IDICSO
Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

ÁREA EMPLEO Y POBLACIÓN

Metodología de análisis de panel de variables categóricas

PRIMERA PARTE

por **Héctor Maletta***

Buenos Aires, DIC/2002

* **MALETTA, Héctor.** Lic. en Sociología, Universidad Católica Argentina. Doctor en Economía, Universidad de Bologna (Italia). Docente de carreras de grado y posgrado, Facultad de Ciencias Sociales, Universidad del Salvador (USAL). Investigador Principal, Área Empleo y Población, IDICSO, USAL. Consultor de organismos internacionales, especialmente la FAO, pero también otros como el FIDA y el BID. Consultor internacional en casi todos los países de América Latina y en algunos países de África y Asia.

BREVE HISTORIA DEL IDICSO. Los orígenes del IDICSO se remontan a 1970, cuando se crea el "Proyecto de Estudio sobre la Ciencia Latinoamericana (ECLA)" que, por una Resolución Rectoral (21/MAY/1973), adquiere rango de Instituto en 1973. Desde ese entonces y hasta 1981, se desarrolla una ininterrumpida labor de investigación, capacitación y asistencia técnica en la que se destacan: estudios acerca de la relación entre el sistema científico-tecnológico y el sector productivo, estudios acerca de la productividad de las organizaciones científicas y evaluación de proyectos, estudios sobre política y planificación científico tecnológica y estudios sobre innovación y cambio tecnológico en empresas. Las actividades de investigación en esta etapa se reflejan en la nómina de publicaciones de la "Serie ECLA" (SECLA). Este instituto pasa a depender orgánica y funcionalmente de la Facultad de Ciencias Sociales a partir del 19 de Noviembre de 1981, cambiando su denominación por la de Instituto de Investigación en Ciencias Sociales (IDICSO) el 28 de Junio de 1982.

Los fundamentos de la creación del IDICSO se encuentran en la necesidad de:

- Desarrollar la investigación pura y aplicada en Ciencias Sociales.
- Contribuir a través de la investigación científica al conocimiento y solución de los problemas de la sociedad contemporánea.
- Favorecer la labor interdisciplinaria en el campo de las Ciencias Sociales.
- Vincular efectivamente la actividad docente con la de investigación en el ámbito de la facultad, promoviendo la formación como investigadores, tanto de docentes como de alumnos.
- Realizar actividades de investigación aplicada y de asistencia técnica que permitan establecer lazos con la comunidad.

A partir de 1983 y hasta 1987 se desarrollan actividades de investigación y extensión en relación con la temática de la integración latinoamericana como consecuencia de la incorporación al IDICSO del Instituto de Hispanoamérica perteneciente a la Universidad del Salvador. Asimismo, en este período el IDICSO desarrolló una intensa labor en la docencia de post-grado, particularmente en los Doctorados en Ciencia Política y en Relaciones Internacionales que se dictan en la Facultad de Ciencias Sociales. Desde 1989 y hasta el año 2001, se suman investigaciones en otras áreas de la Sociología y la Ciencia Política que se reflejan en las series "Papeles" (SPI) e "Investigaciones" (SII) del IDICSO. Asimismo, se llevan a cabo actividades de asesoramiento y consultoría con organismos públicos y privados. Sumándose a partir del año 2003 la "Serie Documentos de Trabajo" (SDTI).

La investigación constituye un componente indispensable de la actividad universitaria. En la presente etapa, el IDICSO se propone no sólo continuar con las líneas de investigación existentes sino también incorporar otras con el propósito de dar cuenta de la diversidad disciplinaria, teórica y metodológica de la Facultad de Ciencias Sociales. En este sentido, las áreas de investigación del IDICSO constituyen ámbitos de articulación de la docencia y la investigación así como de realización de tesis de grado y post-grado. En su carácter de Instituto de Investigación de la Facultad de Ciencias Sociales de la Universidad del Salvador, el IDICSO atiende asimismo demandas institucionales de organismos públicos, privados y del tercer sector en proyectos de investigación y asistencia técnica.

