sometida a lo largo de su vida a las tasas de mortalidad por edad vigentes hoy, pero no puede aplicarse directamente para estimar la cantidad de años que van a vivir los niños nacidos hoy. La mayoría de los modelos sobre "historias de eventos" se basan en este concepto de cohorte teórica.

Un tercer enfoque postula un proceso mixto. Los sujetos, de acuerdo a este enfoque, están guiados por dos procesos diferentes y simultáneos: por una parte, ciertas probabilidades homogéneas y constantes de transición, es decir una matriz markoviana de probabilidades de transición; y por otro lado, un proceso de shocks aleatorios que lo llevan a dar respuestas diferentes a las que corresponderían al proceso markoviano. El problema del analista es distinguir los verdaderos cambios de estado del sujeto (supuestamente gobernados por un proceso de Markov) de los cambios aparentes o momentáneos inducidos por esos factores aleatorios.

Para poder tratar con mayor rigor estos problemas es necesario pasar más allá de los datos de panel a fin de postular un modelo teórico sobre los procesos no observables que tienen lugar en el nivel de los individuos, y que estarían generando los datos observados. Estos modelos teóricos, por lo general, presuponen procesos que operan en forma **continua**, aun cuando el panel se repita a intervalos discretos en el tiempo. A ese tema se dedica el capítulo siguiente de este texto. Antes de ello, la sección que sigue se dedica a la evaluación de los datos de panel en términos de equilibrio o desequilibrio de corto plazo.

3.7.2. Equilibrio y desequilibrio de corto plazo

Hay procesos que a priori se supone que no tienen ninguna tendencia definida, sino que se mantienen estables o casi estables. Esto significa que el proceso se encuentra permanentemente en equilibrio o muy cerca del equilibrio. En ese caso, cualquier par de períodos que se elija daría más o menos los mismos resultados. En otros casos existen **tendencias** de largo plazo o **movimientos cíclicos** que van modificando la distribución de la población, y por lo tanto los resultaos serían distintos según el par de períodos que se escoja. Por ejemplo, esto podría ocurrir en el caso del desempleo: en un período de expansión económica posiblemente sean más los desempleados que encuentran trabajo que las personas ocupadas que pierden su empleo, y a la inversa en un período de recesión. En un país cuyo sistema educativo está aumentando su eficacia, la proporción de desertores tenderá a disminuir con el tiempo, y por lo tanto la probabilidad de desertar observada hoy posiblemente sobreestime las deserciones escolares de mañana. La probabilidad de morir a determinada edad, registrada en un cierto período, posiblemente disminuya con el tiempo del mismo modo, si en el país hay una tendencia al descenso de las tasas de mortalidad.

Un panel con solo dos observaciones permite poner a prueba la hipótesis de que el proceso se encuentra ya en equilibrio. Si así fuese, **las frecuencias marginales deberían ser constantes** en los dos períodos, de modo que el número total de sujetos en cada uno de los estados debería ser invariable (o casi invariable) en los sucesivos períodos. Obviamente no es imprescindible que las frecuencias marginales en t y en $t+1$ sean **exactamente** iguales. Cualquier ajuste empírico debe contemplar la posibilidad de errores aleatorios, producidos por fluctuaciones al azar de los propios sujetos o por errores de medición. Por tal razón, verificar si hay equilibrio equivale a rechazar la hipótesis nula de que las dos distribuciones marginales provienen de poblaciones diferentes entre sí, a un cierto nivel de significación estadística. Esta prueba se realiza usualmente con los tests estadísticos más habituales como el χ^2 (chi cuadrado). Para ello se consideran las frecuencias de $t+1$ como frecuencias "observadas" y las de t como "esperadas". Si la variable tiene k categorías, la

evaluación se realiza con el test χ^2 con ***k-1*** grados de libertad, al nivel de significatividad elegido (usualmente p=0.95 o bien p=0.99. Dado que la hipótesis nula es que ambas distribuciones son **diferentes**, ella es rechazada cuando χ^2 es **inferior** al valor crítico, y aceptada cuando es superior a él (lo contrario de lo que ocurre usualmente en este tipo de prueba estadística). La prueba χ^2, sin embargo, presupone una distribución normal de errores de muestreo, y por lo tanto puede ser usada solamente en muestras al azar suficientemente grandes. Para muestras pequeñas o no aleatorias se deberían usar tests no paramétricos.

4. Procesos continuos con variables categóricas

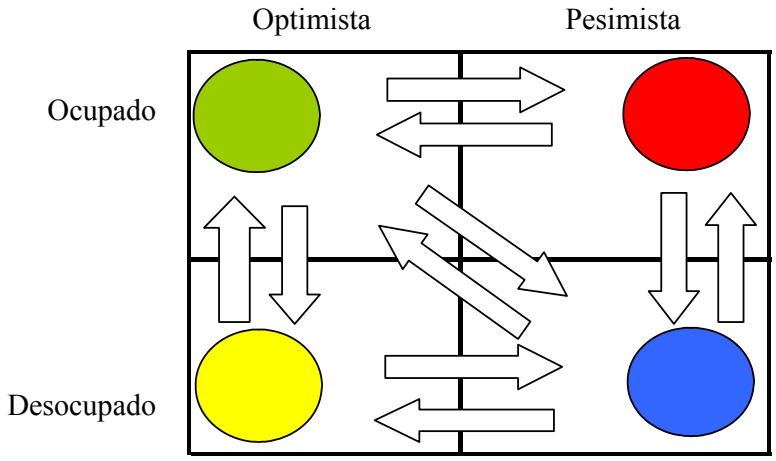
Los modelos de Markov se basan en variables de tipo categórico o cualitativo, y por lo tanto suponen que existen varios estados claramente distintos (es decir, **estados discretos**). Al mismo tiempo esos modelos markovianos únicamente registran el estado de las unidades de análisis en determinados momentos separados por un intervalo. El proceso de Markov explica los cambios ocurridos entre dos (o más) rondas, de modo que **también el tiempo es tratado de hecho como una variable discreta**. No hace referencia al período intermedio.

Hay procesos que tienen, sin duda, un carácter discreto. La agricultura, por ejemplo, es un proceso anual: en cierto momento se siembra, y en cierto momento se cosecha. No hay una producción continua ni una siembra continua. Lo mismo pasa con el progreso de los estudiantes a través de los grados de la educación formal: en un cierto mes comienzan las clases, y en cierto mes termina el año escolar y se otorgan certificados de aprobación de cada nivel de enseñanza. No hay un proceso permanente de ingreso de alumnos ni un proceso permanente de egreso y graduación, sino que esos eventos ocurren sólo en determinados momentos del año (dejamos aquí de lado los traslados de alumnos de una escuela a otra, que pueden producirse en cualquier momento, y contemplamos solo el sistema en su conjunto).

Pero no todos los procesos ocurren en fases discretas. En muchos casos el proceso subyacente sólo puede ser concebido como continuo. Por ejemplo, el proceso de conseguir empleo o perderlo es un proceso continuo: en cada instante puede haber individuos que pasan de ocupados a desocupados o viceversa, que deciden entrar en la fuerza de trabajo o salir de ella. Los niños pueden pasar de sanos a enfermos, o viceversa, en forma continua. Si el panel registra el estado de los sujetos en dos momentos cualesquiera ***t*** y ***t+k*** debe suponerse que la diferencia que se registra es el resultado neto de todos los cambios de estado ocurridos en ese intervalo.

Los cambios agregados que se observan entre dos rondas son el resultado de una acumulación de cambios individuales. En tal sentido, el modelo se refiere primariamente a los procesos que ocurren a nivel individual, cuyos parámetros, a menudo inobservables, son estimados a partir de los datos del panel; posteriormente, las proporciones y magnitudes de los flujos agregados que se acumulan en los períodos empíricos, de longitud arbitraria, entre dos rondas del panel, se explican a partir de los parámetros del modelo. Estos flujos individuales pueden ocurrir a intervalos discretos (por ejemplo, podrían ser datos anuales sobre inscripción y graduación de alumnos), o bien en forma continua.

En este último caso, en cada instante del tiempo habrá individuos que pasan de un estado al otro, en diferentes proporciones. Si hay sólo dos variables, cada una con dos estados posibles, el siguiente gráfico ilustra las cuatro posibles combinaciones de valores, y los ocho diferentes flujos que pueden presentarse.



Supongamos que la dimensión vertical representa la situación de empleo (ocupado o desocupado), y la dimensión horizontal representa una actitud psicológica cualquiera (por ejemplo optimismo o pesimismo sobre su propio futuro). Un individuo situado en el casillero superior izquierdo (ocupado y optimista) tiene en todo momento una cierta probabilidad de pasar al casillero inferior izquierdo (desocupado y optimista) o al casillero superior derecho (ocupado y pesimista). También puede pasar al casillero inferior derecho (desocupado y pesimista). Similares transiciones pueden ocurrir a partir de cualquiera de los otros casilleros. Dado que se trata de procesos continuos, en todo momento habrá algunos individuos que pierden o recuperan su empleo, e individuos que cambian de actitud. La proporción de individuos a los que les ocurre cada una de esas transiciones por unidad de tiempo se denomina **tasa de transición**, y puede ser calculada para cualquier intervalo de tiempo. Cuando el intervalo se reduce a un instante, se la denomina **tasa instantánea de transición**.

Las tasas de transición **no son probabilidades**. En efecto, pueden tener valores superiores a la unidad, pues en una unidad de tiempo los individuos pueden tener en promedio más de una transición. Por ejemplo, supongamos un grupo de trabajadores eventuales que consiguen empleo o no lo consiguen en cada uno de los días de la semana (cosa que a veces ocurre, por ejemplom, con los estibadores portuarios y otros). Hay entonces una rapidísima rotación laboral, de modo que los individuos en general toman empleos y los pierden con gran frecuencia. La situación del individuo en dos encuestas sucesivas tomadas en los días t y $t+h$ ya no representa una transición directa o una permanencia en el mismo estado. Cada individuo podría haber cambiado de situación varias veces. La cantidad de veces que un individuo pasa de ocupado a desocupado por unidad de tiempo podría ser superior a 1, y si eso se cumple para todos los individuos, la tasa de transición (que se refiere al total de esos trabajadores) sería superior a 1, lo cual sería imposible si se tratara de probabilidades. Las probabilidades varían entre 0 y 1. Las tasas de transición varían entre cero e infinito.

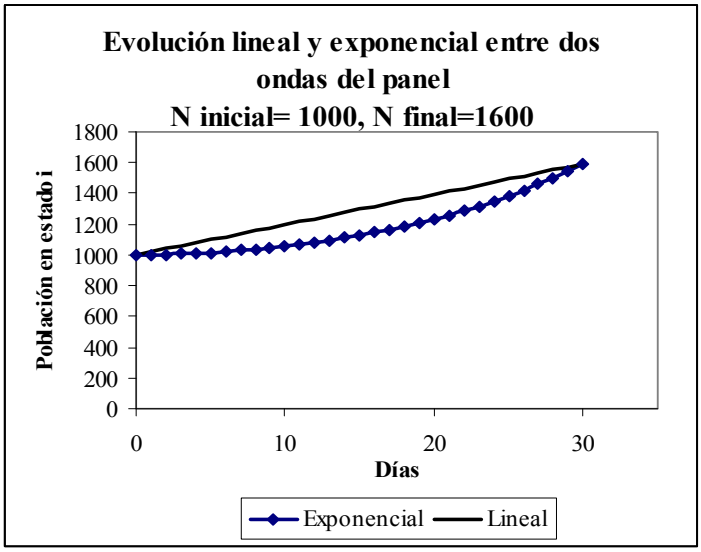
4.1. Tasas instantáneas de transición

La arbitraria longitud del intervalo entre rondas hace que la estimación de la tasa de transición entre rondas encierre un elemento de arbitrariedad. Si la medimos por semana tendremos valores más bajos que si la medimos por mes o por año. Obviamente, al aumentar la longitud del intervalo es probable que aumente la proporción de sujetos que ha cambiado su estado, pero ello no significa que se haya modificado la probabilidad subyacente de cambiar de estado, o la velocidad con que los individuos cambian de estado. Por ejemplo, la proporción de sujetos casados que se divorcian antes de la ronda siguiente será mayor cuanto más largo sea el intervalo entre las rondas, lo mismo que la cantidad de solteros que contrae matrimonio. Si se prolonga el intervalo la tasa bruta de transición entre solteros, casados y divorciados (proporción de casos que cambia de estado) parecerá aumentar, aunque la tasa **anual** o **mensual** de divorcios y casamientos no haya cambiado en absoluto.

Para superar esta arbitrariedad se puede comenzar pensando en la **trayectoria** de las magnitudes a lo largo del período intermedio entre dos rondas. Si en abril hubo 1000 desocupados, y en octubre hubo 2000, la trayectoria puede haber seguido cualquier patrón: pudo haber aumentado hasta 3000

en agosto para luego descender hasta la cifra registrada en octubre, o por el contrario pudo haber bajado de 1000 a 500 y luego subir, o cualquier otra evolución concebible. En ausencia de otros datos, usualmente se suele suponer una evolución **regular o monótona**, es decir un aumento o descenso continuo durante el intervalo. La forma de esa evolución puede ser rectilínea o curvilínea.

En el gráfico siguiente aparece así la cantidad de sujetos en el estado ***i*** en dos rondas (1 y 2), y dos posibles trayectorias de los cambios ocurridos durante el intervalo intermedio. Supóngase que el tiempo se mide en unas unidades de medida cualesquiera, por ejemplo días, y que las dos rondas están separadas por un intervalo ***k*** equivalente a una cierta cantidad de días, por ejemplo un mes (***k*** = 30 días). El panel ha revelado que en un cierto estado ***i*** había 1000 personas en la primera ronda, y 1600 en la ronda siguiente realizada 30 días después de la primera. A partir de allí se puede hipotetizar la trayectoria de esa cifra en el período intermedio.



El gráfico muestra dos posibles trayectorias de esta población a lo largo del mes, entre su estado inicial y su estado final. El aumento ha ocurrido durante ese mes, pero se desconoce cómo se distribuyó a lo largo del mes. Para estimar la trayectoria usualmente (a falta de otros datos) se supone un cambio gradual. Este puede consistir en un **número constante** de aumento por día (interpolación lineal) o una **proporción constante** de aumento por día (interpolación exponencial).

En la interpolación lineal se supone que el número de sujetos en el estado ***i*** aumenta de manera lineal, con un **incremento fijo en términos absolutos** por unidad de tiempo. En la interpolación exponencial se supone que el número de sujetos en el estado ***i*** aumenta en una **proporción fija** por unidad de tiempo. Las ecuaciones que representan ambas trayectorias a lo largo de un intervalo de longitud ***h*** son las siguientes, que estiman el número de sujetos en el estado ***i*** en cualquier instante intermedio ***t+k*** del intervalo total ***t+h*** , donde ***k<h***.

Incremento lineal: $$N_i^{t+k} = N_i^t + bk = N_i^t + \left[\frac{N_i^{t+h} - N_i^t}{h}\right]k \qquad \text{(Ec.11)}$$

Incremento exponencial: $$N_i^{t+k} = N_i^t(1+c)^k = N_i^t\left[\sqrt[h]{\frac{N_i^{t+h}}{N_i^t}}\right]^k \qquad \text{(Ec. 12)}$$

Los dos parámetros fundamentales son, en el caso lineal, el incremento fijo ***b*** por unidad de tiempo, que en nuestro ejemplo es simplemente el incremento mensual absoluto dividido por el número de días en el mes; y en el caso exponencial el factor diario acumulativo de incremento ***1+c***, que es igual a la raíz 30-ésima del incremento relativo mensual para meses de 30 días (en otros meses será 28 o 31). Estas ecuaciones generan la cantidad estimada de personas en el estado ***i*** en cada uno de los días del mes transcurrido (***k*** = 1, 2, 3, ... , 30). En el gráfico precedente las dos trayectorias arrojan valores diferentes de N_i a lo largo del mes, especialmente hacia la mitad del mes cuando

ambas líneas están más separadas. Esta diferencia será más significativa cuanto más grande sea el incremento mensual. Cuando la diferencia entre la cifra inicial y la final es relativamente pequeña, las dos trayectorias tenderán a ser muy coincidentes entre sí. Si el incremento entre rondas es más acentuado, la trayectoria exponencial tenderá a apartarse más perceptiblemente de la lineal.

Por supuesto, no sabemos usualmente si la trayectoria fue lineal, exponencial o de otro tipo. El número de personas en el estado ***i*** pudo haber aumentado en forma gradual, o pudo haber oscilado ampliamente encima y abajo del valor inicial. Puede haber estado constante durante casi todo el intervalo, cambiando poco antes del final; o podría haber seguido cualquier otra trayectoria intermedia. Cuanto más breve sea el intervalo entre las observaciones, más improbables serán las trayectorias con grandes variaciones u oscilaciones, y más cercana será la trayectoria lineal con la exponencial.

El intervalo entre dos rondas de una encuesta de panel es en general arbitrario. En un proceso de cambio continuo no hay en principio nada especial en un intervalo de un mes, o de cuatro, seis u ocho meses.[17] Los parámetros de un modelo teórico sobre un proceso subyacente de cambio que opera en forma continua a nivel de los individuos no pueden depender de la elección arbitraria de un intervalo: ninguna teoría significativa sería capaz de incorporar como parámetro teórico la probabilidad de quedar desocupado en un lapso de, por ejemplo, ocho meses. En cambio, muchos modelos teóricos se basan en el supuesto de un **proceso continuo**, que opera en cada instante del tiempo: los cambios operados en un intervalo entre rondas puede considerarse como el **efecto acumulado de un proceso continuo** que estuvo funcionando todo el tiempo durante todo ese intervalo.

Para formular un modelo de cambio continuo, con estados discretos y tiempo continuo, se necesita definir una magnitud llamada "intensidad de transición" o "tasa instantánea de transición" (véase Coleman 1964b, cap. 3 a 5). La intensidad de transición entre dos estados ***i*** y ***j*** se denota como q_{ij}. Las tasas instantáneas se definen a partir de las tasas de transición acumuladas durante un cierto período de longitud ***h***, denotadas como $u_{ij}^{t,t+h}$.

La tasa **instantánea** de transición q_{ij} puede ser definida a partir de la tasa de transición acumulada a lo largo de un intervalo, **suponiendo un flujo exponencial a lo largo del tiempo.** Denotando como $u_{ij}^{t,t+k}$ la tasa proporcional de transición desde ***i*** hacia ***j*** acumulada a lo largo de un intervalo de longitud ***h***, la tasa por unidad de tiempo será $u_{ij}^{t,t+k} / h$. Las tasas instantáneas de transición desde ***i*** hacia ***j*** se obtienen como límite cuando ***h*** tiende a cero:

$$\lim_{h\to 0} \frac{u_{ij}^{t,t+h}}{h} = q_{ij} \quad (i \neq j) \qquad \text{(Ec.13)}$$

Dado que ***h*** es un intervalo de tiempo, es fácil advertir que las q_{ij} son las derivadas de las tasas $u_{ij}^{t,t+h}$ respecto al tiempo:[18]

$$\lim_{h\to 0} \frac{u_{ij}^{t,t+h}}{h} = \frac{du_{ij}^{t,t+dt}}{dt} = q_{ij} \qquad (i \neq j) \qquad \text{(Ec.14)}$$

Esto significa que la tasa instantánea de transición es igual al incremento de la tasa de transición durante el intervalo ***h***, cuando ***h*** aumenta en una magnitud infinitesimal. En el caso de la tasa de transición q_{ii}, que implica permanecer en el estado ***i***, la tasa instantánea de transición es simplemente la suma (cambiada de signo) de los flujos hacia otros estados:

---

[17] Esto es lo usual, sobre todo en procesos continuos pero no siempre el intervalo es arbitrario, por supuesto. La elección del período no es arbitraria en aquellos procesos que no ocurren en forma continua sino a saltos, en intervalos discretos, como por ejemplo el avance de los estudiantes en un sistema escolar, organizado en grados anuales, donde los cambios (pasar de grado o graduarse) ocurren una vez por año, en fechas determinadas, y no en forma continua. Lo mismo ocurre en la producción agrícola donde las cosechas ocurren con ciclos anuales.

[18] Esta es una propiedad conocida de los procesos estocásticos de este tipo (Doob 1953:239, Coleman 1964b: 129-130).

$$q_{ii} = -\sum_{j \neq i} q_{ij} \qquad \text{(Ec. 15)}$$

Esto es así porque q_{ii} equivale al total de sujetos que estaba en el estado ***i*** menos la suma de todos los sujetos que pasaron desde ***i*** a otros estados. La suma de las tasas q_{ij} que parten del estado ***i*** hacia distintos estados ***j*** puede expresarse como 1 menos la tasa q_{ii}.

$$\lim_{h \to 0} \frac{1 - u_{ii}^{t,t+h}}{h} = \frac{-du_{ii}}{dt} = \sum_{j \neq i} q_{ij} \qquad \text{(Ec. 16)}$$

de donde la definición de q_{ii} se deduce inmediatamente. El uso de estas tasas instantáneas facilita el tratamiento de los paneles cuyos intervalos de espaciamiento no son uniformes. También permite estimar la incidencia de **eventos intermedios** ya que se supone que operan de manera continua durante el intervalo entre rondas. La comparación de las probabilidades de transición entre distintos pares de rondas sucesivas no puede hacerse legítimamente cuando las rondas han sido realizadas con diferentes intervalos. La estimación de las tasas instantáneas que se supone han generado esas transiciones acumuladas a lo largo del intervalo permite superar esa dificultad, ya que las refiere a un intervalo infinitesimal indepediente de la longitud empírica del intervalo entre rondas del panel.

El cambio de un contingente N_i por unidad de tiempo puede, de modo similar, ser expresado en términos de las tasas instantáneas de variación. Supóngase que hay sólo dos estados, ***i*** y ***j***, y que el estado ***j*** es un estado "terminal", de modo que no hay ningún flujo desde ***j*** hacia ***i***. En este caso se tiene para un intervalo infinitesimal de tiempo ***dt***:

$$\frac{dN_i^t}{dt} = -q_{ij} N_i^t \qquad \text{(Ec. 17)}$$

Integrando esta derivada en el intervalo desde ***t*** hasta ***t+h*** se puede deducir la expresión que da el contingente en el estado ***i*** en un momento cualquiera ***t+h*** en función del contingente inicial en el momento ***t*** y la tasa instantánea de transición.

$$N_i^{t+h} = N_i^t e^{-q_{ij} h} \qquad \text{(Ec. 18)}$$

Esta expresión, sin embargo, refleja el flujo de salida desde el estado ***i*** pero suponiendo que no hay un correlativo flujo de entrada. Si además existe un flujo desde ***j*** hacia ***i***, la derivada del contingente en el estado ***i*** es dada por la siguiente ecuación:

$$\frac{dN_i^t}{dt} = q_{ji} N_j^t - q_{ij} N_i^t \qquad \text{(Ec. 19)}$$

La integración de esta función sobre el intervalo (***t, t+h***), tomando en cuenta el hecho de que $N_j = N\text{-}N_i$, arroja una expresión para el contingente esperado en el estado ***i*** en el momento ***t+h*** tomando en cuenta los flujos de entrada y salida entre los estados ***i*** y ***j***.[19]

$$N_{ei}^{t+h} = N \frac{q_{ji}}{q_{ji} + q_{ij}} + \left[N_i^t - N \frac{q_{ji}}{q_{ji} + q_{ij}} \right] e^{-h(q_{ji} + q_{ij})} \qquad \text{(Ec. 20)}$$

El segundo miembro de la derecha tiende a cero si ***h*** tiende a infinito, y por lo tanto el valor de largo plazo (que es el valor de equilibrio del sistema) equivale al primer miembro:

$$\lim_{h \to \infty} N_{ei}^{t+h} = N_{ei}^* = N \frac{q_{ji}}{q_{ji} + q_{ij}} \qquad \text{(Ec. 21)}$$

[19] La derivación de esta expresión, que aquí se omite por brevedad, puede encontrarse en Coleman, 1964b, apéndice 2 del cap. 4, p.131.

donde el asterisco en el símbolo *N** indica el tamaño **esperado** del contingente N_i en la situación de equilibrio. Dado que con dos estados posibles será $N_i+N_j=N$, se ve fácilmente que el contingente esperado de equilibrio en el estado ***j***, es decir N^*_{ej}, viene dado por:

$$N^*_{ej} = N\left[1 - \frac{q_{ji}}{q_{ji} + q_{ij}}\right] = N\frac{q_{ij}}{q_{ji} + q_{ij}} \qquad \text{(Ec. 22)}$$

Estas expresiones para calcular la magnitud de equilibrio de los contingentes en cada estado (o su proporción respecto a *N*) también podrían deducirse a partir de la expresión de la derivada de N_i, si se la hace igual a cero, pues en un estado de equilibrio el contingente en cualquier estado debería permanecer constante a lo largo del tiempo:

$$q_{ji}N^*_{ej} - q_{ij}N^*_{ei} = 0 \qquad \text{(Ec. 23)}$$

A partir de esta ecuación, por simple transposición de términos y recordando que $N_j=N-N_i$, se obtienen las mismas expresiones anteriores (Ec. 21 y 22) para N^*_{ei} y para N^*_{ej}.

4.2. Estimación empírica de las intensidades de transición

Dado que las intensidades de transición q_{ij} son tasas instantáneas, que se definen en función del límite de las tasas de transición durante un intervalo cuando ese intervalo tiende a cero, su cálculo empírico no es inmediato. Se analiza primero el caso de una variable dicotómica, donde el cálculo es relativamente más fácil, y luego el caso de una variable con más de dos categorías.

4.2.1. Variables dicotómicas

Supóngase que hay dos ondas de panel realizadas en las fechas ***t*** y ***t+h***. La aplicación de un modelo que postula un proceso continuo de cambio entre los dos estados de una variable dicotómica exige calcular las tasas instantáneas de transición q_{ij} y q_{ji}. Un **cálculo aproximado** de esas tasas instantáneas de transición para el caso de una variable dicotómica puede basarse en la ecuación 18:

$$N_i^{t+h} = N_i^t e^{-q_{ij}h}$$

Pasando el factor N_i^t como divisor al primer miembro y aplicando logaritmos naturales se obtiene:

$$ln\frac{N_i^{t+h}}{N_i^t} = -q_{ij}h$$

Pasando *h* ahora como divisor al primer miembro se puede estimar inmediatamente:

$$q_{ij} = \frac{lnN_i^t - lnN_i^{t+h}}{h} \qquad \text{(Ec. 24)}$$

Esta estimación, sin embargo, es sólo aproximada o gruesa. En primer lugar, considera los flujos desde ***i*** hacia ***j*** pero no los flujos en sentido contrario. Incluye por lo tanto los sujetos que en el momento ***t*** eran miembros del estado ***i*** y que en el momento ***t+h*** estaban en el estado ***j***, pero no incluye los que estaban en ***j*** en el momento ***t*** y fueron hallados en ***i*** al llegar el momento ***t+h***. En segundo lugar, tampoco incluye las "transiciones revertidas" o "rebotes", es decir los que pasaron de ***i*** a ***j*** durante ese mismo intervalo pero luego **volvieron a pasar** a ***i*** antes de terminar el intervalo.

Si todos los sujetos tienen las mismas probabilidades de transición **en cada instante**, entonces algunos de los que cambiaron de estado durante el período intermedio probablemente retornen a su estado inicial antes que llegue la próxima ronda. Podrían incluso cambiar varias veces de estado a lo largo del período, antes de llegar al estado en que se encontraban en el momento ***t+h***. La fórmula anterior contempla sólo el "primer salto" y no los "saltos secundarios" en la trayectoria de cada sujeto. Unas tasas instantáneas de transición q_{ij} que reflejan un proceso continuo deben ser estimadas teniendo en cuenta estas posibilidades, pues de otro modo subestimarían el flujo de ***i***

hacia ***j*** al no incluir aquellas personas que "retornaron", es decir que hicieron ese cambio ***i*** → ***j*** pero luego volvieron a ***i***.

Para realizar una estimación exacta es necesario introducir algunos ajustes en la notación. Así como anteriormente se había definido N_{ej}^{t+h} como cantidad esperada de sujetos que están en el estado ***j*** en el momento ***t+h***, ahora definiremos $N_{eij}^{t,t+h}$ como la cantidad esperada de sujetos que, habiendo estado en el estado ***i*** en el momento ***t***, se encontrarán en el estado ***j*** en el momento ***t+h***. En la ecuación 20 se encontró una expresión del número esperado de personas en un estado cualquiera ***i*** en el momento ***t+h***:

$$N_{ei}^{t+h} = N\frac{q_{ji}}{q_{ji}+q_{ij}} + \left[N_i^t - N\frac{q_{ji}}{q_{ji}+q_{ij}}\right]e^{-h(q_{ji}+q_{ij})}$$

Para mayor claridad se transponen y reagrupan algunos términos:

$$N_{ei}^{t+h} = N\frac{q_{ji}}{q_{ji}+q_{ij}}\left(1-e^{-h(q_{ji}+q_{ij})}\right) + N_i^t e^{-h(q_{ji}+q_{ij})} \qquad \text{(Ec. 25)}$$

Esta ecuación muestra más claramente que el número total esperado en el estado ***i*** en el momento ***t+h***, se compone de dos subgrupos: los que en el tiempo ***t*** estaban en el estado ***j***, y acabaron en el estado ***i***, que se denotan como $N_{eji}^{t,t+h}$ y corresponden al primer término en el miembro de la derecha en la ecuación 25; y por otra parte los que desde el inicio estaban en el estado ***i*** donde reaparecen en el momento ***t+h***, que se denotan como $N_{eii}^{t,t+h}$ y corresponden al último término a la derecha en la ecuación 25. Si la ecuación 25 se formula para estos dos subgrupos separadamente, resultan las ecuaciones 26 y 27 que para mayor claridad referiremos a los dos estados 0 y 1 de la variable dicotómica, ya que este procedimiento sólo se aplica en esa clase de variables.

$$N_{e00}^{t+h} = N_0^t\frac{q_{10}}{q_{10}+q_{01}}\left(1-e^{-h(q_{10}+q_{01})}\right) + N_0^t e^{-h(q_{10}+q_{01})} \qquad \text{(Ec. 26)}$$

$$N_{e10}^{t+h} = N_1^t\frac{q_{10}}{q_{10}+q_{01}}\left(1-e^{-h(q_{10}+q_{01})}\right) + 0 \qquad \text{(Ec. 27)}$$

La ecuación 26 puede ser dividida por N_0^t en ambos miembros, convirtiéndose en una expresión para u_{00}, mientras que la ecuación 27 puede ser dividida por N_1^t obteniéndose una expresión de u_{10}. Luego se resta la segunda de la primera, con lo cual se obtiene:

$$u_{00} - u_{10} = e^{-h(q_{10}+q_{01})} \qquad \text{(Ec. 28)}$$

Aplicando logaritmos y despejando $q_{10}+q_{01}$ resulta:

$$q_{10}+q_{01} = -\frac{\ln(u_{00}-u_{10})}{h} \qquad \text{(Ec. 29)}$$

Substituyendo las ecuaciones 28 y 29 en la ecuación 27 se obtiene:

$$u_{10} = \frac{-hq_{10}}{\ln(u_{11}-u_{01})}\left[e^{-h(q_{01}-q_{12})}\right] \qquad \text{(Ec. 30)}$$

Advirtiendo que $u_{00} = 1 - u_{01}$, recordando la ecuación 28 y despejando q_{10} y q_{01} se obtiene:

$$q_{10} = \left(\frac{u_{10}}{u_{01}+u_{10}}\right)\left(\frac{-\ln(1-u_{01}-u_{10})}{h}\right) \qquad \text{(Ec. 30)}$$

$$q_{01} = \left(\frac{u_{01}}{u_{01} + u_{10}} \right) \left(\frac{-\ln(1 - u_{01} - u_{10})}{h} \right) \qquad \text{(Ec. 31)}$$

Las u_{ij} que figuran en las ecuaciones 30 y 31 son cantidades directamente estimadas a partir de los flujos obtenidos empíricamente en el panel:

$$u_{10} = \frac{N_{10}}{N_1^t} \qquad\qquad u_{01} = \frac{N_{01}}{N_0^t}$$

De esta manera se pueden obtener estimaciones de las tasas instantáneas de transición entre los dos estados de una variable dicotómica, a partir de los datos de dos rondas del panel, suponiendo un proceso continuo de flujos y reflujos a lo largo del período intermedio entre las dos rondas. Estas estimaciones no serán por lo general iguales a las estimaciones aproximadas (que no toman en cuenta los reflujos) que pueden ser obtenidas por medio de la ecuación 24.

Si se dispone de un panel con varias rondas, la elección de un determinado par de rondas para basar este cálculo sería arbitraria. La adopción de un modelo teórico con tasas de transición constantes implica postular que las probabilidades empíricamente estimadas con cada par de ondas son sólo diferentes **estimaciones muestrales de las mismas probabilidades subyacentes**, de modo que se podría tomar como base no ya un par de ondas determinadas sino **el promedio de todos los pares adyacentes de ondas**.

En otras palabras, el flujo N_{ij} se obtendría como promedio de los flujos observados entre las ondas 1 y 2, entre las ondas 2 y 3, etc., siempre que se piense que las condiciones de contexto no han cambiado, y que por lo tanto el mismo proceso está operando en todos los períodos. Esta hipótesis puede corroborarse analizando la significatidad de las diferencias entre las tablas de rotación univariadas obtenidas para los distintos períodos 1-2, 2-3, 3-4, etc. Si el mismo proceso subyacente determina los flujos de todos los períodos, entonces las probabilidades observadas en todos ellos no deberían diferir mucho entre sí. En caso que difieran significativamente, el analista puede escoger entre analizar los subperíodos separadamente (por ejemplo períodos de auge económico por un lado y períodos de recesión por otro, si se trata de un panel sobre empleo y desempleo), o bien desarrollar un modelo más complejo donde las tasas de transición no sean constantes sino que varíen de acuerdo a la evolución de determinados factores.

Los resultados anteriores se refieren solo al caso de un atributo dicotómico donde los únicos estados posibles son ***i*** y ***j***. En la sección siguiente se ofrece un procedimiento de tipo iterativo para calcular las tasas q_{ij} en el caso general de variables discretas que pueden tener más de dos categorías.

4.2.2. Variables politómicas

Cuando la variable tiene más de dos categorías existen **múltiples trayectorias** que pueden haber determinado el flujo neto observado en el panel. El contingente N_{ij} incluye los que pasaron directamente de ***i*** a ***j*** pero también los que atravesaron el estado intermedio ***k***, más los que atravesaron los estados intermedios ***k*** y ***g***, más los que pasaron primero de ***i*** a ***k*** pero luego volvieron a ***i*** y finalmente pasaron a ***j***, e innumerables posibilidades más. Habrá sujetos que cambiaron "sin escalas", otros "con una escala intermedia", "con dos escalas", etc. El número de "escalas" en la trayectoria de cada individuo que haya pasado en definitiva de ***i*** a ***j*** puede variar desde el salto directo hasta diversas trayectorias más complicadas, con 1, 2, 3 o más estados intermedios, y cada uno de estos grupos puede haber pasado por diferentes trayectorias de dos escalas, o diferentes trayectorias de tres escalas, etc. Esto hace que no exista una fórmula directa de cálculo de q_{ij} en este caso, y que las estimaciones deban ser obtenidas por medio de una serie indefinida de iteraciones o aproximaciones sucesivas.

Si las fechas de dos rondas de panel son ***t*** y ***t+h***, y la variable investigada tiene ***m*** categorías, conociendo las tasas instantáneas de transición q_{ij} se podría estimar el vector fila N_{t+h} de cantidades **esperadas** de sujetos en los ***m*** estados en el momento ***t+h*** a partir del vector fila N_t de las cantidades

iniciales de sujetos en los diferentes estados en el momento ***t***, y de la matriz **Q** de tasas instantáneas de transición q_{ij}, mediante un sistema de ecuaciones cuya notación matricial es la siguiente:

$$N_{t+h} = N_t e^{Qh} \qquad \text{(Ec. 32)}$$

Si ambos miembros, es decir cada elemento de los vectores N_t y N_{t+h}, se dividen por N, el mismo sistema de ecuaciones se refiere al vector P de las probabilidades de estado. Conociendo las q_{ij} se podrían estimar las probabilidades de estado del momento ***t+h*** a partir de las probabilidades de estado del momento ***t***:

$$P_{t+h} = P_t e^{Qh} \qquad \text{(Ec. 32 bis)}$$

El producto Qh es el producto de un escalar (h) por cada elemento de la matriz Q, de modo que Qh es una matriz con elementos hq_{ij}. El exponencial e^{Qh} equivale al límite de la serie infinita siguiente:

$$e^{Qh} = 1 + Qh + \frac{Q^2h^2}{2!} + \frac{Q^3h^3}{3!} + \frac{Q^4h^4}{4!} + \ldots \qquad \text{(Ec. 33)}$$

Los denominadores del tipo ***n!*** representan el factorial de ***n***, es decir ***n!***=1×2×3×...×(***n***-1)×***n***. Si se conociera la matriz Q de tasas de transición instantáneas, la ecuación 33 proveería la matríz e^{Qh} que, con la ecuación 32, permitiría estimar el vector N_{t+h} a partir del vector inicial N_t.

