MÉTODOS CUANTITATIVOS DE LAS CIENCIAS SOCIALES APLICADOS A LOS ESTUDIOS URBANOS Y REGIONALES

519.5
L45m Lemelin, André
Métodos cuantitativos de las ciencias sociales aplicados a los estudios urbanos y regionales / tr. Gay Benoit Frutel. -- Puebla, Pue. : Benemérita Universidad Autónoma de Puebla, Dirección General de Fomento Editorial, 2004.
450 p. ; 21 cm.

ISBN 968-863 793 9

Ciencias Sociales – Métodos estadísticos. 2. Estadística matemática I. t.

MÉTODOS CUANTITATIVOS DE LAS CIENCIAS SOCIALES APLICADOS A LOS ESTUDIOS URBANOS Y REGIONALES

ANDRÉ LEMELIN
INRS-Urbanisation, Culture et Société

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
Dirección General de Fomento Editorial

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA
Enrique Agüera Ibáñez
Rector
Armando Valerdi y Rojas
Secretario General
Lilia Cedillo Ramírez
Vicerrectora de Extensión y Difusión de la Cultura
Ricardo Escárcega Méndez
Director Editorial

Responsabilidad científica:
Andre.lemelin@ucs.inrs.ca
INRS Urbanisation, Culture et Société
ISBN : 2-89575-032-7
Deposito legal: 2003
Bibliothèque nationale du Québec
Bibliothèque nationale du Canada

Con la Colaboración, para la adaptación a los países de habla hispana, de:
Carlos de Castilla Jiménez

Traducción de la versión original en francés: Gay Benoit Frutel

Primera edición, 2004
ISBN 968-863 793 9

Dirección General de Fomento Editorial
2 Norte 1404
Tel. 2 46 85 59
Puebla, Pue.

Impreso y hecho en México
Printed and made in Mexico

ÍNDICE

PREFACIO

Este libro se basa en las clases que imparto desde 1992 en el marco del programa conjunto INRS-UQAM en estudios urbanos. Es justamente esta gran experiencia adquirida a lo largo de estos años, en interacción constante con estudiantes de ciclos superiores, en su mayoría de orientación "cualitativa", la que impulsó el desarrollo de la obra, dándole, así, un toque original. Me propuse lograr un informe de la materia que sea perfectamente riguroso, apoyándome más en la lógica y en la epistemología que en el formalismo matemático. Sin embargo, se exhiben los enunciados matemáticos, casi siempre acompañados de ejemplos numéricos, pues, mi experiencia me enseñó que una buena comprensión de la lógica y de los procedimientos de cálculos no asegura, al estudiante, la capacidad de traducir con aplomo el simbolismo matemático en operaciones numéricas.

Con mis esfuerzos pedagógicos busqué aligerar el método cuantitativo para los jóvenes investigadores, quienes son naturalmente más inclinados hacia el método cualitativo. Sin embargo, me interesa también los estudiantes que tienen gusto y aptitud para los métodos cuantitativos pero suelen descuidar el análisis fundamental y crítico por dejarse absorber por aspectos técnicos.

De hecho pude constatar en más de una ocasión, que hasta los estudiantes en economía son, a menudo, poco conscientes

de los límites de la medición aunque estén muy familiarizados con los métodos econométricos. Lo que les propongo en este libro es un antídoto contra esta tendencia.

Por último, desearía que este libro proporcione una caja de herramientas de análisis de datos a todos los que investigan en ciencias sociales así como a algunos investigadores más experimentados. Hoy en día estas herramientas son indispensables para la investigación aplicada. Y, aunque los ejemplos estudiados sean por lo general del dominio de los estudios regionales y urbanos, los métodos presentados son igualmente pertinentes en geografía, en ciencias políticas, en sociología.

Este libro contiene cuatro partes de alguna manera independientes. La primera, "Cantidad y medición", trata de la naturaleza de la medición, su alcance y sus límites en el caso particular de las ciencias sociales. La reflexión crítica se apoya en la presentación detallada de algunas de las herramientas básicas en análisis de datos: manipulación de tablas de contingencias, análisis de descomposición, construcción de números índices, medición de la concentración (índice de Gini), medición de la disimilitud. Sin embargo, más allá del aprendizaje técnico, el objetivo deseado de esta parte es abrir pistas para la reflexión crítica con el fin anhelado de convertir al estudiante en un lector lúcido y, eventualmente, en un investigador competente capaz de ser crítico y autocrítico, y de interpretar los resultados de la investigación con sumo cuidado.

La segunda parte del libro, titulada "El papel de la estadística en ciencia social", se centra en la lógica y el estatus epistemológico de la inducción estadística. No pretendo presentar el total de las herramientas estadísticas que un joven investigador debe poseer, para esto no faltan muy buenos manuales fáciles de consultar; más bien el objetivo deseado es conseguir que el estudiante adquiera un dominio de los principios fundamentales que le permitirán usar adecuadamente méto-

dos más avanzados o más especializados pero también convertirse en un lector crítico al momento de enterarse de investigaciones basadas en métodos estadísticos.