IDICSO

Departamento de Comunicación

Email: idicso@yahoo.com.ar

Web Site: http://www.salvador.edu.ar/csoc/idicso

TABLA DE CONTENIDOS COMPLETA

PRIMERA PARTE

TABLA DE CONTENIDOS

METODOLOGÍA DE ANÁLISIS DE PANEL DE VARIABLES CATEGÓRICAS

PRIMERA PARTE

1. Introducción al análisis de panel

1.1. El desarrollo de los estudios de panel

Se denomina "estudio de panel" a la **recolección de información sobre una pluralidad de unidades de análisis en varios instantes del tiempo**. Esta situación 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 de naturaleza **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** en los cuales la información es **sincrónica** o **cotemporal**, y se refiere a un mismo instante o período.

La aplicación tradicional del análisis de panel fueron las encuestas de seguimiento, y uno de los primeros ejemplos fueron las encuestas pre-electorales realizadas por Lazarsfeld, Berelson y otros en la década del cuarenta, publicadas luego bajo los títulos **The people's choice** (Lazarsfeld y otros, 1948) y **Voting** (Berelson y otros, 1954). Otra aplicación frecuente de los datos de panel son los estudios experimentales o cuasi-experimentales del tipo "antes-después" sobre el efecto de la publicidad, un modelo muy frecuente en el campo de los estudios de marketing; uno de los primeros, sin embargo, no tuvo carácter comercial, pues 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, realizado en el marco de un estudio más amplio de las tropas norteamericanas dirigido por Samuel Stouffer, y publicado luego de terminado el conflicto con el título **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) en dos momentos del tiempo, y usando como principal instrumento la **comparación de porcentajes**.

Posteriormente 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 el desarrollado por James S. Coleman en su **Introduction to Mathematical Sociology** (Coleman

1964b) y 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 verdadero cambio de las variables subyacentes y las meras variaciones aleatorias de las respuestas. Asimismo en 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), Hagenaars (1994) y Vermunt (1997). El análisis de procesos de Boudon (1967, cap. VII-IX) parte de los modelos basados en cadenas de Markov y en los aportes de Coleman pero además aplica sus "coeficientes de dependencia" (los que en inglés se llaman "path coefficients" y provienen del análisis de regresión) a los datos de panel.

En las últimas décadas, desde mediados de los años sesenta en adelante, se llevaron a cabo diversos estudios de seguimiento de muy largo plazo, en los cuales una misma muestra de sujetos es seguida a lo largo de muchos años, como ocurre por ejemplo con 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, entre muchos otros (Becketti y otros, 1988). 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. Una interesante introducción metodológica a este tipo de estudios es la breve monografía de Scott Menard (1991), así como los textos de von Eye (editor, 1990), Diggle y otros (1994) y Bijleveld y otros (1998).

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 (editoras, 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 y Clogg, 1994; Berkane, 1997, especialmente el artículo de Arminger 1997; Nesselroade 1997; y Hagenaars y McCutcheon, 2002).

Paralelamente se desarrolló una tradición analítica diferente en el marco de la Econometría, referida a situaciones en que se dispone de varias series económicas 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 variables de intervalo, aunque también se han aplicado las misma técnicas econométricas a encuestas de hogares, e incluso se han adaptado los procedimientos para incluir variables categóricas en los tratamientos estadísticos, 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 Nerlove (2000) y en el importante texto de Mátyás y Sevestre, 1996. Otros importantes textos 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 recientes 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.

Recientes desarrollos teóricos 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 Verbeke y otros (editores, 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.

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. Este tipo de enfoque busca identificar factores o variables subyacentes inobservables, correlacionadas con las variables manifiestas, capaces de explicar la covariación de estas últimas en el tiempo.