Lamentablemente por lo general no se dispone de la matriz Q. Pero estas dos ecuaciones pueden también ser usadas en sentido contrario, **para estimar *Q*** sobre la base de dos rondas del panel que hayan suministrado una tabla de rotación, es decir, que haya provisto estimaciones muestrales de los flujos de transición entre estados, N_{ij} así como los contingentes situados en cada estado en la primera y segunda ronda, N_t y N_{t+h}. Consideremos para ello sólo una parte de la población del panel, a saber, los N_i^t individuos que en el momento ***t*** se encontraban en el estado ***i***. Cuando sólo esos sujetos son considerados, el vector N_t está compuesto de ceros, excepto para el estado ***i*** donde figuran N_i^t individuos:

$$N_t = \begin{bmatrix} 0 & 0 & \ldots & N_i^t & \ldots & 0 & 0 \end{bmatrix}$$

Correlativamente, si siempre nos referimos a la población que estaba en ***i*** en el momento inicial, el vector N_{t+h} contendrá sólo aquellos sujetos que en el momento ***t*** se encontraban en el estado ***i***, agrupados según el estado en que se encontraban en ***t+h***. En otras palabras, los elementos del vector N_{t+h} son los contingentes que antes hemos denominado $N_{eij}^{t,t+h}$, esto es, la cantidad esperada de sujetos que, habiendo estado en ***i*** en el momento ***t***, se encuentran en el estado ***j*** en el momento ***t+h***:

$$N_{t+h} = \begin{bmatrix} N_{ei1}^{t,t+h} & \ldots & N_{eij}^{t,t+h} & \ldots & N_{eim}^{t,t+h} \end{bmatrix}$$

Cada elemento de este vector, según surge de las ecuaciones 32 y 33, es una serie infinita cuyos primeros términos son:

$$N_{eij}^{t,t+h} = N_i^t \left[\delta_{ij} + hq_{ij} + \frac{h^2}{2!}\sum_a q_{ia}q_{aj} + \frac{h^3}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + \ldots \right] \qquad \text{(Ec. 34)}$$

Aquí el símbolo δ_{ij} es la *delta de Kronecker*, que vale 1 si es $i=j$ y vale 0 si es $i \neq j$. Si la variable tiene ***m*** estados, habrá ***m*** series de este tipo para cada uno de los ***m*** estados iniciales, es decir que habrá en total $\boldsymbol{m^2}$ series que permiten estimar otros tantos flujos del tipo $N_{eij}^{t,t+h}$. Si cada ecuación se divide por N_i^t se obtiene una formulación equivalente para las probabilidades de transición entre el momento ***t*** y el momento ***t+h***:

$$r_{eij}^{t,t+h} = \delta_{ij} + hq_{ij} + \frac{h^2}{2!}\sum_a q_{ia}q_{aj} + \frac{h^3}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + \ldots \qquad \text{(Ec. 34 bis)}$$

Si se estiman las probabilidades de transición $r_{ij}^{t,t+h}$ a partir de los datos empíricos, estos resultados pueden usarse para estimar iterativamente las tasas de transición q_{ij}. Para ello, en primer lugar, se

adopta por simplicidad la convención que el intervalo entre las ondas del panel es constante, y se tomo como unidad de medida del tiempo: ***h=1***. Esto significa que todos los numeradores del tipo h^u en la ecuación 34 resultan iguales a 1 (esta simplificación puede levantarse sin dificultad si se desea usar otra unidad de medida del tiempo, o si los intervalos entre observaciones son desiguales). La ecuación 34 se divide por N_i^t y se transponen términos para despejar q_{ij}. Dado que la tasa q_{ii} equivale a la suma de flujos hacia otros estados cambiada de signo (Ecuación 14), ella no necesita ser estimada, por lo cual este procedimiento solo se aplica a las tasas entre estados diferentes ($i \neq j$), por lo cual el término con la delta de Kronecker se anula. En definitiva obtenemos para cada una de las ***m(m-1)*** tasas de transición entre estados diferentes una expresión en serie de este tipo:

$$q_{ij} = u_{ij}^{t,t+1} - \frac{1}{2!}\sum_a q_{ia}q_{aj} - \frac{1}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + \ldots \qquad (i \neq j) \qquad \text{(Ec. 35)}$$

El primer término en el miembro de la derecha es la probabilidad de que un individuo situado en el estado ***i*** en el momento ***t*** aparezca en el estado ***j*** en el momento ***t+1***, y puede ser estimada mediante los datos empíricos como se indica en la versión siguiente de la ecuación:

$$q_{ij} = \frac{N_{ij}^{t,t+1}}{N_i^t} - \frac{1}{2!}\sum_a q_{ia}q_{aj} - \frac{1}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + \ldots \qquad (i \neq j) \qquad \text{(Ec. 35 bis)}$$

A partir de esta formulación en serie se puede usar un **procedimiento iterativo** para estimar las tasas q_{ij} con cualquier nivel de precisión. En la primera iteración se estiman todas las tasas correspondientes a estados diferentes ($i \neq j$), **en función únicamente del primer término de la serie**:

$$q_{ij}^{(1)} = u_{ij}^{t,t+1} = \frac{N_{ij}^{t,t+1}}{N_i^t} \qquad \text{(Ec. 36)}$$

Estos valores iniciales son seguramente exagerados, porque aquí la tasa **instantánea** de transición se estima como idéntica a la probabilidad de transición **a lo largo de un intervalo dado**. Pero estas primeras aproximaciones sólo se utilizan para calcular una segunda aproximación $q_{ij}^{(2)}$ de todas las tasas aplicando los valores de la ecuación 35 a las series del miembro de la derecha en la ecuación 34. Se usan todos los términos de la serie que se consideren necesarios, pero usualmente sólo algunos son importantes. Las estimaciones así obtenidas, $q_{ij}^{(2)}$, se usan a su vez para valorizar nuevamente la ecuación 34 con los nuevos valores de las tasas, a fin de obtener una tercera aproximación $q_{ij}^{(3)}$. En general, la aproximación de orden ***r+1*** se basa en el cálculo de la ecuación 34 usando la aproximación precedente de orden ***r***:

$$q_{ij}^{(r+1)} = \frac{N_{eij}^{t,t+1}}{N_i^t} - \frac{1}{2!}\sum_a q_{ia}^{(r)}q_{aj}^{(r)} - \frac{1}{3!}\sum_a\sum_b q_{ia}^{(r)}q_{ib}^{(r)}q_{bj}^{(r)} + \ldots \qquad \text{(Ec. 37)}$$

Nótese que en expresiones como $q_{ij}^{(u)}$ el superíndice no denota un exponente, sino el orden de la iteración de que se trata. No es q_{ij} elevada a la potencia ***u*** sino que se trata de la ***u***-ésima iteración o aproximación sucesiva del valor de la tasa.

Para la aplicación de este método iterativo hay que decidir hasta qué término de la serie se utiliza, y hasta qué número de iteraciones se llega. Para ello se puede establecer un umbral mínimo de variación para cada aspecto. Se añaden términos de la serie hasta que el último término incorporado tenga un valor absoluto inferior a un cierto umbral, por ejemplo ±0.01 o bien ±0.001, lo cual quiere decir que ese término contribuye al valor de q_{ij} con sólo un centésimo o un milésino, en más o en menos. Del mismo modo, se realizan iteraciones hasta que la última iteración no modifique ninguna tasa de transición en un valor absoluto superior a un umbral preestablecido, por ejemplo en más de ±0.001. Asimismo, si el modelo teórico o las características de la variable prescriben que algunas q_{ij} sean necesariamente nulas (por ejemplo, que nadie pase de casado a soltero), este cálculo iterativo puede asignar a priori un valor cero a ciertas tasas, y estimar las restantes a partir de esa asignación.

En tal caso, todos los términos referentes a las trayectorias donde intervenga esas tasas nulas tendrán un valor igual a cero.

La condición para que se pueda usar este método es que la serie iterativa utilizada para la estimación sea una serie convergente. La serie es convergente siempre que no haya una correlación inversa entre las dos rondas del panel, esto es, cuando los que permanecen en el mismo estado son menos que los que pasan a otros estados ($N_{ii} < N_{ij}$ para algún estado ***j***). En la mayor parte de los datos de panel los procesos de cambio son suficientemente lentos como para que entre dos rondas la mayor parte de los sujetos permanece en su estado inicial, pero pueden darse casos en que ello no suceda. En tales casos las diferencias entre iteraciones sucesivas no irá disminuyendo sino aumentando, impidiendo llegar a una solución aceptable.

Sin embargo, es probable que cuando la serie no es convergente haya de todas maneras un número óptimo de iteraciones. Ese número óptimo se obtiene del siguiente modo: después de obtener los resultados de cada iteración, usando datos de dos rondas sucesivas, se utilizan las ***q*** resultantes para estimar los contingentes de otra ronda, por ejemplo la tercera o la cuarta, y se evalúa el ajuste entre estas proyecciones y la realidad observada en el panel. Para evaluar la diferencia entre la proyección y la observación se puede usar la medida χ^2. El grado de ajuste por lo general aumenta al añadir los primeros términos de la serie, hasta que en cierta fase comienza a estancarse o deteriorarse, indicando que la iteración debe detenerse allí.

Un método alternativo, basado en los mismos principios pero de cálculo menos engorroso, también está disponible. De la ecuación 32, si se transponen los factores, se deduce:

$$U_h = e^{Qh} \qquad \text{(Ec. 38)}$$

donde $\boldsymbol{U_h}$ es la matriz de probabilidades de transición entre el momento ***t*** y el momento ***t+h***. Tomando logaritmos la ecuación 38 se puede expresar en la forma siguiente:

$$\ln U_h = Qh \qquad \text{(Ec. 39)}$$

El logaritmo de un número positivo se puede expresar en una serie infinita de potencias:

$$\ln x = (x-1) - \frac{1}{2}(x-1)^2 + \frac{1}{3}(x-1)^3 - \ldots \qquad \text{(Ec. 40)}$$

Del mismo modo se puede expresar el logaritmo de una matriz de números positivos:

$$\ln U_{\mathrm{h}} = (U_h - I) - \frac{1}{2}(U_h - I)^2 + \frac{1}{3}(U_h - I)^3 - \ldots \qquad \text{(Ec. 41)}$$

La matriz ***I*** es la matriz identidad. La diferencia $\boldsymbol{U_h - I}$ es la misma matriz $\boldsymbol{U_h}$ pero con los elementos de la diagonal principal, $\boldsymbol{u_{ii}}$, reemplazados por $\boldsymbol{u_{ii}-1}$. La serie converge si para todos los estados resulta ser $\boldsymbol{u_{ii}>0.5}$, es decir, si menos de la mitad de los sujetos de cada estado pasa a otro estado al cabo de un período de longitud ***h***. Este procedimiento, sin embargo, requiere que se calcule el valor de **todas** las tasas ***q*** sin que algunas de ellas sean fijadas en cero (o en otro valor cualquiera) por razones teóricas, como era posible hacer en el otro procedimiento.[20]

4.3. Trayectorias indirectas de corto plazo

El panel permite observar trayectorias aparentes, o cambios netos de estado, es decir, las posiciones de un mismo sujeto en diferentes momentos del tiempo. Pero ya hemos visto que estas "fotografías" del **estado** de los sujetos en el momento de cada ronda del panel no son equivalentes al registro de los **eventos** que ocurren entre una ronda y otra. Un sujeto que fue encontrado en los estados ***i*** y ***j*** en dos rondas sucesivas puede haber pasado (una o más veces) por cualquier estado ***k*** durante el intervalo entre ambas rondas, sin que esos eventos intermedios fuesen registrados por el panel.

[20] Véase. Coleman 1964b, cap. 4 y 5, donde se desarrolla extensamente este tipo de modelos y procedimientos de estimación.

La estimación de las tasas instantáneas de transición alcanzada en las secciones anteriores no permite conocer la trayectoria intra-período de cada uno de los individuos, pero sí permite estimar la frecuencia de determinadas trayectorias (Coleman 1964b, pp.183-184). El enfoque más sencillo para este propósito consiste en utilizar la ecuación 34:

$$N_{eij}^{t,t+h} = N_i^t \left[\delta_{ij} + hq_{ij} + \frac{h^2}{2!}\sum_a q_{ia}q_{aj} + \frac{h^3}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + \ldots \right]$$

Esta ecuación permite calcular el número de sujetos que comienzan en el estado ***i*** y acaban en el estado ***j***, e incluye aquellos que realizan un solo cambio de estado a la tasa q_{ij}, y también aquellos que han pasado por un estado intermedio ***a***, o por dos estados intermedios ***a*** y ***b***, etc. Los sucesivos términos de la serie van divididos por sucesivos factoriales, es decir por 2!=2, por 3!=6, por 4!=24, por 5!=120, etc., y su denominador es el producto de un número creciente de fracciones inferiores a 1, de modo que en general cada término será menor que el anterior y su magnitud disminuirá rápidamente.

Ahora consideremos una determinada secuencia de estados entre el estado inicial ***i*** y el estado final ***j***, por ejemplo ***iabcj***. Por ejemplo, en una variable con ***m*** categorías se podría considerar la secuencia 23141, que es una de las formas de empezar en el estado ***i***=2 y terminar en el estado ***j***=1. Los sujetos que recorren esta trayectoria han realizado cuatro cambios de estado, a saber: 2→3, 3→1, 1→4 y 4→1. Esto significa en principio que aquellos términos de la serie que no correspondan a esta trayectoria deben ser considerados como nulos e iguales a cero. Esto incluye por ejemplo los primeros cuatro términos de la serie, ya que ellos corresponden a trayectorias con menos de cuatro cambios de estado. El primer término que no es nulo es el término que incluye el producto de cuatro tasas q_{ij}, es decir:

$$\frac{h^4}{4!}\sum_a\sum_b\sum_c q_{ia}q_{ab}q_{bc}q_{cj} + \ldots$$

Sin embargo, no todos los términos de la triple sumatoria se conforman al patrón establecido, sino sólo uno, es decir el término definido por $q_{23}q_{31}q_{14}q_{41}$. Esto significa que una posible estimación del número de personas que recorre esa trayectoria es:

$$N_{23141}^{t,t+h} = N_2^t \left[\frac{h^4}{4!} q_{23}q_{31}q_{14}q_{41} \right]$$

Se podría pensar que esta ecuación suministra directamente la estimación del número de sujetos que recorre esta trayectoria. Sin embargo, hay que añadir también aquellos que realizan estos mismos cambios de estado pero en algún momento el "cambio" consiste en que **permanecen** en su mismo estado. Por ejemplo, considérese la secuencia $q_{23}q_{31}q_{11}q_{14}q_{41}$, que es un producto de cinco factores (tasas de transición) incluyendo un factor de permanencia en el estado 1, es decir q_{11}. Esta secuencia también cumple con la misma trayectoria, del mismo modo que lo hace $q_{22}q_{22}q_{23}q_{23}q_{31}q_{14}q_{41}$, donde el sujeto primero permanece dos veces en el estado 2, y luego realiza sus cambios de estado a 3, a 1, a 4 y finalmente a 1. Esto significa que además de la secuencia básica $q_{23}q_{31}q_{14}q_{41}$ habrá **secuencias con permanencias**. Esto incluirá algunas secuencias de cinco transiciones, esto es, con cuatro cambios y una permanencia, como $q_{23}q_{31}q_{11}q_{14}q_{41}$. En esta secuencia, el sujeto permanece en el estado 1 despues de estar en el estado 3 y antes de pasar al estado 4. Estas secuencias serán cinco en total, incluyendo los cuatro cambios incluidos en la trayectoria y una permanencia en un mismo estado, que puede ser cualquiera de los cinco que figuran en la trayectoria; aparte de la ya mencionada secuencia, ellas son $q_{22}q_{23}q_{31}q_{14}q_{41}$, $q_{23}q_{33}q_{31}q_{14}q_{41}$, $q_{23}q_{31}q_{14}q_{44}q_{41}$ y $q_{23}q_{31}q_{14}q_{41}q_{11}$, Hay también algunas secuencias de seis transiciones como $q_{22}q_{22}q_{23}q_{31}q_{14}q_{41}$ o bien $q_{22}q_{23}q_{23}q_{31}q_{14}q_{41}$ (hay un total de 14 secuencias con seis tasas q_{ij}, que incluyen los cuatro cambios de estado obligatorios de la trayectoria más **dos** permanencias). También hay algunas posibles secuencias con tres permanencias, es decir con siete tasas de transición, y así sucesivamente, aunque estos términos de orden superior rápidamente se tornarán insignificantes en su magnitud.

Trayectorias muy complicadas como éstas son de dudosa utilidad práctica. Pero puede ser importante tener una idea de la posible incidencia de algunas trayectorias breves de mayor interés. Por ejemplo, en un estudio sobre desempleo podría ser importante saber qué proporción de las personas que aparecen "ocupadas" en ambas rondas atravesaron una fase intermedia de desempleo, con una secuencia básica $q_{12}q_{21}$ que puede incluir trayectorias con permanencias, como $q_{12}q_{22}q_{21}$ o bien $q_{11}q_{12}\ q_{21}$.

La incidencia porcentual de una determinada trayectoria no indica **cuáles** sujetos han recorrido esa trayectoria. En el ejemplo anterior, si hay N_{11} sujetos que aparecen ocupados en ambas rondas, y se determina una cierta proporción estimada de esos sujetos que habría tenido algún período intermedio de desocupación, no se podría identificar precisamente cuáles sujetos han tenido esa experiencia.

Tampoco estos resultados permiten, de por sí, estimar **cuánto tiempo** han pasado los sujetos en cada fase de la trayectoria. La mayor parte de los probemas reales involucra cierta "fricción", que implica un cierto tiempo para cada cambio de estado. Nadie puede enviudar y volverse a casar en un instante, y normalmente quedar desocupado y volver a encontrar empleo también toma cierto tiempo. Más aún, cada uno de los cambios de estado puede tener su propio nivel de "fricción"; por ejemplo, una persona casada puede enviudar en un instante, de modo que ahí la "fricción" es muy poca, pero difícilmente se case nuevamente sino después de un razonable intervalo, de modo que enviudar tiene menos fricción que volver a casarse (los estados de los cuales se sale con mucha fricción suelen llamarse **viscosos** o **pegajosos**).

Pese a esto, de todos modos una estimación gruesa podría partir del supuesto de una fricción uniforme y una distribución equidistante de las fases a lo largo del intervalo considerado: si las dos rondas están separadas por seis meses, y hay un grupo de sujetos cuya trayectoria implica tres cambios de estado en ese lapso, se puede asumir que esos cambios ocurren cada dos meses **en promedio**. Más precisamente, se podría asumir que el primero de los tres cambios de estado ocurre en un momento variable cuya fecha promedio se sitúa en la mitad del primer bimestre después de iniciado el intervalo de seis meses; el segundo cambio de estado ocurre también en una fecha variable que en promedio se sitúa en la mitad del segundo intervalo bimestral, es decir a los tres meses de la primera ronda; y el último cambio ocurre en una fecha variable que en promedio se sitúa a la mitad del último bimestre, es decir a los cinco meses desde la primera ronda (un mes antes de la segunda). Estos supuestos no impiden que los cambios ocurran en cualquier fecha, ya que sólo postulan una fecha **promedio** para cada uno de ellos. Ese promedio puede luego usarse para otros cálculos, aun cuando su naturaleza es esencialmente especulativa y no tiene ningún basamento empírico concreto.

Dado que las personas que recorren cada trayectoria ***iab... j*** son sólo una fracción del total de personas que empezó en el estado ***i*** y acabó en el estado ***j***, los cuales de por sí son sólo una fracción del total de sujetos en el panel, es muy posible que el número o el porcentaje de sujetos que se estima que recorre una trayectoria compleja resulte muy pequeño, y por lo tanto estadísticamente poco significativo (a menos que la cantidad de personas en el panel sea realmente muy grande). Por consiguiente, este análisis de **"trayectorias presuntas"** sólo podría hacerse para los flujos más numerosos, y restringido a las trayectorias con pocas fases, pues de otro modo los resultados no indicarían realmente nada en términos estadísticos.

5. Incertidumbre de respuesta

Cuando un sujeto aparece en un estado diferente la segunda vez que se lo entrevista, ello puede deberse a que haya cambiado efectivamente de estado, o a que haya algún error o imprecisión en sus respuestas o en el registro de las mismas (en la primera vez, en la segunda, o en ambas).

Para investigar este problema en un modelo univariado se requeiría un modelo que pueda discriminar entre el auténtico cambio de estado de los sujetos, y los cambios aparentes que resultan de fluctuaciones aleatorias en las respuestas suministradas por los sujetos. Este problema supone distinguir entre el **estado subyacente o latente** de los sujetos, que es inobservable, y su **estado manifiesto** o respuesta observable. En cada estado latente posible, el sujeto no da necesariamente

una determinada respuesta, sino que tiene diferentes **probabilidades** de dar las distintas respuestas posibles. La respuesta particular que se observe en una determinada ronda del panel es sólo una de las que podría haber producido, y además, una misma respuesta manifiesta puede ser producida por sujetos en diferentes estados subyacentes o internos.

5.1. El problema de la incertidumbre de respuesta

La respuesta registrada no necesariamente debe coincidir siempre con el estado subyacente del sujeto. De cada 100 portadores de HIV puede siempre haber alguno que resulte negativo en el examen serológico, y en cambio de cada 100 individuos sanos puede aparecer alguno cuyo examen indique HIV positivo. Estas situaciones se conocen en la investigación médica como "falsos negativos" y "falsos positivos". Del mismo modo, si una pregunta se utiliza como indicador de una actitud subyacente (por ejemplo preguntas que busquen inferir la existencia de una actitud o ideología autoritaria) es posible que algunos sujetos autoritarios de vez en cuando contesten negativamente a esa pregunta, por distintas razones aleatorias (por no haberla comprendido bien, o para ocultar sus creencias, o porque el encuestador formuló la pregunta incorrectamente, etc.), y en cambio algunos sujetos que no son autoritarios podrían ocasionalmente contestar positivamente por similares razones. Estos "falsos positivos" y "falsos negativos" pueden deberse a fluctuaciones aleatorias del estado del sujeto, o a errores aleatorios del procedimiento de medición. Los médicos los controlan mediante la repetición del análisis, pues la probabilidad de que el error se cometa dos veces es más pequeña, y también mediante la aplicación de procedimientos "complementarios" es decir, aplicando dos análisis diferentes (p.ej. tomografía y ecografía) para ver si concuerdan o difieren entre sí.

Estas incertidumbres se plantean en estudios de corte transversal. Por ejemplo un sujeto autoritario puede aparecer como no autoritario, o viceversa. También pueden darse en estudios de panel: un sujeto puede aparecer cambiando de autoritario a tolerante, o viceversa, aunque en realidad el cambio no haya ocurrido en sus actitudes subyacentes.

La posibilidad de que existan estos casos de "falsos positivos" o "falsos negativos" de tipo "dinámico", es decir, que las respuestas puedan variar por razones aleatorias independientemente del verdadero cambo en el estado de los sujetos, es sumamente importante cuando se analizan paneles. A diferencia de los estudios transversales, el panel permite, hasta cierto punto, distinguir el auténtico cambio de las meras fluctuaciones aleatorias.[21]

Supongamos que 1000 sujetos elegidos al azar son encuestados en dos ocasiones acerca de su opinión política. Existen dos posibles opciones, A y B. Los resultados han sido los siguientes.

	Opinión en Octubre		
Opinión en Septiembre	**A**	**B**	**Total**
A	450	113	563
B	120	317	437
Total	570	430	1000

Evidentemente hay un considerable cambio de opinión a nivel individual entre las dos encuestas: un total de 233 personas (un 23.3% de la muestra) cambiaron de opinión (120 en un sentido y 113 en el otro). Las proporciones marginales permanecieron bastante estables, pero no obstante se registra algún cambio en ellas: como resultado de los flujos de A hacia B y viceversa, los sujetos con la opinión A aumentan de 563 en Septiembre a 570 en Octubre, un aumento neto de 1.23% en relación al contingente que en Septiembre declaró la opinión A.

El analista puede estar interesado en saber si esto implica una auténtica tendencia a favor de la opinión A, o si se puede atribuir a otros factores. Un enfoque posible sería ver si 1.23% es una

[21] Este problema, consistente en distinguir el cambio del estado subyacente y el cambio (que puede ser aleatorio) en las respuestas manifiestas de los sujetos, no es el mismo problema de la incertidumbre estadística de las estimaciones, vinculado al error de muestreo, y que usualmente se ataca con las "pruebas estadísticas de hipótesis", sobre lo cual véase más bien el capítulo 10.

diferencia estadísticamente significativa, lo que se podría evaluar con una prueba de significación apropiada como chi cuadrado o la *t* de Student. Pero aparte de que esa es una cuestión diferente, de hecho ese enfoque podría no ser adecuado: en realidad no hubo cambio solo en 1.23% de los casos, sino en un 23.3%, casi una cuarta parte, que cambiaron de opinión en uno u otro sentido. Si bien el resultado neto es pequeño, los cambios fueron mucho más voluminosos. Podría suceder que los flujos entre A y B fuesen significativos, pero no así el resultado neto. Esa cuestión por el momento la dejamos de lado.

La pregunta del analista podría ser otra: tratar de identificar los procesos subyacentes que están ocurriendo en cada contingente inicial, es decir, entre los que tenían cada opinión en el primer mes. La pregunta ahora sería por ejemplo: Entre quienes tenían inicialmente la opinión A, ¿en qué medida los cambios observados reflejan auténticos cambios de opinión, y en qué medida reflejan otros factores no controlados (por ejemplo, cambios en la disposición de los sujetos para dar a conocer verídicamente sus opiniones)?

Si los individuos manifestaran siempre en forma fidedigna su opinión, es evidente que los cambios observados en un 23.3% de la muestra, o ese cambio neto de 1.23%, representarían verdaderos cambios en el estado de opinión de la muestra. Pero podrían estar representando solo un cambio en las opiniones **declaradas**, sin representar necesariamente un cambio en las opiniones subjetivas.¿Cómo decidir cuál es la interpretación más probable?

Un posible enfoque es la aplicación de un modelo de Markov. Si hay una tendencia a ocultar o no declarar verídicamente los cambios de las propias opiniones, por ejemplo, esa tendencia debería manifestarse en todas las rondas sucesivas. Supongamos que las probabilidades de transición se obtuviesen a partir de esta tabla, considerando el cambio manifiesto como reflejo fiel de un cambio subyacente. Si la probabilidad de cambiar de opinión por unidad de tiempo fuese constante, es decir si se tratase de un proceso de Markov, el número esperado de personas con opinión A **en la ronda siguiente** volvería a aumentar. Este aumento a favor de A continuaría hasta alcanzar un estado de equilibrio, en que los dos flujos se compensen. Aplicando la regla anteriormente establecida, el equilibrio se alcanzaría cuando la proporción entre los que sostienen ambas opiniones (N_A/N_B) equivalga al cociente entre las probabilidades de cambiar de opinión en ambos sentidos ($\boldsymbol{u_{BA}/u_{AB}}$). En este caso esas probabilidades de transición son $\boldsymbol{u_{BA}}$=120/437=0.2746, y $\boldsymbol{u_{AB}}$=113/563=0.2007, y su cociente es 0.2746/0.2007=1.37. Con 1000 casos en total, los contingentes de equilibrio (redondeados) serían: N_A=578 y N_B=422. Dada la cercanía entre los valores de equilibrio y los observados en los dos primeros períodos, se percibe que el equilibrio se alcanzaría en sólo una o dos rondas adicionales.

Ahora bien, en este caso la divergencia entre lo observado y el equilibrio es muy pequeña. Alguien podría especular que quizá la distribución **está** en equilibrio desde el inicio, y que las diferencias observadas son meras fluctuaciones producidas por variaciones accidentales de la opinión declarada, debido a cambios en el estado de ánimo de los sujetos o en la forma en que se formuló la pregunta. Esta hipótesis en realidad abre todo un nuevo horizonte de investigación, ya que distingue entre el estado "verdadero" de los sujetos y su respuesta observable, e implica suponer que los individuos no expresan necesariamente su verdadera opinión. Es posible concebir la respuesta de los individuos como función de dos factores: su "verdadero" **estado interno** (A o B) y un factor aleatorio de "error" o **incertidumbre de respuesta** que en algunos casos hace que los individuos den respuestas que no corresponden a su verdadera opinión. Un individuo que siempre ha tenido la opinión A podría dar ocasionalmente la respuesta B por razones variadas (para hacer una broma al encuestador, para expresar su malhumor momentáneo, por no haber entendido la pregunta, o cualquier otra razón similar). Puede ser tambien que el encuestador ocasionalmente anote erróneamente las respuestas de uno u otro individuo, por razones aleatorias. Esta posibilidad pudo estar presente en ambas rondas.

Esto equivale a pensar en la existencia de dos procesos simultáneos: uno de ellos hace que el sujeto adopte la opinión A o la opinión B **en su fuero interno**. El otro proceso hace que al ser encuestado **manifieste (**o quede registrada**)** una opinión, que puede ser su verdadera opinión o la contraria. ¿Es

posible distinguir los verdaderos cambios de opinión subjetiva (cambios en la **variable latente**) de los cambios que sólo afectan a las respuestas (la **variable manifiesta**)? En el ejemplo numérico anterior, ¿es el aumento de opiniones a favor de A (de 563 a 570) la manifestación de una tendencia real en las opiniones subjetivas de los encuestados, o una mera fluctuación de las respuestas manifiestas que no refleja cambios en la variable de fondo?

Estas discrepancias entre la respuesta y la verdadera opinión no son, estrictamente, errores de muestreo en el sentido ordinario. Para empezar, la muestra de individuos es la misma en ambas oportunidades. En todo caso podría suponerse que las dos rondas de entrevistas constituyen una "muestra de fechas", y que los sujetos podrían dar diferentes respuestas según la fecha en que fueran entrevistados, pero esa argumentación es poco convincente. Lo que aquí se analiza es algo más que un problema de muestreo: es un problema real: ¿Mantienen estos individuos su misma opinión, o la han cambiado? ¿Dicen la verdad o mienten? Se necesita no ya un modelo de los errores casuales de muestreo, sino un modelo teórico acerca de cómo se conectan las variables "latentes" (no observadas en forma directa), por ejemplo la opinión política subjetiva, y las variables "manifiestas" que son objeto de observación, por ejemplo las respuestas registradas por el entrevistador. Entre ambas puede haber una relación **determinista** o **probabilista**, y esto se añade al problema general de la variabilidad de muestreo. Aun cuando no hubiese errores de muestreo podría haber aún una relación estocástica entre opinión y respuesta.

Esa relación solo se puede detectar si se tienen al menos **tres** rondas de observación. Si solo se han observado dos períodos, el problema del error estadístico no se puede separar del problema de las variables manifiestas y latentes. En ese caso, como en el ejemplo precedente, sólo cabe la aplicación de un test estadístico, como el χ^2, para ver si la diferencia entre las frecuencias observadas y las de equilibrio es una diferencia estadísticamente significativa. Si no es significativa, no se podría rechazar la hipótesis nula de que el estado de la opinión pública no ha cambiado.

Cuando se dispone de un mayor número de períodos cabe aplicar otros enfoques. El cambio observado puede ser el resultado neto de dos tipos de factores: algún auténtico cambio en el estado verdadero ("latente") de los sujetos, y algún error aleatorio en la declaración o registro de los datos. Puede haber así dos procesos simultáneos: un proceso de cambio subyacente, gobernado por ejemplo por un modelo de Markov, y además ciertos factores aleatorios que hacen fluctuar la respuesta de algunos sujetos. En tal caso es importante identificar qué porción de los cambios observados se debe a los procesos de cambio "verdaderos" gobernados por el modelo subyacente (de Markov o de otro tipo), y qué parte se debe a factores aleatorios.

Este tipo de problema no necesariamente se limita al caso de la expresión de opiniones subjetivas, sino que puede referirse también a variables "objetivas" cuya medición dé lugar a posibles fluctuaciones en los registros aun cuando no haya producido ningún cambio en la variable subyacente, como es el caso de los falsos positivos o falsos negativos en los exámenes médicos. En este caso el origen de las fluctuaciones no son los cambios de humor del entrevistado, sino las fluctuaciones aleatorias del aparato medidor o la imprecisión intrínseca del indicador que se utiliza en el estudio.

Lo característico en esta problemática es que se introduce una distinción entre la variable latente o subyacente que se pretende medir, y la variable manifiesta o indicador que es efectivamente medida a través de la investigación, y se presupone una relación no determinista sino estocástica entre la variable subyacente y el indicador. En este enfoque, cuando un sujeto escucha la pregunta no manifiesta necesariamente su estado de opinión subjetivo, pues su respuesta, además de depender de su opinión subjetiva, se ve afectada por un elemento aleatorio (por ejemplo cambios en su estado de ánimo) que a veces lo lleva a manifestar una opinión diferente a la que realmente sostiene. Por ejemplo, puede suceder que en cada ronda un 2% de los sujetos con opinión A, y quizá un 1% de los sujetos con opinión B, den una opinión contraria a la real debido a este factor subjetivo. Los sujetos que lo hacen no tienen por qué ser siempre los mismos, ya que la probabilidad de hacerlo afecta a todos por igual. También puede haber errores similares debido a variaciones aleatorias en la conducta del encuestador, del codificador, o del personal encargado del ingreso de los datos en la

base de datos (cambios en la forma en que se formula la pregunta, errores en la clasificación de las respuestas, errores al registrar la respuesta, etc.).

Las discrepancias entre respuesta y opinión que estamos considerando aquí no son de carácter **sistemático** sino **aleatorio**. Una discrepancia sistemática aparecería por ejemplo en un país donde no reine mucha libertad de opinión: algunas personas pueden tener temor de manifestar opiniones opositoras al gobierno, de modo que sistemáticamente las **respuestas** opositoras tenderán a subestimar la magnitud de las verdaderas **opiniones** opositoras. Asimismo, si la vigilancia de los espías gubernamentales se fuese acentuando a medida que se acercan las elecciones, la tendencia a ocultar las opiniones opositoras podría aumentar sistemáticamente con el tiempo. El temor a las represalias gubernamentales (ya sea en forma fija o con tendencia a aumentar) sería así un **factor sistemático de distorsión**, que opera siempre en el mismo sentido. Si así fuese, el factor en cuestión debería ser identificado e incorporado explícitamente en el análisis. Pero aquí, para mayor simplicidad, se supone que los factores sistemáticos ya han sido incorporados, y se hace referencia solo a factores puramente aleatorios, que operan en cualquier dirección y sobre cualquier individuo sin ningún patrón determinado. Esos factores aleatorios, por su propia naturaleza, no tienen una tendencia: a veces opera a favor de A, a veces a favor de B, sin una preferencia o sesgo sistemático.

La idea general entonces para distinguir entre un verdadero cambio en el estado de los sujetos y las fluctuaciones aleatorias en las respuestas consiste en verificar si los cambios son acumulativos. Si lo son, probablemente se trata de un proceso efectivo de cambio; en cambio, si los cambios entre la segunda y tercera ronda tienden a cancelar los cambios ocurridos entre la primera y la segunda, probablemente se trata de variaciones aleatorias. Un método para ello consiste en estimar las probabilidades de transición a partir de dos rondas adyacentes, realizadas por ejemplo en los momentos ***t=1*** y ***t=2***, y luego recalcularlas para un período más largo, por ejemplo con la primera y la cuarta ronda, realizadas respectivamente en las fechas ***t=1*** y ***t=4***.

	Onda *t*=2	
Onda *t*=1	**Respuesta A**	**Respuesta B**
Respuesta A	u_{AA}^{12}	u_{AB}^{12}
Respuesta B	u_{BA}^{12}	u_{BB}^{12}
	Onda *t*=4	
Onda *t*=1	**Respuesta A**	**Respuesta B**
Respuesta A	u_{AA}^{14}	u_{AB}^{14}
Respuesta B	u_{BA}^{14}	u_{BB}^{14}

Consideremos primero dos hipótesis extremas: una donde solo hay movimiento real hacia el equilibrio, y otra donde solo hay variaciones aleatorias. En la primera hipótesis, supongamos que el proceso inicialmente estaba en equilibrio; por lo tanto no habría probabilidades de transición hacia el equilibrio, y **solo operarían factores aleatorios**. En el segundo, se supone que el proceso inicialmente no estaba en equilibrio, y que no hay cambios aleatorios; en este caso los cambios se deberán solo a la marcha hacia el equilibrio **sin la intervención de factores aleatorios.** Estos son, por supuesto, dos extremos teóricos: en la mayoría de los casos ocurrirá alguna situación intermedia.

Si el proceso subyacente inicialmente estuviese en equilibrio, y las discrepancias observadas respecto al equilibrio se debiesen **sólo a factores aleatorios**, entonces dos períodos cualesquiera diferirían solamente por factores aleatorios, sin que el tiempo transcurrido tenga ninguna relevancia porque los cambios aleatorios no se acumulan; el sistema subyacente sigue en estado de equilibrio, y las diferencias entre cualquier par de rondas se deben a factores aleatorios. Si fuese así, las probabilidades basadas en las rondas 1 y 4 no tendrían que ser muy distintas de las basadas en las rondas 1 y 2 (o entre cualquier otro par de ondas) ya que se deberían únicamente a fluctuaciones aleatorias. En otras palabras, las probabilidades de transición calculadas sobre los cambios de estado entre ***t=1*** y ***t=2*** deberían ser muy similares a las probabilidades de transición calculadas a partir de los cambios de estado entre ***t=1*** y ***t=4***, o más genéricamente entre ***t*** y ***t+h***.

En cambio, si el proceso no se hallaba inicialmente en equilibrio, y no hubiera **ningún efecto aleatorio**, las variaciones observadas en las distribuciones marginales de las distintas rondas reflejarían únicamente un verdadero cambio conducido por un modelo de Markov, y en ese caso las probabilidades basadas en las rondas 1 y 4 deberían equivaler al **efecto acumulado** de las probabilidades basadas en 1 y 2 aplicadas repetidamente. La proporción de casos en el estado *i* en la primera ocasión es p_i^1, y en general para la ocasión *t* es p_i^t Si el proceso es de Markov, y si no hubiera efectos aleatorios, de acuerdo a la ecuación 8 la proporción esperada en el estado *i* en la cuarta ocasión, tres rondas después de la primera, debería ser (aproximadamente):

$$p_i^4 \approx p_{ei}^4 = p_i^1 (U_{12})^3 \qquad \text{(Ec. 42)}$$

Aquí $\boldsymbol{U_{12}}$ es la matriz de probabilidades de transición basadas en las rondas 1 y 2; $(U_{12})^3$ es dicha matriz elevada al cubo, es decir esas mismas probabilidades aplicadas tres veces para llegar de la ocasión 1 a la ocasión 4; por su parte p_i^t representa la proporción de sujetos en el estado ***i*** en la ocasión ***t***, y p_{ei}^t es la proporción **esperada** en el estado ***i*** en la ocasión ***t,*** todo ello bajo la condición de que el proceso subyacente fuese exclusivamente un proceso de Markov, sin efectos extraños o aleatorios. En este caso, la matriz de probabilidades de transición entre la primera y la cuarta ocasión sería igual al cubo de la matriz de probabilidades entre las dos primeras ocasiones, implicando que las mismas probabilidades (observadas entre las dos primeras ocasiones) se aplican tres veces sucesivas:

$$U_{14} = (U_{12})^3 \qquad \text{(Ec. 43a)}$$

En la hipótesis anterior, es decir si el proceso estuviese inicialmente en equilibrio, y por lo tanto los cambios de las distribuciones marginales entre las dos primeras rondas no reflejasen auténticos cambios del estado latente de los sujetos sino solo fluctuaciones aleatorias, entonces el cruce de las respuestas dadas en ***t***=1 y en ***t***=4 debería arrojar aproximadamente los mismos resultados que el cruce entre ***t*** y ***t***+1, porque las fluctuaciones aleatorias no se acumulan a lo largo del tiempo. Algunas probabilidades serían levemente superiores o inferiores, debido a factores aleatorios, pero no mostrarían ninguna tendencia acumulativa. Se observaría aproximadamente lo siguiente:

$$U_{14} \approx U_{12} \qquad \text{(Ec. 43b)}$$

La ecuación (43a) difiere claramente de la ecuación (43b), pues en la primera la matriz U_{12} aparece elevada al cubo (o en general a la potencia ***h*** si se trata de la transición entre la ronda ***t*** y la ronda ***t+h***), mientras en la 43b las probabilidades de transición 1-4 son aproximadamente iguales a las probabilidades 1-2, difiriendo solo por fluctuaciones aleatorias. En el primer caso, el proceso no estaba inicialmente en equilibrio, y se va moviendo hacia él; en el segundo, el proceso arranca en un estado de equilibrio, y las variaciones son solo aleatorias.