El análisis de regresión llena el contenido de la tercera parte del libro. Como una herramienta muy versátil del análisis multivariado, se usa con frecuencia en ciencias sociales. Se convierte, así, en una herramienta indispensable para cada investigador. Además, este método puede servir de referencia para abordar métodos más especializados y avanzados. Otra vez, en este caso, se privilegia el enfoque epistemológico.

La cuarta parte del libro se titula "El análisis cuantitativo de datos cualitativos". Es una presentación de métodos propios del análisis de variables categóricas, lo cual es frecuente en ciencias sociales. Se tratan tres temas: el análisis de las tablas de contingencias, la aplicación de la regresión múltiple al análisis de la varianza (variables independientes categóricas) y, finalmente, los modelos a variable dependiente cualitativa (logit y otros).

Para terminar, deseo agradecer a todas las personas que de varias maneras contribuyeron a la realización de esta obra. Agradezco en primer lugar a mi amiga Judith Chaffee, profesora en la Facultad de Economía de la BUAP, quien apoyó el proyecto con tanta energía y generosidad. Agradezco también al Mtro. Carlos de Castilla Jiménez, de la Facultad, quien, además de contribuir para mejorar el libro, me hizo conocer aspectos desconocidos de la cultura mexicana. Finalmente, quiero mencionar el trabajo diligente del estudiante Philippe Rivet para la última versión, así como la ayuda que de cuando en cuando y de improviso me aportó Elena Pou, secretaria del Grupo Interuniversitario de Montreal. Y a todos los demás que no me es posible mencionar aquí, ¡gracias!

INTRODUCCIÓN A LA PRIMERA PARTE

Esta parte del curso trata de un problema difícil: la medición. Problema difícil, aún más si consideramos el área de las ciencias sociales donde los fenómenos estudiados son complejos y no se prestan fácilmente al método experimental.

Un científico debe enfrentarse a este problema difícil tanto por ser un "consumidor" como un productor de investigación.

Como lector, el científico debe mantener alerta su sentido crítico al enterarse de los resultados de la investigación; igual si algunos errores en los datos son lo suficientemente obvios para ser detectados con cierta facilidad, otros requieren de un examen más técnico para ser encontrados. Para poder mostrar eso, tuvimos que incluir algunas fórmulas matemáticas y ejemplos cifrados, lo que vino a complicar la lectura de esta obra que, en un principio, pretende ser muy accesible.

Al momento de dedicarse a la investigación aplicada, el investigador acaba forzosamente por cuantificar, o sea: medir. Esto no es tan sencillo y requiere de una reflexión sobre la medición como parte esencial de la metodología:

- ¿Qué queremos medir (definición del concepto y de sus dimensiones)?
- ¿Cuáles indicadores se pueden o se quieren usar? (¡Qué delicado puede ser contestar esta interrogante!)
- ¿Cuál es el modelo subyacente a la medición?

- ¿Hasta que punto la medición depende del “juicio del investigador”? (Esto no es del todo inaceptable en cuanto se respete las exigencias de transparencia)
- ¿Cuáles son los límites de la medición así como el margen de incertidumbre de los resultados?

Por lo tanto y a través del aprendizaje técnico de algunas herramientas de análisis cuantitativos, el abrir caminos de reflexión crítica es el principal objetivo de esta primera parte. Se invita al estudiante a seguir estos caminos a veces abruptos para convertirse en un lector advertido y, si acaso, en un investigador competente quién sepa practicar la crítica y la autocrítica así como interpretar los resultados de la investigación con toda la prudencia que requiere el rigor científico.

CAPÍTULO 1-1
EL ENFOQUE CUANTITATIVO Y LA MEDICIÓN[*]

Cuantitativo se opone a cualitativo, no tanto porque los dos enfoques sean mutuamente exclusivos, pues más bien son complementarios (para más información sobre los debates ideológicos y metodológicos sobre los enfoques cualitativos y cuantitativos, vea Gilles, 1994, "Introducción", p.1-9). Sin embargo, es en su definición que los dos términos se oponen: algo es cuantitativo cuando se puede medir. Con más precisión, la cantidad se define como la propiedad de algo que se puede medir o contar, de algo susceptible de crecimiento o disminución.

Pero, ¿en qué nos interesa la medición en el método científico en el caso particular de las ciencias sociales? Además, ¿qué es medir?

1-1.1 LA OPERACIONALIZACIÓN DE LOS CONCEPTOS: INDICADORES, VARIABLES Y MEDICIÓN

En ciencias como también en ciencias sociales se formulan las teorías y las hipótesis con la ayuda de conceptos y de relaciones entre conceptos. Un concepto es una idea, una repre-

[*] Referencias: Gilles (1994, Introducción y caps. 1 y 2); Bryman et Cramer, 1990, p. 61-74; Blalock (1972, cap. 2); Lazarsfeld (1971).

sentación mental abstracta y general de un ser, una manera de ser, o de una relación; es, por decir así, un átomo del pensamiento. Gilles (1994, p. 15) resalta el proceso (la operación) creador del concepto, explicando su comprensión y fijando su extensión.[1] Según él, un concepto es "una construcción del pensamiento resultado de una operación (proceso) mediante el cual se individualizan los rasgos que permiten relacionar objetos diferentes o distinguir objetos que, de otra manera, son similares", es decir, mediante el cual se definen criterios que permiten determinar si un objeto u otro forma parte de la extensión de un concepto.