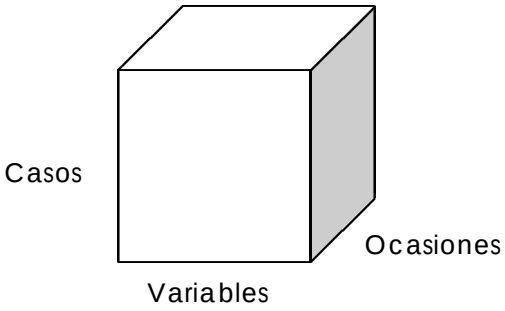
En la presente introducción metodológica no se cubren todos los aspectos del vasto campo de los estudios longitudinales. Se dedica preferente atención a una subcategoría: los **estudios de panel** donde predominan las **variables de tipo categórico**, y con un **número muy limitado de fechas de observación a lo largo del tiempo**. El ejemplo clásico son las encuestas repetidas sobre la misma muestra de respondentes, realizadas a determinados intervalos. 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 ***k*** rondas del panel, y en cada ronda se reemplaza ***1/k*** hogares (***k*** típicamente es igual a cuatro rondas del panel). 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). Cada fila representa una **unidad de análisis**, cada columna una **variable**, y cada celdilla contiene el **valor** de una variable para una determinada unidad de análisis. Pero en realidad hay tres dimensiones y no dos en un conjunto de datos: las **unidades de análisis** sobre las cuales versan los datos; las **variables** que son medidas u observadas en dichas unidades de análisis, y los **períodos** o momentos del tiempo en que se realizan las observaciones. La matriz de datos originada por una encuesta se refiere a **un solo período**, y está formada por **n** casos con **m** variables. La matriz originada por un estudio multivariado longitudinal (por ejemplo los datos econométricos de un país determinado) se refiere a **un solo caso** y está formada por **t** períodos y **m** variables. El análisis de panel añade una tercera dimensión (el tiempo), de modo que hay una matriz de datos para cada uno de los momentos o períodos de observación. 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] 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.

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	
Encuestas repetidas	Diferentes muestras en cada ronda
Series retrospectivas	Datos recogidos en una sola oportunidad acerca de diversos momentos del pasado
Panel	Seguimiento de los mismos casos en dos o más rondas
Panel rotativo	En cada ronda se reemplaza rotativamente una proporción ***1/k*** de los casos, de modo que cada caso permanece en el panel por ***k*** rondas.
Registro continuo	Datos sobre los mismos casos, registrados de manera continua.

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**

[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.

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, con muestras diferentes cada vez, 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).

Un paso más adelante en cuanto a estudios longitudinales está constituido por las "series 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 embarazo, matrimonio, divorcio, etc.). También son de este tipo las "historias ocupacionales" 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 las encuestas retrospectivas es la dudosa confiabilidad de las respuestas que descansan únicamente en la memoria de los entrevistados.

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. 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.

Es difícil mantener los mismos sujetos durante muchas rondas de un estudio de panel, pues los sujetos evidencian fatiga o bien reducen la calidad de sus respuestas. Para contrarrestar este factor se usan frecuentemente los **paneles rotativos**, en los cuales cada sujeto permanece en la muestra durante un cierto número de rondas, tras de lo cual es reemplazado. Normalmente en cada ronda hay un cierto número de casos que están siendo entrevistados por primera vez, otros que están en su segunda ronda, otros en la tercera, y así sucesivamente hasta aquellos que están contestando la encuesta por última vez. Por ejemplo en la Encuesta Permanente de Hogares de la Argentina los hogares permanecen en la muestra durante dos años, lo cual en general significa cuatro rondas pues la encuesta tradicionalmente ha sido realizada dos veces por año, aunque en algunas ocasiones se realizaron tres rondas por año.