Nótese que con estos dos casos extremos estamos tratando de **explicar** el proceso mediante el **supuesto** de que hay un proceso subyacente de tipo Markov, donde los movimientos netos solo ocurren cuando se parte de una situación de desequilibrio. **Suponiendo** que hay un proceso de Markov se puede aislar el efecto de ese proceso markoviano separándolo del efecto de eventuales fluctuaciones aleatorias, del mismo modo que en la regresión lineal se **supone** una relación lineal y ello permite separar la relación lineal y los errores aleatorios. Si no queremos suponer un proceso de Markov podríamos suponer cualquier otro proceso (del mismo modo que si no queremos postular una relación lineal podemos postular una relación cuadrática o de otro tipo), y aun en ese caso podríamos evaluar quizá si los datos son compatibles con un puro proceso de cambio (de acuerdo al modelo adoptado) o más bien implican un escenario de meras fluctuaciones aleatorias.

En líneas generales, por otra parte, la lógica subyacente del presente enfoque es que si entre las ondas 1 y 2 las distribuciones marginales no varían significativamente, se podría lanzar la hipótesis de que se trata de un proceso de Markov que se encuentra cercano al equilibrio. Si fuese así, las diferencias entre las distribuciones marginales en las ondas 1 y 2 se deberían solamente a factores

aleatorios; a su vez, si fuese así, la distribución marginal tampoco debería variar significativamente en las ondas 3 y 4. En cambio, si las distribuciones marginales de las dos primeras ondas fuesen muy diferentes, se concluiría que **si el proceso es markoviano** entonces el mismo no se encontraba en equilibrio en la primera onda; bajo esos supuestos, al menos parte de los cambios serían genuinos (movimientos markovianos hacia el equilibrio).

Todo esto, obviamente, presupone que el proceso subyacente sea markoviano, y además le otorga un privilegio conceptual a las primeras dos ondas, que en realidad podrían no ser representativas: quizá comenzando con las ondas 3 y 4 se llegaría a conclusiones distintas. En cualquier caso, para distinguir cambios "genuinos" de fluctuaciones aleatorias, es preciso definir de algún modo los que serían cambios genuinos; en este caso, serían genuinos los cambios dictados por un proceso de Markov. Toda esta línea de análisis está condicionada por estos supuestos implícitos: que el proceso es markoviano, y que se usen ciertas ondas (por ejemplo las dos primeras) para obtener una estimación de las probabilidades de transición. Si se levantan esos supuestos sería posible analizar el problema bajo otras hipótesis o modelos.

En el marco de la presente exposición, no obstante, el único tipo de proceso subyacente que se considera son los procesos de Markov, tal como en muchos análisis de regresión solo se considera la regresión lineal, pero el enfoque tiene potencialmente una validez más general.

Las dos hipótesis consideradas representan, pues, un caso donde no hay cambio (markoviano) auténtico sino solo fluctuaciones aleatorias (43b), y otro caso en el cual solo hay cambio impulsado por el proceso markoviano sin fluctuaciones aleatorias (43a). Las probabilidades de transición observadas entre las ocasiones 1 y 4 pueden aproximarse a una o a otra de estas dos situaciones extremas. Lo más probable es que se registre una situación intermedia, donde el sistema inicial no estaba en equilibrio (y por lo tanto hay cambios auténticos en las sucesivas rondas), y donde también hay algunos factores aleatorios. Si así fuese, las proporciones marginales finales p_i^{t+h} (por ejemplo las de t=4) frecuentemente resultarán estar entre las proporciones iniciales o sus cercanías (lo cual implicaría que se cumple la condición 43b) y las proporciones esperadas por el puro proceso de Markov, indicadas en la ecuación 42, y que implican la condición 43a. Las proporciones observadas pueden estar más cerca de uno de los extremos, o aproximadamente equidistantes. Ello podría ser evaluado en forma aproximada o cualitativa comparando dos estimaciones de las probabilidades de transición entre el inicio y el final de la trayectoria (por ejemplo, entre las ondas 1 y 4):

(1) La tabla de rotación **observada** entre las rondas 1 y 4, o en general 1 y ***h*** (donde ***h>2***),
(2) La tabla de rotación **esperada** entre las rondas 1 y 4 (o bien 1 y ***h***), la que refleja solo los efectos acumulativos de un proceso de Markov, cuyas probabilidades de transición fueron calculadas sobre la base de las rondas 1 y 2.

Si las probabilidades de transición derivadas de (1) y (2) son muy similares entre sí, y se presupone un proceso markoviano, se deduciría que no hay mucha incertidumbre de respuesta o variación aleatoria, sino solo el proceso de Markov. Si en cambio difieren entre sí puede deducirse que existe un componente aleatorio, mayor o menor según la magnitud de esa diferencia (o bien, que el proceso subyacente no es markoviano). El siguiente ejemplo puede ilustrar este tipo de situación. Las dos primeras ondas de una encuesta de panel arrojan la siguiente tabla de rotación de frecuencias.

	Onda 2		
Onda 1	**Respuesta A**	**Respuesta B**	**Total**
Respuesta A	400	400	800
Respuesta B	100	700	800
Total	500	1100	1600

De esta tabla se pueden estimar las probabilidades de transición:

	Onda 2	
Onda 1	**Respuesta A**	**Respuesta B**
Respuesta A	0.500	0.500
Respuesta B	0.125	0.875

A partir de estas cifras se pueden estimar las frecuencias **esperadas** para las ondas 3 y 4, aplicando las probabilidades de transición observadas entre las ondas 1 y 2. Primero se estiman las frecuencias esperadas en la onda 3 a partir de la onda 2, y luego las de la onda 4 a partir de las cifras **esperadas** de la onda 3. Los resultados son los siguientes.

	Onda 3 esperada		
Onda 2	**Respuesta A**	**Respuesta B**	**Total**
Respuesta A	250	250	500
Respuesta B	137	963	1100
Total	387	1213	1600

	Onda 4 esperada		
Onda 3 esperada	**Respuesta A**	**Respuesta B**	**Total**
Respuesta A	194	193	387
Respuesta B	152	1061	1213
Total	346	1254	1600

Ahora bien, las respuestas efectivamente obtenidas en las ondas 3 y 4 arrojaron los siguientes resultados:

	Onda 3 observada		
Onda 2	**Respuesta A**	**Respuesta B**	**Total**
Respuesta A	320	180	500
Respuesta B	160	860	1100
Total	480	1120	1600

	Onda 4 observada		
Onda 3 observada	**Respuesta A**	**Respuesta B**	**Total**
Respuesta A	240	240	480
Respuesta B	200	920	1120
Total	440	1160	1600

De acuerdo a los datos observados **se da sin duda un proceso de cambio**: el número de sujetos en la respuesta A va disminuyendo (de 800 iniciales a 500 en la segunda ronda, a 480 en la tercera y 440 en la cuarta); pero al mismo tiempo **ese cambio claramente no concuerda con las previsiones del proceso de Markov estimado a partir de las dos primeras ondas**, según el cual los casos con respuesta A habrían bajado de 800 a 500 y luego a 387 y 346. Evidentemente, la proporción de respuestas A disminuye mucho más lentamente que lo que se podría prever a partir de las dos primeras rondas. Ello significa que en algunas de las rondas (incluyendo las dos primeras) las respuestas estuvieron afectadas (en más o en menos) por otros factores, incluyendo factores aleatorios.

Ante esta situación, y sin descartar la posibilidad de tener en cuenta **otras variables** que pudieran explicar esa discrepancia, una de las explicaciones puede ser que junto al proceso de Markov con probabilidades de transición constantes hay también **factores aleatorios** que influyen en la evolución de las cifras. El problema ahora radica en medir la magnitud del efecto de esos factores, a fin de poder aislar el efecto del proceso de Markov intrínseco y eventualmente el efecto de otras variables, separándolo del efecto de factores aleatorios. La forma de separar o aislar el proceso de auténtico cambio separándolo de la incertidumbre de respuesta se discute en la sección siguiente.

5.2. Análisis del cambio con incertidumbre de respuesta

Cuando hay incertidumbre de respuesta, la medición directa del cambio a partir de las variaciones en las respuestas puede no dar resultados confiables. En esta sección se establecen bases para medir el cambio en el estado de los sujetos, una vez despejado el efecto de la incertidumbre de respuesta.

Hasta ahora se han concebido los datos de panel como un proceso de cambio de los sujetos, de un estado ***i*** consistente en dar la respuesta ***i***, a un estado ***j*** consistente en dar la respuesta ***j***. Ahora vamos a distinguir el **estado subyacente** del sujeto por un lado, y por otro lado las **respuestas manifiestas** que ese sujeto proporciona. Se supone que entre ambos aspectos no hay una relación determinista sino probabilística. Para modelizar el estado del sujeto en una forma sencilla, y sin que este artificio sea intrínsecamente necesario, vamos a suponer que el cambio de un estado a otro no es algo que le ocurre al sujeto como tal, sino algo que le ocurre a un conjunto de **elementos de respuesta,** o **factores de respuesta**, de los cuales a cada individuo le corresponde un gran número. Pueden concebirse como elementos subjetivos o bien como estímulos externos, todos ellos afectando al respondente. Cada uno de estos elementos puede estar condicionado a producir determinada respuesta ***i*** en una variable con ***m*** categorías, o puede estar en un estado "no condicionado" que no lo inclina a dar ninguna respuesta en particular. Esto significa que no se concibe que **el individuo** esté en un determinado estado, sino que son estos hipotéticos "elementos" los que están en uno u otro estado. Estos elementos están sujetos a un proceso estocástico de cambio continuo de estado, según ciertas tasas instantáneas de transición q_{ij} del mismo modo que anteriormente se concebía al individuo. Podemos imaginar estos elementos como "asesores" de los individuos que deben responder a la encuesta, suponiendo que cada individuo tiene una gran cantidad de "asesores", cada uno con una opinión propia, pero que cambian de opinión permanentemente a ciertas tasas q_{ij}. El individuo toma su decisión en un momento dado según el estado predominante de opinión de sus "asesores." Este modelo se aplica particularmente para modelizar las actitudes subjetivas de los individuos y sus respuestas manifiestas ante estímulos externos, pero puede aplicarse en general a toda clase de unidades de análisis y a toda clase de variables.[22]

Si la variable observable que interesa es una variable categórica con ***m*** categorías, por ejemplo una pregunta con ***m*** respuestas posibles, supondremos que, para cada individuo, **cada elemento de respuesta** (cada "asesor") puede estar en ***m*** estados diferentes. Cada uno de esos estados condiciona la producción de una cierta respuesta ***i*** (***i***=1, 2, ... ***m***)**.** En un momento dado **cada individuo *k*** poseerá una cierta cantidad w_{1k} de elementos que están en el estado 1, es decir, condicionados a la respuesta 1, otra cantidad w_{2k} de elementos que están en el estado 2, y así sucesivamente para los ***m*** estados posibles.[23] En ese momento, la probabilidad de que ese individuo produzca la respuesta ***i*** será:

$$P_k(i) = \frac{w_{ik}}{\sum_{j=1}^{m} w_{jk}} \qquad \text{(Ec. 44)}$$

Como el sujeto tiene elementos en todos los estados, aunque con diferente cantidad en cada uno de ellos, en cada instante los sujetos distribuirán sus respuestas entre las ***m*** respuestas posibles de acuerdo a la distribución de sus elementos entre esos ***m*** estados. No se puede afirmar con certeza qué respuesta dará cada individuo, sino sólo las probabilidades que tiene de dar diferentes respuestas. Si el mismo individuo fuese interrogado varias veces, es posible que en cada ocasión dé diferentes respuestas aun cuando el estado de sus elementos no haya cambiado en absoluto. Igualmente, si fuese entrevistado un grupo de individuos cuyos elementos internos estén distribuidos de igual modo entre los diferentes estados, aún así no todos darían la misma respuesta.

---

[22] Esta clase de modelos fue introducido originariamente como un modelo de aprendizaje por Estes y Burke 1955, y retomado con alcance más amplio por Coleman (1964a) y también en Coleman 1964b, cap.12 y 13.

[23] Para mayor simplicidad se dejan de lado los casos "sin respuesta"; implícitamente suponemos que no existen, o que entre los ***m*** estados de los elementos puede haber un estado no condicionado, en el cual no se favorece ninguna respuesta en particular, de modo que una de las ***m*** categorías posibles de respuesta es la ausencia de respuesta (lo que en las encuestas suele ser clasificado como "No sabe/No responde"), y sólo habría ***m***-1 categorías "válidas" aparte de la ausencia de respuesta. El problema general de los datos no válidos merece un tratamiento más amplio: véase por ejemplo Little y Rubin, 1987; o también Ahlo, 1990. Para el caso específico de los datos faltantes en los estudios de panel puede verse Alderman y otros, 2001.

Los cambios de estado de estos elementos hipotéticos son cambios **continuos**, gobernados por **tasas instantáneas de transición** q_{ij}. Debe remarcarse que estas no son tasas de transición de los individuos, sino de los "elementos de respuesta" o "asesores internos" de cada individuo. En un intervalo infinitesimal ***dt*** la probabilidad de que **un elemento** pase del estado ***i*** al estado ***j*** será igual a $q_{ij}dt$. Llámese v_{ik}^t a la probabilidad de que un elemento del individuo ***k*** esté en el estado ***i*** en el momento ***t***. El cambio en esas probabilidades de estado de los elementos vendrá dado **para cada individuo *k*** por un conjunto de ecuaciones como el siguiente:

$$\frac{dv_{1k}^t}{dt} = q_{11k}v_{1k}^t + ... + q_{i1k}v_{ik}^t + ... + q_{m1k}v_{mk}^t$$

.....

$$\frac{dv_{ik}^t}{dt} = q_{1ik}v_{ik}^t + ... + q_{iik}v_{ik}^t + ... + q_{mik}v_{mk}^t \qquad \text{(Ec. 45)}$$

.....

$$\frac{dv_{mk}^t}{dt} = q_{1mk}v_{1k}^t + ... + q_{imk}v_{ik}^t + ... + q_{mmk}v_{mk}^t$$

Dado que las q_{ikj} son aquí tasas instantáneas de transición **de los elementos de un individuo, e**n principio estas tasas se conciben como específicas para cada individuo, por lo cual se las denota como q_{ijk} y se interpretan como la tasa instantánea a la cual los elementos del individuo ***k*** pasan del estado ***i*** al estado ***j***. Todos los elementos de un individuo se suponen afectados por las mismas tasas instantáneas de transición. En las aplicaciones usuales no es posible calcular estas probabilidades separadamente para cada individuo; en cambio, se supone además que **todos los individuos están afectados por las mismas tasas de transición de sus elementos**, de modo que $q_{ijk} = q_{ij}$ para todo individuo ***k*** perteneciente a una determinada población o subpoblación. En cambio la distribución de los elementos, es decir, las proporciones de elementos de un individuo que en un momento dado están en cada estado, v_{ik}^t, son específicas a cada individuo, por más que (cuando ello no introduce ambigüedad) el subíndice ***k*** puede ser omitido indicándolas simplemente como v_i^t.

La solución del sistema de ecuaciones diferenciales (45) para un determinado individuo ***k*** es:

$$\begin{aligned} v_{1k}^{t+h} &= v_{1k}^t g_{11}^h + v_{1k}^t g_{21}^h + ... + v_{1k}^0 g_{m1}^h \\ &..... \\ v_{ik}^{t+h} &= v_{1k}^t g_{1i}^h + v_{1k}^t g_{2i}^h + ... + v_{1k}^0 g_{mi}^h \\ &..... \\ v_{mk}^{t+h} &= v_{1k}^t g_{1m}^h + v_{1k}^t g_{2m}^h + ... + v_{1k}^0 g_{mm}^h \end{aligned} \qquad \text{(Ec. 46)}$$

Las cantidades g_{ij}^h , que se suponen **iguales para todos los individuos** al igual que las q_{ij}, son **probabilidades de transición de los elementos** de un individuo, y son funciones de las tasas de transición instantánea q_{ij} y del intervalo ***h***; son análogas a las probabilidades de transición de individuos (u_{ij}^h) de los modelos más simples sin incertidumbre de respuesta. Estas cantidades g_{ij}^h son específicas para cada intervalo ***h***, e indican la probabilidad de que un **elemento** que se encontraba en el estado ***i*** en el momento ***t*** se encuentre en el estado ***j*** en el momento ***t+h***. En otros términos, permiten estimar la distribución de los elementos de un individuo en el momento ***t+h***, es decir las proporciones de estado de los elementos (v_{ik}^{t+h}) vigentes en el momento ***t+h***, a partir de las proporciones v_{ik}^t vigentes en el momento ***t***. La función que relaciona los valores g_{ij}^h con las tasas q_{ij} es una serie cuyos primeros términos son, como se vio antes en la ecuación 34 bis:

$$g_{ij}^h = \delta_{ij} + hq_{ij} + \frac{h^2}{2!}\sum_a q_{ia}q_{aj} + \frac{h^3}{3!}\sum_a\sum_b q_{ia}q_{ab}q_{bj} + ... \qquad \text{(Ec. 47)}$$

donde el primer término (δ_{ij}) es la delta de Kronecker, igual a 1 si ***i=j*** e igual a cero si ***i≠j***. Si v_{ik}^t es la probabilidad de que un **elemento** del sujeto ***k*** esté en el estado ***i*** en el momento ***t***, y todos los elementos de cada individuo analizados están gobernados por las mismas tasas de transición $\boldsymbol{q_{ijk}}$, entonces la proporción v_{ik}^t de elementos del sujeto ***k*** que están condicionados a la respuesta ***i*** será también la probabilidad de que ese individuo proporcione la respuesta ***i***:

$$v_{ik}^t = P_k(i) = \frac{w_{ik}}{\sum_{j=1}^{m} w_{jk}} \qquad \text{(Ec. 48)}$$

Este esquema "microdecisional", que representa lo que ocurre con cada individuo, debe ser usado para explicar la distribución de respuestas en la población, es decir las cantidades N_i^t de sujetos que suministran la respuesta ***i*** en un determinado momento ***t***, y que representan una proporción p_i^t del total *N* de sujetos. Del mismo modo, las tasas de transición **de los elementos** de cada individuo ***k***, es decir $\boldsymbol{q_{ijk}}$, deben ser usadas para explicar las probabilidades de transición de los *N* **individuos** entre diferentes estados, u_{ij}^h.

Para encontrar una solución a este problema es preciso adoptar alguna hipótesis sobre el origen de la variabilidad de las respuestas entre los diferentes individuos. Habría que tomar una decisión sobre la forma en que pueden variar las respuestas de diferentes individuos en un momento dado. Hay dos **posibilidades extremas:** la variabilidad puramente **"intra-individual"** (dentro de los elementos de los individuos) y la variabilidad puramente **"inter-individual"** (entre diferentes individuos).[24]

Si la única fuente de variabilidad es **entre individuos,** en cada momento cada individuo está totalmente determinado a favor de una respuesta específica; en otras palabras en cada momento ***t*** hay una cantidad de sujetos cuya respuesta es ***i*** (donde ***i***=1, 2, ... ***m***). Para cada individuo, en ese momento la probabilidad de dar una de las respuestas será igual a 1, y la probabilidad de dar cualquiera de las otras respuestas será 0, de modo que en definitiva una proporción p_i^t de sujetos dará la respuesta ***i***. Este es el supuesto adoptado en las secciones anteriores, donde la "respuesta" de cada sujeto se identificaba con su "estado" de manera determinista. Hay sujetos en diferentes estados, y esta diferencia entre los individuos se traduce en sus diferentes respuestas. Cualquier cambio en las respuestas reflejará cambios en el estado de los sujetos. La coincidencia entre el estado verdadero de los individuos y sus respuestas observables radica en que la probabilidad de una respuesta es igual a uno, y la probabilidad de otras respuestas es igual a cero.

En la hipótesis contraria, de una pura **variabilidad intra-individual,** hay variabilidad **dentro de los individuos**, pero **todos los sujetos son iguales**, de modo que en cada momento del tiempo todos los sujetos distribuyen sus respuestas según una distribución probabilística dada por la ecuación 47;

---

[24] Los términos "variabilidad intra-individual" y "variabilidad inter-individual" son solo ilustrativos. Surgen de la situación típica en la cual los sujetos son individuos, y se mide una actitud, un rasgo psicológico o una disposición subjetiva, que no son directamente observables, pero que son inferidos a través de respuestas objetivas ante determinadas preguntas. Pero el problema general que diferencia entre los estados latentes y las mediciones manifiestas se aplica también a otras clases de situaciones. En un caso hay variabilidad a nivel de los "elementos" de cada "caso", en el otro hay variabilidad entre "casos". Por ejemplo, el estado latente puede ser "pobreza" y los datos manifiestos pueden ser la presencia o ausencia de determinadas condiciones de bienestar como una vivienda adecuada, servicios sanitarios o acceso a la educación de los niños. Una persona puede ser "pobre" a pesar de tener vivienda adecuada, así como puede ser "no pobre" y sin embargo vivir en una vivienda inadecuada. La dualidad entre estado subyacente y respuesta surge en cualquier caso en el cual una dimensión subyacente es medida a través de indicadores observables imperfectos. Si los sujetos de estudio fuesen grupos humanos, y la variable fuese una variable relacionada con la conducta grupal (como por ejemplo la elección democrática entre varias alternativas), los "elementos" podrían ser los individuos que los componen, los cuales pueden cambiar de estado (es decir de opinión) subjetivamente y así determinar la respuesta del grupo colectivamente considerado. Así que en general no se trata de "variables psicológicas y sociológicas" sino de variables latentes y manifiestas.

dentro de todos los individuos hay una homogénea probabilidad ***p(i)*** de producir la respuesta ***i***, de modo que una proporción v_i^t de los sujetos producirá cada una de las respuestas ***i*** (donde ***i***=1, 2, ... ***m***). Si esta es la situación, entonces todas las diferencias en las respuestas de los sujetos se deben a incertidumbre de respuesta, y no a diferencias reales entre los sujetos, y por lo tanto todos los cambios observados en las respuestas reflejan sólo incertidumbre de las mismas, y no verdaderos cambios de los sujetos.

Obviamente, puede haber también una situación intermedia, en la cual haya por una parte procesos estocásticos a nivel de los elementos de cada individuo, y por otro cierta variabilidad también entre los diferentes individuos. El problema entonces radica en poder separar la variabilidad "psicológica" de la "sociológica", o más genéricamente la variabilidad entre los elementos componentes de cada sujeto, y la variabilidad entre sujetos. De este modo podrá decirse qué parte de los cambios observados corresponde a cambio "verdadero", y qué parte se origina en la incertidumbre de respuesta.

En primer lugar se puede notar que **para cada individuo** la secuencia de distribuciones de sus elementos, es decir el vector-fila V_i^t para diferentes fechas ***t***=0, 1, 2, ..., viene dado en forma similar a la ecuación 32 bis:

$$V_{t+h} = V_t e^{Qh} \qquad \text{(Ec. 49)}$$

donde e^{Qh} es una matriz que es, como en la ecuación 33, es la suma de una serie matricial exponencial cuyos primeros términos son los siguientes:

$$e^{Qh} = 1 + Qh + \frac{Q^2h^2}{2!} + \frac{Q^3h^3}{3!} + \frac{Q^4h^4}{4!} + ... \qquad \text{(Ec. 50)}$$

Cada término de la matriz e^{Qh} representa aquí la proporción de **elementos** de un individuo que se encontraban en el estado ***i*** en el momento ***t*** y en el estado ***j*** en el momento ***t+h***, es decir las probabilidades g_{ij}^h. La ecuación 49 predice la distribución **de los elementos de cada individuo** en el momento ***t+h*** a partir de la distribución en el momento ***t***, usando para ello las probabilidades de transición **de los elementos**. El problema es que así como se desconoce a distribución V de los elementos, también se desconocen sus probabilidades de transición g_{ij}^h. Sólo se conocen las respuestas manifiestas producidas por los individuos. El problema empírico consiste en usar los datos obtenidos sobre una muestra **de individuos** para estimar las tasas instantáneas de transición **de los elementos** dentro de cada individuo.

Las ecuaciones 46 a 50 permiten relacionar las tasas inobservables de transición instantánea de los elementos de cada individuo, q_{ij}, con las probabilidades de que un determinado sujeto ***k*** produzca las diferentes respuestas 1, 2, ..., ***i***,... ***m***, es decir la proporción v_i^t de sus elementos que se encuentran en ese momento condicionados a favor de producir la respuesta ***i***. La tarea consiste en relacionar estas probabilidades intra-individuales v_i^t, con la proporción p_i^t de personas que efectivamente producen la respuesta ***i***, y relacionar también las probabilidades inobservables g_{ij}^h de **transición de elementos** en un intervalo de longitud ***h***, con las probabilidades observables u_{ij}^h de **transición de individuos**, las cuales indican la probabilidad de que un individuo que dio la respuesta ***i*** en el momento ***t*** produzca la respuesta ***j*** en el momento ***t+h***. Por haber admitido que hay incertidumbre de respuesta, las proporciones observadas p_i^t de sujetos que dan la respuesta ***i*** no puede tomarse como una estimación de la proporción v_i^t de elementos intra-individuales que se encuentran condicionados a dar la respuesta ***i***, ni tampoco las probabilidades u_{ij}^h de transición de los individuos, pueden tomarse como estimaciones de las probabilidades g_{ij}^h de transición de los elementos en un intervalo de longitud ***h***.

Si la variable tiene ***m*** categorías de respuesta, el estado de cada individuo en un momento cualquiera ***t*** está caracterizado por el estado de sus elementos, los cuales se distribuyen en las proporciones v_i^t para cada individuo. La proporción de elementos de un individuo que están en el estado ***i*** es v_i^t, y esa proporción varía de un individuo a otro, situándose entre 0 y 1. El conjunto de *N* individuos en el momento ***t*** puede representarse entonces mediante una distribución multivariada de frecuencias relativas con ***m-1*** dimensiones. Se consideran ***m-1*** dimensiones porque una de las categorías se obtiene por diferencia, dado que la suma de las v_i^t es igual a uno. Integrando esa distribución en todas sus dimensiones en el intervalo de 0 a 1 se llega a la solución (véase Coleman 1964a, pp.19-23). Si se tienen datos de tres momentos 0, 1 y 2, entonces para cada fila de g_{ij}^h se puede formular un sistema de ecuaciones que relaciona las **proporciones de flujo de los individuos,** $p_{ij}^{t,t+h}$, con las **probabiidades de transición de los elementos**, g_{ij}^h. El siguiente sistema explica las probabilidades de flujo entre los momentos 0 y 2, para todos aquellos flujos que terminan en el estado ***j***:

$$
\begin{aligned}
p_{1j}^{02} &= p_{11}^{01} g_{1j}^h + p_{12}^{01} g_{2j}^h + \ldots + p_{1m}^{01} g_{mj}^h \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
p_{ij}^{02} &= p_{i1}^{01} g_{1j}^h + p_{i2}^{01} g_{2j}^h + \ldots + p_{im}^{01} g_{mj}^h \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
p_{mj}^{02} &= p_{m1}^{01} g_{1j}^h + p_{m2}^{01} g_{2j}^h + \ldots + p_{mm}^{01} g_{mj}^h
\end{aligned}
\qquad \text{(Ec. 50)}
$$

Debe remarcarse que las p_{ij} en estas ecuaciones son **proporciones de flujo**, es decir, cada flujo N_{ij} dividido por el número total de sujetos *N*. No deben confundirse con las **probabilidades de transición de individuos en el intervalo h,** es decir r_{ij}^h, que son el cociente del flujo N_{ij} entre ***t*** y ***t+h***, sobre el total de personas que en el momento inicial ***t*** estaban en el estado ***i***, es decir N_i^t. En la tabla de rotación, las p_{ij} son proporciones o porcentajes **sobre el total de la tabla**, mientras las u_{ij}^h son proporciones o porcentajes **sobre el total de la fila**.

Este sistema de ***m*** ecuaciones permite estimar las ***m*** probabilidades de transición entre los elementos correspondientes a la columna ***j***, es decir g_{ij}^h. Se puede repetir el procedimiento para cada una de las ***m*** categorías de la variable, formando así ***m*** sistemas de ecuaciones como el sistema (50), y resolviendo así todas las probabilidades de transición entre elementos, g_{ij}^h. En realidad, los sistemas necesarios son ***m***-1, porque el total de las probabilidades debe ser igual a uno, de modo que la última se obtiene por diferencia. En términos matriciales, este conjunto de ***m-1*** sistemas de ***m*** ecuaciones puede denotarse por el siguiente producto matricial:

$$P(0,2)=P(0,1)G(h) \qquad \text{(Ec. 51)}$$

donde *P0,1)* es la matriz de las proporciones p_{ij}^{01} en el intervalo que va desde ***t***=0 hasta ***t***=1, separados por un intervalo de longitud ***h***; y en forma análoga *P(0,2)* es la matriz de las cantidades p_{ij}^{02} para el intervalo de longitud ***2h*** entre ***t***=0 y ***t***=2. La solución para la matriz *G(h)*, que contiene todas las probabilidades g_{ij}^h de transición entre estados de los elementos en un intervalo ***h***, implica obtener la matriz inversa de *P(0,1)*:

$$P(0,1)^{-1}\, P(0,2) = G(h) \qquad \text{(Ec. 52)}$$

Estas estimaciones de g_{ij}^h, a su vez, pueden ser usadas para estimar las probabilidades de estado de los elementos, v_i^t, mediante las ecuaciones (46), y las tasas instantáneas de transición de los elementos, q_{ij}, a partir de las ecuaciones (47). Este último paso sigue un camino similar al

desarrollado para el caso general sin incertidumbre de respuesta, y procede mediante un método iterativo.

Usando similares ecuaciones una vez estimadas las probabilidades *G(h),* se puede obtener la distribución de equilibrio de los individuos entre los diferentes estados. Las ecuaciones básicas son de la forma:

$$p_i^* = p_1^* g_{1i}^h + p_2^* g_{2i}^h + ... + p_m^* g_{mi}^h \qquad \text{(Ec. 53)}$$

Por transposición esto se convierte en ecuaciones de la forma:

$$0 = p_1^* g_{1i}^h + p_2^* g_{2i}^h + ... + p_i^* (g_{ii}^h - 1) + ... + p_m^* g_{mi}^h \qquad \text{(Ec. 54)}$$

Este sistema se debe plantear para ***m-1*** categorías de la variable, pues la última se obtiene por diferencia debido a que la suma de todas ellas debe ser igual a 1. Esto significa que hay ***m***-1 ecuaciones independientes como la ecuación 54, más otra que es simplemente $\sum p_i^* = 1$ con las cuales se pueden estimar la distribución de equilibrio de los individuos entre las diferentes respuestas manifiestas, p_i^*, a partir de las tasas de transición entre los elementos internos de los mismos individuos.

5.3. Incertidumbre de respuesta en presencia de cambio

En la sección anterior se obtuvieron estimaciones de las tasas instantáneas de transición entre los elementos internos de cada individuo, y otras magnitudes inobservables, a partir de las tablas observadas de rotación de los individuos, bajo el supuesto de que las respuestas manifiestas están sujetas a una distribución probabilística, de modo que dos sujetos idénticos (o distintas entrevistas con el mismo sujeto cuando es interrogado más de una vez) pueden producir respuestas diferentes. Ese tratamiento, sin embargo, no permitió **cuantificar la incertidumbre de respuesta**. Ese es el objetivo de esta sección.

La incertidumbre de respuesta podría concebirse como una tabla de rotación entre dos respuestas producidas **sin separación en el tiempo**, es decir, sin tiempo suficiente para que ocurran cambios de estado. Las celdillas de esa hipotética tabla de rotación entre ambas respuestas no separadas en el tiempo contendrán unas probabilidades (o frecuencias relativas) p_{ij}^{00}, que representan la proporción de individuos que dan la respuesta ***i*** y la respuesta ***j*** en dos experimentos simultáneos sin cambio en la posición subyacente de los sujetos. A menudo en los tests psicológicos se incluyen dos versiones muy similares de la misma pregunta o estímulo, y los resultados se usan para medir la **confiabilidad** de la pregunta: una pregunta totalmente confiable debería dar la misma respuesta en las dos versiones, de modo que todos los casos deberían situarse en la diagonal principal de la tabla de rotación; las proporciones p_{ij}^{00} serían iguales a cero para todas las celdillas ubicadas fuera de la diagonal principal (celdillas definidas por ***i≠j***). Si existe alguna incertidumbre de respuesta, algunas de esas celdillas fuera de la diagonal principal albergarían casos, y la confiabilidad de a respuesta sería menos que perfecta.[25] Las proporciones p_{ij}^{00} son, por lo tanto, una medida de la incertidumbre de respuesta, y en especial para los casos en que ***i≠j***.

Si bien la obtención empírica directa de las proporciones p_{ij}^{00} es difícil, y muchas veces imposible, existe la posibilidad de aplicar los procedimientos anteriormente descritos a fin de **estimar el valor esperado** de p_{ij}^{00}. Aparte de esas incertidumbres de respuesta en el momento t=0, también se pueden estimar las mismas proporciones para el instante 1, es decir p_{ij}^{11} o en el instante 2, o en

[25] Estos estímulos repetidos son difíciles de aplicar en el caso de las encuestas porque el sujeto recuerda su primera respuesta y eso influye sobre la segunda; pero pueden ser aplicados con más facilidad cuando se trata de respuestas fisiológicas o de otro tipo donde no infuya mayormente la conciencia del sujeto, o bien cuando se usan dos versiones del indicador que en principio no parecen referirse al mismo tema. En algunos estudios psicológicos sobre actitudes, o variables subyacentes similares, es frecuente el uso de versiones diferentes del mismo indicador, que además aparecen mezcladas con otros indicadores irrelevantes para el tema, a fin de "despistar" al respondente.

general en el instante ***t,*** o sea p_{ij}^{tt}. Para ello es necesario estimar las tasas instantáneas de transición q_{ij} y las probabilidades de transición g_{ij}^{h} sobre la base de dos períodos, para luego estimar las incertidumbres de respuesta en el tercero. Las ecuaciones 50 y 51 relacionan las rotaciones entre los períodos 0 y 1 con las rotaciones entre 0 y 2, utilizando las probabilidades g_{ij}^{h} que transforman las respuestas dadas en el momento 1 en las respuestas producidas en el momento 2. Del mismo modo se podría plantear una similar relación para el período que va del momento 0 al momento 1:

$$P(0,1)=P(0,0)G(h) \qquad \text{(Ec. 53)}$$

La matriz de incertidumbre de respuesta P(0,0) puede ser obtenida en una forma similar a la indicada en la ecuación 52 para calcular *G(h)*, usando esta vez la inversa de la matriz *G(h)* que se supone ya calculada anteriormente:

$$P(0,0)= P(0,1)\ G(h)^{-1} \qquad \text{(Ec. 54)}$$

Las estimaciones de *G(h)* obtenidas anteriormente mediante la relación entre *P(0,1)* y *P(0,2)* son usadas aquí para estimar *P(0,0*). Es obvio que en ausencia de incertidumbre de respuesta sería *P(0,0)=**I***. En ese caso la matriz de incertidumbres de respuesta sería una matriz identidad, pues los que dieron la respuesta ***i*** en el momento ***t*** tendrían una probabilidad igual a uno de volver a dar la misma respuesta en el momento ***t+h*** o en el momento ***t-h***. Probablemente eso no sea así: si las proporciones de flujo fuera de la diagonal principal son diferentes de cero, es decir si *P(0,0)≠**I***, la diferencia permitiría estimar la incertidumbre de respuesta en el momento ***t***=0. De modo similar se puede estimar la matriz de incertidumbre de respuesta para el momento 1, es decir:

$$P(1,1)=P(1,2)[G(h)]^{-1} \qquad \text{(Ec. 55)}$$

Para estimar las incertidumbres del momento final ***t***=2 hay una complicación, pues se necesitan las **probabilidades retrospectivas** g_{ij}^{h*} que indican la probabilidad de que los elementos que estaban en ***j*** en el momento ***t+h*** hayan estado en ***i*** en el momento ***t***. Las probabilidades retrospectivas de transición de elementos se calculan por un procedimiento similar (conceptualmente, una integración hacia atrás desde ***t+h*** hasta ***t***, en lugar de integrar desde ***t*** hasta ***t+h***; véase Coleman, 1964a, pp.57-59). Estas probabilidades integran una matriz *G*(h)* con la cual se puede obtener la matriz de incertidumbres de respuesta del momento ***t***=2:

$$P(2,2)=P(1,2)[G^{*}(h)]^{-1} \qquad \text{(Ec. 56)}$$

Este procedimiento permite obtener para cada momento analizado (o incluso para momentos **τ** cualesquiera del pasado o del futuro) una estimación de la matriz de incertidumbre de respuesta *P(**τ**, **τ**)*. Este resultado puede ser el punto de partida de varios tipos de análisis.

Por ejemplo, la suma de todos los elementos de *P(**τ**, **τ**)* situados fuera de la diagonal principal, dividida por el número ***m*** de categorías de la variable, proporciona una medida global, z_t, con valores que varían entre 0 y 1, del grado de incidencia de la incertidumbre de respuesta en una determinada ronda del panel realizada en el momento ***t***:

$$z_t = \frac{1}{m}\sum_{j\neq i} p_{ij}^{tt} \qquad \text{(Ec. 57)}$$

El valor complementario ***1- z_t*** puede considerarse como un indicador de la **certidumbre** de respuesta. Si no hay incertidumbre de respuesta la matriz *P(**τ**, **τ**)* es una matriz identidad, cuyos elementos no-diagonales suman cero, mientras que los elementos de la diagonal principal son todos iguales a 1, y por lo tanto suman ***m***. En ese caso será z_t=0. Un valor de z_t superior a cero sugiere que los datos podrían ser explicados mediante la combinación de un proceso de Markov gobernando los cambios en los elementos latentes, y un cierto grado de incertidumbre en la respuesta manifiesta. En forma similar, si se define una matriz **D=[*I* -- *P(**τ**, **τ**)*]^2**,** la suma de sus elementos Σd_{ij} tiene una distribución χ^2 con ***m(m-1)*** grados de libertad, que permite evaluar si la incertidumbre de respuesta es estadísticamente significativa.

6. Modelos multivariados de panel con variables categóricas

Hasta ahora hemos considerado principalmente modelos **univariados**, en los cuales se trata de una sola variable y sus cambios a lo largo del tiempo. Los **modelos multivariados de panel** tratan de postular **relaciones entre diferentes variables**, observadas en diferentes fechas, a fin de **explicar** los datos observados. Las hipótesis en que se basan esos modelos son usualmente hipótesis **causales**, las cuales postulan que existe algún proceso de **influencia de una variable sobre otra**, de modo que un cambio en una variable **induce** cambios en otra. La postulación de relaciones causales es un complejo problema epistemológico, cuyas dificultades a menudo se proyectan sobre los problemas prácticos del investigador; en esta presentación, sin embargo, esta problemática no será considerada.[26] En los procesos multivariados no interesa solamente estudiar los cambios en la ubicación de los sujetos entre diferentes valores de la misma variable, sino la relación entre diferentes variables a lo largo del tiempo, estudiada a través de los datos de panel.

6.1. Problemas del análisis de la causación

El análisis de panel es particularmente adecuado para el análisis causal, porque la causación **opera en el tiempo**. Los efectos siempre aparecen en fecha **simultánea o posterior** a las causas, nunca en una fecha anterior. El presente puede ser causado por el presente o por el pasado, pero nunca por el futuro. Cuando dos variables varían en forma concomitante, la que ocurra primero es la que puede ser considerada como causa, y la que ocurre después como el efecto (obviamente, también puede darse el caso de que ambas obedezcan a una causa común, sin influirse mutuamente). En los estudios de sección transversal, donde todas las observaciones corresponden a la misma fecha, la determinación del orden causal de las variables es a menudo un problema. En los estudios de panel existen medios para evitar la ambigüedad, ya que se cuenta con observaciones en varios momentos a lo largo del tiempo.