Ejemplo:

> "El consumo de los hogares crece al mismo tiempo que el ingreso". Esta proposición contiene los conceptos "consumo de los hogares" e "ingreso". Las palabras "crecen al mismo tiempo que" expresan una relación entre estos dos conceptos.

Para relacionar las proposiciones teóricas con la realidad, o para confrontar las hipótesis con la observación, es necesario, por medio de indicadores, "operacionalizar" los conceptos, o sea establecer una relación entre los conceptos y la realidad observable. Podemos definir los indicadores como "signos, comportamientos o reacciones observables de manera directa que permiten detectar las magnitudes de un concepto al nivel de la realidad" (Gilles, 1994, p. 27). Las dimensiones son los diferentes componentes de un concepto (Gilles, 1994, p. 24). Regresaremos más tarde sobre esta noción de dimensión.

[1] Se define la extensión lógica como el conjunto de objetos concretos o abstractos a los cuales se aplica un concepto, una proposición (total de casos cuando es verdadero) o una relación (total de sistemas que la verifican). La extensión de un concepto se opone a la comprensión, la cual es el conjunto de caracteres de un concepto. Por ejemplo, el concepto hombre tiene una menor extensión pero una mayor comprensión que mamífero.

Por lo tanto, “operacionalizar” un concepto es asociarle uno o más indicadores, los cuales permiten distinguir con exactitud las variaciones observadas en la realidad con relación al concepto. Distinguir las variaciones significa medir. Así que la operacionalización de un concepto implica medir. Y como los indicadores sirven para medir las variaciones, las medidas asociadas a los conceptos se llaman *variables*.

Notemos que esta relación entre la operacionalización y la medición existe tanto en el enfoque cualitativo como en el enfoque cuantitativo, puesto que el enfoque cualitativo requiere también que se clasifiquen y cuenten los sujetos lo que constituye una operación de medición, como lo veremos después. Respeto a esto, Gilles escribe: “De manera general, los métodos conocidos como cualitativos (historia de vida, análisis de relato, observación participante, entrevista en detalle, estudio de casos,...) usan también la estadística con fines descriptivos” (1994, p. 3). Observa que “los métodos cualitativos y los datos cualitativos no proceden con la misma lógica. Los primeros se basan en una concepción humanista, hermenéutica o interpretativa de las ciencias sociales que se convierten en estas condiciones en ciencias humanas. En cuanto a lo segundo, se entienden como datos que requieren el uso exclusivo de algunas técnicas estadísticas conocidas como “robustas” (1994, p. 3, nota 5). Pronto veremos cómo es posible medir de un cierto modo las propiedades cualitativas; por este motivo, no es absurdo querer hablar del análisis cuantitativo de datos cualitativos (cuarta parte de esta obra).

Refiriéndose al esquema clásico de Lazarsfeld (1971), Gilles (1994, p. 24) resume lo arriba descrito de esta manera: operacionalizar es “someter, por medio del análisis, los conceptos a un proceso para transformarlos en dimensiones, luego en indicadores con el fin de poder observarlos, medirlos y cuantificarlos.”

Ejemplo:

Para operacionalizar el concepto “consumo del hogar”, o sea para medir el consumo de un hogar, se puede tomar el monto declarado en una encuesta hecha a hogares contestando una pregunta como: “La semana pasada, ¿cuánto fue el gasto de cada miembro del hogar?”

Sin embargo, pudiéramos también medir el consumo de un hogar haciendo el cálculo de la diferencia entre sus ingresos y la suma de los impuestos por ingresos y el monto ahorrado en un año.

En general, varios indicadores pueden usarse para ilustrar un concepto. En la investigación la elección de los indicadores es de suma importancia. Los indicadores escogidos deben ser válidos y confiables.

- Un indicador es válido cuando mide bien lo que se quiere medir, o sea cuando da cuenta de las variaciones con relación al concepto que representa. Para determinar la validez de un indicador, debemos, obviamente, definir con claridad el concepto.
- Un indicador es fiable o confiable cuando las variaciones de la medida son variaciones verdaderas.

Ejemplo:

El monto gastado la semana pasada no es, quizás, una medición válida del consumo ya que este monto tal vez incluye los gastos en capital (inversión residencial o inmobiliaria), cuando la definición del concepto teórico “consumo” excluye el pago de adquisición de bienes duraderos.

La repuesta con relación a la suma gastada la semana pasada no es, quizás, fiable porque puede ser que la persona que contestó el cuestionario no estaba enterada de los gastos de los demás miembros del hogar.

Una variable es lo que resulta de la aplicación de un indicador a un conjunto de objetos. Por lo tanto se define una va-

riable por lo que se quiere medir (el concepto), por la manera de medir (el indicador) y por su campo de aplicación (los objetos a los cuales se aplica la medición).