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

En cada ronda hay un 25% de la muestra que está contestando por primera vez, 25% por segunda vez, 25% por tercera vez, y 25% que contestan por cuarta y última vez. Estos modelos de panel rotativo permiten que si en cada ronda se reemplazan **M** casos, que representan por ejemplo un 25% del total, en todos los períodos haya una muestra de **4M** casos entrevistados, y que al mismo tiempo se cuente 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.

Esta rotación planificada de las muestras no tiene nada que ver con el **desgranamiento no planificado** (*attrition*). De una onda a otra hay siempre algunos sujetos que "desaparecen" de la muestra por distintas razones: muerte, emigración, 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" sin causa aparente, 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. Un fenómeno similar son las "entradas tardías", es decir, sujetos que se añaden a la encuesta en fecha posterior a la planificada, como ocurre con las personas que se incorporan a los hogares de la muestra después de la primera ronda en que dichos hogares fueron incluidos.[3]

Los estudios de **registro continuo** son habitualmente los que se basan en datos de registro, como por ejemplo los datos extraidos 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 los datos sobre movimientos de cada cuenta o transacciones de cada cliente que pueden poseer los bancos u otras empresas, o los datos de inventario de mercadería de las empresas comerciales. En algunos casos se trata de datos fechados de manera continua pero **registrados a intervalos**, de modo que los eventos intermedios están sujetos a un **registro retrospectivo**. De todas maneras, si la frecuencia de los registros es alta, y la longitud del intervalo es breve, los datos se pueden considerar en la práctica como registrados en forma continua.

En este punto conviene distinguir diferentes clases de información en relación a la dimensión temporal, y de acuerdo al tipo de variables. Cuando se trata de variables categóricas o cualitativas, básicamente hay que distinguir entre la

[3] Sobre el fenómeno del desgranamiento (*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).

información referida al **estado** de una variable en un **momento** determinado, y la información referida a los **eventos** (cambios de estado) experimentados por una variable a lo largo de un **período** determinado. Los principales tipos de análisis cuando se trata de variables **categóricas** son los siguientes.

Tipos de análisis longitudinal con variables categóricas	
Serie de cortes transversales	Se mide el **estado** de los sujetos en diferentes fechas, y se compara el estado actual 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 fechados**

Los estudios de panel consisten en una **serie de cortes transversales** de la misma muestra, que sólo proveen información sobre el estado de las unidades de análisis en el momento de cada 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 a reconstruir la evolución de esa situación desde la anterior entrevista realizada seis meses antes.

En esos casos, los estudios de panel permiten **comparar estados**, y por lo tanto proveen 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, aunque 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 de tipo retrospectivo).

También es frecuente que, si se registra un **cambio de situación**, es decir un **evento**, no se registre el momento o **fecha exacta de la transición** (la fecha, o mejor dicho la última fecha, en que el sujeto quedó desocupado), lo cual puede haber ocurrido en cualquier momento dentro del intervalo entre las dos rondas.

En muchos estudios de panel se incluyen variables referidas al período intermedio (por ejemplo: ¿cuántas veces perdió su empleo en los últimos seis meses?) lo cual 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: los sucesos ocurridos "en los últimos seis meses" si la encuesta se realiza en octubre no son los mismos que se registrarían si la encuesta se realizase agosto o en otro momento. 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.

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 están interesados 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. 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 analizan períodos muy largos con muchas entrevistas a lo largo del mismo, y los procesos que se investigan en general son de larga maduración, la serie resultante puede a menudo ser considerada, para fines prácticos, 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 "va-

riaciones de valor", que no ocurren por saltos entre estados discretos sino por una variación gradual a lo largo de un continuo.

1.4. La dimensión temporal de las variables

Un panel observa a la misma muestra en varias "ondas", también llamadas "cortes temporales" o "rondas". Cada ronda o corte temporal puede referirse a un **período** (entre dos fechas) o a una **fecha** determinada: se utiliza la palabra período de manera genérica, entendiéndose que el "período" puede ser de duración instantánea. Es importante distinguir el **período (o fecha) de recolección de datos** y al **período (o fecha) de referencia de la observación**. 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. En esa 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**. 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.