Uno de los principales problemas que surgen en este tema, y que también se presenta en el caso de paneles con variables cuantitativas, es el carácter **arbitrario** de las fechas elegidas para realizar las rondas del panel, las cuales no tienen por qué coincidir con los intervalos relevantes para estudiar efectos causales. Hay sin duda procesos causales que operan a intervalos fijos; por ejemplo, la cantidad de niños en cuarto grado en el primer mes del año escolar es una función directa de los resultados finales del año escolar anterior, registrados tres o cuatro meses antes. En cambio, entre el primer y cuarto mes de clases no habría cambios comparables. Las fechas en que ocurren los cambios, y las demoras entre causas y efectos, están determinadas por el proceso mismo.

En cambio hay otros procesos que operan de manera prácticamente continua, sin fechas predeterminadas para dar saltos de un estado a otro. Los cambios en la situación laboral de las personas ocurren en cualquier momento, y obedecen a factores que pueden haber ocurrido inmediatamente antes, o un mes antes, o tres meses antes. La situación laboral de las personas encuestadas en la ronda de abril de una encuesta de empleo no tiene por qué explicar o predecir la situación laboral de esas personas en octubre, pues los sucesos desencadenantes de esta última pueden perfectamente haber ocurrido en los meses intermedios. Por otra parte, el período necesario para que las causas produzcan sus efectos puede ser un período variable, de modo que algunas personas sufren el efecto antes que otras. Al momento de la segunda ronda algunos efectos de la situación de abril no se habrán todavía producido, otros ya habrán ocurrido pero habrán sido eclipsados por eventos ulteriores, y la situación en la segunda ronda será entonces la resultante de una variedad de causas, distribuidas en el tiempo a lo largo de los meses precedentes, sin que sea posible aislar específicamente el efecto que tuvo la situación registrada en abril. Este problema se relaciona con el problema general, examinado antes, que implica reflejar un **proceso continuo** a través de un panel con observaciones realizadas en momentos discretos espaciados en el tiempo.

---

[26] Hay una vasta bibliografía sobre causalidad en epistemología y en disciplinas específicas como la economía y la sociología. Sólo para indicar algunas lecturas: Popper [1934], cap.III; Simon (1952 y 1987), Bunge (1979), Davis (1985); Schuster (1982). Para la aplicación de modelos causales en las ciencias sociales puede verse Blalock (1964, 1969, 1985). Véase también Eerola 1994.

Dejando de lado por ahora este problema, supóngase de todas maneras un modelo causal con dos variables X e Y. Estas variables pueden ser concebidas, a los fines de este análisis, como variables cuantitativas, o bien como variables dicotómicas codificadas con cero para la ausencia y uno para la presencia de determinada característica o atributo. Si fuesen variables politómicas de tipo nominal u ordinal el modelo debería adaptarse, pero no es necesario introducir esa complicación ahora pues no altera la esencia del asunto.

Las relaciones causales entre estas variables pueden ser **sincrónicas** o **diacrónicas**. En una relación causal sincrónica, el estado de una variable en el momento ***t*** influye sobre el estado de otra variable en el mismo momento ***t***. Por ejemplo, la temperatura **actual** del agua influye sobre el estado **actual** del agua (líquido, sólido o gaseoso). En una relación causal diacrónica, el estado de una variable en una fecha (o período) ***t-k*** influye sobre el estado de la misma o de otra variable en una fecha (o período) posterior ***t***. El intervalo entre las fechas ***t-k*** y ***t***, en este caso, debe ser suficientemente largo para que el efecto se produzca, y no tan largo que ese efecto se esfume o diluya antes de ser registrado.

La causación muchas veces es concebida como un proceso que **necesariamente** involucra el paso del tiempo, de modo que la causación estrictamente sincrónica, en realidad, no podría existir. Cualquier influencia necesita algún tiempo para transmitirse, aunque ese tiempo sea muy breve. Toda causación, entonces, es diacrónica. Sin embargo, para fines prácticos muchos efectos **aparecen** en una forma sincrónica con sus causas porque son registrados o medidos al mismo tiempo. Si bien hay seguramente algún intervalo entre la causa y el efecto, ambos varían en estrecha concomitancia, por lo cual pueden ser considerados como sincrónicos.

Otro aspecto de los procesos causales es que ellos pueden ser unidireccionales o de causación recíproca. En un modelo unidireccional, **X** provoca cambios en **Y**, pero los cambios de **Y** no tienen ninguna influencia (directa o indirecta) sobre **X**. En un modelo de causación recíproca ocurre lo contrario. Cuando un modelo de causación recíproca es diacrónico las relaciones son más claras: $X_t \rightarrow Y_{t+1} \rightarrow X_{t+2} \rightarrow Y_{t+3} \rightarrow$**...** En un modelo de causación recíproca de tipo sincrónico las dos variables se influyen mutuamente sin que medie el paso del tiempo, lo que dificulta la identificación de cada influencia. En un proceso diacrónico de causación recíproca, donde se asume que el "intervalo de causación" que transcurre entre la causa y el efecto es aproximadamente igual al intervalo entre las observaciones, el estado de las variables en el momento ***t*** dependerá del estado de ambas variables en el momento ***t-1*** y de factores aleatorios no controlados (ε). Si se tratase de variables cuantitativas y la relación entre ellas fuese lineal, los vínculos causales diacrónicos y recíprocos podrían representarse a través del siguiente sistema de ecuaciones:

$$Y_t = \beta_{XY} X_{t-1} + \beta_{YY} Y_{t-1} + \varepsilon_Y \qquad \text{(Ec. 58)}$$

$$X_t = \beta_{XX} X_{t-1} + \beta_{YX} Y_{t-1} + \varepsilon_X \qquad \text{(Ec. 59)}$$

En este modelo causal no hay causación sincrónica, es decir, procesos causales que operen en el mismo período: los únicos efectos previstos son de un período sobre el siguiente. La introducción de efectos sincrónicos recíprocos daría lugar a las siguientes ecuaciones, donde el valor de cada variable en ***t*** no depende solamente de los valores en ***t-1*** sino también del valor de la otra variable en el mismo período ***t:***

$$Y_t = \beta_{XY} X_{t-1} + \beta_{YY} Y_{t-1} + \alpha_{XY} X_t + \varepsilon_Y \qquad \text{(Ec. 60)}$$

$$X_t = \beta_{XX} X_{t-1} + \beta_{YX} Y_{t-1} + \alpha_{YX} Y_t + \varepsilon_X \qquad \text{(Ec. 61)}$$

El esquema de las ecuaciones 60 y 61 corresponde a un modelo muy genérico que suele denominarse "modelo cruzado con retardos" (*cross-lagged model*: véase por ejemplo Finkel 1995, pp.24-31). Los vínculos causales **directos** involucrados en estas ecuaciones (dejando de lado los factores aleatorios) son los siguientes:

Orden temporal del vínculo	Vínculo	Coeficiente
Diacrónico	$X_{t-1} \rightarrow Y_t$	β_{XY}
	$X_{t-1} \rightarrow X_t$	β_{XX}
	$Y_{t-1} \rightarrow X_t$	β_{YX}
	$Y_{t-1} \rightarrow Y_t$	β_{YY}
Sincrónico	$X_t \rightarrow Y_t$	α_{XY}
	$Y_t \rightarrow X_t$	α_{YX}

Esta clase de modelos supone un proceso causal que opera en forma **discreta.** Se mide la causa en el momento ***t-1*** y el efecto en el momento ***t***, y lo que sucede entre esos dos momentos es irrelevante o ajeno al modelo. Sin embargo, en la mayor parte de los procesos causales la influencia causal opera en forma **continua**, y no entre puntos discretos del tiempo. Por tal razón es preferible conceptualizar el proceso causal como un proceso continuo si bien las observaciones empíricas sólo ocurren a intervalos discretos. Esto implica formalizar las relaciones entre variables como un proceso continuo, en el cual cada cambio en una de ellas determina un cambio en otra u otras. En el caso de las variables categóricas ese proceso es un **proceso continuo de cambios de estado** en una variable dependiente determinado por el estado de otra u otras variables, o por el cambio de estado de otras variables. En ese enfoque, la distinción entre cambios sincrónicos y diacrónicos pierde importancia.

En un estudio transversal o sincrónico, cuando se observan correlaciones entre tres o más variables a menudo se utilizan los coeficientes de correlación parcial (Blalock 1964) o de regresión parcial o "path coefficients", también llamados "coeficientes de dependencia" (Boudon 1967) para seleccionar el conjunto de vínculos causales que mejor se ajusta a los datos. Por ejemplo, si la correlación de X_{t-1} con Y_t controlando Z_{t-1} es más grande que la de Z_{t-1} con Y_t controlando X_{t-1} podría suponerse que X tiene mayor influencia que Z sobre la variable Y. En un temprano intento en el mismo sentido, Pelz y Andrews (1964) intentaron aplicar ese método a los estudios de panel a fin de establecer "prioridades causales" de una manera sistemática; sin embargo, ese enfoque en general se ha revelado erróneo, porque la correlación obedece no sólo a esas variables sino a factores aleatorios, y puede por lo tanto conducir a resultados equívocos o absurdos. En cambio las propuestas metodológicas para el análisis causal con variables categóricas se ha orientado más bien en dirección a la identificación de procesos continuos, ya sea mediante el enfoque inaugurado por Coleman (1964, 1991) o mediante modelos log-lineales (véase por ejemplo Hagenaars 1990 o el capítulo 11 de Agresti 1990), si bien estos últimos no han tenido un desarrollo comparable todavía en lo que se refiere a estudios de panel.

6.2. Procesos causales continuos con variables categóricas

Supóngase que la variable de interés es ***Y***, una variable categórica con ***m*** categorías. Cada individuo o unidad de análisis está sujeto permanentemente a la posibilidad de cambiar de estado. Su **tasa instantánea de transición** de un estado a otro es, en la notación que se ha usado anteriormente, q_{ij}. En cualquier pequeño intervalo de tiempo ***dt***, la probabilidad de un sujeto ubicado en el estado ***i*** de pasar al estado ***j*** es $q_{ij}dt$. Del mismo modo, la probabilidad de un sujeto ubicado en el estado ***j*** de pasar al estado ***i*** es $q_{ji}dt$. Por lo tanto la tasa de variación de la probabilidad de estar en el estado ***i*** sería igual a la suma algebraica de estas probabilidades de entrada y salida. Cuando hay sólo dos estados 0 y 1 la tasa de variación de p_1 sería:

$$\frac{dp_1}{dt} = q_{01}p_0 - q_{10}p_1 \qquad \text{(Ec. 62)}$$

La expresión para p_0 sería similar aunque los términos de la derecha aparecerían cambiados de signo:

$$\frac{dp_0}{dt} = -q_{01}p_0 + q_{10}p_1 \qquad \text{(Ec. 63)}$$

Esto significa que la tasa de variación de p_0 es simplemente la tasa de variación de p_1 cambiada de signo:

$$\frac{dp_0}{dt} = -\frac{dp_1}{dt} \qquad \text{(Ec. 64)}$$

Esto es obvio: las entradas en un estado por unidad de tiempo deben ser iguales a las salidas del otro estado por unidad de tiempo. Para variables con más de dos categorías las ecuaciones sufren una transformación trivial ya que hay que sumar todos los posibles estados con que puede tener intercambios el estado ***i***:

$$\frac{dp_i}{dt} = \sum_j (q_{ji} p_j - q_{ij} p_i) \qquad \text{(Ec. 65)}$$

En este caso, claramente, el cambio en la probabilidad de estar en el estado ***i*** equivale a la diferencia entre las entradas en ese estado y las salidas desde ese estado.

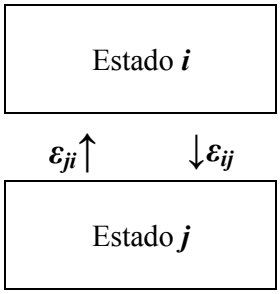
6.2.1. Cambio sin factores causales explícitos

Si la variable ***Y*** fuese independiente de toda otra variable, es decir, si no operase sobre ***Y*** ninguna influencia causal sistemática, los cambios de estado serían eventos puramente aleatorios. En todo instante algunos individuos pasarían de ***i*** a ***j***, mientras que otros pasarían de ***j*** a ***i*** (donde tanto ***j*** como ***i*** pueden variar de 1 a ***m***). Cuáles individuos sufrirían estos cambios de estado sería algo intrínsecamente impredecible, ya que todos tienen la misma probabilidad de sufrirlos. En esta situación de ausencia de factores causales sistemáticos, las tasas instantáneas de transición equivalen al efecto de una multitud de factores aleatorios no identificados, que pueden resumirse en el símbolo ε_{ji} que significa "influencias aleatorias que favorecen el cambio desde el estado ***j*** hacia el estado ***i***". El estado ***j*** aquí incluye **todos** los estados, incluso el propio estado de origen, ***i***, de modo que ε_{ji} incluye aquellas influencias aleatorias que provocan el **cambio** de estado del sujeto desde ***i*** hacia algún otro estado, o sea ε_{ji} (para $i \neq j$), y también las influencias aleatorias que hacen **permanecer** al sujeto en el estado ***i***, es decir ε_{ii}**.** Para cada estado ***j*** esto significa que $q_{ji} = \varepsilon_{ji}$ y para el conjunto de estados ***j***:

$$\sum_j q_{ji} = \sum_j \varepsilon_{ji} = \varepsilon_i \qquad \text{(Ec. 66)}$$

donde ε_i representa el conjunto de efectos aleatorios que influyen sobre la probabilidad de que un sujeto se encuentre en el estado ***i*** en un momento cualquiera. Entre dos estados ***i*** y ***j*** habrá en cada instante un flujo al azar en ambas direcciones, determinado por el conjunto de factores aleatorios ε_{ij} que provocan cambios desde el estado ***i*** hacia el estado ***j***, y por los factores aleatorios ε_{ji} que provocan cambios desde ***j*** hacia ***i***.

Cambios en el atributo Y por factores aleatorios



Si los factores aleatorios que operan a favor del ingreso en cada estado están balanceados con los factores que impulsan a salir de él, es decir si es $\varepsilon_{ij} = \varepsilon_{ji}$, la probabilidad de que un sujeto esté en un determinado estado será constante, y la cantidad **esperada** de sujetos en cada estado también será constante, en cuyo caso el proceso se encuentra en un **equilibrio estable**. Sin embargo, aun en un equilibrio estable habrá individuos que pasan aleatoriamente de un estado a otro. Por ejemplo, aunque la tasa de desempleo permanezca estable en un 6%, siempre habrá algunas personas ocupadas que pierden su empleo y algunas personas desocupadas que encuentran trabajo, debido a

factores aleatorios, aun cuando no se identifique ningún factor que explique por qué precisamente esos individuos pierden el empleo y esos otros lo encuentran. En otro ejemplo, aunque en cada árbol siempre haya determinado número de pájaros, en todo momento existe la posibilidad de que algunos pájaros vuelen hacia otro árbol, y otros pájaros se asienten en éste árbol, de manera totalmente aleatoria, manteniendo en promedio el mismo número de pájaros por árbol.

6.2.2. Factores causales

Los efectos aleatorios pueden concebirse como una expresión sintética que refleja el efecto de una **multitud indeterminada de causas** que operan sobre cada sujeto, cada una con un efecto particular a favor o en contra de la probabilidad de que ese sujeto se encuentre en el estado ***i***, influyendo para mantenerlo en su estado actual o impulsándolo a pasar a otro estado. De ese universo de factores, el análisis multivariado aísla algunos factores X_1, X_2, etc., para considerarlos como "causas" o "variables independientes", con el resto de los factores subsumido en el residuo aleatorio ε_i. De este modo se efectúa una partición de q_{ij} en dos partes: una de ellas determinada por estos factores identificados como causas, y el resto determinado por factores aleatorios. Para cada individuo ***k*** y para cada tasa q_{ij} se tendrá entonces:

$$q_{ij}^{k} = \beta_{1ijk} X_{1k} + \beta_{2ijk} X_{2k} + ... + \beta_{gijk} X_{gk} + ... + \varepsilon_{ijk} \qquad \text{(Ec. 67)}$$

donde los símbolos tienen el siguiente significado:

q_{ij}^{k} = Tasa instantánea de transición del estado ***i*** al estado ***j*** para el individuo ***k***.

β_{gijk} = Efecto del atributo X_g a favor de que el individuo ***k*** cambie de ***i*** a ***j***.

X_{gk} = Valor del atributo Xg para el individuo ***k***.

ε_{ijk} = Efectos aleatorios debidos a otros factores no considerados explícitamente, a favor de que el individuo ***k*** pase del estado ***i*** al estado ***j***.

Suponiendo que se ha estimado previamente la tasa de transición q_{ij}, la introducción de factores explicativos X_i permite **particionar** la tasa de transición en varias porciones: una que responde a un factor X_1, otra que responde a un factor X_2, etc., y un residuo no explicado que se supone aleatorio.

6.3. Efectos causales en un corte transversal

En esta sección se exploran los modelos causales aplicables entre variabes categóricas **en un corte transversal**, sin introducir todavía la dimensión temporal, es decir, sin aplicar diseños longitudinales. La introducción de otras variables categóricas como posibles variables independientes que explicarían causalmente la variable ***Y*** implica pensar que las tasas de transición entre distintos estados de la variable Y serán diferentes para cada individuo, según los valores que ese individuo tenga en las variables *X*. La tasa de transición laboral que afecta a un individuo, por ejemplo, podría ser diferente según su sexo, su nivel educativo, su experiencia laboral anterior u otros factores relevantes.

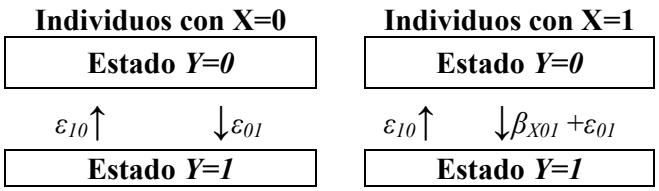
6.3.1. Variables dicotómicas con un solo factor independiente

Comenzaremos con **variables dicotómicas** pues como se ha visto la generalización a variables con múltiples categorías no ofrece mayores dificultades. La variable dependiente *Y* es una variable dicotómica, como por ejemplo la intención de votar por un determinado candidato, que puede tener sólo dos estados, codificados con 0 y 1. Se introduce un solo factor causal, la variable dicotómica *X* que consiste en la presencia o ausencia de determinado atributo (por ejemplo conocer o no conocer las propuestas del candidato en cuestión). La variable *X* está también codificada con 0 (ausencia del atributo) y 1 (presencia del atributo), y se supone que los códigos han sido elegidos de tal modo que la presencia del atributo *X* favorece el ingreso en el estado 1 de la variable dependiente *Y,* entonces las tasas de transición de *Y*=0 a *Y*=1 para el sujeto ***k*** serán:

Tasas de transición entre Y=0 e Y=1

Individuos con X=0:	$q_{01}^{k} = \varepsilon_{01k}$	$q_{10}^{k} = \varepsilon_{10k}$	(Ec. 68)
Individuos con X=1:	$q_{01}^{k} = \beta_{X01k} + \varepsilon_{01k}$	$q_{10}^{k} = \varepsilon_{10k}$	(Ec. 69)

La presencia del atributo X modifica el flujo entre Y=0 e Y=1



En este modelo teórico, la presencia del atributo *X* modifica el flujo desde ***i*** hacia ***j***, pero su ausencia no tiene ningún efecto especial. En otros términos, en ausencia de *X* los individuos se mueven entre Y=0 e Y=1 impulsados únicamente por factores aleatorios; pero en presencia de *X*=1 hay un incremento (o decremento) del flujo desde Y=0 hacia Y=1 representado por el coeficiente β_{X01} (que puede ser negativo o positivo). Estos flujos podrían ser estimados a partir de una tabla cruzada de *X* con nuestra variable dependiente *Y* aunque no se cuente con datos de panel, **si se asume que la tabla representa el estado de equilibrio**. Supóngase por ejemplo que se tiene la siguiente tabla:

	Atributo Y		
Atributo X	**Y=0**	**Y=1**	**Total**
X=0	N_{00} = 300	N_{01} = 200	500
X=1	N_{10} = 100	N_{11} = 400	500
Total	400	600	1000

Según se vio anteriormente para tablas de rotación, en un estado de equilibrio los flujos de entrada y salida en cada estado de *Y* deben estar balanceados. Las frecuencias de cada celdilla interior se denominan N_{ij} donde ***i*** representa uno de los valores de X mientras que ***j*** representa uno de los valores de Y**.** Ahora bien, la tabla observada empíricamente puede o no representar una situación de equilibrio. Se ignora si ella variaría en caso de ser observada en algún otro momento. Pero en ausencia de información específica, y si la observación se tomó en un momento cualquiera, es posible asumir que el proceso se encontrase en esa situación. **En una situación de equilibrio** deben cumplirse las siguientes igualdades:

Para los individuos con valor X=0: $\varepsilon_{01}N_{00} = \varepsilon_{10}N_{10}$ (Ec. 70)

Para los individuos con valor X=1: $(\beta_{X01} + \varepsilon_{01})N_{01} = \varepsilon_{10}N_{01}$ (Ec. 71)

Esto permite estimar el valor relativo, pero no absoluto, de los coeficientes. Nótese que $\varepsilon_{01}/\varepsilon_{10}=N_{10}/N_{00}$. Denominando p_{01} a la proporción de personas en el estado 1 de la variable dependiente *Y* respecto al total de personas en el estado 0 de la variable X (200/500 en la tabla precedente), y p_{11} a la proporción en el estado 1 de *Y* respecto al total de personas con *X*=1 (400/500 en la tabla), rápidamente se deducen las siguientes equivalencias:

$$p_{01} = \frac{\beta_{X01} + \varepsilon_{01}}{\beta_{X01} + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 72)}$$

$$p_{11} = \frac{\varepsilon_{01}}{\beta_{X01} + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 73)}$$

Se puede fácilmente obtener el valor relativo de cada coeficiente respecto a la suma de los tres, usando como ejemplo los valores numéricos del ejemplo anterior:

Efecto de X hacia el estado Y=1

$$\frac{\beta_{X01}}{\beta_{X01} + \varepsilon_{01} + \varepsilon_{10}} = \frac{p_{01} - p_{11}}{1 - p_{11}} = \frac{0.40 - 0.80}{0.20} = -2 \qquad \text{(Ec. 74)}$$

Efectos aleatorios hacia el estado Y=1

$$\frac{\varepsilon_{01}}{\beta_{X01} + \varepsilon_{01} + \varepsilon_{10}} = \frac{p_{11}(1 - p_{01})}{1 - p_{11}} = \frac{0.80(1 - 0.40)}{0.20} = 2.4 \qquad \text{(Ec. 75)}$$

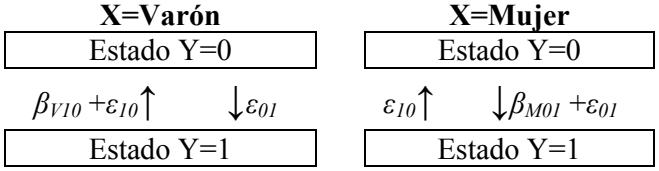
Efectos aleatorios hacia el estado Y=0

$$\frac{\varepsilon_{10}}{\beta_{X01} + \varepsilon_{01} + \varepsilon_{10}} = 1 - p_{01} = 1 - 0.40 = 0.60 \qquad \text{(Ec. 76)}$$

Los tres efectos, naturalmente, suman uno: –2.0 + 2.4 + 0.60 = 1.0. Cada uno de los componentes puede ser visto como aquella parte de la variación en la proporción de personas en el estado Y=1 que se debe a cada una de estas tres causas: el efecto de X=1, y los efectos aleatorios en ambas direcciones. Según este ejemplo, los efectos aleatorios más poderosos son aquellos que influyen a las personas para que pasen del estado 0 al estado 1, es decir ε_{01} = 2.4. Los efectos aleatorios en la dirección contraria son cuatro veces más pequeños ε_{10} =0.60. Y además, el efecto del atributo X consiste en inhibir fuertemente el flujo desde Y=0 hacia Y=1, ya que su valor es β_{X01} = – 2.0.

En este ejemplo, *X* tiene un efecto sobre *Y* cuando vale *X*=1, pero no cuando *X*=0. Esto es aceptable cuando se trata de variables que consisten en la **ausencia o presencia** de un atributo, como por ejemplo tener o no tener educación superior. Es posible, sin embargo, que el atributo *X* tenga efecto **en ambos sentidos**. Esto ocurre con aquellas variables cuyas categorías no representan presencia o ausencia sino **diferentes posibilidades cualitativas**. Por ejemplo, si *X*=Sexo, podría ser que los individuos de uno de los sexos tengan mayor tendencia a fluir desde ***i*** hacia ***j*** mientras los individuos del otro sexo tienen mayor tendencia a fluir desde ***j*** hacia ***i***, como se indica en el siguiente diagrama. Aquí se usa implícitamente el artificio de dividir la variable "Sexo" en dos variables dicotómicas conjugadas: "Ser varón" y "Ser mujer". Se supone que β_{V10} opera cuando "Ser varón"=1 y "Ser mujer"=0, y que β_{M10} actúa en el caso opuesto, cuando "Ser varón"=0 y "Ser mujer"=1.

Efectos dobles: Una de las categorías del atributo X aumenta el flujo desde 1 hacia 0, y la otra categoría tiene el efecto opuesto



En este caso, como se ve, hay dos efectos de dirección opuesta. El valor *X*=V (varón) incrementa uno de los flujos mientras el valor *X*=M (mujer) incrementa el otro. En este caso, una tabla cruzada sincrónica es insuficiente para estimar el valor (absoluto o relativo) de los cuatro coeficientes. La estimación sólo sería posible con datos de panel, como se verá luego. Sin embargo, aun en el caso de un corte transversal resulta posible la estimación si se utiliza algún supuesto simplificador. Por ejemplo, podría adoptarse la hipótesis de que el efecto β_{V10} de ser varón es de igual magnitud que el efecto contrario β_{M01} de ser mujer. Si $\beta_{V01}=\beta_{M10}=\beta$, los efectos relativos pueden estimarse en forma similar al caso anterior. Las proporciones de varones y mujeres serían:

$$p_{01} = \frac{\beta + \varepsilon_{01}}{\beta + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 77)}$$

$$p_{11} = \frac{\varepsilon_{01}}{\beta + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 78)}$$

De estas equivalencias se desprende fácilmente:

Efecto de ser varón hacia Y=1, o de ser mujer hacia Y=0

$$\frac{\beta}{\beta + \varepsilon_{01} + \varepsilon_{10}} = p_{V1} - p_{M1} = 0.40 - 0.80 = -0.40 \qquad \text{(Ec. 79)}$$

Efectos aleatorios hacia Y=1

$$\frac{\varepsilon_{01}}{\beta + \varepsilon_{01} + \varepsilon_{10}} = p_{11} = 0.80 \qquad \text{(Ec. 80)}$$

Efectos aleatorios hacia Y=0

$$\frac{\varepsilon_{10}}{\beta + \varepsilon_{01} + \varepsilon_{10}} = 1 - p_{01} = 1 - 0.40 = 0.60 \qquad \text{(Ec. 81)}$$

Si la variable *Y* fuera la intención de votar (1) o no votar (0) por un cierto candidato, la importancia relativa del flujo aleatorio entre los dos estados (que opera para ambos sexos por igual) es de 0.80 desde 0 hacia 1, y de 0.60 desde 1 hacia 0. El efecto del sexo es doble: a los varones les reduce el flujo desde 0 hacia 1, y a las mujeres les reduce el flujo desde 1 hacia 0. En otros términos, ser

varón obstaculiza la decisión de pasar a tener una intención positiva de voto, mientras que ser mujer dificulta la decisión de pasar a tener una intención negativa. Estos resultados, por supuesto, dependen de la validez del supuesto de que el efecto de ser varón es de igual magnitud que el efecto de ser mujer, aunque de sentido opuesto.

Como se ve, ambos modelos arrojan diferentes resultados numéricos, y no hay en los datos de corte transversal nada que permita escoger uno de ellos, ni que permita decidir si los dos efectos del sexo son iguales o diferentes. Si el marco teórico subyacente indica que el atributo *X* corresponde a la presencia o ausencia de algo, se impone la primera solución: el atributo *X* tiene influencia cuando está presente, y no la tiene cuando está ausente. En cambio, si el modelo teórico asume que *X*=1 indica la presencia de algo, mientras *X*=0 no indica la ausencia de ese algo sino la presencia de otra cosa, como por ejemplo cuando *X*=Sexo, ambas categorías de *X* pueden tener efectos propios, de direcciones opuestas y (por simplicidad) de igual magnitud, entonces el segundo modelo sería el apropiado. Sólo con datos longitudinales o de panel se podría obtener en los datos algún elemento que permita decidir cuál de los dos modelos teóricos es el apropiado en un caso particular, así como la magnitud relativa de los dos efectos del sexo (el efecto de ser varón y el efecto de ser mujer).

6.3.2. Análisis transversal multivariado con dicotomías

El análisis precedente, con un factor causal dicotómico, puede ser extendido al caso más general en que hay **varios** factores causales, siempre de tipo dicotómico. Posteriormente se generaliza este enfoque a variables categóricas con más de dos categorías. Supóngase que se tiene una variable dependiente dicotómica ***Y*** y tres factores independientes dicotómicos ***X***, ***W*** y ***Z***. Por ejemplo, con ***Y***=Intención de votar al candidato ***A***, los factores podrían ser ***X***= Educación universitaria; ***W***=Información sobre las propuestas del candidato; y ***Z***=Simpatizante del partido político del candidato A. Cada factor podría tener efecto solamente cuando está presente (es decir, cuando su valor es 1), o podría tener efectos diferenciados en sus dos valores (un efecto cuando vale 0, y otro efecto cuando vale 1). En otras palabras, puede tener un efecto "simple" o "doble". Por el momento propondremos el caso de un efecto "simple". Supondremos también que los tres factores son **aditivos**: el efecto de cada uno se suma al efecto de los demás sin **interacción**. Esto conduce al modelo siguiente:

$$q_{01:xwz} = \beta_X X + \beta_W W + \beta_Z Z + \varepsilon_{01} \qquad \text{(Ec. 82)}$$

$$q_{10} = \varepsilon_{10} \qquad \text{(Ec. 83)}$$

La **primera ecuación del modelo** dice que cada subgrupo de sujetos tendrá diferente tasa de transición según su combinación de valores en las variables independientes. La notación $q_{01:xwz}$ indica la tasa de transición de 0 a 1 de aquellos individuos con ciertos valores en X, W y Z. Los sujetos con nivel universitario (X=1), informados sobre el candidato (W=1) y simpatizantes del partido (Z=1) tendrán una tasa de transición desde "No tener intención de votar al candidato A" hacia "Tener intención de votarlo" equivalente a: $\boldsymbol{q_{01:111}=\beta_X+\beta_W+\beta_Z+\varepsilon_{01}}$. Los que no tengan estudios universitarios pero tengan información sobre el candidato y simpatía partidaria tendrán $\boldsymbol{q_{01:011}=\beta_W+\beta_Z+\varepsilon_{01}}$. Los que sean simpatizantes del partido ***Z*** pero carezcan de estudios superiores y desconozcan las propuestas del candidato tendrán $\boldsymbol{q_{01:001}=\beta_Z+\varepsilon_{01}}$. Y lo mismo para otras combinaciones. Si tienen cero en las tres variables, su tasa será $\boldsymbol{q_{01:000}=\varepsilon_{01}}$. La **segunda ecuación del modelo** sostiene que la posibilidad de pasar de 1 a 0 en la intención de voto es la misma para todos ($\boldsymbol{q_{10}=\varepsilon_{10}}$): la ausencia de X, W y Z no tiene ningún efecto especial.

Si no se dispone de datos de panel sino que se cuenta solamente con datos transversales, no se tiene información directa sobre la cantidad de sujetos que cambian de opinión, sino sólo una "fotografía" tomada en un momento dado. Si se supone que ese momento representa una situación de equilibrio, es decir que los flujos hacia y desde ***Y***=1 están compensados entre sí, es posible tener, como antes, una estimación de la importancia relativa de cada coeficiente (aunque no su magnitud absoluta).

En forma similar al caso de un solo factor independiente, las proporciones **esperadas** en el estado 1 de la variable dependiente Y pueden ser expresadas en función de los coeficientes del modelo. Para ello la notación será análoga a la indicada para las tasas de transición: $\boldsymbol{p_{111}}$ es la proporción que está en el estado 1 de Y (tiene intención de votar al candidato A) entre aquellos sujetos que tienen

respuesta positiva en las tres variables independiente XWZ, y en general $\boldsymbol{p_{xwz}}$ es la proporción que piensa votar por el candidato entre aquellos que tienen una cierta combinación de valores de X,W y Z. Añadiremos además un asterisco para indicar que se trata del valor esperado que debería tener en condiciones de equilibrio si los datos respondieran al modelo. Para algunas de las combinaciones resulta por ejemplo:

$$p^*_{000} = \frac{\varepsilon_{01}}{\beta_X + \beta_W + \beta_Z + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 84)}$$

$$p^*_{100} = \frac{\beta_X + \varepsilon_{01}}{\beta_X + \beta_W + \beta_Z + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 85)}$$

$$p^*_{111} = \frac{\beta_X + \beta_W + \beta_Z + \varepsilon_{01}}{\beta_X + \beta_W + \beta_Z + \varepsilon_{01} + \varepsilon_{10}} \qquad \text{(Ec. 86)}$$

Del mismo modo se pueden formular las otras combinaciones (001, 010, 101, 110 y 011). El numerador es la suma de los coeficientes activos en cada combinación de valores de X, W y Z, y el denominador es la suma de todos los coeficientes. Para simplificar la notación, denominaremos b_X al cociente de β_X sobre la suma de todos los coeficientes, y de igual manera definiremos b_W, b_Z, e_{01} y e_{10}, de modo que la suma de los cinco coeficientes es igual a la unidad. Estos coeficientes estandarizados pueden ser estimados por una variante del método de mínimos cuadrados, es decir, minimizando la suma (elevada al cuadrado) de las diferencias entre las proporciones observadas *p* y las proporciones esperadas *p** incluidas en las precedentes ecuaciones, es decir minimizando $\sum(p_{xwz} - p^*_{xwz})^2$. Tal como demuestra Coleman (1964b, pp.196-197), esto conduce a expresiones muy simples para los coeficientes estandarizados ***b*** y ***e***. En este caso se estiman cuatro de ellos y se estima el último por diferencia ya que su suma es igual a uno. Los tres coeficientes $\boldsymbol{b_X}$, $\boldsymbol{b_W}$ y $\boldsymbol{b_Z}$ se estiman como promedio de las diferencias entre las proporciones observadas en presencia y en ausencia de cada variable:

$$b_X = \frac{(p_{100} - p_{000}) + (p_{110} - p_{010}) + (p_{101} - p_{001}) + (p_{111} - p_{011})}{4} \qquad \text{(Ec. 87)}$$

$$b_W = \frac{(p_{010} - p_{000}) + (p_{110} - p_{100}) + (p_{011} - p_{001}) + (p_{111} - p_{101})}{4} \qquad \text{(Ec. 88)}$$

$$b_Z = \frac{(p_{001} - p_{000}) + (p_{101} - p_{100}) + (p_{110} - p_{010}) + (p_{111} - p_{110})}{4} \qquad \text{(Ec. 89)}$$

El coeficiente estandarizado de efectos aleatorios hacia el estado 1, es decir e_{01}, resulta ser:

$$e_{01} = \frac{2p_{000} + p_{100} + p_{010} + p_{001} - p_{111}}{4} \qquad \text{(Ec. 90)}$$

El último coeficiente estandarizado, e_{10}, se obtiene por diferencia:

$$e_{10} = 1 - b_X - b_W - b_Z - e_{01} \qquad \text{(Ec. 91)}$$

Este procedimiento puede generalizarse fácilmente a un número cualquiera de factores dicotómicos aditivos. Si existen ***m*** atributos dicotómicos independientes, la fórmula general de un coeficiente estandarizado b_X para un atributo independiente cualquiera ***X*** será:

$$b_X = \frac{\sum_{c=1}^{c=2m-1}(p_{xc} - p_c)}{2^{m-1}} \qquad \text{(Ec. 92)}$$

donde "***c***" representa una determinada combinación de los demás factores (además de *X*) que intervienen en el modelo. En esta notación, $\boldsymbol{p_{xc}}$ es la proporción de sujetos en el estado 1 de la variable dependiente *Y*, dentro del total representado por aquellos que ostentan una determinada

combinación ***c*** de valores de los otros factores y además tienen valor *X*=1. En cambio, $\boldsymbol{p_c}$ es la proporción en el estado *Y*=1 para aquellos sujetos que tienen la misma combinación ***c*** pero para quienes el atributo *X* vale cero. Si hay ***m*** factores dicotómicos independientes habrá en total 2^m proporciones $\boldsymbol{p_c}$ y $\boldsymbol{p_{xc}}$. Para cada atributo ***X*** habrá una cantidad de pares de proporciones a comparar equivalente a la mitad de 2^k, es decir $2^k/2 = 2^{k-1}$ (en el caso de tres atributos X, W y Z esas diferencias son cuatro, es decir 2^{3-1}). Para el coeficiente estandarizado e_{01} correspondiente a los efectos aleatorios que estimulan el pasaje del estado 0 al estado 1 la fórmula general es la siguiente:

$$e_{01} = \frac{1}{2^k}\left\{(k+1)p_{000}+(k-1)\sum_{i=1}^{k} p_i+(k-3)\sum_{i=1}^{k}\sum_{j=i+1}^{k} p_{ij}+(k-5)\sum_{i=1}^{k}\sum_{j=i+1}^{k}\sum_{h=j+1}^{k} p_{ijh}+...+[k-(2k-1)p_{ijh..k}\right\} \qquad \text{(Ec.93)}$$

En esta expresión la notación p_i corresponde a la proporción en el estado Y=1 en todas aquellas combinaciones en que sólo un atributo independiente ***i=X, W, ..., Z*** es positivo; p_{ij} representa todas las combinaciones en que **dos** factores independientes son positivos; p_{ijh} corresponde a combinaciones con tres factores positivos, etc. Esta formidable expresión no es demasiado complicada en los casos más frecuentes, pues no suele haber más de tres o cuatro factores independientes. Para un solo factor independiente es $e_{01}=p_0$. Para el caso de dos factores la expresión se reduce a:

$$e_{01} = \frac{3p_{00} + p_{10} + p_{01} - p_{11}}{4} \qquad \text{(Ec. 94)}$$

Si hay tres factores independientes dicotómicos la expresión es la siguiente, que simplificando (multiplicando y dividiendo por 2) se reduce a la ecuación 81:

$$e_{01} = \frac{4p_{000} + 2p_{100} + 2p_{010} + 2p_{001} - 2p_{111}}{8} \qquad \text{(Ec. 95)}$$

Con cuatro factores independientes y por aplicación de la misma fórmula (83) el coeficiente estandarizado de efectos aleatorios e_{01} es:

$$e_{01}=\frac{1}{16}\{5p_{0000}+3p_{1000}+3p_{0100}+3p_{0010}+3p_{0001}+p_{1100}+p_{1010}+p_{1001}+p_{0110}+p_{0101}+p_{0011}+p_{1110}+p_{1101}+p_{1011}+p_{0111}-3p_{1111}\} \qquad \text{(Ec. 96)}$$

Mediante este enfoque, por lo tanto, se puede cuantificar el efecto de cualquier número de factores independientes de tipo dicotómico sobre una variable dependiente que también es dicotómica. En este desarrollo se ha supuesto que los efectos ocurren sólo en uno de los valores de los factores (típicamente cuando el atributo está presente y la variable vale 1), pero las fórmulas se podrían adaptar con facilidad al caso en que hay efectos dobles. De hecho, los modelos donde ***X***, como el sexo, tiene efectos tanto cuando es ***X***=0 como cuando es ***X***=1, se formalizan subdividiendo el atributo en dos, por ejemplo "Ser varón" y "Ser mujer", cada uno de ellos con un efecto específico en direcciones contrarias. De este modo el modelo de efectos dobles equivale a introducir dos atributos en lugar de uno. En la sección siguiente, y siempre bajo el supuesto de datos de corte transversal, se analiza el caso de las variables cualitativas con más de dos categorías.