Finalmente, es importante observar la diferencia entre una *variable* y los diferentes valores que puede tomar.

Ejemplos:

> En una encuesta a una muestra de hogares se pregunta lo siguiente: "¿La semana pasada, cuánto gastó el total de personas que compone el hogar?". La repuesta es una variable que toma valores diferentes para cada hogar de la encuesta.

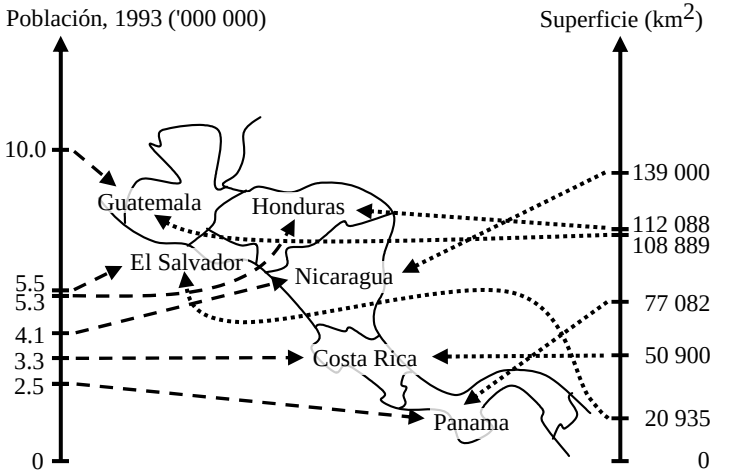
1-1.2 ¿QUÉ ES LA MEDICIÓN?

Medir es comparar. Pero ¿qué más? En el lenguaje cotidiano, medir se define como: "la evaluación de una dimensión comparándola con otra del mismo tipo y tomada como unidad" (*Dictionnaire Larousse de la langue française*, Cédéron, 1996). Veremos que esta definición es muy limitante. *Encyclopaedia Britannica* (Cédéron, 1998) propone: "measurement: the process of associating numbers with physical quantities and phenomena". En el mismo sentido, Gilles (1994, p. 34) dice: "Medir es establecer una correspondencia entre un conjunto constituido por el fenómeno a medir y un conjunto de números que se escogen en función del tipo de fenómeno." Estas dos últimas definiciones, aunque menos limitantes, son sin embargo incompletas en cuanto no especifican las condiciones que debe cumplir una correspondencia numérica para constituir una medida. Es el tema de la teoría de la medición.

Para nuestros fines, guardaremos esta definición: una correspondencia constituye una medición en cuanto permite *comparar* dos *objetos cualesquiera* con relación a una *propiedad dada*.

Por ejemplo, supongamos que la propiedad que queremos medir sea la superficie. Los países, los cuartos para dormir y los pañuelos son objetos para los cuales la propiedad "superficie" está definida. Una correspondencia debe permitir comparar la superficie de dos países, de dos pañuelos, o lo mismo de un país y un pañuelo. Pero, ¿qué es comparar? En el caso de la medición, comparar es determinar si dos objetos son o no son parecidos con relación a una propiedad escogida. Y en caso de que no lo sean, cuál de los dos objetos tiene la propiedad medida en un grado más elevado que el otro.

Medir es una relación....



Fuente de datos: Facultad Latino Americana de Ciencias Sociales FLACSO, Sede Costa Rica, San José, Costa Rica

Podemos formalizar lo que precede de esta manera. Tomemos *A* y *B*, dos objetos cualesquiera (dos países, por ejemplo) que tienen una propiedad a medir (la superficie, por ejemplo). Una medición asocia un número a cada objeto tanto como una función matemática; así, es natural representar la medición del mismo modo. En estas condiciones, conve-

nimos que *f(A)* representa la superficie de *A*, y *f(B)* la superficie de *B*. La comparación examina las relaciones siguientes:

$f(A) = f(B)$
$f(A) \neq f(B)$
$f(A) < f(B)$
$f(A) > f(B)$

Una medición es una correspondencia que permite, por lo menos en una de las relaciones arriba indicadas, determinar si es verdadera o falsa. En el caso de la superficie, se puede establecer una correspondencia entre cada país y el número de kilómetros cuadrados contenidos en sus fronteras, o entre cada pañuelo y la fracción de kilómetros que cubre. Cuando se comparan las cifras de la correspondencia, se puede determinar si es cierto que $f(A) = f(B)$ (*A* y *B* tienen la misma superficie), o $f(A) \neq f(B)$ (*A* y *B* no tienen la misma superficie), o $f(A) < f(B)$ (*A* es más pequeño que *B*), o $f(A) > f(B)$ (*A* es más grande que *B*).

En este ejemplo de la superficie, la medición permite determinar el valor de verdad (cierto o falso) de las cuatro relaciones $=$, $\neq$, $<$ y $>$. Sin embargo, la definición de la medición no exige que se pueda determinar el valor de verdad de las cuatro relaciones. Como ejemplo, supongamos que la propiedad examinada sea la nacionalidad. Se podría definir la correspondencia como sigue:

$f(X) = 0$ si la persona *X* es de nacionalidad costarricense;

$f(X) = 1$ si la persona *X* es de otra nacionalidad centroamericana;

$f(X) = 2$ en los demás casos.