Hay 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, **sin dimensión temporal** (en personas, unidades monetarias, años, etc.) aunque sí con una **referencia temporal** (corresponden a una fecha o período determinado). Otras variables 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 "unidades de medida 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 en su unidad de medida (ingresos **mensuales**, medidos en unidades monetarias **por mes**). Esta distinción corresponde a la distinción entre variables de stock o estado y variables de flujo en economía.

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?" posiblemente significa "¿Estuvo Ud ocupado [en algún momento] durante los últimos siete días?" lo cual es completamente distinto a la pregunta "¿Estuvo Ud ocupado [todo el tiempo] durante los últimos siete días?" En cualquiera de los dos casos la respuesta se mide en unidades sin referencia temporal (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 (por ejemplo, se la podría considerar ocupada con 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 es una variable de flujo.

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. Las variables de stock se miden en cantidades de su propia unidad de medida. 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. Se registra en un momento dado, pero se mide en sus propias unidades de medida en una determinada referencia temporal pero 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 llamamos **M** a una unidad de medida genérica, y **T** al 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**.

Escribimos "las variables de flujo **más sencillas**" porque puede haber variables de dimensionalidad más complicada como las que reflejan **cambios en el flujo**. Por ejemplo, la **aceleración** (aumento de la velocidad), que se mide, digamos, en kilómetros por hora adicionales por segundo, o en forma abreviada $\mathbf{M/T^2}$. 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 variación o tasa anual de incremento del producto bruto tiene dimensionalidad $\mathbf{M/T^2}$ (incremento por año de los bienes producidos anualmente). En la siguiente tabla se suministran algunos ejemplos de variables de stock y de flujo.

Variables de estado o de stock **(referidas a un momento dado)**	**Unidad de medida**
Población	Personas
Capital fijo existente	Unidades monetarias
Existencias de mercadería	Unidades físicas
Hijos nacidos vivos en toda la vida	Niños nacidos vivos
Patrimonio hogareño	Unidades monetarias
Existencias ganaderas	Cabezas de ganado
Variables de flujo **(referidas a un período determinado)**	
Nacimientos	Niños nacidos vivos por año
Defunciones	Personas por año
Producto bruto	Unidades monetarias por año
Unidades vendidas	Unidades por año (u otro periodo)
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

Cuando las variables son expresadas en forma **relativa**, se abandonan las unidades naturales **M** (dólares, toneladas, metros), pero **no se abandona la dimensión temporal**. Por ejemplo, la **tasa de crecimiento** del producto bruto tiene dimensión $\mathbf{1/T^2}$, mientras que en cambio el **crecimiento** del producto bruto en términos absolutos tiene dimensionalidad $\mathbf{M/T^2}$ (dolares/año adicionales por año).[4]

En el caso de las variables categóricas en los estudios de panel, generalmente los **estados** se expresan en variables de stock o de estado, y miden la ubicación de los sujetos en las distintas categorías **en un momento dado**. En cambio los

[4] Una discusión de la dimensionalidad de las variables económicas puede hallarse en Lange 1964.

eventos se suelen medir como variables de flujo (eventos por unidad de tiempo). Por ejemplo, la condición ocupacional de las personas activas (ocupado o desocupado) es una variable de estado que 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.

Aparte del tipo de variables que se mide, los estudios longitudinales pueden diferir también en la forma en que tratan la variable tiempo. En la forma habitual de 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 del tiempo puede concebirla como una variable continua).

Cuando se trata, en cambio, de una gran cantidad de momentos muy cercanos entre sí, como ocurre por ejemplo con 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 cuando 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. 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.