6.3.3. Variables politómicas

Debe tenerse que en cuenta que una variable politómica con ***m*** categorías puede ser reemplazada por ***m***-1 variables dicotómicas tipo "dummy" (codificadas con 1 y 0). Por ejemplo la variable "Estado civil" con las categorías 1. Soltero, 2.Casado, 3.Divorciado y 4.Viudo podría ser sustituida por tres variables dicotómicas que podrían ser "Casado", "Divorciado" y "Viudo", de modo que cada estado civil tendría la siguiente configuración de valores:

Recodificación de una politomía por medio de variables dicotómicas			
	Variables dicotómicas "dummy"		
Estado civil	**"Casado"**	**"Divorciado"**	**"Viudo"**
1. Soltero	0	0	0
2. Casado	1	0	0
3. Divorciado	0	1	0
4. Viudo	0	0	1

En esta codificación se ha tomado el estado civil "Soltero" como referencia, creando variables dummy para todos los demás estados civiles; naturalmente podría haberse elegido cualquiera de los estados como referencia, no necesariamente "Soltero". En el presente ejemplo, cuando las tres dummy valen 0, el individuo es soltero. Cuando alguna de ellas vale 1, el individuo tiene algún otro estado civil. Para cada individuo, o bien las tres dummies son cero, o bien solo una de las dummies es igual a uno, ya que las dummies se excluyen entre sí. Si se usa este tipo de recodificación, las indicaciones formuladas para múltiples factores dicotómicos también pueden usarse para factores cualitativos con más de dos categorías.

Cuando abandonemos el campo del corte transversal para analizar datos de panel, veremos que el análisis causal se puede hacer directamente sobre variables politómicas, ya que los efectos causales se calculan a partir de las tasas instantáneas de transición, y éstas pueden ser calculadas para variables politómicas sin ninguna dificultad mediante la ecuación 35 bis.

6.3.4. Interacción entre factores

Hasta ahora se ha supuesto que los efectos de los diferentes factores son **aditivos** e independientes entre sí. Sin embargo, es frecuente que el efecto de un factor varíe según el valor que asuma otro factor, en cuyo caso se registra una **interacción** entre esos factores. El segundo factor puede **reforzar** o **atenuar** el efecto del primero, y viceversa. Los factores pueden **potenciarse** mutuamente o pueden tender a **suavizar** o **cancelar** mutuamente sus efectos.

Usualmente, la existencia de interacción se representa mediante el producto de los factores interactuantes. Esto significa que además de los términos individuales de cada factor, como por ejemplo $\beta_X X$ y $\beta_Z Z$ puede haber un término como $\beta_{XZ} XZ$, el cual operaría sólo si los dos factores X y Z tienen el valor 1. Esta situación tiene una solución muy fácil, pues equivale a la presencia de una tercera variable, digamos V=XZ, que valdría 1 si tanto X como Z valen 1, y valdría 0 en los demás casos. Si se desea introducir un efecto de interacción, por lo tanto, sólo es necesario crear una variable multiplicativa como V=XZ, e incluirla en el modelo como un factor aditivo adicional. Si hay más de dos factores independientes puede haber también interacciones entre tres o más variables.

Si ese coeficiente β_{XZ} resulta positivo, significaría que tanto el efecto de X como el de Z sufren un incremento en caso que ambos factores valgan 1. Si el efecto individual de X o Z es del mismo signo que el efecto de la interacción, entonces la interacción **intensifica** o refuerza el efecto de la variable considerada aisladamente. Si el efecto univariado y el efecto de interacción son de diferente signo, entonces la interacción tiende a contrarrestar o **atenuar** el efecto individual de la variable involucrada. Es posible que los efectos de intensificación o atenuación ocurran no sólo por la combinación de un valor 1 en X y un valor 1 en Z, sino por cualquier otra combinación de valores de esas variables. Pero en ese caso, esos factores no estarían actuando de acuerdo al concepto de presencia o ausencia de un atributo, sino a través de efectos "dobles" en que cada categoría tiene efectos especiales, de modo que habría que desdoblarlo como se explicó antes. Por ello se puede pensar que todas las interacciones son del tipo multiplicativo, y que valen 1 sólo si todas las variables interactuantes valen también 1.

En conclusión, se dispone de una adaptación multivariada del lenguaje de tasas instantáneas de transición que permite el análisis de relaciones causales entre una variable dicotómica dependiente y varios factores causales dicotómicos o politómicos (estos últimos pueden tener categorías ordenadas o constituir una escala simplemente nominal sin orden intrínseco). Este enfoque permite identificar la importancia relativa de los efectos de los factores y de los efectos aleatorios, aunque no su valor absoluto, porque en los estudios de corte transversal no hay observaciones directas sobre los cambios de estado que ocurren entre las unidades de análisis, como sí las hay en el caso de los estudios longitudinales. En la sección siguiente se extiende este análisis al caso de los estudios de panel.

6.4. Procesos causales continuos con datos de panel

La extensión a los estudios de panel de los métodos anteriormente expuestos para cortes transversales no ofrece muchas dificultades. Dado que se trata de procesos modelizados como continuos, la brecha temporal entre las rondas del panel no implica necesariamente suponer una demora (*lag*) en-

tre la causa y el efecto. La disponibilidad de datos de panel tiene como principal beneficio la **medición directa de los flujos entre estados**, en situaciones de equilibrio o de desequilibrio, sin necesidad de suponer que los datos de una cierta fecha correspondan a una situación de equilibrio.

6.4.1. Planteo general

Existen varias posibilidades en cuanto al modelo causal que podría vincular ciertas variables en un modelo causal. La más obvia es la que concierne el número de factores causales, desde modelos bivariados (una variable dependiente y una independiente) hasta modelos con varios factores causales operando al mismo tiempo.

Otra distinción importante concierne al carácter constante o variable (en el tiempo) del atributo X. Un atributo como el sexo es fijo a lo largo del tiempo, mientras uno como el conocimiento de las propuestas de un candidato puede variar entre una y otra ronda del panel. En algunos casos el cambio ocurre sólo a unas pocas unidades de análisis a lo largo del intervalo entre las rondas, de modo que para la mayor parte de los sujetos el factor puede considerarse constante, y el grupo que cambia podría ser estadísticamente poco significativo, como ocurre por ejemplo con la variable "estado civil", ya que los que efectúan por semestre cada cambio de estado civil (de casado a divorciado, de soltero a casado, etc.) son un porcentaje muy bajo del total. En esos casos a menudo el factor se considera como constante, y se descartan aquellos sujetos que hayan cambiado de categoría durante el intervalo, salvo cuando los contingentes que sufrieron cambios sean de un tamaño estadísticamente suficiente para que las estimaciones no tengan demasiado margen de error.

En muchos casos en que el cambio de X es un evento relativamente poco frecuente, difícilmente un sujeto sufra más de una transición en un intervalo entre dos rondas que sea razonablemente breve (por ejemplo encuestas trimestrales, semestrales o anuales). En cambio hay otros posibles factores causales cuyos cambios podrían ocurrir varias veces durante el intervalo, como por ejemplo el tener o no tener una elevada presión arterial.

La constancia o variabilidad del factor independiente hace que el modelo causal incluya o no el **cambio en la variable independiente** como causa del cambio en la variable dependiente. La variable dependiente en este caso sería por ejemplo la tasa de transición de la variable Y desde el estado 0 hacia el estado 1. El factor causal que modifica esa tasa, en el caso de una constante como el sexo, sería precisamente el valor de la variable independiente, por ejemplo el sexo del sujeto, en el sentido de que los varones pasarían de 0 a 1 con mayor o menor intensidad que las mujeres. En el caso de un factor causal variable, puede haber dos clases de efectos sobre Y: el nivel inicial de X donde estaba ubicado el sujeto, y el cambio sufrido (o no sufrido) por X durante el intervalo entre las dos rondas. En el primer caso el modelo es: $X \rightarrow \Delta Y$ mientras en el segundo el modelo es $(X,\Delta X) \rightarrow \Delta Y$. Supongamos que Y=Intención de votar por A, y X=Percepción del sujeto sobre su propia situación económica, donde X tiene los valores 0=Buena y 1=Mala. Los cambios en la intención de votar por el candidato A, es decir ΔY, dependerán del **nivel** de la percepción inicial y de los **cambios** sufridos por la percepción de su situación que hayan experimentado los sujetos en los últimos tiempos (algunos sujetos habrán mantenido la misma percepción de su situación como "Buena" o "Mala", otros pasaron de "Mala" a "Buena" y otros a la inversa). En el primer caso habrá que observar las diferencias en las tasas de transición entre las subpoblaciones de sujetos situadas en dos estados fijos (por ejemplo varones comparados con mujeres, o personas con información comparadas con personas sin información acerca de las propuestas del candidato); en el segundo caso, en que X puede variar a lo largo del tiempo, se observan las diferencias en las tasas de transición entre las subpoblaciones que corresponden a cuatro flujos entre estado en t y estado en $t+h$: Buena-Buena, Mala-Mala, Mala-Buena y Buena-Mala.

Otra distinción importante que debe hacerse aquí, en cuanto a las características del proceso causal involucrado, es entre procesos causales **unidireccionales** y procesos **recíprocos**. En los primeros, X influye sobre Y pero Y no influye sobre X. En el segundo caso, los dos atributos se influyen mutuamente. Además, hay variables dicotómicas que reflejan la presencia o ausencia de algo, como por ejemplo la variable "Haber visto previamente el aviso publicitario" en un estudio de mercado. El aviso publicitario tiene efecto si fue visto, y no tiene ningún efecto en caso contrario. Hay otras

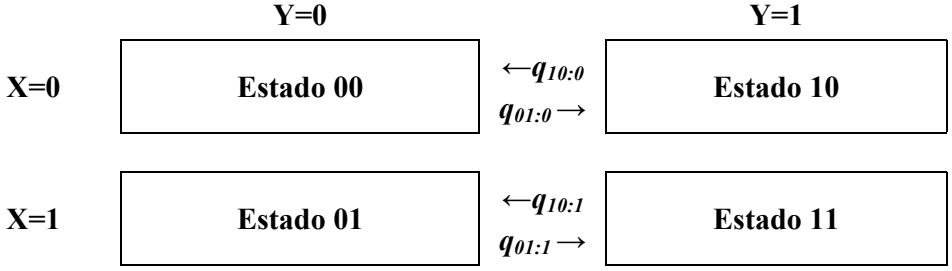
variables que reflejan dos situaciones diferentes, cada una de las cuales puede tener efectos específicos, como por ejemplo la variable "Sexo": puede haber un efecto entre los varones y otro efecto (de igual o de opuesta dirección) entre las mujeres. En el primer caso, el valor 1 tiene efecto y el valor 0 no lo tiene, mientras en el segundo caso hay efectos del valor 1 y efectos del valor 0. La siguiente tabla sintetiza las opciones que se abren para la especificación de modelos causales.

Opciones metodológicas para el análisis causal de variables dicotómicas

Opciones metodológicas		Significado
Número de factores causales	Uno	Un solo factor causal
	Varios	Varios factores causales
Dirección causal	Unidireccional	X opera sobre Y, pero Y no influye sobre X
	Recíproca	X influye sobre Y, e Y influye sobre X
Tipo de efectos	Por presencia	X=1 tiene efecto; X=0 no tiene efecto
	Por categoría	Ambas categorías tienen efectos específicos
Variabilidad de los factores en el tiempo	Factores constantes	Y responde al **valor** de X
	Factores variables	Y responde al **valor** y a los **cambios de valor** de X

6.4.2. Un factor constante con efecto simple unidireccional

Comenzaremos con un proceso causal **unidireccional**, con **un solo factor**, con efectos simples (por **presencia**), donde X es **constante** en el tiempo. Los sujetos pueden cambiar de estado en la variable Y, pero no cambian su valor de X. Esos flujos entre ambos estados de la variable Y se supone que están gobernados por tasas de transición $q_{ij:x}$, es decir, tasas de transición específicas según el valor que el individuo tenga en la variable X. Del mismo modo se denota como "estado 00" al estado Y=0 bajo la condición de que los sujetos pertenezcan al grupo X=0, y de igual modo otros estados. Dado que X no varía, sólo hay transiciones entre los dos estados de Y:



El análisis de las relaciones causales entre X e Y implica, por una parte, calcular las tasas de transición, y en segundo lugar, **particionar** las tasas de transición entre estados de Y distinguiendo los factores aleatorios del influjo específico del factor X. En este modelo simple, si el factor X es un atributo cuya presencia incrementa o reduce la tasa de transición desde Y=0 hacia Y=1, sería $q_{01}=\varepsilon_{01}$ y $q_{23}=\beta_X+\varepsilon_{01}$. El efecto de la presencia de X, es decir el coeficiente β_X, puede ser positivo o negativo, según tenga como resultado acelerar o retardar los cambios de estado desde Y=0 hacia Y=1. Un razonamiento similar puede aplicarse a los flujos en sentido contrario, que para mayor claridad se indican con la letra griega α. Sería $q_{10}=\varepsilon_{10}$ y por otro lado $q_{32}=\alpha_X+\varepsilon_{10}$. Por lo tanto en este caso simple, con un solo factor independiente de tipo dicotómico, los coeficientes se encuentran fácilmente. Los efectos corresponden a la **diferencia en las tasas de transición** en presencia o ausencia del factor:

$$\beta_X = q_{23} - q_{01} \quad \text{(Ec. 97)}$$

$$\alpha_X = q_{32} - q_{10} \quad \text{(Ec. 98)}$$

En otras palabras, el efecto de la presencia de X se traduce en la diferencia de las tasas instantáneas de transición de Y cuando X está presente y cuando X está ausente. El efecto β_X de la presencia del

atributo X sobre los flujos en una dirección (de Y=0 a Y=1) puede ser diferente del efecto α_X que gobierna los flujos opuestos (de Y=1 a Y=0).

6.4.3. Varios factores constantes con efecto simple unidireccional

En el caso de un modelo causal multivariado donde hay varios factores causales involucrados, los efectos de un atributo sobre la variable dependiente son estimadas por el **promedio de las diferencias en las tasas de transición** que difieren sólo en la presencia o ausencia de ese factor. Por ejemplo, supóngase que los dos factores causales son el sexo y el nivel socioeconómico, ambos considerados constantes en el tiempo, y la variable dependiente es la opinión política, que puede variar entre Y=1 (votar por el partido A) e Y=0 (no votar por el partido A). Los sujetos pasan del estado Y=1 al estado Y=0, o viceversa, según las tasas instantáneas de transición q_{10} y q_{01}. Estas tasas varían según el valor que tengan los factores independientes. Sea $q_{01:10}$ la tasa de transición de Y=0 a Y=1 cuando el sexo es X=1 y el nivel socioeconómico es Z=0. La correspondiente tasa para el sexo opuesto sería $q_{01:00}$. Estas dos tasas se refieren al mismo nivel socioeconómico (Z=0) y difieren en el sexo. Asimismo se definen $q_{01:11}$ y $q_{01:01}$ para el caso que el nivel socioeconómico sea Z=1. Las ecuaciones que particionan las tasas de acuerdo a los factores que inciden en ellas son las siguientes:

$$q_{01:00} = \varepsilon_{01} \quad \text{(Ec. 99.1)}$$
$$q_{01:10} = \beta_x + \varepsilon_{01} \quad \text{(Ec. 99.2)}$$
$$q_{01:01} = \beta_z + \varepsilon_{01} \quad \text{(Ec. 99.3)}$$
$$q_{01:11} = \beta_x + \beta_z + \varepsilon_{01} \quad \text{(Ec. 99.4)}$$
$$q_{10:00} = \varepsilon_{10} \quad \text{(Ec. 99.5)}$$
$$q_{10:10} = \alpha_x + \varepsilon_{10} \quad \text{(Ec. 99.6)}$$
$$q_{10:01} = \alpha_z + \varepsilon_{10} \quad \text{(Ec. 99.7)}$$
$$q_{10:11} = \alpha_x + \alpha_z + \varepsilon_{10} \quad \text{(Ec. 99.8)}$$

El impacto de un factor independiente, por ejemplo el sexo, se mide por la diferencia entre las tasas de transición para varones y para mujeres, promediando la diferencia que se obtiene entre los sujetos de nivel socioeconómico alto la que presentan los sujetos de nivel socioeconómico bajo. La tasa $q_{01:10}$ es la tasa de transición de Y=0 a Y=1 cuando el sexo es X=1 y el nivel socioeconómico es Z=0. La correspondiente tasa para el sexo opuesto sería $q_{01:00}$. Estas dos tasas se refieren al mismo nivel socioeconómico (Z=0) y difieren en el sexo. Por lo tanto su diferencia ($q_{01:10} - q_{01:00}$) indica el efecto del sexo entre aquellas personas con ese nivel socioeconómico Z=0. Del mismo modo se puede calcular esa diferencia por sexo para el otro nivel socioeconómico, ($q_{01:11} - q_{01:01}$). Dado que el modelo es aditivo y sin interacción entre los factores causales, el influjo del sexo tendría que ser el mismo (salvo perturbaciones aleatorias) en los dos casos. Por lo tanto, en este caso con dos factores se estima el valor del efecto del sexo como **promedio** de esas dos diferencias:

$$\beta_X = \frac{1}{2}\sum_{k=0}^{1}(q_{01:1k} - q_{01:0k}) = \frac{(q_{01:10} - q_{01:00}) + (q_{01:11} - q_{01:01})}{2} \quad \text{(Ec. 100)}$$

En este caso con sólo dos factores dicotómicos hay que promediar sólo dos diferencias: la que se obtiene con Z=0 y la que existe cuando Z=1. En el caso general en que hay k factores independientes, la fórmula es similar, aunque la cantidad de diferencias que se deben promediar es mayor. Los k factores independientes dicotómicos dividen la población en 2^w subpoblaciones. En cada una de ellas rige una tasa de transición de Y=0 a Y=1. Tómese ahora uno de los factores independientes, X_h. Llámese $q_{01:hz}$ a la tasa de transición de Y=0 a Y=1 en la subpoblación caracterizada por el valor X_h=1 y por una determinada combinación Z de valores de los otros factores. Del mismo modo, llámese $q_{01:h'z}$ a la tasa correspondiente cuando X_h=0. En cada una de las dos subpoblaciones determinadas por X_h=0 y X_h=1 hay a su vez 2^{w-1} subpoblaciones correspondientes a diferentes combinaciones Z de valores de los otros $k-1$ factores independientes. En cada una de estas 2^{k-1} subpoblaciones se puede medir la diferencia entre la tasa de transición $q_{01:hz}$ cuando la variable X_h=1 y la misma tasa ($q_{01:h'z}$) cuando la variable X_h=0. Hay por lo tanto que promediar 2^{k-1} diferencias para obtener una estimación del efecto de X_h. La diferencia $q_{01:hz} - q_{01:h'z}$

indica el efecto del atributo X_h sobre la tasa de transición de Y=0 a Y=1 **cuando los demás factores tienen la combinación de valores *Z***, que puede variar desde ***Z***=000...0 hasta ***Z***=111....1. El **efecto promedio** del atributo X_h, por lo tanto se estima como promedio de las 2^{k-1} diferencias obtenidas para las 2^{k-1} combinaciones ***Z*** de los otros ***k-1*** factores:

$$\beta_h = \frac{1}{2^{k-1}} \sum_z (q_{01:hz} - q_{01:h'z}) \qquad \text{(Ec. 101)}$$

Del mismo modo se obtiene el efecto de otros factores X_k, X_g, ... X_k. De este modo se obtiene un conjunto de ***k*** coeficientes β_h (donde ***h***=1, 2, ... ***k***) que miden el impacto de cada uno de los ***k*** factores sobre la tasa de transición q_{01}.

El efecto aleatorio ε_{01} de otros factores no identificados que influyen para que los sujetos pasen del estado Y=0 al estado Y=1 se estima como promedio de los varios residuos obtenidos al restar de cada tasa de transición la suma de efectos de los factores. Denominaremos $q_{01:Z}$ a la tasa de transición desde Y=0 a Y=1 en la subpoblación *Z* caracterizada por los valores (0,1) de ***k*** factores, y llamaremos B_Z a la suma de los coeficientes β_h de todos los factores con valor 1 en la subpoblación *Z*. Existen 2^k subpoblaciones Z. Entonces la estimación del efecto aleatorio sobre $q_{01:Z}$ es:

$$\varepsilon_{01} = \frac{1}{2^k} \sum_Z (q_{01:Z} - B_Z) \qquad \text{(Ec. 102)}$$

Del mismo modo se estiman los coeficientes α_h que miden el impacto de los ***w*** factores sobre la tasa de transición desde Y=1 hacia Y=0, es decir q_{10}, y el coeficiente de influencias aleatorias en la misma dirección, ε_{10}. Como resultado se obtienen todos los coeficientes necesarios para estimar el efecto de los **k** factores (todos ellos constantes en el tiempo) en el proceso estocástico que mueve a los sujetos entre los estados 0 y 1 de la variable Y, y que en el caso de dos factores causales están reflejados en las ecuaciones 99.1 a 99.7.

Este enfoque requiere como primer paso el cálculo de todas las tasas de transición: las tasas globales q_{01} y q_{10}, y las tasas "parciales" $q_{01:hz}$, $q_{01:h'z}$, $q_{10:hz}$, y $q_{10:h'z}$ para cada valor 0 y 1 de cada uno de los ***k*** factores, y para cada una de las 2^{k-1} combinaciones ***Z*** de los otros ***k-1*** factores. Una vez que se dispone de todas estas tasas de transición se puede estimar el efecto de cada factor sobre las mismas. Esto puede ser muy laborioso si los factores independientes son muchos, aun para un número de factores que parezca razonable de acuerdo a otros criterios. Con tres factores independientes hay que calcular en total $2x2^3=16$ tasas de transición parciales. Con cuatro factores la cifra asciende a $2x2^4=32$ tasas. La fórmula general de las tasas de transición entre los dos estados de una variable dicotómica es la de las ecuaciones 30 y 31:

$$q_{10} = \left(\frac{u_{10}}{u_{01} + u_{10}} \right) \left(\frac{-\ln(1 - u_{01} - u_{10})}{h} \right)$$

$$q_{01} = \left(\frac{u_{01}}{u_{01} + u_{10}} \right) \left(\frac{-\ln(1 - u_{01} - u_{10})}{h} \right)$$

donde ***h*** es la longitud del intervalo entre rondas, y u_{ij} son las probabilidades de transición entre estados $u_{ij} = N_{ij} / N_i^t$ (flujo de ***i*** a ***j*** en el intervalo ***h*** sobre población inicial en el estado ***i***). En el caso de las tasas de transición globales se aplica esta fórmula sobre el total de casos en el panel. Para las tasas de transición "parciales" se efectúa el mismo cálculo sobre cada una de las subpoblaciones correspondientes. La necesidad de trabajar con una gran cantidad de subpoblaciones, algunas de las cuales serán probablemente muy pequeñas, hace que este enfoque requiera una muestra de considerable tamaño para poder medir el efecto de cada uno de los factores controlando todos los demás. Este problema puede ser importante. Por ejemplo, supóngase que el panel abarca 4000 sujetos, y que se estudian simultáneamente cuatro factores dicotómicos, de modo que hay que calcular tasas de transición parciales en $2^3=8$ subpoblaciones. El tamaño promedio de esas subpoblaciones será 4000/8=500 sujetos, pero algunas de ellas pueden tener un número mucho

menor de casos. Para cada tasa de transición se debe considerar una tabla 2x2 donde cada uno de los cuatro flujos (00, 01, 10 y 11) tendrá en promedio 500/4=125 casos, pero en muchas oportunidades la tabla misma puede tener menos de 500 casos, y algunos de los flujos pueden tener menos de la cuarta parte de los casos. Es sumamente fácil que alguno de esos flujos sea muy pequeño, tal vez menos de 10 o menos de 5 casos. Para que la estimación sea más o menos confiable debería haber al menos unos 20-30 casos involucrados en cada flujo. Es cierto que la estimación final es un promedio de varias diferencias, de modo que los errores de muestreo de algunas de ellas tenderán a compensarse con los errores de sentido opuesto en otras, pero de todos modos los márgenes de error de tales estimaciones corren serio peligro de ser altos.

6.4.4. Varios factores constantes con efecto doble unidireccional

La siguiente "complejización" de este enfoque consistirá en permitir que los factores tengan efectos "dobles". Hasta ahora, el valor X=0 no tenía ningún efecto, y el valor X=1 tenía un efecto β_x sobre la tasa $\boldsymbol{q_{01}}$ y un efecto α_x sobre la tasa $\boldsymbol{q_{10}}$. Ahora supondremos que X=0 también tiene un efecto. Llamaremos β_{x1} al efecto de X=1 sobre la tasa $\boldsymbol{q_{01}}$ y β_{x0} al efecto de X=0. Del mismo modo llamaremos α_{x1} al efecto de X=1 sobre la tasa $\boldsymbol{q_{10}}$, y α_{x0} al efecto de X=0.

Adviértase que aquí no sólo se supone que ambos valores de X tienen efectos sobre Y, sino también que cada valor de X tiene un efecto sobre $\boldsymbol{q_{01}}$ y otro efecto sobre $\boldsymbol{q_{10}}$, de modo que hay cuatro efectos para identificar y estimar. Las ecuaciones serían del siguiente tipo (en el ejemplo hay sólo dos factores independientes X y Z):

$$q_{01:00} = \beta_{X0} + \beta_{Z0} + \varepsilon_{01} \qquad \text{(Ec. 104)}$$

$$q_{01:01} = \beta_{X0} + \beta_{Z1} + \varepsilon_{01} \qquad \text{(Ec. 105)}$$

$$q_{01:10} = \beta_{X1} + \beta_{Z0} + \varepsilon_{01} \qquad \text{(Ec. 106)}$$

$$q_{01:11} = \beta_{X1} + \beta_{Z1} + \varepsilon_{01} \qquad \text{(Ec. 107)}$$

En este caso es evidente que no se puede estimar cada coeficiente por el conveniente método de medir la diferencia de las tasas de transición que incluyan o no incluyan cada factor. Por ejemplo, en el caso de X, la diferencia de las ecuaciones 106 y 104 sería:

$$q_{01:10} - q_{01:00} = \beta_{X1} - \beta_{X0} \qquad \text{(Ec. 108)}$$

Se puede así estimar la diferencia entre estos coeficientes, pero no cada uno por separado. Lo mismo ocurriría en otras diferencias similares, como por ejemplo si a la ecuación 97 se le resta la ecuación 96, donde la diferencia de tasas de transición sería igual a $\boldsymbol{\beta_{z1} - \beta_{z0}}$. En todos los casos resulta imposible calcular separadamente el efecto del valor 1 y el efecto del valor 0, sino sólo la diferencia entre ambos, que equivale al **efecto neto** del factor. Ese efecto neto, por supuesto, puede ser positivo o negativo según predomine el efecto del valor 1 o el efecto del valor 0. Debe interpretarse por ejemplo como la variación neta que sufriría la tasa $\boldsymbol{q_{01:00}}$ si el factor *Z* pasase de 0 a 1. Dado que sólo se puede calcular el efecto neto de cada factor $\boldsymbol{h}$, que podemos denominar simplemente $\boldsymbol{\beta_h}$, el modelo con efectos dobles se reduce a un modelo de efectos netos que es indistinguible del modelo de efectos simples anteriormente analizado. Por lo tanto en lo sucesivo nos referiremos únicamente a efectos simples, en el entendido que los resultados pueden también ser interpretados como referidos al saldo neto de efectos dobles.

6.4.5. Factores variables en el tiempo

Hasta ahora hemos supuesto que los factores causales no varían durante el intervalo entre las rondas del panel. En esta sección se estudian las modificaciones que debe sufrir el modelo cuando esos factores pueden variar al mismo tiempo que varía la variable dependiente. Puede ser que la variable dependiente no influya sobre los factores (efectos unidireccionales) o que pueda a su vez influir sobre ellos (efectos recíprocos).

En un modelo multivariado con efectos recíprocos, distintas variables se influyen mutuamente. Puede haber algunas que no son determinadas por otras variables del modelo, y por lo tanto operan como variables **exógenas**, mientras otras obedecen a factores internos al modelo y son por lo tanto

endógenas. En estos modelos, la distinción entre variables dependientes e independientes es de hecho reemplazada por la distinción entre variables endógenas y exógenas. Una variable es exógena cuando ninguna otra variable influye sobre ella, sino sólo factores aleatorios externos al modelo. Si una variable endógena sufre cambios a lo largo del tiempo, esos cambios son atribuidos por el modelo exclusivamente a factores aleatorios externos, mientras que las variaciones en una variable endógena se explican no sólo por factores aleatorios sino también por la influencia de otras variables.

Para tratar esta situación supondremos un caso sencillo con dos variables endógenas (Y, Z) y una tercera variable que puede ser endógena o exógena (X), todas ellas dicotómicas codificadas con 0 y 1. En un momento dado cada unidad o individuo está caracterizado por un vector [XYZ] que varía entre [000] y [111], con un total de ocho estados posibles. El primer paso consiste en partir de la tabla de rotación para dos momentos en el tiempo, ***t*** y ***t+1***, a fin de calcular las tasas instantáneas de rotación entre esos ocho estados. La notación que se usará será la siguiente:

$q_{XYZ:xyz}$: Tasa instantánea de transición del estado [XYZ] al estado [xyz]. Por ejemplo:

$q_{000:001}$=Tasa instantánea de transición del estado [000] al estado [001], lo que implica que X e Y permanezcan en el valor 0 mientras Z pasa de 0 a 1.

En total hay 8 x 8 = 64 tasas instantáneas de transición en este modelo. Cada una de las que involucran algún cambio de estado puede ser hallada mediante la serie exponencial de la ecuación 35 bis. Las tasas de la diagonal principal (que implican permanencia en el mismo estado, como por ejemplo $q_{000:000}$) equivalen a la suma de las otras tasas de la misma fila, cambiada de signo. Hay por lo tanto 8 x 7 = 56 tasas de transición linealmente independientes.

Una vez calculadas las tasas instantáneas de transición se procede a particionarlas en sus componentes. Si se supone que los cambios en X son exclusivamente exógenos, ellos obedecerán únicamente a efectos aleatorios, mientras los cambios en Y, Z son en parte exógenos y en parte debidos a la influencia de las mismas variables X, Y y Z. Los efectos que deben ser estimados si X es exógena son los siguientes:

α_{XY} = Efecto de X=1 sobre el desplazamiento de Y=1 a Y=0

α_{XZ} = Efecto de X=1 sobre el desplazamiento de Z=1 a Z=0

α_{YY} = Efecto de Y=1 sobre el desplazamiento de Y=1 a Y=0

α_{YZ} = Efecto de Y=1 sobre el desplazamiento de Z=1 a Z=0

α_{ZY} = Efecto de Z=1 sobre el desplazamiento de Y=1 a Y=0

α_{ZZ} = Efecto de Z=1 sobre el desplazamiento de Z=1 a Z=0

β_{XY} = Efecto de X=1 sobre el desplazamiento de Y=0 a Y=1

β_{XZ} = Efecto de X=1 sobre el desplazamiento de Z=0 a Z=1

β_{YY} = Efecto de Y=1 sobre el desplazamiento de Y=0 a Y=1

β_{YZ} = Efecto de Y=1 sobre el desplazamiento de Z=0 a Z=1

β_{ZY} = Efecto de Z=1 sobre el desplazamiento de Y=0 a Y=1

β_{ZZ} = Efecto de Z=1 sobre el desplazamiento de Z=0 a Z=1

Si X es endógena hay que añadir parámetros que reflejen el efecto de Y y Z sobre X (otros seis parámetros). Además de estos efectos de las variables entre sí, hay que estimar los siguientes efectos aleatorios:

ε_{X10} = Efectos aleatorios sobre el desplazamiento de X=1 a X=0

ε_{X01} = Efectos aleatorios sobre el desplazamiento de X=0 a X=1

ε_{Y10} = Efectos aleatorios sobre el desplazamiento de Y=1 a Y=0

ε_{Y01} = Efectos aleatorios sobre el desplazamiento de Y=0 a X=1

ε_{Z10} = Efectos aleatorios sobre el desplazamiento de Z=1 a X=0

ε_{Z01} = Efectos aleatorios sobre el desplazamiento de Z=0 a X=1

Todos estos efectos α, β y ε pueden ser negativos, nulos o positivos. En total, como se ve, si X es exógena hay 18 parámetros a estimar con un total de 56 ecuaciones linealmente independientes. Si X fuese también endógena habría 24 parámetros a estimar con 56 ecuaciones, y por lo tanto la situación no variaría fundamentalmente. El mismo parámetro figurará en varias ecuaciones, de modo que en principio habrá varias estimaciones posibles de cada parámetro, y por lo tanto hay que usar un método de mínimos cuadrados, minimizando los errores derivados de usar una u otra estimación. En líneas generales esto significa que los efectos α y β se estiman como un promedio de las diferencias entre las $\boldsymbol{q}$ con presencia y con ausencia de determinados factores. Los efectos aleatorios se estiman luego en función de esos resultados, como ya se mostró anteriormente en el caso del análisis multivariado de corte transversal. Sería excesivo detallar la partición de todas las tasas de transición, pero como ejemplo se incluyen las siguientes:

$$\boldsymbol{q_{000:001} = \varepsilon_{Z01}} \quad \text{(Ec. 109)}$$

$$\boldsymbol{q_{100:101} = \varepsilon_{Z01} + \beta_{XZ}} \quad \text{(Ec. 110)}$$

$$\boldsymbol{q_{110:111} = \varepsilon_{Z01} + \beta_{XZ} + + \beta_{YZ}} \quad \text{(Ec. 111)}$$

$$\boldsymbol{q_{110:101} = \varepsilon_{Y10} + \varepsilon_{Z01} + \alpha_{XY} + \alpha_{YY} + \beta_{XZ} + \beta_{YZ}} \quad \text{(Ec. 112)}$$

La ecuación 109 parte del estado [000] con todas las variables en cero. Dado que se postulan efectos simples, del tipo "ausencia-presencia", donde cada atributo ejerce influjo causal cuando está presente, y no lo ejerce en ausencia, entonces la tasa de transición $\boldsymbol{q_{000:001}}$, que implica un desplazamiento de la variable Z de 0 a 1, está determinada únicamente por factores aleatorios no identificados. En cambio en la ecuación 110 la tasa $\boldsymbol{q_{100:101}}$ implica el mismo cambio en Z, de 0 a 1, pero en presencia del atributo X, y por lo tanto obedece a los mismos factores aleatorios más la influencia $\boldsymbol{\beta_{XZ}}$ de X=1. En el caso de la ecuación 111, se parte del estado [110] para pasar al estado [111], y entonces el cambio en Z de 0 a 1 está condicionado por el influjo de X y de Y a través de los coeficientes $\boldsymbol{\beta_{XZ}}$ y $\boldsymbol{\beta_{YZ}}$. Por último, en el caso de la ecuación 112, se producen dos cambios: Z pasa de 0 a 1 mientras Y pasa de 1 a 0, al pasar los sujetos involucrados desde [110] a [101]. En este caso hay cuatro influencias de las variables: los efectos $\boldsymbol{\beta_{XZ}}$ y $\boldsymbol{\beta_{YZ}}$ que desplazan Z de 0 a 1, y los efectos $\boldsymbol{\alpha_{XY}}$ y $\boldsymbol{\alpha_{YY}}$ que contribuyen a desplazar Y desde 1 hacia 0, junto con los dos efectos aleatorios $\boldsymbol{\varepsilon_{Y10}}$ y $\boldsymbol{\varepsilon_{Z01}}$. En forma similar se especifican las ecuaciones correspondientes a la partición de todas las demás tasas de transición según los efectos que operen sobre ellas.

La fórmula para calcular los distintos parámetros α y β es simplemente el promedio de todas las diferencias de las $\boldsymbol{q}$ caracterizadas por la presencia y ausencia del respectivo factor. Por ejemplo:

$$\alpha_{XY} = \frac{\sum_m \sum_i \sum_n (q_{11k:i0n} - q_{01k:i0n})}{8} \quad \text{(Ec. 113)}$$

donde los subíndices $\boldsymbol{i}$, $\boldsymbol{k}$ y $\boldsymbol{n}$ pueden valer 0 o 1. Las ocho diferencias que se incluyen son:

($q_{111:101} - q_{011:101}$); ($q_{111:100} - q_{011:100}$), ($q_{110:101} - q_{010:101}$); ($q_{110:\ 100} - q_{010:100}$); ($q_{111:001}$-$q_{011:001}$); ($q_{111:000}$-$q_{011:000}$); ($q_{110:001} - q_{010:001}$); y ($q_{110:000} - q_{010:000}$). Dentro de cada paréntesis, el primer término contiene X=1, y el segundo X=0, y esa es la única diferencia entre ambos términos. Fórmulas similares se utilizan para el resto de los parámetros α y β.

En cuanto a los efectos aleatorios, una vez estimados los parámetros α y β se procede de modo similar, promediando aquellas diferencias entre tasas en que un efecto aleatorio esté presente y ausente.

Para que estas estimaciones tengan significación estadística es conveniente que cada uno de los flujos involucrados entre un estado [$\boldsymbol{ijk}$] y otro estado [$\boldsymbol{lmn}$] esté representado por un número adecuado de casos en la muestra. No existe una regla general al respecto, pero es razonable que todos los flujos tengan al menos 20-30 casos, para que las diferencias del tipo expresado en la ecuación 113 no estén afectadas por mucho error de muestreo. Sin embargo Coleman (1964b:201-213) demuestra que el factor más importante es el tamaño **total** de la muestra, y que un fuerte desbalance entre el tamaño de dos de los flujos que se comparan no tendría en general un impacto significativo.