Entonces:

$f(A) = f(B)$ significa que la persona *A* y la persona *B* son de la misma nacionalidad (en esta particular clasificación);

f(A) ≠ *f(B)* significa que la persona *A* y la persona *B* no son de la misma nacionalidad.

Por el contrario, las relaciones *f(A)* < *f(B)* y *f(A)* > *f(B)* no tienen ningún significado. Sin embargo, la correspondencia constituye una medición dentro de lo aceptado: es la medición de la nacionalidad. De cierta manera, se puede, por consiguiente, medir propiedades cualitativas.

Observación: los valores numéricos de la correspondencia no tienen ningún significado y son totalmente arbitrarias. Se podría hasta definir la correspondencia en términos de símbolos que no sean números. Por ejemplo, hubiéramos podido definir:

f(X) = 'CR' si la persona X es de nacionalidad costarricense;

f(X) = 'CA' si la persona es de alguna otra nacionalidad centroamericana;

f(X) = 'OT' en los demás casos.

1-1.3 ESCALAS DE MEDICIÓN Y TIPOS DE VARIABLES

Mientras admitimos que se pueden medir propiedades cualitativas como la nacionalidad de una persona, la medición de tales propiedades parece imperfecta comparada con la medición de propiedades como la superficie o el ingreso. De hecho, en el caso de propiedades como la nacionalidad, no tiene significado determinar si *f(A)>f(B)* o *f(A)<f(B)* es verdadero o falso cuando sí lo tiene para la superficie de un territorio o el ingreso de un hogar; en este último caso la medición es más completa.

Es la razón por la cual distinguimos varios tipos de variables según la escala de medición asociada.[2]

[2] Se encuentra una clasificación similar en Legendre y Legendre (1998, p. 28 y siguientes).

1. Variables categóricas.
2. Variables ordinales.
3. Variables de intervalo.
4. Variables racionales.

1-1.3.1 Variables categóricas

Las variables categóricas ("nominal" en inglés) resultan de la aplicación de una escala de medición que permite solamente determinar las relaciones = y ≠. El valor que toma una variable categórica indica la categoría a la cual pertenece un individuo; por lo tanto, una variable categórica permite clasificar los individuos en grupos. Distinguimos:

- Variables dicotómicas: dos categorías posibles.
- Variables politómicas: más de dos categorías.

Ejemplos:

Sexo (hombre / mujer): Variable categórica dicotómica.

Nacionalidad: Variable categórica politómica (cuando distinguimos más de dos nacionalidades).

Es posible reemplazar una variable politómica por varias variables dicotómicas. Además, algunos métodos de análisis lo exigen. Como ejemplo, consideremos una variable politómica de nacionalidad:

NAT = 0 si la persona X es de nacionalidad costarricense;

NAT = 1 si la persona es de cualquier otra nacionalidad centroamericana;

NAT = 2 en los demás casos.

Podemos reemplazar esta variable por dos variables dicotómicas como:

COR = 1 si la persona X es ciudadana de Costa-Rica; y COR = 0 si no.

CAM = 1 si la persona es ciudadana de un país de América Central otro que Costa-Rica y *CAM* = 0 si no.

Pregunta: ¿por qué solamente dos variable dicotómicas cuando la variable politómica puede tomar tres valores?

1-1.3.2 Variables ordinales

Las variables ordinales resultan de la aplicación de una escala de medición que permite determinar las cuatro relaciones =, ≠,<, y >. Por lo tanto, los valores que toma una variable ordinal para diferentes individuos permite arreglar los individuos en un orden creciente o decreciente con relación a una propiedad medida. Distinguimos los órdenes débiles –incompletos, por clases de equivalencia– y los órdenes completos.

Ejemplos:

- Número de puntos obtenidos en un test de aptitudes (orden completo: si dos sujetos obtienen el mismo número de puntos, la medida indica que poseen el mismo grado de aptitudes con relación al test).
- Variable definida por: 1 si el estudiante aprueba un examen dado y 0 si reprueba (orden débil: si dos estudiantes aprobaron no significa que son de fuerzas iguales).

Las medidas ordinales se definen "con aproximación a una transformación monótona creciente", es decir que no cambia la medida en caso que se le aplique a la variable una transformación matemática siempre y cuando no se cambie el orden numérico de los valores. Por ejemplo, podríamos sustituir el número de puntos por el logaritmo del número de puntos, o por el cuadrado del número de puntos, o podríamos agregar un millón de puntos a todos los sujetos.

1-1.3.3 Variables de intervalo

Las variables de intervalo son similares a las variables ordinales pero, además de permitir de arreglar los individuos en un orden creciente o decreciente, permite comparar las diferencias entre individuos.