Estas reflexiones indican que hay que distinguir entre el carácter discreto o continuo **de las observaciones** (que por lo general son discretas pero si son muy frecuentes podrían considerarse como continuas), 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), la **operacionalización del tiempo** como una variable discreta o continua, y 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. Los cambios de estado

observados en las variables categóricas en dos observaciones puede modelizarse como un proceso continuo, donde diferentes individuos cambian en diferentes momentos a lo largo de ese intervalo, o 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 basan en el análisis de variaciones discretas, aunque las variables de observación sean intrínsecamente variables de intervalo; 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.

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.

1.5. Paneles de datos cualitativos

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 son **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 los conceptos de **estados, eventos**, y **procesos**. Los estados son las varias categorías de una variable de tipo cualitativo en las cuales puede resultar clasificado un sujeto o unidad de análisis en un momento determinado. 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 "ocupado" y "desocupado", un evento sería el paso de una situación a otra (encontrar empleo, o quedar sin trabajo). 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** es simplemente el pasaje de su estado en la primera ronda a su estado en la segunda, pero si el proceso suibyacente 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), 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", y que "mañana" 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".

Variables tan inocentes como la edad o la antigüedad en el empleo son 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 trabajando 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 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).

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 se puede preguntar sobre la cantidad de cabezas de ganado que cada unidad productiva poseía al 30 de junio último, que no necesariamente 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 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 típica variable categórica que usualmente describe el **estado actual** de los sujetos en un momento dado, y que puede incluir 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, esas categorías no son exhaustivas. Si se parte en el momento inicial con sujetos situados en alguno de estos cuatro estados civiles, en el segundo momento algunos sujetos podrían estar, por ejemplo, **muertos**, o "**No vivos**". Eso añadiría un quinto "estado", que consiste en "no estar vivo". Aun cuando ese estado no contenga ningún caso en el primer período, podría albergar algunos en el segundo. 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".[5] Esta complicación no es significativa para los presentes propósitos, por lo cual se deja de lado por el momento. También se dejan de lado los individuos "perdidos", "desgranados" o "desertores", que no reaparecen la segunda vez por haberse mudado, por rechazar la segunda entrevista o por cualquier otra razón, así como los individuos de "aparición tardía", que no fueron encuestados la primera vez pero aparecen con información en la segunda ronda.

[5] 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.

A partir de los estados civiles iniciales, y dejando de lado por simplicidad la posibilidad de nacimiento, fallecimiento, aparición tardía o desaparición del respondente, los **eventos posibles** serían los siguientes:

a) Casamiento de un soltero	**Pasaje del estado 1 al estado 2**
b) Casamiento de un divorciado	**Pasaje del estado 3 al estado 2**
c) Casamiento de un viudo	**Pasaje del estado 4 al estado 2**
d) Disolución del vínculo por divorcio	**Pasaje del estado 2 al estado 3**
e) Disolución del vínculo por muerte del cónyuge	**Pasaje del estado 2 al estado 4**

Otros eventos posibles, si se incluyeran las otras posibilidades, serían "Nacer", "Morir", "Aparecer" y "Desaparecer". Es fácil advertir, de todos modos, que esta lista no incluye todas las combinaciones. Se excluyen eventos como el pasaje del estado 1 al estado 3, o del estado 4 al estado 1. Estas exclusiones obedecen a dos clases de razones. Algunos eventos son directamente **imposibles**, mientras que otros son posibles sólo si se admite la existencia de **eventos intermedios** no registrados. Los eventos (a), (b), (c), (d) y (e) mencionados precedentemente son en realidad los únicos **eventos básicos** que pueden suceder con la variable Estado Civil, pero entre dos rondas del panel puede haber algunos **cambios de estado** que impliquen **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

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. 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.[6] Todos los flujos desde otros

[6] Aquí el concepto de soltero equivale a "nunca casado". En inglés, por lo mismo, se distingue entre "sin pareja" (*single*) y "nunca casado " (*never married*). Los "sin pareja" (*singles*) incluyen solteros, viudos y divorciados.

estados civiles al estado de soltería son entonces, por definición, inexistentes e imposibles.