7. Variables latentes en estudios de panel

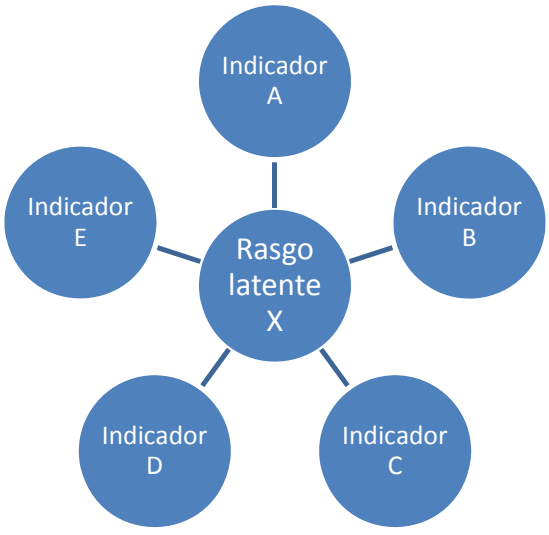
En el análisis de la incertidumbre de respuesta se hizo la distinción entre el estado subyacente de los sujetos y sus respuestas manifiestas. Ese estado subyacente fue modelizado como un conjunto de "elementos" que podían estar en varios estados, y cuya distribución condicionaba probabilísticamente al sujeto para producir determinadas respuestas con mayor o menor probabilidad. Esa noción puede ser ampliada mucho más si se piensa que cualquier dato observable podría considerarse como un mero indicador de alguna "verdadera" variable que permanece latente e inobservable.

Existen diferentes enfoques que procuran identificar y medir algunas variables latentes a partir de variables manifiestas. Cuando las variables manifiestas son de tipo cuantitativo, el modelo clásico es el **análisis factorial**. Cuando las variables manifiestas son atributos dicotómicos, el enfoque más importante es el **análisis de estructura latente**, que incluye modelos de **clases latentes,** modelos con variables latentes de tipo **ordinal**, y modelos variables latentes con **escala de intervalo**. Otros enfoques parecidos para variables cualitativas son el **escalamiento multidimensional** (*multidimensional scaling,* o **MDS**), y varios esquemas análogos como en análisis de correspondencias múltiples, el análisis de homogeneidad, y el análisis de componentes principales para variables categóricas. En este capítulo se revisan los principales conceptos del **análisis de estructura latente**, sin desarrollar el método como tal.

7.1. El principio de independencia local

Cuando dos variables están correlacionadas entre sí, un viejo principio de la lógica dice que hay dos razones que pueden explicar esa correlación: puede ser que una de las variables sea la causa de la otra, o puede ser que ambas sean efecto de una causa común. En el capítulo anterior se han examinado situaciones en que hay dos o más variables que se influyen causalmente **entre sí**. Pero hay casos en que esto no es así. En muchas situaciones de investigación las distintas variables no tienen relaciones causales entre sí. Por ejemplo, si en una prueba psicológica se formulan diversas preguntas al paciente, es difícil pensar que las respuestas a una pregunta sean la causa o el efecto de las respuestas a otra pregunta. De hecho en las aplicaciones prácticas se suele alterar el orden de las preguntas para estar seguros que las respuestas de los pacientes a la pregunta **B** no varíen en función de sus respuestas a una pregunta **A** anterior.

La interpretación más lógica es que la correlación observada entre las respuestas **A, B, ..., Z** en una prueba psicológica es efecto de algún rasgo psicológico subyacente, **X**, que influye a los pacientes para que respondan de cierta manera a todas esas preguntas. Habría así una relación causal de **X** con **A,** de **X** con **B**, etc., pero no habría relación causal alguna entre las variables observables **A**, **B, ..., Z** Como lo sugiere el diagrama, las distintas variables observables son sólo **indicadores** o síntomas observables que reflejan un rasgo subyacente.



Esta concepción guarda mucha relación con el concepto de **correlación espuria**. Cuando dos variables X e Y están correlacionadas, se dice que su correlación es espuria cuando sólo refleja el influjo de una tercera variable Z. Esto significa que **controlando Z la correlación entre X e Y desaparece**. Supóngase por ejemplo dos preguntas X e Y cuyas respuestas están relacionadas en la forma que muestra la siguiente tabla.

Caso 1: Asociación entre dos variables			
	Pregunta Y		
Pregunta X	0=No	1=Sí	Total
0=No	350	220	570
1=Sí	200	320	520
Total	550	540	1090

En esta tabla hay evidentemente una asociación directa entre las dos preguntas. Gran parte de los casos están concentrados en la diagonal principal (670 casos, dos tercios del total). Esta correlación se consideraría **espuria** si existe alguna otra variable Z tal que la correlación desaparezca dentro de cada una de sus categorías. Supóngase una variable Z con la cual aparezcan (por ejemplo) los resultados que se vuelcan en la tabla siguiente, compuesta por dos subtablas.

Caso 2: Asociación entre dos variables explicada por una tercera			
	Pregunta Z = 0 = No		
	Pregunta Y		
Pregunta X	0=No	1=Sí	Total
0=No	50	100	150
1=Sí	150	300	450
Total	200	200	600
	Pregunta Z = 1 = Sí		
	Pregunta Y		
Pregunta X	0=No	1=Sí	Total
0=No	300	120	480
1=Sí	50	20	80
Total	350	140	490

En la subtabla superior, con Z=0, no existe asociación entre X e Y: la proporción entre respuestas positivas a la pregunta X es igual para quienes tienen Y=0, para quienes tienen Y=1 y para el total de esa subtabla. En todos los casos la proporción entre X=0 y X=1 es de uno a tres. En la subtabla inferior la situación es similar: la proporción es de seis a uno en las tres columnas. Del mismo modo ocurre con las filas: en la subtabla superior las proporciones son 1:2 (por ejemplo 50:100) y en la inferior son de 10:4 (por ejemplo 50:20 o 300:120). La aparente correlación que aparecía en la tabla total sin discriminar Z era una correlación espuria, pero dentro de cada una de las dos clases de sujetos determinadas por la variable Z existe independencia entre X e Y. Si se denominara φ_{XY} al coeficiente "phi" de correlación entre X e Y, se observaría que $\varphi_{XY} > 0$. Si la correlación fuese inversa, sería $\varphi_{XY} < 0$. Pero una vez controlada la variable Z resultaría un coeficiente de correlación parcial $\varphi_{XY.Z} = 0$.

Dado que en este caso existe correlación global entre X e Y, pero no existe correlación entre ellas en cada uno de los valores de Z, se dice que existe **independencia local** entre X e Y a pesar de no existir **independencia global** entre ellas.

7.2. Análisis de la estructura latente

En una investigación donde se hayan registrado las tres variables X, Y y Z, la situación del Caso 2 en el ejemplo anterior serviría para descartar la hipótesis causal X→Y, o la hipótesis Y→X, y más bien se afirmaría la hipótesis de que Z es la causa de las otras dos, por ejemplo en un modelo X←Z→Y. En la situación del Caso 1, en cambio, no hay ninguna variable Z que haya sido registrada y que conduzca a ese tipo de interpretación. Si el investigador tiene la hipótesis de que entre X e Y no hay ninguna relación causal directa, y no encuentra ninguna variable Z que

"explique" la correlación observada entre X e Y, podría concluir que (contra sus hipótesis) una de esas variables es realmente la causa de la otra.

Se puede imaginar, sin embargo, una tercera explicación. Supóngase que entre X e Y no hay ninguna relación causal, pero sin embargo están correlacionadas. Si bien no hay ninguna variable observable Z que pueda explicar esa correlación, se sabe o se presume que puede haber **alguna hipotética variable W de naturaleza inobservable** que, si pudiese ser medida y fuese introducida en el análisis, conduciría a una situación de independencia local entre X e Y una vez controlada esa variable **W**. El análisis de estructura latente busca precisamente esa variable puramente hipotética. En otras palabras, **trata de asignar a cada sujeto un valor de W, eligiendo esos valores de modo tal que cuando se controla W aparezca una situación de independencia local de X e Y.** Esa variable W es totalmente hipotética, y no corresponde a ninguna magnitud observable. Puede concebirse como una variable dicotómica análoga a X e Y, o bien como una variable politómica, o incluso como una variable ordinal o de intervalo. El único requisito es que controlando W desaparezca la asociación entre X e Y.

Si se pudiese encontrar esa variable hipotética, se podría pensar que esa variable (mensurable sólo indirectamente a través de sus efectos sobre X e Y) es simplemente un artificio matemático, o bien podría pensarse que W es alguna variable subyacente, inobservable pero real, que "explica" la correlación de las variables manifiestas. Por ejemplo, si X e Y son preguntas relacionadas con algún rasgo psicológico, como la inteligencia o la agresividad, podría pensarse que la variable W que agrupe los sujetos en tal forma que X e Y no estén correlacionadas dentro de cada una de sus categorías, debe ser necesariamente una variable que agrupe los sujetos en función precisamente de ese rasgo psicológico subyacente. Si las dos preguntas se relacionan con la agresividad (X e Y podrían plantear la opción entre una reacción violenta y una reacción pacífica ante diferentes situaciones de la vida), aquellos sujetos que tengan una W (agresividad) más alta tenderán a elegir la alternativa más agresiva tanto en X como en Y, mientras los sujetos que tengan una W (agresividad) más baja tenderán a elegir la alternativa contraria. Entre los sujetos con W alta no habría mucha correlación entre X e Y, y tampoco la habría entre los sujetos con W baja.

Del mismo modo, si X e Y son pruebas de rapidez mental o inteligencia, se concebiría la variable W como una medición de la rapidez mental o inteligencia de los individuos, una cualidad esencialmente inobservable. Los sujetos con elevada W (inteligencia) tenderían a responder correctamente a las pruebas X e Y, mientras los sujetos con W (inteligencia) más baja tenderían a no responder correctamente. Entre los sujetos con elevada inteligencia no habría mucha correlación entre respuestas correctas en X y en Y, ni tampoco entre los sujetos con poca inteligencia.

La variable subyacente W podría tener mucha o poca influencia causal sobre las variables manifiestas. Si la correlación entre X e Y se mantiene inalterada después de haber controlada la variable W, se deduce que W es irrelevante. Más exactamente, si la correlación se mantiene después de haber calculado **la mejor variable W posible**, se deduce que esa correlación no es explicable satisfactoriamente por **ninguna** variable subyacente que actúe sobre X e Y. La correlación de W con X y con Y sería igual a cero, y por tanto W no sería capaz de explicar la correlación observada entre X e Y. En el otro extremo, podría ser W esté **perfectamente** correlacionada con X y con Y. En tal caso, **todos** los que tengan W alta o positiva elegirán la respuesta más agresiva en X y en Y, y todos los que tengan W baja o negativa optarán por la alternativa más pacífica tanto en X como en Y.

Estas situaciones extremas de correlación nula o de correspondencia perfecta, sin embargo, rara vez se observan. Lo más frecuente es una situación intermedia. Cuánto más estrecha sea la asociación entre los indicadores observables, más fuerte será la correlación de la variable subyacente con dichos indicadores. La correlación entre los indicadores pocas veces es perfecta. Es probable que siempre haya algunos sujetos "discordantes" (que den respuestas agresivas en X y no agresivas en Y, o viceversa; que respondan correctamente a la prueba de inteligencia X pero no a la prueba Y, o viceversa). Aun controlando W seguirían existiendo esos casos discordantes. Como esas discordancias son en principio aleatorias no habría correlación entre ellas: elegir la respuesta "errónea" en X no iría asociado con una respuesta "errónea" en Y, o viceversa.

Es posible, sin embargo, que una vez computado el efecto de W se explique una parte de la correlación anteriormente observada entre X e Y, pero que todavía subsista alguna correlación residual entre las variables observables. Esta correlación residual sería independiente de W, de modo que sólo podría ser explicada por una segunda variable subyacente, digamos T. Si ese proceso continúa sería posible identificar todo un conjunto de variables subyacentes, una verdadera "estructura latente" que determina las correlaciones observadas en los datos.

Si bien la variable subyacente no puede ser medida directamente, ni se conocen sus unidades de medida, es posible asignarle un significado cualitativo a sus distintos valores. Volvamos al caso "perfecto" en que W explica prácticamente toda la correlación entre las variables observables. Habría independencia local de las variables observables entre los sujetos agrupados en cada valor o categoría de W. En el caso de una W dicotómica relacionada con la agresividad, seguramente una de las categorías de W mostraría mayor cantidad de respuestas "agresivas", por lo que sería natural pensar que ese valor de W corresponde a "alta agresividad subyacente", y la otra categoría a una "baja agresividad subyacente". Lo mismo pasaría si se concibe a W como una variable de intervalo, y la proporción de respuestas positivas a X e Y aumenta en función directa del valor de W. En ambos casos, la variable W podría considerarse como una medición indirecta de un rasgo psicológico intrínsecamente inobservable, como la agresividad o la inteligencia, a partir de su capacidad para producir independencia local entre los indicadores observables X e Y. Si W se concibe como una variable continua, se puede suponer (sin pérdida de generalidad) que su media es igual a cero y su desviación estándar igual a 1, y con ese supuesto (u otro similar) se pueden asignar valores de W a cada uno de los sujetos.

En los problemas reales de este tipo generalmente no se usan sólo dos indicadores observables como X e Y, sino muchos más, pues cuanto mayor sea el número de indicadores observables mutuamente correlacionados, habrá mayor posibilidad de inferir correctamente la ubicación subyacente de los individuos. Por ello las pruebas y escalas psicológicas usan un número elevado de items (preguntas) para que el índice compuesto con ellos, aunque no sea necesariamente obtenido por análisis de estructura latente, resulte más confiable. De hecho, por razones puramente algebraicas los modelos más simples de estructura latente necesitan **al menos tres** indicadores, y por razones de confiabilidad usualente se usan muchos más que tres.

Un modelo de estructura latente **para datos de panel** supone que los cambios ocurridos en los sujetos a lo largo del tiempo son el efecto de alguna variable subyacente, junto con factores aleatorios residuales. Las características y valores individuales de la variable subyacente no se infieren solamente a partir de la correlación transversal entre valores de las variables observables en una onda determinada, sino la correlación de las **variaciones** de esas variables a lo largo del tiempo. Cada trayectoria o combinación secuencial de valores de las variables observables se asociará con diferentes valores de la variable subyacente.

En el presente texto no se desarrolla en detalle el modelo de estructura latente, que por otra parte es sólo uno de los varios enfoques matemáticos emparentados entre sí que se pueden aplicar a esta situación. En varios textos mencionados en la lista de referencias bibliográficas se pueden encontrar desarrollos detallados sobre este tema.[27]

[27] Por ejemplo en Lazarsfeld & Henry, 1968; asimismo véase von Eye & Clogg, 1994; Berkane, 1997, especialmente el artículo de Arminger 1997; y Hagenaars y McCutcheon, 2002. Otro enfoque posible para el análisis de variables latentes es el análisis factorial, aplicado principalmente a variables de intervalo pero que puede también ser generalizado a otros tipos de variable, principalmente las dicotómicas. El análisis factorial dinámico es la principal técnica que aplica análisis factorial a datos de panel: véase Geweke (1977) y Nesselroade 1997, así como varios trabajos recientes de Mario Forni y colaboradores que se mencionan en las Referencias Bibliográficas. Para la aplicación de modelos log-lineares a datos de panel véase Hagenaars (1990, 1994).

8. Datos de panel y análisis de supervivencia

8.1. Características generales

A veces se usan los datos de panel, o en general los estudios longitudinales que siguen a los mismos sujetos a través del tiempo, para examinar las probabilidades de que ocurra (o no ocurra) algún evento después de diferentes períodos de tiempo. Por ejemplo, se puede hacer el seguimiento de pacientes después de una determinada operación quirúrgica, para examinar la probabilidad de que se mueran, o que tengan una recaída, o cualquier otro evento de interés. El resultado crucial en esos casos es la *tasa de supervivencia* de los sujetos a partir del inicio del proceso; esa tasa equivale a la probabilidad de que hasta un período *t* todavía no les haya ocurrido el evento de interés. Por ejemplo, se pueden tratar de examinar las probabilidades de supervivencia de pacientes con trasplante de corazón: después de un año se han muerto algunos; entre los que han llegado al año, algunos mueren antes de llegar al segundo año; otros antes de completar el tercero, y así sucesivamente (si se prolonga el estudio lo suficiente, terminan por morirse todos).

El objetivo más importante de esos estudios consiste en determinar *factores de riesgo*, es decir variables que pueden acelerar o retardar el evento en cuestión. La probabilidad de ocurrencia de un accidente cerebrovascular (ACV), por ejemplo, aumenta con la edad, pero a cualquier edad el riesgo es mayor entre las personas con mayores factores de riesgo (colesterol en la sangre, sobrepeso, fumadores, falta de actividad física, etc.). Los factores de riesgo afectan las chances de ocurrencia del evento, y el análisis de supervivencia investiga el efecto cuantitativo de cada factor de riesgo (o de la interacción entre ellos).

Los materiales básicos para los análisis de supervivencia son datos de los sujetos obtenidos en diferentes períodos sucesivos. El único elemento común es un evento (positivo o negativo) que en cada período puede ocurrir o no, y que (si ocurre) puede ocurrir con mayor o menor retardo a partir del momento en que se inicia el estudio. Ejemplos de aplicaciones del análisis de sobrevivencia podrían ser los siguientes:

- Sobrevivencia de personas desde el nacimiento hasta la muerte (tablas de mortalidad).
- Sobrevivencia de una persona hasta sufrir una recaída después de una enfermedad, tratamiento, o intervención quirúrgica.
- Probabilidad de casarse (o separarse, o enviudar) a diferentes edades.
- Persistencia de una sustancia en el organismo, o en el ambiente, después de su ingesta o liberación.
- Duración de aparatos o artefactos desde la fabricación hasta que fallan (vida útil).
- Asistencia continuada de los niños a la escuela, desde la matriculación en el primer grado hasta que dejan los estudios.
- Permanencia de los trabajadores en condición de ocupados o desocupados.
- Duración de las situaciones de pobreza por ingresos en los hogares.

8.2. Riesgos simples y riesgos múltiples

En su forma más simple el análisis de sobrevivencia analiza una sola clase de evento, y ese evento es "terminal", en el sentido de que cualquier unidad afectada por el evento desaparece de la población estudiada y ya no le pueden ocurrir otros eventos en el futuro (por ejemplo, si el evento en cuestión es la muerte del sujeto). En modelos más complejos pueden contemplarse dos o más eventos de distinto tipo, los cuales pueden ser excluyentes o compatibles entre sí. Los modelos de "eventos excluyentes" (también llamados "de riesgos competidores" o *competing risks models*) son aquellos en que un evento excluye la posibilidad de que ocurra el otro. Por ejemplo, en un estudio de pacientes con transplante de corazón, la muerte por problemas cardíacos es el evento de interés, pero si durante el período de observación la persona muere por un accidente de tránsito queda fuera del estudio: resultaría imposible que le ocurra la muerte por problemas cardíacos en el resto del tiempo del estudio; ambas causas de muerte son "riesgos competidores" que se excluyen entre sí.

Un modelo de riesgos *no competidores* podría incluir dos clases de eventos **compatibles entre sí**, como el de casarse y el de ingresar en el desempleo. Una persona soltera y ocupada puede casarse y todavía seguir en riesgo de quedar sin trabajo, o viceversa, ya que ambos riesgos no se excluyen entre sí. El riesgo de maternidad adolescente es compatible con el riesgo de abandono escolar: una adolescente que tenga un hijo puede seguir expuesta al riesgo de abandonar la escuela, y viceversa. Ambos riesgos pueden además **influirse mutuamente**: las chances de abandono escolar pueden acentuarse después de tener un hijo, y viceversa.

8.3. Modelos de análisis de supervivencia

La modelización estadística de esta clase de situaciones requiere alguna simplificación, pues de otro modo exigiría estudios muy específicos y complejos de cada situación para determinar la forma de las relaciones funcionales involucradas, es decir la forma matemática de la curva de sobrevivencia (la probabilidad del evento como función del tiempo) y la forma matemática de los riesgos diferenciales (es decir la forma matemática de las relaciones entre la curva de sobrevivencia y los distintos factores de riesgo). Dado que en la práctica es imposible llevar a cabo esos estudios en cada caso concreto, el análisis se realiza bajo dos modalidades: (a) el análisis "paramétrico" que *asume* una cierta distribución o forma de las curvas probabilísticas involucradas (exponencial, logarítmico-logística, de Weibull, etc.), o (b) el análisis "no paramétrico" o "semi-paramétrico" que no hace supuestos matemáticos tan audaces y se limita a calcular proporciones más básicas y aproximadas.

El enfoque más común es el análisis semi-paramétrico denominado "regresión de Cox" (Cox 1972), el cual estima primero la curva de sobrevivencia de base, y luego *asume* que los factores de riesgo ejercen sus efectos **en forma proporcionalmente igual *en todos los momentos del tiempo***. Esto significa que si un factor de riesgo aumenta la probabilidad del evento en un 20%, ese impacto proporcional se supone que es el mismo en todos los momentos del período de observación. Si un paciente que fue fumador tiene 20% más chances de morirse en el primer año respecto a otro paciente que nunca fumó, se supone que también tendrá 20% más chances en el segundo año, en el tercero, etc.

Ese modelo de Cox se denomina por ello "**modelo de riesgos proporcionales**" (*proportional hazards model*). Ese supuesto puede ser bastante irreal en algunos casos, pues la proporción de riesgo adicional causado por un cierto factor puede aumentar (o disminuir) con el correr del tiempo. Se pueden usar para esos casos otros modelos, o bien el propio modelo de Cox admite una modificación que tiene en cuenta ese aspecto: las "covariadas dependientes del tiempo" (variables predictoras o factores de riesgo cuyo valor varía con el tiempo transcurrido).[28]

8.4. La función de sobrevivencia

Con respecto a cada posible evento, las poblaciones exhiben una cierta **curva de sobrevivencia**, que expresa la proporción de casos **a los cuales todavía no les ocurrió el evento** en diferentes momentos a partir del momento inicial. A partir de ese momento inicial, los sujetos son estudiados hasta llegar a un momento o etapa final del estudio; al llegar ese momento puede haberles ocurrido el evento (y por lo tanto ya han sido contabilizados) o bien pueden haber sobrevivido sin haberles ocurrido el evento todavía; también pueden haber abandonado el estudio antes de esa fecha, perdiéndoseles el rastro. En estos dos últimos casos (individuos que sobreviven hasta el final, o que se les pierde el rastro) se ignora si les ocurre el evento, y la fecha eventual de ocurrencia del mismo; se dice que esos casos están "censurados", pero se usa de todos modos la información disponible acerca de ellos hasta el punto en que permanecieron en el estudio: fueron observados durante un cierto tiempo y no les ocurrió el evento hasta que dejaron de ser observados. El análisis se basa en los sujetos a los cuales les ocurrió el evento, los que llegaron al final del estudio sin haber sufrido el evento, y aquellos que resultaron "censurados" antes del fin del estudio por haber dejado el estudio debido a cualquier causa: todos intervienen en el cálculo de la curva promedio de sobrevivencia.

[28] Véase Allison 1984; Yamaguchi 1991; Klein & Moeschberger 1997; Hosmer & Lemeshow 1999; Blossfeld & Rohmer 2002; Vermunt 1997; Therneau & Grambsch 2000; Box-Steffensmeier & Jones 2004.

La curva de sobrevivencia, a su vez, se ve modificada por diversos **factores de riesgo** que afectan de manera diferencial a diversas sub-poblaciones: la curva de sobrevivencia de un niño en la escuela, o de un trabajador en su empleo, puede ser diferente según la edad, el sexo, las condiciones socioeconómicas, la condición migratoria, el nivel educativo del niño o de sus progenitores, u otros factores. Estos factores de riesgo, sin embargo, no permiten predecir el desenlace de cada caso individual, sino sólo la *probabilidad* de ocurrencia del evento, a diferentes plazos, para cada población o sub-población. Los factores de riesgo no tienen que ser variables categóricas, como el evento: pueden ser también variables cuantitativas o de intervalo, como la edad o los ingresos mensuales.

La noción fundamental del análisis de sobrevivencia es la *función de sobrevivencia*. Esta función es una función no-creciente (y en general una función estrictamente decreciente) del tiempo transcurrido desde un instante inicial, e indica, para cada momento del tiempo considerado, cuál es la probabilidad de que un sujeto haya "durado" o "sobrevivido" hasta ese momento sin que le ocurra el evento al menos por un período de longitud k, desde el momento inicial t_0 hasta el momento t_k. En otras palabras, la función de sobrevivencia da el número de sobrevivientes en el momento t_k como función del número inicial n_0 multiplicado por la probabilidad de sobrevivir hasta t_k:

$$n_{0,k} = n_0 p(k)$$

La probabilidad p en esta expresión es el cociente entre los sujetos a los cuales no les ocurrió el evento entre t_0 y t_k, dividida por la cantidad de sujetos *expuestos al riesgo* (es decir la cantidad n_0 de sujetos que iniciaron el estudio, y a los cuales por supuesto en ese momento inicial t_0 no les había ocurrido el evento. Supongamos que los eventos ocurren entre sucesivos "instantes" t_0, t_k, t_h, etc. Si en el momento inicial t_0 había n_0 sujetos, y en el primer momento intermedio t_k han ocurrido g_k eventos, la cantidad de "sobrevivientes" en t_k a los cuales aún no les ha ocurrido el evento es igual a $n_k=(n_0 - g_k)$. La probabilidad de sobrevivir desde t_0 al menos hasta t_k es $p_k = (n_0 - g_k)/n_0$. Más sucintamente, si el número de sobrevivientes al momento k se define como $n_k=n_0 - g_k$, la probabilidad es $p_k=n_k/n_0$. En el período siguiente, de t_k a t_h, los sujetos expuestos son $n_k=n_0-g_k$, es decir los que no habían sobrevivido hasta el momento k. Entre los momentos t_k y t_h el evento les ocurre a g_h sujetos. La cantidad de sobrevivientes en t_h es $n_h=n_k-g_h$, y la probabilidad de sobrevivencia de t_k a t_h es $p_h=n_h/n_k$. En general en un instante cualquiera t_u será $p_u=n_u/n_{u\text{-}1}$. Esta es la probabilidad de que los sujetos que sobrevivieron hasta u-1 sobrevivan hasta u. En otras palabras, es la probabilidad de estar vivo en el momento u condicionada a haber estado vivo en el momento u-1.

Esta función p_u expresa la probabilidad de sobrevivir *entre dos instantes sucesivos*. A partir de ella se puede definir la *función acumulada de sobrevivencia*, que indicaremos con una S mayúscula. Esta función refleja la probabilidad de sobrevivir desde el momento inicial hasta un momento ulterior cualquiera, lo que implica haber sobrevivido durante todos los momentos intermedios. En teoría los "instantes" pueden ser de longitud infinitesimalmente pequeña, de modo que el tiempo sería continuo, y la probabilidad acumulada sería la integral de la función de sobrevivencia instantánea (período a período). En la práctica, sin embargo, los eventos se miden en intervalos discretos de tiempo, de modo que la función acumulada de sobrevivencia S_t hasta el período t es el producto de las sucesivas probabilidades de sobrevivir a través de los distintos períodos transcurridos desde el punto de origen hasta t:

$$S_u = p_0 \times p_1 \times p_2 \times \ldots \times p_t$$

La función de sobrevivencia es una probabilidad, y varía siempre entre 0 y 1. A ningún sujeto le ha acaecido aún el evento en el instante inicial t_0, de modo que por definición $p_0 = S_0 = 1$. A partir de allí el número (y la proporción) de sobrevivientes no puede aumentar: en cada período solo puede permanecer invariado o disminuir.

8.5. La tasa de ocurrencia

Otra noción básica del análisis de sobrevivencia es la *tasa de ocurrencia* del evento por unidad de tiempo (*hazard rate*).[29] La tasa de ocurrencia *h(t)* de un evento en un período *t* de duración *k*, es el **número de eventos** que se espera que ocurran en ese período. Estrictamente se la define como una tasa instantánea, es decir como el límite del número de eventos en un período de duración *k* cuando la duración del período tiende a cero. Para calcularla se parte de la probabilidad de que ocurra un evento en el período que va del instante *t* al instante *t*+*k*, calculada como el número de eventos ocurridos en ese período, sobre el número de *sujetos en riesgo* al inicio del período. Dado que los intervalos pueden ser de diferente longitud, supongamos que tenemos un intervalo de longitud *k*. La tasa de ocurrencia en ese intervalo se define (para datos agrupados por intervalos) del siguiente modo:

$$\hat{h}(t) = \frac{g_{t,k}/k}{n(t)}$$

En esta fórmula, $g_{t,k}$ es el número de eventos ocurridos entre *t* y *t*+*k*; la longitud del intervalo transcurrido es *k*, que puede estar expresado en cualquier unidad de tiempo (días, meses, años, etc.); y *n*(*t*) es el número de sujetos expuestos al riesgo en ese intervalo, es decir los que iniciaron el intervalo en el momento *t* sin que (hasta ese momento) les hubiese ocurrido el evento. La fórmula que antecede se refiere la tasa de ocurrencia durante un intervalo de longitud mensurable, *k*, pero la tasa de ocurrencia instantánea en un momento *t* se define como el límite de la anterior expresión cuando la longitud del intervalo tiende a cero:

$$h(t) = \lim_{k \to 0} \hat{h}(t) = \lim_{k \to 0} \left[\frac{g_{t,k}/k}{n(t)} \right]$$

La tasa acumulada de ocurrencia, *H*(*t*), es la integral de las tasas *h*(*t*) desde el momento 0 hasta el momento *t*. Tomando la variable *u* para representar el tiempo sería

$$H(t) = \int_0^t h(u)du$$

Si se consideran intervalos de longitud mensurable, esa tasa acumulada se puede aproximar mediante un estimador estadístico (Hosmer & Lemeshow 1999:74-75). Uno de esos estimadores es el de Nelson-Aalen, equivalente a la suma de las ocurrencias por persona expuesta, en los sucesivos intervalos de 0 a *t*. Sobre una serie de intervalos *u* desde *u*=0 hasta *u*=*t* se tiene:

$$\hat{H}(t) = \sum_{u \le t} \frac{g_u}{n_u}$$

Otro estimador muy semejante en sus resultados es el de Peterson, basado en la función de sobrevivencia de Kaplan-Meier:

$$\hat{H}(t) = \sum_{u \le t} -ln\left(1 - \frac{g_u}{n_u}\right)$$

Entre la tasa de ocurrencia *h*(*t*) y la de supervivencia *s*(*t*) o entre sus estimadores hay una relación muy simple (Hosmer & Lemeshow 1999:73): la primera es el logaritmo natural de la segunda (cambiado de signo):

$$h(t) = -ln\,(s_t). \qquad \text{Es decir: } s_t = e^{-h(t)}$$

Similar relación existe entre las tasas **acumuladas** de ocurrencia y de supervivencia:

$$H(t) = -ln\,(S_t). \qquad \text{Es decir: } S_t = e^{-H(t)}$$

---

[29] El nombre de esta tasa en inglés (*hazard rate*) tiene una connotación negativa: un *hazard* es un peligro, un evento posible *de carácter adverso*. Podría traducirse como *tasa de siniestralidad*, tomando un termino técnico de los seguros contra siniestros. Pero en sí misma esta clase de análisis estadístico se refiere a toda clase de eventos, tanto positivos como negativos, por lo cual es más adecuada una designación más neutral, como *tasa de ocurrencia* de los eventos.

A veces se concibe incorrectamente la tasa de ocurrencia como una probabilidad, pero estrictamente no lo es porque puede tener (y tiene frecuentemente) valores superiores a la unidad, si ocurre más de un evento por unidad de tiempo, aun cuando la unidad de tiempo sea muy pequeña. Se asemeja así a las tasas de transición que se han examinado antes.

La tasa de ocurrencia de ciertos eventos puede ser de muchos casos por unidad de tiempo, incluso miles de casos, no importa si la unidad de tiempo considerada es una hora (donde k=1 hora) o un día (donde k=1 día) o cualquier otro período, por ejemplo un mes o un año. La tasa de ocurrencia, entonces, puede ser mayor que uno si el "evento" ocurre muy cuantiosamente o muy frecuentemente, y en particular si le puede ocurrir varias veces a cada sujeto expuesto al riesgo (por ejemplo si la unidad de medida es un día y el evento consiste en algo muy frecuente, que puede suceder muchas veces en un día, por ejemplo pronunciar la palabra "gracias"). Puede ser concebida más correctamente como la *velocidad* con que ocurren los eventos (número de eventos por unidad de tiempo).

La recíproca de $h(t)$, es decir $1/h(t)$, indica el *tiempo esperado* desde el inicio del período hasta la ocurrencia de un evento. Se lo llama habitalmente "tiempo al evento" (*time to event*).

La comparación de tasas de ocurrencia suministra una medida del *riesgo relativo* o *comparativo* entre sujetos con características distintas. Si dos grupos o tipos de individuos tienen tasas de ocurrencia de 0.2 y 0.6, se puede decir que el segundo tiene tres veces más riesgo que el primero de que le ocurra el evento en un período determinado.

La tasa de ocurrencia de los diferentes períodos da origen, como se ha visto, a la tasa *acumulada* de ocurrencia, $H(t)$, la cual refleja la proporción de casos a los cuales les ha ocurrido el evento desde el inicio (t_0) hasta un determinado momento (t). Muy conectada con la tasa acumulada de ocurrencia está naturalmente la tasa acumulada de sobrevivencia (*cumulative survival rate*), $S(t)$. Esta tasa $S(t)$ es la proporción de casos a los cuales (al momento t) aún no les ha ocurrido el evento. Así como la tasa acumulada de ocurrencia (*cumulative hazard rate*) suele ser *creciente* a medida que pasa el tiempo (pues cada vez hay más sujetos a los cuales ya les ha ocurrido el evento), la tasa acumulada de supervivencia es normalmente una función *decreciente* del tiempo (pues cada vez hay una menor proporción de sujetos a los cuales *todavía no les ha ocurrido* dicho evento). Por ejemplo, en un estudio de seguimiento después de la quimioterapia, a medida que pasa el tiempo algunas personas mueren, de modo que la proporción de sobrevivientes tiende a ser cada vez menor, y la proporción de fallecidos cada vez mayor.

En cambio la tasa de ocurrencia por unidad de tiempo, lo mismo que la tasa de supervivencia por unidad de tiempo, no tiene por qué ser creciente ni decreciente. Por ejemplo en las tablas de vida la tasa acumulada de fallecimientos de una cohorte es creciente (y el porcentaje de sobrevivientes es decreciente), pero la tasa específica de mortalidad o de supervivencia *por edad* puede ser creciente o decreciente: la tasa específica de mortalidad por edad es de hecho decreciente desde el nacimiento hasta los 5-8 años, luego se estabiliza por una o dos décadas, y luego comienza a crecer hasta la extinción de la cohorte en la ancianidad.

Ambas funciones son así opuestas entre sí. La tasa de supervivencia en un momento t expresa en forma proporcional el número de sobrevivientes (aun no afectados por el evento) hasta ese momento, y así depende del número inicial de casos menos la suma de las ocurrencias del evento hasta ese momento; la tasa acumulada de ocurrencia es precisamente la suma de las ocurrencias del evento hasta ese momento. Si un 80% de los nuevos aparatos sobrevive al menos tres años sin fallas, el número total de aparatos que fallen hasta los tres años equivaldrá a 20% del número inicial. Sin embargo, habitualmente la tasa acumulada de ocurrencia se expresa en *número de eventos por persona*, mientras que la tasa de supervivencia se expresa como la *proporción de casos* que no han sufrido el evento. La dimensionalidad de la tasa de ocurrencia es "eventos por persona por unidad de tiempo" (E/NT), mientras que la tasa de supervivencia no tiene dimensionalidad excepto que se mide por unidad de tiempo (1/T).

8.6. Regresión de Cox

8.6.1. Supuestos y enfoque general

En forma general, la tasa de ocurrencia de un evento puede ser concebida como una magnitud dependiente de dos clases de factores: el mero paso del tiempo, y el valor de diversas variables condicionantes o *covariadas*:

$$h(t) = f\,(t, X, W, Z, \ldots)$$

Por ejemplo, si se considera la vida útil de un cierto tipo de aparatos, la cantidad de aparatos que fallan dependerá del tiempo que llevan en uso (t), y también de factores como el modelo y marca del aparato (X), la cantidad promedio de horas diarias de funcionamiento (W), la temperatura reinante en el sitio donde está ubicado el aparato (Z), etc.

Las covariadas pueden ser *fijas en el tiempo* (es decir que cada caso mantiene el mismo valor de sus covariadas en todos los momentos del tiempo, como por ejemplo el modelo y marca del aparato) o *dependientes del tiempo* (donde el valor de una covariada para cada caso individual podría variar con el correr del tiempo, por ejemplo las horas acumuladas de funcionamiento, o el promedio de horas diarias de funcionamiento en los últimos 30 días). La variable "edad de la madre al nacer el niño" es una covariada cuyo valor es fijo a medida que el niño crece, mientras que las variables "edad actual del niño" o "nivel educacional alcanzado por el niño" son covariadas cuyo valor aumenta al correr el tiempo.

El modelo de Cox postula que la tasa de ocurrencia en el período t para cada sujeto o grupo de sujetos caracterizado por ciertos valores de las covariadas obedece a dos factores:

1. La tasa de ocurrencia de base o de referencia (*baseline hazard rate*), $h_0(t)$
2. El efecto de las covariadas sobre el riesgo de ocurrencia, que es una función $r(x)$.

La tasa de ocurrencia de base, denotada como $h_0(t)$, depende sólo del tiempo, y afecta a todos los sujetos. Aquellos sujetos cuyas covariadas son todas iguales a cero tienen una tasa de ocurrencia equivalente a la de base. El efecto de las covariadas es una función $r(x)$, que en la regresión de Cox se supone *exponencial* ($e^{bX+\cdots}$) y que difiere de un sujeto (o grupo de sujetos) a otro. Esa función multiplica la tasa de ocurrencia de base por un cierto múltiplo, el cual incrementa o reduce el riesgo básico de ocurrencia del evento representado por $h_0(t)$. La magnitud y dirección del efecto depende de los valores de las covariadas en cada caso particular. La tasa de ocurrencia es así el producto de los dos elementos:

$$h(t) = h_0(t)r(x) = h_0(t)e^{\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \ldots.}$$

En principio, el exponencial podría contener una constante β_0, en esta forma:

$$h(t) = h_0(t)r(x) = h_0(t)e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3 \ldots.}$$

Pero en realidad la constante β_0 no se puede estimar separadamente, y resulta subsumida en $h_0(t)$. Esto es fácil comprobarlo a partir de la ecuación anterior:

$$h(t) = h_0(t)e^{\beta_0+\beta_1 X+\beta_2 W} = h_0(t)e^{\beta_0}e^{\beta_1 X+\beta_2 W} = h_0^*(t)\,e^{\beta_1 X+\beta_2 W}$$

donde $h_0^*(t) = h_0(t)e^{\beta_0}$ es el valor que realmente se puede estimar para la tasa de ocurrencia de base (el asterisco, usado aquí con fines demostrativos, no se usa en la práctica).