Ejemplo:

> La temperatura: si hace –15°C en Montreal, +24°C en San José (Costa Rica) y +18°C en Miami, podemos decir que la diferencia de temperatura es menos grande entre San José y Miami (6°C) que entre Miami y Montreal (33°C). Este tipo de comparación no tiene ningún sentido con una variable ordinal.

De manera formal, las variables de intervalo son el resultado de la aplicación de una escala de medición donde las diferencias entre los valores son también medidas ordinales; la escala de medición permite determinar un valor de verdad para cada una de las relaciones siguientes:

$f(A) - f(B) = f(C) - f(D)$
$f(A) - f(B) \neq f(C) - f(D)$
$f(A) - f(B) < f(C) - f(D)$
$f(A) - f(B) > f(C) - f(D)$

Con una variable de intervalo, el 0 de la escala de medición es arbitrario pero las transformaciones de la escala deben preservar la comparación entre las diferencias. Es el motivo por el cual se define las escalas de intervalo “con aproximación a una transformación lineal”. Por ejemplo, saltamos de la escala Celsius a la escala Fahrenheit por medio de una transformación lineal

$F = 32 + 1.8 \times C$

1-1.3.4 Variables racionales

Las variables racionales (conocidas también como variables proporcionales) son similares a las variables de intervalo, excepto que, en el caso de las variables racionales, existe un cero natural lo que permite que la razón entre dos valores tenga sentido (racional viene de *ratio*, razón).

Ejemplo:

> El ingreso es una variable racional. Si una persona gana $60,000, se puede decir que gana lo doble que una persona que ingresa $30,000. Sin embargo, no tiene sentido pretender que hace dos veces más calor a 20°C que a 10°C (20°C = 68°F y 10°C = 50°F).

De manera formal, las variables racionales son el resultado de la aplicación de una escala de medición donde las fracciones entre los valores son también medidas ordinales. La escala de medición permite determinar un valor de verdad para cada una de las relaciones siguientes:

$$\frac{f(A)}{f(B)} = \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} \neq \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} < \frac{f(C)}{f(D)}$$

$$\frac{f(A)}{f(B)} > \frac{f(C)}{f(D)}$$

Si observamos nuevamente la definición del *Larousse* como “la evaluación de una dimensión comparándola con otra del mismo tipo y tomada como unidad”, podemos constatar, de hecho, que se aplica exclusivamente a las escalas de medición racionales. La definición del *Larousse* es, por lo tanto, restrictiva.

Pasa a menudo que los valores observados de variables racionales o de intervalo se agrupan en clases. Por ejemplo, una variable "ingreso" puede tener la forma siguiente:

ING = 1 si ingreso < 10 000 $

ING = 2 si 10 000 $ ≤ ingreso < 25 000 $

ING = 3 si 25 000 $ ≤ ingreso< 50 000 $

ING = 4 si ingreso ≥ 50 000 $

Una variable de este tipo es una variable ordinal que define un orden débil. El acto de agrupar los valores observados en clases implica transformar una variable racional (o de intervalo) en una variable ordinal de orden débil. Pasamos, así, a una escala de medición más "primitiva" y perdemos informaciones. Por lo tanto, es preferible, en cuanto sea posible, usar los datos en su forma original.

1-1.3.5 Escala de medición y métodos cuantitativos

Existen métodos de análisis cuantitativos adaptados a todos los tipos de variables. Por lo tanto, se pueden usar métodos *cuantitativos* para analizar datos *cualitativos*, en cuanto puedan ser medidos con variables categóricas u ordinales.

1-1.4 TIPOS DE DATOS

Esta parte del curso trata de los métodos cuantitativos de análisis de datos. Sin embargo, la calidad de la análisis depende ante todo de la calidad de los datos analizados Los datos nunca son "perfectos" y el analista competente debe adaptar sus métodos a la calidad de los datos recibidos.

Podemos distinguir tres tipos de datos:

1. Los datos primarios.
2. Los datos secundarios no publicados.
3. Los datos secundarios publicados.

Cada tipo de datos está sujeto a problemas de calidad específicos.

1-1.4.1 Datos primarios (encuestas)

Se hace el control de calidad durante todas las etapas:

- preparación de los instrumentos de adquisición de datos (cuestionarios).
- levantamiento.
- codificación.
- captura, validación, corrección y organización
- evaluación ex post de la calidad

1-1.4.2 Datos secundarios no publicados

Este tipo de datos (por ejemplo, datos sacados de registros de evaluación municipal para el impuesto sobre los bienes inmuebles como el predial) se colectan a menudo para fines administrativos u otros, fines que difieren de los de la investigación: sucede, a menudo, que se definen mal los conceptos o que éstos no son los que buscamos medir (las variables creadas a partir de estos datos no son perfectamente válidas).

El control de calidad de los datos secundarios no publicados ocasiona a menudo problemas parecidos a los encontrados al momento de tratar de datos primarios. Sin embargo, en caso de datos secundarios, el analista no puede, por sí solo, cuidar del control de calidad durante todas las etapas.

1-1.4.3 Datos secundarios publicados

El uso adecuado de datos secundarios publicados exige de tomar en cuenta toda la información pertinente que acompaña los datos (metadatos).