Ciertos cambios de estado civil entre dos diferentes momentos del tiempo son en cambio posibles, pero son cambios de estado que sólo pueden existir si existe un evento intermedio (o más de un evento intermedio). No son eventos **básicos** sino **cambios netos** observables al cabo de un tiempo como resultado de dos o más cambios básicos ocurridos en el período intermedio. Por ejemplo, para pasar de soltero en el tiempo **t** a 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. Si la observación es muy frecuente o se hace mediante registro continuo no existe posibilidad práctica de pasar de soltero a divorciado, o de divorciado a viudo, en dos observaciones sucesivas.

Cuando las observaciones están suficientemente espaciadas en el tiempo, sin embargo, tales cambios aparentes pueden ser registrados, pero en realidad habría siempre un **evento intermedio no registrado**. 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 un evento intermedio: el cambio neto de soltero a divorciado podría reflejar una secuencia no observada como la siguiente: soltero→**casado**→**viudo**→**casado**→divorciado. Es importante distinguir, entonces, entre cambios intrínsecamente imposibles y cambios netos aparentemente imposibles 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. 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.

2. Análisis descriptivo de panel

Es conveniente distinguir entre el simple uso de tabulaciones cruzadas de las variables registradas en diferentes períodos, 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.[7]

Primera ronda	**Segunda ronda**		**Total**
	X=1	**X=2**	
X=1	N_{11}	N_{12}	N_1^t

[7] 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 $N_{\bar{1}\bar{2}}$. Nótese que en nuestra notación los subíndices se refieren a los **valores de la variable** (X=1 y X=2), mientras que aquí los mismos subíndices se refieren a los **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.

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 **en un determinado período de tiempo**.

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

Los 100 individuos que estaban en el estado 1 siguieron en ese estado, y del mismo modo permanecieron en el estado 2 los otros 300 sujetos. Nadie cambió de estado, y por lo tanto la distribución marginal de sujetos 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**) no es 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 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. Esta situación depende de **que ambos flujos sean de igual magnitud**, pero no depende de la magnitud 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			
Primera ronda	**Segunda ronda**		**Total**
	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.

La estabilidad agregada puede no existir al observar dos períodos cualesquiera. La distribución de la variable puede cambiar de un período a otro, y de hecho esa es la situación más frecuente. Por ejemplo:

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 total cambia: 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 esta 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:

$N_{12} = N_{21}$

Cuando hay más de dos estados no es necesario que **cada uno** de los flujos en un sentido se compense con un flujo de igual magnitud y de sentido contrario. La condición anterior se generaliza para **variables con más categorías** en una forma más amplia: sólo es necesario que la **suma** de todos los **flujos de salida** de cada estado se compense con la **suma** de todos los **flujos de entrada** en el mismo estado. En símbolos, habrá estabilidad global agregada si para todo estado **i** se cumple la siguiente condición:

$$\sum_j N_{ij} = \sum_j N_{ji} \quad (i \neq j)$$

Por ejemplo, si hay tres estados que corresponden a las situaciones laborales de los inactivos, ocupados y desocupados, 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 lo mismo para los otros estados. 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.

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 de seres humanos 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 (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.

Primera ronda	Segunda ronda						Total
	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_z" corresponde a "No vivo en el momento de la encuesta", y el estado N_a corresponde a "Ausente o con estado civil ignorado en el momento de la encuesta". La celdilla N_{zs} 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} 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.[8] 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 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 se encuentran compensada por la suma de los flujos de salida respectivos.

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

[8] 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. Si el panel, como suponemos por simplicidad, sólo registra el estado inicial y el estado final, estos casos serán 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 nacidos y muertos entre las dos rondas.