8.6.2. Riesgos relativos

Si se tienen dos individuos (i,j) con diferente valor de las covariadas, la proporción de sus respectivas tasas de ocurrencia sería:

$$\frac{h(t)_i}{h(t)_j} = \frac{h_0(t)e^{b_1 X_i+b_2 W_i}}{h_0(t)e^{b_1 X_j+b_2 W_j}} = \frac{e^{b_1 X_i+b_2 W_i}}{e^{b_1 X_j+b_2 W_j}}$$

Se observa claramente que la relación proporcional entre las tasas de ocurrencia del sujeto *i* y el sujeto *j* **no depende del tiempo**: depende solamente de los valores individuales que ellos tengan en las covariadas (que en este caso son solamente dos, X y W). Normalmente en la regresión de Cox lo que se evalúa son las **razones de riesgo**, es decir el aumento o disminución proporcional en el riesgo para sujetos con una cierta combinación de valores de las covariadas, en comparación con sujetos con una combinación de referencia

Podría usarse como "sujeto de referencia" al sujeto de base, es decir, el sujeto que tiene valor cero en todas las covariadas, pero a veces ello no es realista. Es importante destacar aquí que la tasa de ocurrencia de base corresponde al valor cero de todas las covariadas, y no representa la situación del *sujeto promedio*, salvo si todas las variables tienen una media igual a cero. La tasa de base representa la tasa de ocurrencia para sujetos que tengan *valor cero* en todas las covariadas, sin importar lo que el cero signifique. Si una de las covariadas es la edad al momento inicial, la tasa de base corresponde a personas cuya edad inicial fuese cero (lo cual en general es un caso puramente teórico, pues muy pocos estudios arrancan con recién nacidos). Si otra variable fuese el número de hijos tenidos por la madre del sujeto, la tasa de base corresponde a madres con cero hijos, lo cual sería imposible (pues el estudio versa precisamente sobre uno de sus hijos).

Por ello se suele usar como tasa referencial de ocurrencia la que le corresponde al *individuo promedio*, es decir un individuo cuyas covariadas estén todas en el valor promedio. Pero podría tomarse como referencia cualquier otra configuración de valores de las covariadas. Por ejemplo, en un estudio de riesgo cardiovascular la referencia podrían ser los individuos que no estén excedidos de peso, que hagan ejercicio físico, que no fumen, que no tengan alto nivel de colesterol o diabetes, etc., para evaluar en cuánto se incrementa el riesgo al variar una o varias de esas covariadas.

Si las covariadas son fijas en el tiempo, la proporción entre las tasas de ocurrencia de dos individuos no dependerá del tiempo transcurrido. La proporción entre los riesgos de los dos individuos, es decir el mayor o menor riesgo de *i* respecto a *j* es, por lo tanto, una proporción *constante*. Por esto es que el modelo de Cox se denomina "de riesgos proporcionales". En un ejemplo sencillo: supóngase que, después de haber sufrido un primer infarto, a un fumador se le estima el doble de chances de morir que a un no-fumador; en el modelo de Cox se supone que las chances de morir del fumador serán siempre el doble, sea un mes después del infarto, un año después o cinco años después, pues en la regresión de Cox la proporción entre los riesgos no depende del tiempo, sino que se mantiene entre ellos *una proporción constante*. Este supuesto puede relajarse si se usan *covariadas dependientes del tiempo,* por ejemplo la edad de los pacientes, que aumenta al pasar el tiempo, o los resultados de exámenes periódicos que se les practiquen a partir del primer infarto, que pueden dar resultados distintos en diferentes momentos para cada uno de los pacientes.

La tasa de base es una función creciente del tiempo. Por ejemplo la tasa de mortalidad de las personas que han sufrido un infarto aumenta sostenidamente con el tiempo transcurrido desde que comienzan a ser monitoreados. Asimismo, la probabilidad de encontrar con vida a una persona secuestrada disminuye con el paso del tiempo. Ese riesgo creciente de base, que aumenta con el tiempo, es afectado luego por las características particulares de cada sujeto (o sea por el valor de las covariadas en cada caso).

En la regresión de Cox (si las covariadas son fijas en el tiempo) el efecto del tiempo está completamente separado del efecto de las covariadas. Si todas las covariadas son iguales a cero, el exponencial $e^{b_1X+b_2W}$ se reduce a $e^{b_1 0+b_2 0} = e^0 = 1$, pues todo número elevado al exponente cero es igual a uno. En ese caso la tasa de ocurrencia es igual a la de base: $h(t) = h_0(t)$. Si la tasa de ocurrencia correspondiente a una combinación cualquiera (i) de valores de las covariadas se compara con la tasa de base $h_0(t)$, de la ecuación precedente se desprende:

$$\frac{h(t)_i}{h_0(t)} = \frac{e^{b_1X_i+b_2W_i}}{1} = e^{b_1X_i+b_2W_i}$$

En forma logarítmica esta ecuación puede expresarse en la forma lineal siguiente:

$$\ln\left(\frac{h(t)_i}{h_0(t)}\right) = b_1 X_i + b_2 W_i$$

Esto se aplica perfectamente al caso de covariadas "invariables en el tiempo" como el sexo del sujeto, pero no cuando las covariadas van cambiando de valor a lo largo del tiempo, como por ejemplo si esa covariada es el nivel educativo o la edad del niño cuando se trata de un estudio sobre el riesgo de deserción escolar; el nivel educativo alcanzado probablemente aumenta a medida que pasa el tiempo desde que se matriculó en primer grado, y ello puede influir (positiva o negativamente) en su probabilidad de abandonar la escuela a diferentes edades. La educación del niño (o su edad), en tal caso, sería una *covariada dependiente del tiempo*.

8.6.3. Covariadas dependientes del tiempo

La inclusión de covariadas dependientes del tiempo no altera fundamentalmente el modelo en sus aspectos teóricos, aunque las técnicas de cálculo son más complicadas. Su principal ventaja es que permite incorporar *riesgos no proporcionales*, donde la proporción entre los riesgos de dos individuos puede variar a lo largo del tiempo. Puede así haber una covariada X_k que tenga valores distintos para cada período, como la edad o el nivel educativo de un estudiante, o una covariada que que refleje el saldo adeudado de un préstamo con sus intereses acumulados, es decir el monto prestado original, menos la parte ya devuelta, multiplicado por un factor de interés compuesto: $X_m=K(1+i)^t$ donde K es el saldo pendiente en el momento *t*. Esta covariada es una función **exponencial** del tiempo. Otra forma de covariada dependiente del tiempo sería una función **lineal** del tiempo del tipo $X_m=a+bt$.

El resultado de incluir covariadas dependientes del tiempo es que los riesgos relativos de cada sujeto o grupo de sujetos (respecto a la situación de referencia con covariadas iguales a cero) ya no serán constantes, sino que pueden aumentar o disminuir conforme pasa el tiempo. Puede ser por ejemplo que en los primeros dos años la pertenencia al sexo femenino incremente el riesgo, pero en los años siguientes no lo afecte, o incluso lo reduzca. Los riesgos relativos ya no serían proporcionales, o mejor dicho, la proporción ya no sería constante, e incluso puede revertir la relación entre los riesgos de dos individuos.

Nótese que en este enfoque lo que varía a lo largo del tiempo son los *valores* de las covariadas, pero sus *coeficientes* (*b*) siguen considerándose constantes.[30] Esto implicaría que si, por ejemplo, el riesgo relativo de los varones (respecto de las mujeres) aumenta con la edad, cada año de edad adicional incrementa el riesgo relativo de los varones *en una cantidad fija* (*b*).

8.6.4. Estratificación

Otra variante que admite el modelo de Cox es la *estratificación*. En ese enfoque la tasa de ocurrencia de base (*baseline hazard rate*) o en forma equivalente la curva de sobrevivencia de base (*baseline survival curve*) se calcula por separado en cada estrato, por ejemplo para zonas urbanas y rurales, pero el efecto de los factores de riesgo se calcula utilizando simultáneamente todos los casos. Así se mantiene el análisis conjunto para el impacto de los factores de riesgo, pero se acepta que los riesgos de partida pueden ser diferentes en cada estrato por razones ajenas a los factores de riesgo. Este enfoque se usa a veces para reflejar diferentes factores ambientales en epidemiología (la sobrevivencia de base puede ser diferente entre zonas tropicales y templadas, pero el efecto de las covariadas sería el mismo en ambas zonas).

Por supuesto, también se pueden estimar ambas cosas por separado, realizando un análisis totalmente separado en cada estrato, con curvas de sobrevivencia diferentes y coeficientes de riesgo también diferentes, cuando el número de casos en cada estrato lo permite. En caso de usar, por ejemplo, bases de datos censales, se podrían calcular diferentes regresiones de Cox para cada región del país. Cada uno de esos estratos regionales tendría su propia curva de base y sus propios coeficientes de

[30] El modelo de Cox no contempla *coeficientes* dependientes del tiempo, los cuales están contemplados, en cambio, en el modelo lineal de sobrevivencia de Aalen: véase Aalen (1980, 1989, 1993) y Aalen *et al* (2001).

impacto de los factores de riesgo. En lugar de un solo análisis con estratos regionales se tendrían varios análisis separados, uno por región, entre los cuales diferirían no solo los riesgos de base sino también los efectos de las covariadas (por ejemplo, el incremento del riesgo de desnutrición causado por un menor nivel educativo de la madre podría ser diferente en zonas urbanas y rurales).

Aparte de la regresión de Cox con covariadas dependientes del tiempo, existen otros métodos para analizar riesgos no proporcionales. Uno de ellos es la regresión de Aalen (véase Aalen 1980, 1990, 1993; Aalen *et al* 2001).[31] En esta última, los riesgos relativos (*odds ratios*) no responden a una función exponencial del tipo e^{+bX+cZ} sino a una función lineal del tipo $a+bX+cZ$, lo cual enfrenta en forma directa los casos de riesgos no proporcionales, sin sufrir los problemas computacionales de las covariadas dependientes del tiempo en la regresión de Cox. Sin embargo, la regresión de Aalen no está debidamente implementada en los *softwares* estadísticos estándar y presenta algunos otros problemas técnicos, por lo cual no será usada en el presente contexto.

8.6.5. Interacción de covariadas

Como ya se anticipó en la sección precedente, los modelos de regresión de Cox (como los de Aalen) admiten también la *interacción* entre las variables predictoras. Para introducir interacciones basta con introducir variables que sean función (no lineal) de dos o más predictoras, como por ejemplo XZ, o bien X/Z, o bien X^Z, o cualesquiera otras funciones, en interacciones dobles, triples o de orden superior. Tal como se especificó para el caso de la regresión logística, en principio se prueban modelos sin interacciones, salvo cuando hay bases teóricas para introducirlas, aunque a menudo se introducen luego algunas interacciones razonables para verificar si mejoran significativamente la capacidad predictiva del modelo no interaccional. Las interacciones que no mejoren significativamente el modelo serán descartadas, para mantener los modelos lo más simples que sea posible.

En las bases de datos de panel obtenidas en encuestas con paneles rotativos es habitual que cada sujeto permanezca en la muestra por un número limitado (k) de ocasiones, por ejemplo cuatro rondas. Esto significa que para cualquier evento que se quiera examinar por medio del análisis de supervivencia, todos los sujetos serán "censurados" después de k ocasiones. Por otro lado, esos sujetos son reemplazados gradualmente, de modo que en cada ronda se reemplaza una parte de la muestra. Para tener un número más grande de sujetos observados cuatro veces, habría que tomar sujetos que ingresan y salen de la muestra en diferentes rondas.

Cuando el fenómeno que se quiere analizar es considerado como independiente del tiempo, ello no debería causar ningún problema. Por ejemplo, las condiciones que operan sobre eventos de salud como los accidentes cardiovasculares no se espera que varíen significativamente entre una y otra "generación" de participantes de la encuesta. En tal caso, se puede formar una base de datos conformada por varias "generaciones" de participantes, cada uno de ellos con hasta cuatro observaciones, pero que fueron observados en diferentes fechas ya que entraron y salieron de la muestra en forma escalonada.

Cuando el fenómeno que es objeto del estudio puede tener alguna relación con la fecha del estudio (por ejemplo, el evento de quedarse sin empleo), este procedimiento debería complementarse con alguna medida adicional para controlar ese aspecto. La hipótesis más simple es que los procesos causales intervinientes son siempre los mismos, pero con un efecto de la fecha, que es independiente de los otros factores. Así, por ejemplo, el efecto del nivel educativo sobre la pérdida del empleo se podría suponer que es constante, pero además la probabilidad del evento puede aumentar o disminuir según la fecha en que hayan sido obtenidos los datos. Si cada fecha se identifica con una variable *dummy* (por ejemplo, una variable que vale 1 si se trata del mes de septiembre de cierto año, y vale cero en todas las otras fechas), el coeficiente correspondiente a esa fecha expresaría el efecto (positivo o negativo) de la fecha en cuestión sobre la probabilidad de quedarse sin empleo. Otra po-

[31] Véase también Huffer & McKeague 1987; Lin & Ying 1994; McKeague 1997; Amir & McKeague 2000; Scheike 2001; Gandy & Jensen 2005; Ma *et al* 2006. Cao 2005 compara los métodos de Cox y de Aalen. Existen también modelos aditivo-multiplicativos como el llamado modelo Cox-Aalen y otros más generales: Lin & Ying 1995; Martinussen & Scheike 2002; Scheike & Zhang 2002, 2003; Baldi *et al* 2006.

sibilidad para la captura de estos efectos es la introducción de una covariada dependiente del tiempo, usando por ejemplo alguna variable que refleje la fase del ciclo económico (tasa trimestral de crecimiento del producto bruto, o algún indicador de confianza de las empresas, etc.). De uno u otro modo, esas series de *k* observaciones por sujeto, realizadas en varias fechas a medida que esos sujetos entraban y salían de la muestra, podrían ser acumuladas en una única base de datos, logrando así un número más grande de casos.

Si el análisis se basa en un estudio longitudinal específicamente diseñado para ello, probablemente se cuente con observaciones durante un período largo, como en los estudios médicos de seguimiento donde los pacientes son observados regularmente a lo largo de varios años (en ciertos casos, *muchos* años). En algunos de esos estudios todos los sujetos entraron al mismo tiempo en el estudio, y el estudio terminó también en una fecha determinada (para todos los que aun seguían en él). En otros estudios médicos la fecha de entrada o salida no tiene importancia: un estudio acerca de la sobrevida de personas con trasplante de corazón puede hacerse con pacientes trasplantados en diferentes fechas, sin que haya que controlar la fase del ciclo económico (aunque la temporada del año puede ser importante si los pacientes invernales tienen diferente evolución que los de verano).

Otro ejemplo en el cual la fecha efectiva de los sucesos podría ser ignorada son los llamados datos de *cohorte*. Una cohorte está constituida por los individuos a los cuales les ocurre un evento en un determinado momento. El ejemplo más usual son las *cohortes de nacimiento*: las personas nacidas en un cierto período (como el año 2000) forman una cohorte. Dentro de esa cohorte puede haber pequeñas diferencias temporales (por ejemplo entre los nacidos en diferentes meses del año), pero a los efectos de definir la cohorte esas pequeñas diferencias no se toman en cuenta.También hay cohortes escolares (estudiantes que ingresaron en el año 2000) o matrimoniales (matrimonios que se celebraron en el año 2000).

A veces se pueden usar simultáneamente diferentes cohortes, cuando la fecha inicial no tiene importancia para el tema del estudio. Supongamos que se tienen datos sobre nacimientos (o ingresos a la escuela, o matrimonios) ocurridos en diferentes años (2000, 2001, 2004, 2008), no necesariamente consecutivos. El estudio busca analizar lo que sucede con esas personas (o esas parejas) a medida que pasa el tiempo desde el hecho inicial, observándolos en los años sucesivos al evento. Puede ocurrir así que cada caso haya sido seguido y observado por un número determinado de años (por ejemplo siete años), pero esos años no son los mismos: algunas series de siete años comenzaron en 2000, otras en 2004, etc. Este problema es analizado por Verbeek & Nijman 1992.

Como en otros casos, estos paneles combinados, con cohortes que comienzan en diferentes años, tal como antes hemos mencionado para los paneles de pacientes que fueron operados del corazón en diferentes fechas, pueden ser tratados como datos de panel, pero la validez de las conclusiones dependerá de la posible correlación entre los eventos estudiados y las fechas efectivas de observación. Si se trata de datos socioeconómicos, no será lo mismo si se comienza en un año con alto o con bajo desempleo. Este problema puede controlarse introduciendo variables dummy que identifican el año, o bien introduciendo variables predictoras que reflejen los posibles factores asociados a la fecha (por ejemplo el nivel de desempleo).

9. Ponderación muestral

Muchas veces los datos de las encuestas deben ser **ponderados** si se quieren obtener estimaciones muestrales insesgadas referidas al total de una población. Esto ocurre en particular en aquellas muestras que no son **muestras simples al azar**, sino **muestras complejas al azar**, en las cuales los diferentes casos incluidos en la muestra han sido seleccionados con **diferentes probabilidades de selección.** Esta situación surge por ejemplo cuando se tienen muestras **estratificadas** en las cuales haya una diferente probabilidad de selección en los distintos estratos. Casi todas las encuestas de hogares tienen esa característica. En el caso de la Encuesta Permanente de Hogares de la Argentina, la muestra de cada ciudad tiene una fracción de muestreo diferente, y en algunas ciudades cada zona o sector tiene su propia fracción de muestreo. Además, una vez seleccionadas las unidades primarias de muestreo (por ejemplo ciertos bloques o manzanas de la ciudad), en las muestras complejas hay una segunda instancia en la cual se eligen ciertos hogares o viviendas dentro de cada

bloque o manzana previamente seleccionado, y nuevamente aquí puede haber diferentes fracciones de muestreo según de qué bloque o manzana se trate. En definitiva, en el caso de la EPH argentina y de otras encuestas parecidas, hay una multiplicidad de factores de ponderación según de qué ciudad o unidad menor de muestreo se trate. Cada hogar elegido de la zona A puede estar "representando" a 1000 hogares mientras cada hogar en la muestra de la zona B puede estar representando a 500, de modo que para obtener una estimación insesgada se debe dar a los hogares B un peso equivalente a la mitad del peso de los hogares A.

Esta situación no es muy problemática en los análisis de corte transversal, cuando se examina la encuesta realizada en una fecha u onda determinada. Pero surge un problema cuando se trata de un panel, porque **las ponderaciones de un mismo hogar pueden variar de una onda a la siguiente**. El mismo hogar que representaba 1000 hogares en abril puede representar 1200 en octubre. Si los datos de ambas ondas se consideran conjuntamente, ¿cuál factor de ponderación se deberá usar? Evidentemente no se pueden usar ambos al mismo tiempo: si se cruza una variable de abril con una de octubre, cada caso que figure en la tabla debe ser multiplicado por un factor de ponderación, y la tabla reflejará el resultado de esa multiplicación, que no puede ser de 1000 y de 1200 al mismo tiempo. Puede ser 1000, o 1200, o el promedio de ambos, o alguna otra solución semejante. La elección de la solución apropiada exige una reflexión más detenida.

En los factores de ponderación de este tipo hay en realidad dos aspectos conceptualmente separables. Por una parte, esos factores son factores de **expansión**, que transforman la **escala** o tamaño de los resultados. Si cada caso se multiplica por 1000, por 500 o por la cifra que corresponda, el total ponderado resultará ser quizá de varios millones de personas, cuando la muestra es solo de unos miles. Pero además del efecto expansivo, las ponderaciones tienen un efecto **proporcional**: cuando las ponderaciones de los distintos sujetos son diferentes entre sí, las ponderaciones influyen en el **peso relativo** de los distintos casos de la muestra. En una muestra simple de 1000 para una población de un millón, cada caso representa a 1000 miembros de la población. En muestras complejas un caso representa a 500, otro a 1000, otro a 1200, etc.: cada uno tiene un "peso" distinto.

Puede haber **expansión sin ponderación relativa**, por ejemplo, cuando a todos los casos se les aplica la misma ponderación. Puede haber **ponderación relativa sin expansión** si todos los factores anteriores se dividen por la población total y se multiplican por el tamaño de la muestra, de modo que el total ponderado sea igual al tamaño de la muestra, pero donde algunos individuos pesarán, por ejemplo, 0.80 mientras otros pesarán 1.20.

Las ponderaciones de una onda pueden ser distintas a las de la onda anterior por dos clases de razones: por una parte, por un cambio **absoluto** en la población, y por otra parte, por un cambio en el **peso relativo** de sus diferentes grupos componentes. Por ejemplo, si la población crece 1% por semestre, la muestra de abril corresponde a una población de 1.000.000 y en octubre a una población de 1.010.000. Completamente aparte de este crecimiento global, pueden haber cambiado las ponderaciones relativas de los casos: dos hogares que tenían el mismo peso en abril pueden aparecer con diferente peso en octubre, debido a cambios en la composición de la muestra, independientemente del aumento de la población total estimada. Esto puede suceder, por ejemplo, si uno de los hogares fue elegido en un bloque de 20 hogares en abril, y en un bloque de 25 hogares en octubre, debido a cambios en la cantidad de hogares contabilizados en el bloque.

Ante este panorama, una solución simple cuando se consideran dos ondas consiste en tomar la **ponderación promedio de ambas ondas**. De este modo, la población total resultante sería 1.005.000, correspondiente a la población media del período, y el peso relativo de los casos también aparecería en valores intermedios entre los de mayo y los de octubre: si los pesos de dos hogares estaban entre sí en una relación 2:1 en abril, y en una relación 3:1 en octubre, aparecerán con una relación 2,5:1 en las ponderaciones promedio.

En las secuencias largas de panel, por ejemplo cuando se consideran tres o cuatro ondas simultáneamente, se aplica el mismo principio. Si se consideran cuatro ondas, por las mismas razones antes mencionadas, se debería usar el valor promedio de las cuatro ponderaciones, que remite a la población promedio de esas cuatro ondas (la cual, si las encuestas se realizan en abril y octubre de

dos años consecutivos, correspondería a la población esperada en el mes de enero del segundo año, punto medio de las cuatro fechas). La relación de pesos relativos de dos casos cualesquiera A y B será la relación entre el peso promedio del caso A y el peso promedio del caso B.

Estos ajustes en general tienen poca importancia práctica, porque el peso de cada caso por lo general no varía sustancialmente de una onda a la siguiente, y tampoco la población total estimada experimenta cambios muy significativos, pues la tasa de crecimiento de la población por semestre es bastante baja. En el caso de la Argentina el crecimiento de la población como máximo suele ser del orden del 2.5% anual (alrededor de 1.25% por semestre) en ciudades que reciben flujos inmigratorios fuertes, y oscila en un promedio de 1.5% anual (0.75% por semestre) para el conjunto de zonas urbanas. Esto significa que el uso de las ponderaciones de abril, de octubre o del promedio de ambas no afecta significativamente ningún resultado.

En los párrafos y ejemplos precedentes se ha usado por simplicidad la **media aritmética** de las ponderaciones correspondientes a diferentes ondas de la encuesta. Para ser conceptualmente precisos, sin embargo, se debe tener en cuenta que la población crece de manera exponencial, y por ello correspondería más bien usar la **media geométrica** de las ponderaciones. La media geométrica es siempre un poco inferior a la media aritmética, y supone crecimiento curvilíneo en lugar de rectilíneo entre las fechas consideradas. Sin embargo, cuando los intervalos entre ondas son relativamente breves (un semestre o un año) ambos tipos de promediación arrojarán resultados muy similares.

Por ejemplo, si la tasa anual de crecimiento es 2%, la interpolación exponencial para un semestre sería de 0.995% mientras que la interpolación lineal sería de 1.0%. En una población de un millón de habitantes, la interpolación lineal supone un aumento de 10.000 habitantes por semestre, mientras la exponencial equivale a un crecimiento de 9.950 personas en el primer caso, y de 10.050 en el segundo, es decir 50 personas de diferencia (0.5%) por semestre, que es un margen de error muy pequeño; ese error fácilmente podría ser estadísticamente no significativo, pues la estimación general de un millón de habitantes (que suele ser una proyección a partir del último censo) podría tener ella misma un margen de error mjuy superior, del orden del 2% o 3% (unas 30.000 personas), error que se haría más pronunciado cuanto más alejado en el tiempo esté el último censo. Por ello la diferencia entre interpolaciones lineales y exponenciales, o entre medias aritméticas y geométricas, carece de importancia práctica en el caso de paneles que involucran intervalos relativamente breves y tasas de crecimiento de magnitud pequeña o moderada. Dado que las cifras de población son intrínsecamente imprecisas, estas pequeñas diferencias no son dignas de mucha consideración pues están dentro del margen de error propio de las cifras demográficas. En una proyección de largo plazo, sin embargo, se volverían más importantes.

Una situación un poco distinta a la anterior se presenta cuando se superponen datos de distintas cohortes y se los considera simultáneamente (lo cual sólo tiene sentido para aquellos fenómenos que no presentan variaciones estacionales o cíclicas). Si hay varias cohortes de casos con una secuencia de cuatro entrevistas, que han ingresado en sucesivos trimestres, y todos ellos son considerados como "casos" de una base de datos en que se reúnan cohortes de diferentes períodos, la muestra total estará formada por grupos de casos que ingresaron en la encuesta en diferentes fechas. Si son cuatro cohortes (todas las ingresadas en un cierto año, a razón de una por trimestre, como en la EPH continua de la Argentina) cada cohorte representará aproximadamente un 25% de la muestra total de una determinada onda. Por otro lado, cada uno de los contingentes permanece por seis trimestres, y es entrevistado en los trimestres 1, 2, 5 y 6 a partir de su entrada en la muestra.

Cada una de las cohortes sería ponderada por la ponderación promedio de sus cuatro ondas, que corresponde a la población del punto medio de sus cuatro apariciones. Si un conjunto de hogares fueron incorporados a la muestra en el primer trimestre de 2010, y no desaparecieron precozmente de la muestra, entonces ellos fueron entrevistados en el primer y tercer trimestre de 2010, y en iguales trimestres de 2011. Sus ponderaciones deberían ser el promedio de las ponderaciones que tuvieron en esas cuatro ondas, como se explicó antes, y la población total resultante de aplicar a esa cohorte dichas ponderaciones sería la población que esos hogares representaban al final del *tercer trimestre* de 2010, que es el punto medio del período que va desde enero de 2010 a junio de 2011. Por

otro lado, cada cohorte que ingresa en la muestra de la EPH constituye solo un 25% de la muestra de cada trimestre en que sea encuestada, y por ello representa solo el 25% de la población. Por ello esa cohorte expandida resultaría así equivalente (aproximadamente) al 25% de la población al 30 de septiembre de 2011. Otras cohortes deberán ser expandidas a la población de otras fechas, y en conjunto se expandirán a la población de la fecha que resulte estar en el punto medio de todas.

En resumen, los datos obtenidos de una cohorte a lo largo de sus varias apariciones deben ponderarse con un peso equivalente al **promedio de las ponderaciones que cada familia tuvo en sus distintas apariciones**. Si cada familia aparece k veces, la aplicación de esas ponderaciones estimaría una proporción $1/k$ de la población correspondiente al punto medio del período delimitado por todas las apariciones de esas familias en la muestra.

Análisis de supervivencia de cohortes teóricas. Se discutió en una sección anterior el concepto de "panel virtual" o "panel simulado" conformado por una cohorte teórica de sujetos. Estos sujetos tienen diferentes edades, o diferentes niveles de "antigüedad" en la variable relevante, pero son observados todos ellos al mismo tiempo. Por ejemplo, se calcula la tasa de abandono escolar en el año 2010 para todos los escolares de diferentes edades (o que asisten a diferentes grados en el sistema escolar). Luego se postula una población teórica de 100.000 niños ingresantes a la escuela, sometidos a lo largo de los años subsiguientes (por ejemplo de los 5 a los 18 años) a las mismas tasas de abandono escolar observadas en 2010, aun cuando estas tasas no afectaban a una misma cohorte sino a niños de diferentes edades, es decir pertenecientes en realidad a diferentes cohortes. Así como se calcula la probabilidad de abandono escolar para los miembros de esa cohorte teórica, y los riesgos diferenciales según distintos factores, también pueden definirse cohortes teóricas de otro tipo: de niños que nacen, de parejan que se casan, de personas que ingresan a un empleo, para estudiar la probabilidad de que les ocurra algún evento en algún momento futuro (morir, divorciarse, perder el empleo).

Supongamos que se poseen datos sobre los divorcios ocurridos en un año determinado, y se tienen diversos datos sobre cada pareja y sobre los cónyuges individualmente. Dado que además del dato específico (existencia del divorcio, antigüedad del matrimonio) se pueden conocer otros datos (edad, nivel socioeconómico, condición de empleo de los cónyuges, número de hijos, edad de los hijos, etc.), se pueden identificar los factores de riesgo que incrementan o reducen la probabilidad de divorcio **para los miembros de la cohorte teórica**. Esto es aplicable a la población real, pero con ciertos recaudos y limitaciones, lo mismo que la expectativa de vida en las tablas demográficas o actuariales. Un matrimonio celebrado en 2012 no tiene por qué estar sometido durante los años subsiguientes a las tasas de divorcio observadas en 2010 para matrimonios celebrados en diferentes años del pasado. Las cohortes teóricas proporcionan indicadores de resumen sobre los riesgos (por ejemplo la expectativa de vida, o la duración media de los matrimonios), pero su poder predictivo depende de la estabilidad futura de las tasas subyacentes de mortalidad, de divorcio, o lo que fuese.

10. Incertidumbre estadística

En toda la precedente exposición se ignoraron los problemas derivados de la estimación estadística. Cada coeficiente que se estime (sean los porcentajes de una tabla, las tasas de transición, o los coeficientes de la regresión de Cox) es estimado sobre una muestra, y está afectado por un margen de **error aleatorio**. Aun cuando la base de datos sea un censo, y no una muestra convencional, los datos del censo pueden considerarse como sujetos a errores aleatorios de medición. Esos errores, si son aleatorios, seguirán una distribución normal alrededor del dato estimado por el censo, exactamente igual que en una muestra, ya que el censo realizado es sólo uno de los muchos censos posibles de ese año: pudo haberse hecho el día anterior o el día siguiente, cada hogar pudo haber sido visitado por otro agente censal, las personas que ese día estaban ausentes podrían haber estado presentes en algún otro día de la misma semana, los agentes censales pueden haver cometido errores aleatorios al registrar los datos en el formulario censal (por ejemplo las edades), los centros de cómputo también pueden haber introducido errores al ingresar los datos en las computadoras (por ingreso manual o por lectura óptica), y así sucesivamente. Todos estos son errores aleatorios que requie-

ren un tratamiento estadístico como los datos de una muestra, si bien el tamaño del censo frecuentemente determina que tales márgenes de error aleatorio sean muy pequeños.

Hay además **errores no aleatorios**. Por ejemplo los hogares omitidos o que rechazan las entrevistas podrían no ser similares al promedio de hogares, y por ello su exclusión originaría un sesgo. En las encuestas las ausencias o rechazos son reemplazados con otros casos de la misma zona, pero eso no garantiza que no haya sesgo; por ejemplo, típicamente los hogares más ricos suelen estar subrepresentados en las encuestas, lo mismo que algunos muy marginales (como los hogares sin techo). Usualmente los censos realizan verificaciones posteriores de la omisión censal, y en algunos casos consiguen detectar algunos de estos sesgos, pero habitualmente ello es imposible por lo cual la omisión censal **se suele considerar como aleatoria**: si se verifica un 3% de omisión, todos los resultados se multiplican por 1.03 para estimar el tamaño total correspondiente, como si los omitidos fuesen equivalentes, en promedio, al resto de la población.

El problema con los datos de panel es que los errores no suelen cumplir con una de las características esenciales de los errores aleatorios ocurridos en muestras al azar: los casos u observaciones del panel no son independientes entre sí. Esto se debe a la existencia de observaciones repetidas de los mismos sujetos, de modo que la muestra de observaciones es una muestra jerárquica por conglomerados (*clusters*): aparte de otras características de la muestra (que puede ser estratificada, conglomerada por zonas, etc.) al nivel inferior se seleccionan n individuos que son observados k veces. Este conjunto de k observaciones constituye un conglomerado de observaciones del mismo sujeto, y por supuesto esas k observaciones de cada sujeto están correlacionadas entre sí. La existencia de datos autoconglomerados genera un efecto aleatorio *propio de cada sujeto*, independientemente del error aleatorio general de la muestra de sujetos.

El problema de la estimación de márgenes de error en esquemas muestrales jerárquicos es bastante complicado. Usualmente se aplican modelos multi-nivel donde se distingue entre el error de muestreo generado por la selección de conglomerados (en este caso, la selección de sujetos) y el error generado dentro de cada conglomerado (por la selección de ocasiones y por la autocorrelación de estas observaciones para cada sujeto).

En el caso de la regresión de Cox u otros modelos de supervivencia el tema se simplifica, pues en realidad el "conglomerado de ocasiones" no es exactamente una muestra aleatoria de ocasiones, sino una *secuencia* de observaciones. Para cada individuo lo que importa no es la colección de datos obtenidos, sino la *trayectoria* observada a través de las sucesivas observaciones. Estrictamente hablando, hay una sola trayectoria para cada sujeto, de modo que (desde ese punto de vista) no se genera un conglomerado de observaciones para cada individuo (como ocurriría, por ejemplo, si al mismo sujeto se le mide la tensión arterial en diferentes ocasiones). En el análisis de supervivencia, cada sujeto en realidad genera un solo dato, que es el "tiempo transcurrido hasta el evento". Esa variable puede dar un valor específico, en caso que el evento ocurra después de un tiempo determinado, o bien el sujeto puede terminar el estudio o retirarse del mismo sin que le haya ocurrido aún el evento, en cuyo caso aparece como "censurado". En cualquier caso, se genera solo un dato por individuo, de modo que no se presenta el caso de observaciones autocorrelacionadas que da origen a los "modelos multi-nivel con efectos aleatorios".[32]

Los textos corrientes sobre análisis de supervivencia, anteriormente citados, contienen capítulos específicos sobre las diferentes pruebas estadísticas que pueden usarse para estimar el margen de error de las estimaciones. Dado que la regresión de Cox es un procedimiento *no paramétrico*, sus errores de estimación no puede suponerse que siguen una distribución normal. Más aún, las estimaciones de error en las regresiones y otros procedimientos análogos se basan en la magnitud de los "residuos" o diferencias entre los valores predichos y los valores observados. Dado que en el análisis se supervivencia no se estima un valor observable sino una probabilidad de supervivencia, y la estimación no se realiza mediante el procedimiento de mínimos cuadrados sino por métodos

[32] Sobre modelos multinivel con efectos aleatorios, véanse entre otros textos los de Snijders & Bosker 2012, y Kreft & De Leeuw 1999.

iterativos de "máxima verosimilitud", no es obvio a qué se denomina un "residuo" y qué es lo que hay que evaluar para saber si la estimación es "estadísticamente significativa". Ello se refuerza por el hecho de que generalmente hay casos "censurados" cuyo destino efectivo (la ocurrencia o no ocurrencia del efecto) se ignora.

La conclusión es que "no hay una analogía obvia con el residuo usual, 'observado menos predicho', que se usa en otros modelos regresión" (Hosmer & Lemeshow 1999:198). Esto lleva, en palabras de los mismos autores, a desarrollar "varios diferentes [tipos de] residuos que juegan un papel importante para examinar algún aspecto referente al ajuste del modelo de riesgos proporcionales" (*ibídem*). Algunos procedimientos tienden a evaluar los márgenes de error de las tasas de ocurrencia o de las probabilidades de supervivencia; otras pruebas tratan de comprobar si es estadísticamente válido el supuesto de riesgos proporcionales; otras pruebas estadísticas tratan de establecer el margen de error en la estimación de los coeficientes. En muestras relativamente pequeñas también es importante evaluar la influencia de casos individuales (especialmente los *outliers* que tienen valores "extremos") sobre los resultados obtenidos (esto se analiza comparando los resultados obtenidos incluyendo o no esos sujetos). Los principales paquetes de *software* que incluyen la regresión de Cox ofrecen los principales *tests* estadísticos aplicables.

Una observación final importante es que las estimaciones sobre paneles son siempre afirmaciones probabilísticas, que estrictamente no se refieren a individuos sino a grupos o poblaciones. Si hay un 20% de probabilidad de que una persona ocupada pierda su empleo, de ello no se puede inferir nada específico sobre cada persona individual: dentro de ese mismo grupo habrá un 80% que conservará su empleo, y 20% que lo perderá, pero (en ausencia de otros datos) no se puede saber quién, concretamente, se quedará desocupado. Aun teniendo otros datos (por ejemplo el tipo de empleo y la antigüedad) se puede refinar la estimación de la probabilidad, pero ésta siempre se referirá al **grupo** de personas que comparten esas características; por ejemplo, las personas con menos de un año de antigüedad pueden tener un mayor riesgo de perder el empleo que las personas con 10 o más años de antigüedad, pero dentro de cada uno de esos grupos no es posible identificar las personas individuales que perderán el empleo. Se puede, del mismo modo, prever que un 3% de los discos rígidos fallarán antes de un año, pero no se puede saber de antemano **cuáles** son los discos que fallarán. **Las predicciones probabilísticas no se refieren a sujetos individuales**.

Anexo 1 – Vectores y matrices

Vectores. Los vectores consisten en una fila o columna de números, con dimensiones ***1 × n*** (en el caso de un vector fila con **n** componentes) o bien ***m × 1*** (en el caso de un vector columna con **m** componentes).

Vector fila: $a = [a_1 \quad a_2 \quad a_3]$ Vector columna: $b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$

Transposición de vectores. Los vectores pueden ser **transpuestos**, convirtiendo así un vector fila en un vector columna y viceversa. Un vector transpuesto se indica con un apóstrofo. Así ***b′*** es el vector ***b*** transpuesto. Si ***b*** era un vector columna, ***b′*** será un vector fila, y viceversa.

Matrices. Los vectores pueden ser vistos como una cuadrícula con **una** fila y **varias** columnas, o **una** columna y **varias** filas. Esto significa que los vectores son un caso especial de unas cuadrículas más generales, las matrices, que pueden tener cualquier número de filas y de columnas. Las matrices son disposiciones con ***n*** filas y ***m*** columnas, y por ello se denominan matrices de dimensión ***n × m***. Un vector o matriz con sólo una fila y sólo una columna se denomina un **escalar**, y es un simple número. Una matriz con igual número de filas que de columnas (***n × n***) se denomina **matriz cuadrada**. Una matriz podría considerarse como varios vectoresfila colocados uno debajo del otro, o como varios vectores columna colocados uno al lado del otro. Un vector fila puede ser considerado como una matriz de dimensión ***1 × m***, y un vector columna como una matriz de dimensión ***n × 1***. El siguiente es un ejemplo de una matriz ***C*** con dos filas y tres columnas. Sus elementos c_{ij} se denotan con dos subíndices: el primero indica la fila, y el segundo la columna.

$$C = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \end{bmatrix}$$

Una matriz puede **transponerse** intercambiando filas con columnas. En la transposición de una matriz como la que antecede, la primera fila de ***C*** pasa a ser la primera columna de ***C′*** y del mismo modo el resto de las filas y columnas.

$$C' = \begin{bmatrix} c_{11} & c_{21} \\ c_{12} & c_{22} \\ c_{13} & c_{23} \end{bmatrix}$$

Suma y resta de vectores y matrices. Los vectores y matrices pueden ser sumados, restados y multiplicados entre sí. La suma o resta de dos vectores o de dos matrices equivale a un vector o matriz cuyos elementos son la suma o la diferencia de los elementos correspondientes de los sumandos. Por lo tanto, los vectores o matrices que se suman o restan deben tener exactamente la misma dimensión para que esta operación sea posible. En el siguiente ejemplo se suman dos matrices de igual dimensión. Los elementos de la suma son simplemente la suma de los elementos correspondientes de los sumandos.