- Definiciones y conceptos
- Métodos de cosecha y de compilación
- Evaluación de calidad por el emisor

- Credibilidad da las fuentes.

Además de depender de la calidad de los datos, los errores en el tratamiento de datos anterior a la aplicación de los métodos de análisis comprometen la misma calidad de análisis.

Ejemplos:

> Error de variable durante la extracción de datos de un banco de datos (masa salarial en lugar de salario por hora).
>
> Error de programación durante el apareamiento (la fusión) de dos archivos ("merge"): desdoblamiento (repetición) de algunas observaciones.
>
> Error de fórmulas en un tabulador (direcciones relativas y absolutas...); estos errores son a menudo el resultado de una operación de "copiar-pegar".

Si no encuentra errores en sus datos es que no busca muy bien.

1-1.5 La matriz, estructura fundamental de los datos

Para que los datos sean de utilidad es necesario organizarlos de tal manera que sepamos a qué se refiere cada número. Existen varias maneras de organizar los datos mismos si bien todas deben conformarse a la estructura fundamental de los datos. Esta estructura fundamental es una matriz o una tabla donde, por convenio:

- las columnas suelen corresponder a diferentes variables (características, propiedades, atributos, indicadores, descriptores...);
- las líneas suelen corresponder a diferentes observaciones (casos, individuos, objetos...)

Puede ser el caso que las observaciones se refieran a momentos o periodos sucesivos: hablamos, entonces, de series cronológicas o temporales. De la misma manera, cuando las

observaciones se refieren a diferentes lugares de un conjunto geográfico (países de un continente, ciudades o regiones de un país, colonias de una ciudad, barrios...) podemos hablar de series espaciales. Los datos espaciales no son siempre series completas, o sea que tenga una única observación para cada lugar encontrado en un espacio dado. De todos modos, que sea una serie completa o no, los datos son georeferidos cuando contienen una o varias variables que permiten situar cada una de las observaciones en un espacio geográfico.

La estructura de matriz fundamental se generaliza a más de dos dimensiones[3] cuando algunas variables son categóricas (variables de clasificación). Como ejemplo, supongamos que hayamos realizado una encuesta a una muestra de personas pidiéndoles su profesión entre otras variables. En este caso, el resto de los datos puede ser organizado en varias tablas de dos dimensiones, una por cada profesión. Si se superponen estas tablas, la organización de los datos es un cubo con capas sucesivas que corresponden a las diferentes profesiones, las líneas corresponden a los diferentes encuestados y las columnas a las otras variables. Con más de una variable categórica, es posible imaginar un "hipercubo" de datos de cuatro o más dimensiones. La más apropiada representación mental depende del tipo de análisis que queramos emprender.

Cabe mencionar que la distinción entre observaciones y variables no es totalmente hermética. Puede pasar que las observaciones y las variables se intercambien cuando los observadores corresponden a las diferentes categorías de una variable de clasificación, mientras que las variables son los atributos que se refieren a las diferentes categorías de otras variables de clasificación.[4] Como ejemplo, tomemos el caso

[3] Cuidado: se usa la misma palabra "dimensiones" en un contexto diferente para enseñar las dimensiones de un concepto.

[4] Tal y como lo demuestra el ejemplo que se exhibe más adelante, tal ambivalencia se debe a un primer tratamiento de datos que fueron luego dispuestos en tabla de contingencia o en tabla de análisis de varianza.

de la tabla de números de empleos por ramo de actividad y por ciudad en una región escogida. En estas condiciones, podemos considerar:

- que cada observación corresponda a una ciudad, y las variables sean los números de empleos de los diferentes ramos de actividades en esta ciudad.
- que cada observación corresponda a un ramo de actividad, y que las variables sean los números de empleos de cada ramo en esta ciudad.

Nuevamente, en este caso, la representación mental que privilegiamos depende de los análisis que queremos emprender.

Si regresamos al modelo de organización elemental de una tabla de dos dimensiones, distinguimos dos puntos de vista o tipos de análisis, según nos fijemos en las relaciones entre las observaciones o las relaciones entre las variables (Jayet, 1993, pp. 1-2; Legendre y Legendre, 1998, p. 248, usan una terminología de Cattell, 1952, y separan el análisis "modo R" que analiza las relaciones entre los descriptores, y el análisis "modo Q", que analiza las relaciones entre los objetos). Esta distinción permite clasificar los tipos de análisis y los métodos que se les asocia (ver la tabla que sigue).