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, si existieran, 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, 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 diferentes de los demás, en cuyo caso los resultados del panel podrían tener una distorsión. Esa distorsión sería mayor, por supuesto, si se excluyen todos los desertores y tardíos y se retienen solamente los que fueron encuestados en ambas rondas. En otras palabras, las desercios, 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

La tabla de rotación, estrictamente hablando, 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. Tratándose de estados civiles, esos casos son sumamente improbables, aunque en principio posibles, sobre todo si el período intermedio no es demasiado breve, porque la mayoría de los cambios sucesivos de estado civil no ocurren en forma rápida dentro de plazos breves. Para que esta situación se presente se requiere que el período transcurrido haya sido suficientemente largo para 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 trimestral o semestral) 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 en los tres o seis meses posteriores al 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 de preferencias políticas 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.[9]

Porcentajes de fila

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 qué porcentaje o proporción de los sujetos que estaban en un cierto estado ***i*** en la primera observación aparecen en un estado ***j*** en la segunda observación.

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

En la tabla de rotación de una variable dicotómica estas probabilidades aparecen como proporciones o porcentajes de fila:

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

[9] 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.

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). Más adelante veremos que en efecto los principales modelos teóricos utilizan este enfoque, introduciendo nociones 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 determinadas consecuencias del matrimonio, como el tener hijos, y para ello se averigua la presencia de embarazos o hijos recién nacidos entre las personas que han pasado de soltero a casado entre las dos últimas rondas. Obviamente, este grupo incluye personas que se casaron en diferentes momentos dentro del intervalo intermedio, de modo que la posibilidad de haber tenido un hijo será mayor para los que se casaron al inicio del período, respecto a los que se casaron al final del período. En general, 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.

Porcentajes de columna

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: Hace cinco años, ¿dónde residían hace cinco años los sujetos que ahora viven en cada una de las provincias? 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).[10]

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

Primera ronda (t)	Segunda ronda (t+1)		Total
	X=1	X=2	
X=1	0.40	0.25	0.35
X=2	0.60	0.75	0.65
Total	1.00	1.00	1.00

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)		Total
	X=1	X=2	
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 de flujo, 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 que mida el status de pobreza de los hogares en dos fechas en el tiempo, y que en cada fecha se clasifican los

hogares como "pobres" o "no pobres". Estos datos podrían usarse para determinar cuatro tipos de hogares, que sin muchas exigencias 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.

Porcentajes 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.

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 mayormente 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				
Primera observación	**Recuerda propuestas de A**		**No recuerda propuestas de A**		
	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 **no varía en el tiempo que transcurre entre ambas observaciones.** Otro caso podría ser el de una **variable interviniente**, que puede cambiar de valor **entre las dos observaciones**, como por

ejemplo el conocimiento de las propuestas del candidato sobre el cual versa la encuesta.[11]

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.

Las consideraciones que anteceden pueden servir como una introducción general a los datos longitudinales de panel, y a las tablas y notaciones con que se los representa, antes de analizar los diferentes modelos posibles que pueden ser aplicados a estas situaciones. Estos modelos teóricos, como en otras ramas de la investigación científica, generalmente consisten en la postulación de unos mecanismos inobservables que, en caso de estar funcionando, generarían como consecuencia los datos observados, y por ello se dice que "explican" esos datos.. Por ejemplo, en el caso del panel esos procesos inobservables pueden referirse a los cambios subjetivos de los individuos encuestados, o los sucesos (no registrados) ocurridos entre una y otra ronda, cuya consecuencia observable son las respuestas registradas en sucesivas rondas de la encuesta.

–– CONTINÚA EN LA SEGUNDA PARTE ––

Buenos Aires, DIC/2002

por **Héctor Maletta**
Investigador Principal, Área Empleo y Población, IDICSO, USAL.
Email: hmaletta@fibertel.com.ar

[11] Por lo general el conocimiento de las propuestas de un candidato no puede desaparecer: 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".