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{bmatrix} \qquad \text{(Ec. A.1)}$$

Multiplicación de vectores. El producto de dos vectores es la suma de los productos de cada término de uno con cada término del otro. Para esto se requiere que ambos vectores sean de la misma dimensión, es decir que tengan el mismo número de elementos. El resultado será distinto según el orden en que estén dispuestos los vectores, y según se trate de vectores fila o vectores columna. En el mundo de los vectores y matrices **el orden de los factores altera el producto**. En general, para multiplicar dos vectores o matrices se multiplica **cada fila del primer factor por cada columna**

del segundo. Por ello cuando se trata de vectores, que tienen una sola fila o una sola columna, se requiere que ambos tengan el mismo número de elementos. Consideremos primero el producto de un vector fila por un vector columna.

$$\begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix} \times \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3 + a_4 b_4 = \sum a_i b_j \qquad \text{(Ec. A.2)}$$

En este caso, obviamente, el primer vector tiene una sola fila, y el segundo una sola columna. Se procede a multiplicar la fila por la columna. La operación consiste en multiplicar cada elemento de la fila de la izquierda por el respectivo elemento de la columna de la derecha, y sumar todos esos productos. Dado que hay una sola fila en el vector de la izquierda, y una sola columna en el vector de la derecha, el resultado es una sola suma de productos, es decir un número (un **escalar**).

En el caso inverso las cosas son algo diferentes. Cuando el primer vector es un vector columna, y el segundo un vector fila, se sigue aplicando la regla de multiplicar las filas del primero por las columnas del segundo. Un vector columna tiene ***n*** filas con un solo elemento cada una, y un vector fila tiene una sola fila con **m** elementos (es decir **m** columnas). Cuando se multiplica cada fila del primero por cada columna del segundo, cada par de elementos origina un producto diferente. Al repetir la operación con todas las filas del vector de la izquierda, aplicadas a cada columna del vector de la derecha, esta operación origina una matriz.

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} \times \begin{bmatrix} b_1 & b_2 & b_3 & b_4 \end{bmatrix} = \begin{bmatrix} a_1 b_1 & a_1 b_2 & a_1 b_3 & a_1 b_4 \\ a_2 b_1 & a_2 b_2 & a_2 b_3 & a_2 b_4 \\ a_3 b_1 & a_3 b_2 & a_3 b_3 & a_3 b_4 \\ a_4 b_1 & a_4 b_2 & a_4 b_3 & a_4 b_4 \end{bmatrix} \qquad \text{(Ec. A.3)}$$

Para que la primera operación (ecuación A.2) fuese posible se requería que el número de filas del primer vector (una sola) fuese igual al número de columnas del segundo. El resultado tiene como dimensión el número de filas del primero y el número de columnas del segundo, que en el caso de la ecuación A.2 es uno por uno, es decir un escalar (que puede ser considerado como una matriz con una sola fila y una sola columna). La regla general sobre este punto se puede indicar del modo siguiente:

Dimensión de los vectores	Dimensión del producto
(1 × m)(m × 1)	1 × 1 (escalar)
(1 × m)(n × 1) siendo n ≠ m	Imposible
(m × 1)(1 × n)	m × n (matriz)
(1 × m)(1 × n)	Imposible
(m × 1)(n × 1)	Imposible

En general, el número de columnas del primero debe ser igual al número de filas del segundo, y el producto tiene como dimensiones el número de filas del primero y el número de columnas del segundo. Por esa razón **es imposible multiplicar entre sí dos vectores fila**, y también **es imposible multiplicar entre sí dos vectores columna** (salvo en el caso trivial de dos vectores fila o dos vectores columna **de orden uno**, que serían escalares).

Multiplicación de matrices. En el caso de la multiplicación de dos matrices se procede del mismo modo. La condición necesaria para que la operación sea posible es **que el número de columnas de la matriz de la izquierda sea igual al número de filas de la matriz situada a la derecha**. Una matriz de ***n*** filas por ***m*** columnas se puede multiplicar por otra matriz de ***m*** columnas por ***p*** filas. La matriz resultante tendrá la misma cantidad de filas que la matriz de la izquierda (***n***), y la cantidad de columnas de la matriz de la derecha (***p***).

$$[n \times m] * [m \times p] = [n \times p]$$

La dimensión común ***m*** (el número de columnas de la primera, que debe ser igual al número de filas de la otra) **desaparece en el resultado**, que es una matriz que tiene las otras dos dimensiones (el número de filas de la primera, y el número de columnas de la segunda).

Cada celdilla de la matriz resultante puede considerarse como el producto de una fila de la primera por una columna de la segunda, como ocurría con los vectores en la ecuación A.2:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{bmatrix} = \begin{bmatrix} \sum_j a_{1j}b_{j1} & \sum_j a_{1j}b_{j2} & \sum_j a_{1j}b_{j3} & \sum_j a_{1j}b_{j4} \\ \sum_j a_{2j}b_{j1} & \sum_j a_{2j}b_{j2} & \sum_j a_{2j}b_{j3} & \sum_j a_{2j}b_4 \end{bmatrix} \quad \text{(Ec.A.4)}$$

Con las probabilidades de transición y las proporciones de estado en un modelo de Markov, todos los vectores son del mismo tamaño y todas las matrices son **cuadradas** (igual número de filas y de columnas). Si hay ***k*** estados posibles, la matriz ***R*** tendrá ***k*** filas y ***k*** columnas, y los vectores de probabilidades o proporciones de estado tendrán siempre ***k*** elementos, dispuestos como fila o como columna según convenga. Una matriz cuadrada de ***k*** filas y ***k*** columnas, o un vector de ***k*** elementos, se denomina "de orden ***k***".

Elevar al cuadrado una matriz no es lo mismo que elevar al cuadrado cada uno de sus elementos. Sólo las matrices cuadradas pueden ser elevadas al cuadrado. El cuadrado de una matriz cuadrada **T** es un producto matricial **T x T** que implica multiplicar cada fila de **T** por la respectiva columna. Cada uno de esos productos de la fila ***i*** por la columna ***j*** es un escalar o número, que será el elemento c_{ij} en la matriz resultante. El cuadrado de una matriz cuadrada **T** es otra matriz cuadrada $\mathbf{T^2}$. El elemento c_{ij} de $\mathbf{T^2}$ es el producto de la fila ***i*** por la columna ***j*** de la matriz **T**. Por ejemplo, si **T** tiene tres filas y tres columnas, el elemento c_{23} de $\mathbf{T^2}$ será el producto de la segunda fila por la tercera columna: $c_{23} = c_{21}c_{13} + c_{22}c_{23} + c_{23}c_{33}$. Nótese que elevar la matriz **T** al cuadrado, **T x T**, no es lo mismo que multiplicarla por su matriz transpuesta, **T x T'.** En este último caso, cada fila de **T** es igual a la respectiva **columna** de **T'**.

La matriz identidad. Una matriz cuadrada muy usada es la **matriz identidad**, denotada por **I**. Esta matriz tiene **1** en las celdillas de la diagonal principal, y **0** en las demás celdillas. La siguiente es una matriz identidad de dimensión 4 (cuatro filas y cuatro columnas):

$$I_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

La matriz inversa. A partir de la matriz identidad se define también una importante matriz, la matriz **inversa** de una matriz **cuadrada**, denotada como $\boldsymbol{T^{-1}}$. Dada una matriz cuadrada ***T*** de orden ***k***, es decir una matriz con ***k*** filas y ***k*** columnas, su matriz inversa $\boldsymbol{T^{-1}}$ es aquella matriz del mismo orden que, multiplicada por ***T*,** arroja como resultado la matriz identidad:

$$\boldsymbol{TT^{-1} = I}$$

Esta noción de matriz inversa es intuitivamente similar a la inversa de un número, es decir su recíproca, como 4 y 1/4, que podría expresarse como 4×4^{-1}: en efecto, la inversa de un número tiene la misma propiedad: $4 \times 4^{-1} = 4/4=1$. Un número multiplicado por su inversa es igual a 1. Una matriz cuadrada multiplicada por su inversa es igual a la matriz identidad.

Forma matricial de un sistema de ecuaciones. Supóngase que se tiene un sistema de ecuaciones lineales como el siguiente, que tiene tres ecuaciones y tres variables (las X):

$$a_{11}X_1+a_{12}X_2+a_{13}X_3=b_1$$
$$a_{21}X_1+a_{22}X_2+a_{23}X_3=b_2$$
$$a_{31}X_1+a_{32}X_2+a_{33}X_3=b_3$$

Este sistema puede expresarse de manera más compacta en forma de vectores y matrices:

$$\boldsymbol{Ax = b}$$

$\boldsymbol{A}$ es la matriz cuadrada de coeficientes a_{ij} , mientras que $\boldsymbol{x}$ es un vector columna cuyos elementos son X_1, X_2 y X_3; $\boldsymbol{b}$ es un vector columna con los tres coeficientes libres b_1, b_2 y b_3**.** Los valores de los coeficientes a y b se suponen conocidos. Las incógnitas del sistema son los valores de las X que satisfacen las tres igualdades. La solución del sistema de ecuaciones consiste en despejar el vector $\boldsymbol{x}$ para lo cual se usa la matriz inversa $\boldsymbol{A^{-1}}$ en la forma:

$$\boldsymbol{x = A^{-1}b}$$

Si en vez de usar los vectores columna $\boldsymbol{x}$ y $\boldsymbol{b}$ se utilizaran sus transpuestos $\boldsymbol{x'}$ y $\boldsymbol{b'}$ que son vectores fila, la formulación del sistema en forma matricial sería: $\boldsymbol{x'A = b'}$ y la solución sería $\boldsymbol{x'=b'A^{-1}}$. Esta transposición no altera para nada los términos del problema ni la solución del mismo (excepto que $\boldsymbol{x'}$ es **pre**-multiplicada por $\boldsymbol{A}$ en la forma $\boldsymbol{x'A}$, en tanto que $\boldsymbol{x}$ era **post**-multiplicada en la forma $\boldsymbol{Ax}$. En cualquier caso, para que el sistema tenga solución se requiere que el número de incógnitas sea igual al número de ecuaciones, es decir que $\boldsymbol{A}$ sea una matriz cuadrada del mismo orden que $\boldsymbol{x}$ y que $\boldsymbol{b}$, y además se requiere que las ecuaciones sean independientes entre sí, esto es, que ninguna de ellas pueda obtenerse mediante una combinación de las otras. Esto equivale a requerir que el determinante de la matriz $\boldsymbol{A}$ no sea igual a cero.

Cuando un sistema de ecuaciones tiene la forma precedentemente mostrada, donde cada combinación de variables es igual a un valor b_i, los procedimientos de solución son capaces de calcular el valor absoluto de cada incógnitga, es decir el valor de cada variable X_j. En cambio, si cada combinación de las X en el sistema de ecuaciones es igual a cero en vez de ser igual a b_i, el sistema de ecuaciones se llama *homogéneo*. Estos sistemas resultan ser válidos también cuando se multiplican todas las ecuaciones por una constante cualquiera. Si $b_1X_1+b_2X_2=0$, también se cumplirá $k[b_1X_1+b_2X_2]=0$ para cualquier número k, y por lo tanto si x_1 y x_2 son soluciones de la ecuación, también lo serán kx_1 y kx_2. El valor absoluto de las X estaría indeterminado. Esto significa que las ecuaciones, como tales, no se pueden resolver en la misma forma. Sin embargo, en *todas* las soluciones con *cualquier* número k se mantiene la *proporción* entre X_1 y X_2. Se pueden entonces calcular las **razones o cocientes** entre las soluciones (por ejemplo x_2/x_1, pero no el valor absoluto de cada una de ellas, porque cualquier otro par de valores kx_1 y kx_2 también sería una solución. En esos casos, solo se obtienen soluciones *relativas*, tomando una de las variables como punto de referencia, es decir como unidad. Si se asume arbitrariamente que $X_1=1$, entonces el cociente X_2/X_1 sería igual a X_2, pero esta variable X_2 no estaría medida en sus propias unidades de medida sino en cantidades de X_1. Si las X son cantidades de mercancías y los coeficientes son sus precios, solo se puede obtener el precio *relativo* de X_2 respecto a X_1. Por ejemplo el precio de un kg de naranjas no se expresaría en unidades monetarias, sino solo en una cantidad equivalente de manzanas.

Anexo 2 – Datos de panel en SPSS

Si se tienen las bases de datos de varias rondas de un panel, por ejemplo una encuesta de hogares, para poder formar una base de datos de panel se requiere que cada caso individual (cada hogar y/o cada persona) estén identificadas por el mismo código en todas las rondas. En una base de datos de hogares habrá así (en cada ronda) una variable **identificatoria del hogar** (que puede tener cualquier nombre, por ejemplo NUMHOGAR) que identifica a un determinado hogar. Aparte habrá una serie de variables que contienen características de ese hogar (tipo de paredes o pisos, tamaño del hogar, acceso a electricidad, etc.). En una base de personas puede haber dos variables de identificación: la variable identificatoria del hogar (que figura también en el archivo de hogares) y la variable **identificatoria de la persona**. Esta última puede ser simplemente el número de orden de la persona dentro del hogar (1 para el primer miembro que ordinariamente es el jefe, 2 para el segundo miembro, 3 para tercero, etc.).

Unir o fusionar las matrices de datos de varias rondas equivale a "añadir variables": para cada hogar, por ejemplo, el nuevo archivo contendrá todas las variables de la primera ronda, más todas las variables de la segunda ronda, y así sucesivamente. Esto genera un problema, ya que las variables usualmente tienen el mismo nombre en todas las rondas, y en un archivo no puede haber dos o más variables con el mismo nombre. Por ello, lo mejor es añadirle un sufijo al nombre de cada variable, para indicar la ronda. Supongamos una encuesta de hogares cuyas rondas representan un trimestre; el primer trimestre sería enero-marzo, el segundo abril-junio, y así sucesivamente. Hará falta indicar el año y el trimestre, por ejemplo 2011_1, 2011_2, etc.

En el archivo fusionado las variables **de cada ronda** deben tener en su nombre el sufijo correspondiente. Todas las variables medidas en la ronda del primer trimestre de 2011 tendrán el sufijo 2011_1. Así una variable P28 (pregunta 28) pasará a llamarse P28_2011_1, o bien más brevemente P28_111. La misma variable en el archivo del tercer trimestre de 2011 se llamaría P28_113, y de modo similar las demás. Así la fusión de los archivos de rondas sucesivas comprende dos fases:

1. Añadir un **sufijo de ronda** a cada variable en el archivo de cada ronda (**excepto las variables identificatorias de hogar y/o persona** que deben tener igual nombre en todas las rondas (y cada hogar y persona el mismo valor en todas las rondas).
2. Fusionar los archivos, usando las variables identificatorias como clave de unión.

Añadir sufijos de ronda a las variables

La forma más simple de añadir un sufijo a un conjunto de variables es a través de un comando RENAME VARIABLES donde se enumeran todas las variables involucradas. Suponiendo con fines ilustrativos que haya solo cuatro variables (aparte de la identificación del caso), ese comando sería como el siguiente en la primera ronda de 2011:

```
GET FILE 'encuesta 2011-1.sav'.
RENAME VARIABLES
        (P1 P2 P3 P4 = P1_111 P2_111 P3_111 P4_111).
GET FILE 'hogares 2011-1 renamed.sav'.
```

Esta operación se repite (con las modificaciones del caso) en otros períodos, por ejemplo 2011-2, 2011-3, etc. El comando RENAME VARIABLES tiene que incluir entre paréntesis la lista de variables originales, el signo igual, y la lista de los nuevos nombres de las variables; naturalmente, las dos listas tienen que tener el mismo número de variables, ordenadas del mismo modo.

Si son muchas variables, el comando puede prepararse primero en Word o Excel, copiando la lista de variables con sus nombres originales en una columna, y los nuevos nombres en la segunda columna. Para esta columna de nuevos nombres se puede usar una fórmula de Excel para unir cadenas de caracteres, donde se une el nombre que aparece en la primera columna (por ejemplo P5) con un sufijo escrito en algún otro sitio de la planilla (por ejemplo _111). La función de Excel que une cadenas de caracteres es CONCATENAR(texto1,texto2), donde se debe reemplazar los parámetros con la ubicación de los dos textos que se quiere unir. Por ejemplo, si los nombres originales están en la columna A, en una de las celdillas de la columna B se escribe la fórmula:

	A	B	C
1	PREG01	=CONCATENAR(A1, C1)	_111
2	PREG02		

Si en A1 decía PREG01, y en C1 estaba el sufijo _111, el resultado será PREG01_111. La referencia al sufijo va en la forma C1 porque su ubicación se debe considerar constante al copiar la fórmula. Esta fórmula se copia para todas las celdas de la columna B, situadas a la derecha de los nombres originales, para producir los nuevos nombres de todas las variables (excepto las que sirvan como código identificatorio de los casos).

Una vez creadas esas dos columnas con los viejos y nuevos nombres de las variables, ellas pueden ser copiadas en una hoja de sintaxis de SPSS, para integrar el comando RENAME VARIABLES, que usualmente será muy largo.

Hay la posibilidad en el SPSS de desarrollar métodos abreviados o recursivos, evitando el engorro de transcribir la lista de variables y crear los nuevos nombres. Eso se puede lograr mediante scripts en lenguaje Basic o programas en lenguaje Python, pero esa posibilidad no se desarrolla aquí.

Fusión de archivos

Una vez rebautizadas todas las variables de los diversos archivos, que además deben estar **todos ordenados de acuerdo al mismo criterio** (por ejemplo, por número de hogar y/o persona), se puede proceder a la fusión de los archivos de distintas ondas.

Los archivos de sucesivas ondas se fusionan "horizontalmente", añadiendo para cada individuo las variables medidas en ondas subsiguientes. Si se fusionan archivos de varias rondas, algunos individuos tendrán información solamente en algunas rondas, y otros en otras rondas. Cada individuo tendrá información solo en aquellas ondas en las cuales haya sido entrevistado. En el resto de las ondas todas sus variables aparecerán en blanco.

La sintaxis del SPSS para la fusión sería como la siguiente. Por simplicidad se pone como ejemplo la fusión de solo tres archivos de ondas sucesivas, con cuatro variables cada uno. Se supone aquí que son variables de hogar, y que cada hogar es identificado por un código único, registrado en la variable COHOGAR.

```
MATCH FILES
 /FILE 'hogares 2011-1 renamed.sav'
/ FILE 'hogares 2011-2 renamed.sav'
/ FILE 'encuesta 2011-3 renamed.sav'
BY CODHOGAR.
```

El aspecto general del archivo luego de la fusión será como el siguiente (tres variables por onda):

Codhogar	P1_111	P2_111	P3_111	P1_112	P2_112	P3_112	P1_113	P2_113	P3_113
1									
2									
3									
4									
5									
6									
7									
…..									

En el caso de los archivos de personas en una encuesta de hogares se procede de modo similar. Naturalmente, en ese caso la clave de unión será el código único de cada persona, que puede ser una variable por sí misma o la combinación del código de hogar con el código de cada persona dentro del hogar. Es importante verificar que cada persona tiene consistentemente **el mismo código en todas las ondas**, aunque cada persona puede figurar en unas rondas y no en otras.

REFERENCIAS BIBLIOGRAFICAS

AALEN, Odd O., 1980. "A model for non-parametric regression analysis of counting processes." **Lecture Notes in Statistics**. No.2, pp.1-25.

AALEN, Odd O., 1989. "A linear regression model for the analysis of life times." **Statistics in Medicine** 8: 907-925.

AALEN, Odd O., 1993. "Further results in the non-parametric linear regression model in survival analysis." **Statistics in Medicine** 12:1569-1588.

AALEN, Odd O.; Ornulf BORGAN & Harald FERKJAER, 2001. **Covariate adjustment of event histories estimated from Markov chains: The additive approach**. University of Oslo, Dept of Mathematics, Statistical Research Report No.2, Disponible en www.math.uio.no/eprint/stat_report/2001/02-01.ps.

AGRESTI, Alan, 1990. **Categorical data analysis**. New York, John Wiley & Sons.

AHLO, Juha M., 1990. "Adjusting for non-response bias using logistic regression," **Biometrika** 77(3), pp. 617-624.

AIGNER, D.J & A.S. GOLDBERGER, Eds., 1977. **Latent variables in socio-economic models,** North Holland, Amsterdam.

ALDERMAN, Harold; Jere R. BEHRMAN; Hans-Peter KOHLER; John A. MALUCCIO & Susan COTTS-WATKINS, 2001. "Attrition in longitudinal household survey data**", Demographic Research** (www.demographic-research.org), Vol.5.

ALLISON, Paul D., 1984. **Event history analysis: Regression for longitudinal event data**. Thousand Oaks, California, Sage Publications, Quantitative Applications in the Social Sciences Series No.46.

ARMINGER, Gerhard, 1997. "Dynamic Factor Models for the Analysis of Ordered Categorical Panel data." En: Berkane 1997:177-194.

ARMINGER, Gerhard; Clifford C. CLOGG & M. SOBEL, 1995. **Handbook of statistical modeling for the social and behavioral sciences**, New York: Plenum Press.

BALDI, Ileana; Giovannino CICCONE; Antonio PONTI; Stefano ROSSO; Roberto ZANETTI & Dario GREGORI (2006). "An application of the Cox-Aalen model for breast cancer survival." *Austrian Journal of Statistics* 35 (1): 77–88.

BALTAGI, Badi H., 1995. **Econometric analysis of panel data.** London, J.Wiley and Sons.

BECKETTI, Sean; William GOULD; Lee LILLARD & Finis WELCH, 1988, "The Panel Study of Income Dynamics after Fourteen Years: An Evaluation," **Journal of Labor Economics** Vol. 6, pp. 472-492.

BERELSON, Bernard B.; Paul F. LAZARSFELD & W. McPHEE, 1954. **Voting**. Chicago, University of Chicago Press.

BERKANE, Maia (editora), 1997. **Latent variable modeling and applications to causality**. New York, Springer-Verlag.

BERKANE, Maia (editora) 1997. **Latent Variable Modeling and Applications to Causality**. New York: Springer Verlag.

BIJLEVELD, Catrien C. J. H.; John VANDERKAMP & Leo J. Th. VAN DE KAMP, 1998. **Longitudinal data analysis: Designs, models and methods.** Newbury Park**,** Sage Publications.

BLALOCK, Hubert M. (editor), 1985. **Causal models in the social sciences**. Segunda edición. New York, Aldine De Gruyter.

BLALOCK, Hubert M., 1964. **Causal inferences in nonexperimental research**. The University of North Carolina Press, Chapel Hill, North Carolina.

BLALOCK, Hubert M., 1969. **Theory construction**. Englewood Cliffs, N.J., Prentice Hall.

BLOSSFELD, Hans-Peter & Gotz ROHWER, 2002. **Techniques for event history modeling: New approaches to causal analysis**. 2nd edition. Erlbaum, New Jersey.

BOUDON, Raymond, 1967. **L'analyse mathématique des faits sociaux**. Paris, Plon.

BOX-STEFFENSMEIER, Janet M. & Bradford S. JONES, 2004. **Event history modeling: A guide for social scientists**. Cambridge University Press, Cambridge, UK.

BRYK, Anthony S. & Stephen W. RAUDENBUSCH, 1992. **Hierarchical linear models**. Newbury Park, Sage Publications (segunda edición: 2000).

BUNGE, Mario, 1979. **Causality and modern science.** New York, Dover Publications.

CAO, Huiling, 2005. *A comparison between the additive and multiplicative risk models*. Tesis de maestría, Facultad de Ciencias e Ingeniería, Universidad de Laval, Montréal, Canada. Disponible en http://newton.mat.ulaval.ca/theses/H-Cao_05.pdf.

COLEMAN, James, 1964a. **Models of change and response uncertainty**. Englewood Cliffs (New Jersey), Prentice-Hall.

COLEMAN, James, 1964b. **Introduction to mathematical sociology**. Glencoe, Free Press.

COLEMAN, James, 1968. "The mathematical study of change", en Hubert M. Blalock & Ann B. Blalock (editores), **Methodology in social research**, New York, McGraw-Hill.

COLEMAN, James, 1991**. Longitudinal data analysis.** New York, Basic Books.

COLLINS, Linda M. & A. G. SAYER (Editores.), 2001. **New methods for the analysis of change.** Washington, DC: American Psychological Association.

COX, D.R. 72). "Regression models and life tables". **Journal of the Royal Statistical Society – Series B**, Vol. 34, pp.187-202.

DAVIS, James A., 1985. **The logic of causal order**. Thousand Oaks (California), Sage Publications, Quantitative Applications in the Social Sciences Series No.55.

DEATON, Angus, 1985. "Panel data from the time series of cross-sections." **Journal of Econometrics,** 30, pp. 109-126.

DIGGLE, Peter J., Kung-Yee LIANG & Scott L. ZEGER, 1994. **The analysis of longitudinal data.** Oxford, Oxford University Press, Clarendon Press.

DOOB, J.L., 1953. **Stochastic processes**, New York, J.Wiley and Sons.

EEROLA, Mervi, 1994. **Probabilistic causality in longitudinal studies.** New York, Springer-Verlag.

ESTES W. & C. J. BURKE, 1955. "Application of a statistical model to simple discrimination learning in human subjects", **Journal of Experimental Psychology**, pp. 81-88.

FINKEL, Steven E., 1995. **Causal analysis with panel data**. Thousand Oaks (California), Sage Publications, Quantitative Applications in the Social Sciences Series No. 105.

FIREBAUGH, Glenn, 1997. **Analyzing repeated surveys**, Thousand Oaks (California), Sage Publications, Quantitative Applications in the Social Sciences Series No. 115.

FITZGERALD, John, Peter GOTTSCHALK, & Robert MOFFITT, 1998, "An Analysis of Sample Attrition in Panel Data". **The Journal of Human Resources** 33 (2): 251-99.

FORNI, Mario & Marco LIPPI, 2001. "The generalized dynamic factor model: representation theory". **Econometric Theory**, Vol.17, pp.1113-1141.

FORNI, Mario, Marc HALLIN, Marco LIPPI & Lucrezia REICHLIN, 1999. "The generalized dynamic factor model: identification and estimation". Center for Economic Policy Research (CEPR), Discussion Paper Series, No. 2338. Publicado luego en la **Review of Economics and Statistics**, Vol 82 (2000), pp.540-554.

FORNI, Mario; Marc HALLIN, Marco LIPPI, & Lucrezia REICHLIN, 2002. "The generalized dynamic factor model: one-sided estimation and forecasting". Center for Economic Policy Research (CEPR), Discussion Paper No.3432, London.

FREEDMAN Deborah; Arland THORNTON; Donald CAMBURN; Duanne ALWIN & Linda YOUNG-MARCO, 1988. "The life history calendar: a technique for collecting retrospective data", en Clifford C. CLOGG, (editor), **Sociological Methodology 1988**, Washington DC, American Sociological Association, pp.37-68.

GALTUNG, Johann, 1964. **Teoría y métodos de la investigación social**, Buenos Aires, EUDEBA, 2 vols.

GANDY, Axel & Uwe JENSEN, 2005. On goodness-of-fit tests for Aalen's additive risk model. *Scandinavian Journal of Statistics*, Vol. 32, pp. 425–445

GEWEKE, J., 1977. "The dynamic factor analysis of economic time series." Incluido en D.J. AIGNER & A.S. GOLDBERGER, 1977.

GONG, Xiaodong & Arthur VAN SOEST, 2001. "Wage Differentials and Mobility in the Urban Labor Market: A Panel Data Analysis for Mexico". Documento de trabajo de IZA (Instituto de Estudios Laborales), Bonn. http://www.iza.org/publications/dps/.

HAGENAARS Jacques A. & A. L. McCUTCHEON, 2002. **Applied Latent Class Analysis**. Dordrecht: Kluwer.

HAGENAARS, Jacques A. & Allan L. McCUTCHEON (editores), 2002. **Applied latent class analysis.** Cambridge (UK), Cambridge University Press.

HAGENAARS, Jacques A., 1990. **Categorical longitudinal data: Loglinear panel, trend and cohort analysis**. Newbury Park, Sage Publications.

HAGENAARS, Jacques A., 1994. "Latent variables in log-linear models of repeated observations". En von Eye y Clogg 1994), pp.329-352.

HAMERLE, A., & G. RONNING, 1995. "Panel Analysis for Qualitative Variables," en ARMINGER et al. (1995), pp. 401-451.

HAND, David & Martin CROWDER, 1996**. Practical longitudinal data analysis.** Chapman & Hall Texts in Statistical Science Series, Londres, Chapman & Hall/CRC Press.

HECKMAN, J. 1981, "Statistical Models for Discrete Panel Data," en MANSKI &. McFADDEN (eds.), 1981, pp. 114-178.

HECKMAN, J., & B. SINGER, eds., 1982. **Econometric analysis of longitudinal data**, número especial del **Journal of Econometrics**, vol. 18, pp. 1-190.

HECKMAN, J., & B. SINGER, eds., 1985. **Longitudinal analysis of labor market data,** Cambridge (UK), Cambridge University Press.

HOSMER, David W. Jr. & Stanley LEMESHOW, 1999. **Applied Survival Analysis: Regression modeling of time to event data**. New York:John Wiley & Sons.

HSIAO, Cheng, 1986. **Analysis of panel data.** Cambridge Univ.Press, reimpresión 1999.

HUFFER, Fred W. & Ian W. McKEAGUE, 1987. Survival analysis using additive risk models. Department of Statistics, Florida State University, Report No. M-756.

HYMAN, Herbert, 1963. **Survey design and analysis.** New York, Free Press.

INDEC, 2003. *La nueva Encuesta Permanente de Hogares de Argentina. 2003.* Buenos Aires: http://www.indec.gob.ar/nuevaweb/cuadros/4/Metodologia_EPHContinua.pdf. Instituto Nacional de Estadística y Censos.

KISH, Leslie, 1987. **Statistical design for research**. New York, John Wiley & Sons.

KLEIN, John P. & Melvin L. MOESCHBERGER, 1997, **Survival analysis: Techniques for Censored and Truncated Data**. Springer, New York.

KREFT, Ita & Jan DE LEEUW, 1999. **Introducing multilevel analysis**. London, Sage.

KYRIAZIDOU, E., 1997. "Estimation of a Panel Data Sample Selection Model," **Econometrica**, vol. 65, pp. 1335-1364.

KYRIAZIDOU, E., 1999. "Estimation of dynamic panel data sample selection models," disponible en http://www.econ.ucla.edu/kyria/dsele/dsele.pdf.

LANCASTER, T., 1990. **The econometric analysis of transition data**, Cambridge (UK), Cambridge University Press.

LANGE, Oskar, 1964. **Introducción a la econometría**, México-Buenos Aires, Fondo de Cultura Económica.

LAZARSFELD Paul F. & Nell W. HENRY, 1968. **Latent structure analysis.** Boston, Houghton Mifflin.

LAZARSFELD Paul F., 1961. "The algebra of dichotomous systems", en SOLOMON Henry (editor), **Studies in item analysis and prediction.** Stanford University Press, Stanford, California, 1961.

LAZARSFELD Paul F., 1965. "Repeated observations on attitude and behavior items", incluido en STERNBERG, Saul; V.CAPECCHI; T.KLOEK; & C.T.LEENDERS (editores), **Mathematics and social sciences,** Paris, Mouton & Co., 1965.

LAZARSFELD, Paul F., B. BERELSON & H.GAUDET, 1948. **The people's choice.** New York, Columbia Univ. Press.

LILLARD, Lee A. & Constantijn W.A. PANIS, 1998, "Panel Attrition from the Panel Study of Income Dynamics," **The Journal of Human Resources**, Vol. 33 No.2, pp.437-457.

LIN, D.Y. & Z.YING, 1994. "Semiparametric analysis of the additive risk model." **Biometrika** 81:61–71.

LIN, D.Y. & Z.YING, 1995. Semiparametric analysis of general additive-multiplicative hazard models for counting processes. *Annals of Statistics* 23:1712–1734.

LITTLE, Roderick & Donald RUBIN, 1987. **Statistical analysis with missing data**,New York, John Wiley & Sons.

MA, Shuangge; Michael R. KOSOROK & Jason P. FINE, 2006. Additive risk models for survival data with high-dimensional covariates. *Biometrics*, Vol. 62, pp. 202–210.

MAGNUSSON, David, Georg RUDINGER & Lars R. BERGMAN (eds.), 1994. **Problems and methods in longitudinal research : Stability and change.** European Network on Longitudinal Studies on Individual Development. Cambridge (UK), Cambridge University Press.

MALETTA, Hector, 1970. **Estructura latente de sistemas dicotómicos**, Buenos Aires, Editorial Nueva Visión.

MANSKI, C. and D. McFADDEN (eds.), 1981. **Structural analysis of discrete data with econometric applications**, MIT Press, London, 114-178.

MÁTYÁS, Laszlo, & Patrick SEVESTRE, 1996. **The econometrics of panel data: A handbook of the theory with applications,** segunda edición revisada. Dordrecht: Kluwer Academic Publishers.

McCULLOCH, Charles E. & Shayle Robert SEARLE, 2000. **Generalized, linear, and mixed models.** Chichester (UK), J. Wiley & Sons.

McKEAGUE, Ian W., 1997. Aalen's additive risk model. En: S. Kotz and C. B. Read, editores, *Encyclopedia of Statistical Sciences*, Update Volume 1 (1997):1–6, New York:John Wiley & Sons.

MENARD, Scott, 1991. **Longitudinal research**. Thousand Oaks, California, Sage Publications, Quantitative Applications in the Social Sciences Series No.76.

MEREDITH, W., & HORN, J., 2001. "The role of factorial invariance in modeling growth and change", en L. M. COLLINS & A. G. SAYER (Eds.), 2001.

MERTON, Robert K & Paul F. LAZARSFELD (editores), 1950. **Studies in the scope and method of *The American Soldier***. Glencoe, Free Press.

NERLOVE, Marc, 2000. "An Essay on the History of Panel Data Econometrics". Documento de trabajo, Department of Agricultural and Resource Economics, University of Maryland. http://www.arec.umd.edu/mnerlove/mnerlove.htm.

NESSELROADE, John R., 1994. "Exploratory factor analysis with latent variables and the study of processes of development and change". En von Eye y Clogg (1994), pp.131-154.

PELZ, Donald C. & Frank M. ANDREWS, 1964. "Causal priorities in panel study data". **American Sociological Review**, pp.836-848.

PLEWIS, Ian, 1985. **Analysing change: Measurement & explanation using longitudinal data.** Chichester (UK), J. Wiley & Sons.

POPPER, Karl R., [1934], **Logik der Forschung**. Viena. Versión inglesa actualizada: **The logic of scientific discovery,** Londres, Routledge, 1992. Traducción castellana, versión no

actualizada: **La lógica de la investigación científica**, Madrid, Tecnos, 1973 (basada en la versión inglesa de 1959).

PRADHAN, M. & A. VAN SOEST, 1997, "Household labor supply in urban areas of Bolivia." **Review of Economics and Statistics,** vol.79, pp. 300-310.

RAO, B. Bhaskara (editor), 1994. **Cointegration for the applied economist**. Londres, MacMillan Press Ltd.

SAMAJA, Juan, 1995. **Epistemología y metodología,**, Buenos Aires, Eudeba, 1995.

SCHEIKE, Thomas H. & Mei-Jie ZHANG, 2002. An additive-multiplicative Cox-Aalen regression model. *Scandinavian Journal of Statistics* Vol. 29, pp. 75–88.

SCHEIKE, Thomas H. & Mei-Jie ZHANG, 2003. "Extensions and applications of the Cox-Aalen survival model." **Biometrics***,* Vol. 59, pp.1036–1045.

SCHEIKE, Thomas H., 2001. "A generalized additive regression model for survival times." **The Annals of Statistics***,* Vol. 29, No. 5, 1344–1360.

SCHUSTER, Felix G., 1982. **Explicación y predicción**, Buenos Aires, CLACSO.

SIMON, Herbert A., 1952. "On the definition of the causal relation". **The Journal of Philosophy**, vol. 49, pp.517-528.

SIMON, Herbert A., 1987. "Causality in economic models", en John Eatwell y otros (editores), **The New Palgrave: A dictionary of economics**, Londres, MacMillan Press Ltd.

SNIJDERS, Tom A. & Roel J. BOSKER, 2012. **Multilevel analysis**. Segunda edición. Los Angeles, Sage Publications.

STOUFFER, Samuel et al., 1949. **The American Soldier**. Princeton University Press.

THERNEAU, Terry M. & Patricia M. GRAMBSCH, 2000. **Modeling survival data**. New York: Springer Verlag.

TISAK, J. & W. MEREDITH, 1990**.** "Longitudinal factor analysis"**,** en VON EYE (ed.) 1990**.**

VAN DEN BERG, Gerard J. & Maarten LINDEBOOM, 1998. "Attrition in Panel Survey Data and the Estimation of Multi-State Labor Market Models," **The Journal of Human Resources**, Vol. 33, No.2, pp. 458-478.

VERBEEK, M. & T.E. NIJMAN, 1992. "Can cohort data be treated as genuine panel data?" **Empirical Economics**, 17, pp.9-23.

VERBEEK, M., 1996. Pseudo panel pata. En: L. Matyas & P. Sevestre (compiladores), **The econometrics of panel data: A handbook of the theory with applications**, segunda edición, pp. 280-292. Vol. 33 de la colección: **Advanced Studies in Theoretical and Applied Econometrics**, Dordrecht (Holanda), Elsevier/Kluwer Academic.

VERBEKE, Geert; G. MOLENBERGHS; P. BICKEL; & P. DIGGLE (Editores), 2000. **Linear mixed models for longitudinal data**, New York, Springer-Verlag.

VERMUNT, Jeroen K., 1997. **Log-linear models for event histories**. Thousand Oaks, California, Sage Publications.

VON EYE Alexander & Clifford C. CLOGG (eds.), 1994. **Latent variables analysis: Applications for developmental research**. Tousand Oaks, California, Sage Publications.

VON EYE, Alexander (editor), 1990**. Statistical methods in longitudinal research.** 2 Vols. Boston, Academic Press.

WHO, 2006b. *WHO Child Growth Standards: Methods and development*. Geneva: World Health Organization. http://www.who.int/entity/childgrowth/standards/velocity/tr3_velocity_report.pdf.

YAMAGUCHI, Kazuo, 1991. **Event history analysis**. Newbury Park, California: Sage.

ZABEL, Jeffrey E., 1998, "An Analysis of Attrition in the Panel Study of Income Dynamics and the Survey of Income and Program Participation with an Application to a Model of Labor Market Behavior," **The Journal of Human Resources,** Vol. 33, No.2.

ZEISEL, Hans, 1966. **Dígalo con números**, México, Fondo de Cultura Económica.

ZILIAK, James P. & Thomas J. KNIESNER, 1998, "The importance of sample attrition in life cycle labor supply estimation." **The Journal of Human Resources**, Vol. 33, No.2.