Punto de vista "horizontal" entre las variables

- Para resumirlas, combinar varias variables en una construcción de números índices
- Comparar dos variables: medición de la similitud/disimilitud
- Estudiar las relaciones de dependencia
  - Entre dos variables: correlación, regresión simple
  - Entre una variable dependiente y varias variables independientes: regresión múltiple y otros métodos multivariados

Punto de vista “vertical”: entre las observaciones o los objetos

- Caracterizar la distribución de una variable: medición de la desigualdad o de la concentración, métodos estadísticos univariados
- Estudiar las relaciones entre las observaciones: medición y modelización del crecimiento de las series temporales, análisis de la autocorrelación (temporal o espacial)
- Comparar dos objetos: medición de la similitud/disimilitud

CAPÍTULO 1-2
LA INTERPRETACIÓN DE LAS MAGNITUDES

Más allá del problema de la medición, surge el problema de la interpretación de las magnitudes, o sea del sentido que se da a los números. Trataremos de tres temas. Primero, examinaremos dos técnicas numéricas de uso frecuente para facilitar la interpretación de las magnitudes: la construcción de una medición relativa y el análisis de descomposición. En los dos casos, se ilustrará el método con una técnica de gran manejo en las ciencias regionales y en estudios urbanos. Los límites de estas herramientas serán el objeto de un especial énfasis. Luego, hablaremos de la medición del crecimiento o, de manera más general, del modo de resumir la evolución de una magnitud en el tiempo.

1-2.1 MEDICIONES RELATIVAS: EL EJEMPLO DEL COCIENTE DE LOCALIZACIÓN

- En 1600, Inglaterra contaba con una población de aproximadamente, cinco millones de habitantes.[5] ¿Se

[5] Braudel (1979, p. 49) compara Inglaterra con Francia (que tenía entonces 20 millones) y concluye que Francia estaba superpoblada, dado que "Si ambos países hubieran crecido al ritmo promedio del mundo, Inglaterra tendría 40 millones de habitantes hoy, y Francia 160", lo que es muy dife-

puede decir que, en ese entonces, Inglaterra estaba densamente poblada?

- En Berlín, en los años 1800, una familia destinaba 44.2% de sus ingresos para comprar pan.[6] ¿Era normal para la época?
- En el siglo XV, y en el siglo XVI hasta el año de 1543, el precio del trigo traducido en horas de trabajo de mano de obra equivalía a menos de 100 horas el quintal (1 quintal = 100 kg), pues arriba de 100 hasta 1883 aproximadamente.[7] ¿Qué significa esto con relación al nivel de vida?

Dicho de otra manera, "¿es mucho?": es la pregunta que nos hacemos, por lo general, al enterarnos de una cifra en un campo que no nos es familiar. Como vemos en los ejemplos citados arriba, no son las unidades de medición las que causan el problema; es la falta de puntos de referencia. Por lo tanto, si medir es comparar, la interpretación de las magnitudes requiere de una "metacomparación", o sea, de una comparación con una magnitud que tiene sentido para el observador con el fin de tener una perspectiva de los datos y de entender el orden de magnitud de las cifras.

La representación gráfica con comparación es, con seguridad, el método más usado para dar al observador un punto

---

rente de las cifras reales : en 2001, Francia contaba con 59.6 millones de habitantes y el Reino Unido 58.9 (PNUD, 2003).

[6] Braudel (1979, p. 142). El total de los alimentos representa 72.7% del presupuesto. Entonces el pan cuenta por 60.8% de los gastos alimenticios de la familia, "una proporción enorme dado el precio relativo de las cereales", El autor hace la comparación con los gastos alimenticios del Parisino en 1788 y 1854: "El trigo, primera fuente de energía, no logra más que el tercer rango de los gastos, después de la carne y del vino (cada uno 17% de los gastos totales)" (pp. 143-144).

[7] Braudel (1979, p. 145) explica: "Un trabajador cumple *aproximadamente* 3 000 horas de trabajo cada año; su familia (4 personas) come *aproximadamente* 12 quintales por año... Pasar el límite de 100 horas por un quintal, siempre es grave; pasar las 200 da la alerta; a las 300, es la hambruna".

de referencia y permitirle interpretar las magnitudes. Sobre este tema, se invita al lector a consultar el pasaje "Del buen y mal uso de las gráficas" en Wonnacott y Wonnacott (1992, pp. 61-69).

De igual manera, es importante conocer el campo de dominio de una magnitud para poder interpretarla correctamente. Regresaremos sobre este tema cuando examinamos las mediciones de desigualdad y las mediciones de similitud/disimilitud.

Sin embargo, es a menudo necesario formalizar más adelante esta metacomparación construyendo una medición relativa, o sea, la razón entre dos valores. Es lo que hace el cociente de localización.

1-2.1.1 El cociente de localización[*]

También conocidos como *índices de concentración relativa*, los cocientes de localización son mediciones relativas de la importancia relativa del empleo en una rama de actividad en una ciudad o una región[8]. Por lo tanto, se aplican a datos de una tabla de empleo por rama y por ciudad o región. Mostramos, aquí, un ejemplo ficticio:

Empleo por zona y por rama

Rama	B1	B2	B3	Total
Zona				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

[*] Referencias: Page-Patton, 1991, ch. 14; Polèse, 1994, pp. 128-129.

[8] Pertenecen a la categoría de lo que Jayet (1993, p. 18) llama los "indicadores de especificidad".

Este tipo de tabla se llama una tabla de contingencia (vea Gilles, 1994, sección 6.3). Queremos ser capaces de contestar a preguntas como las que siguen: "¿48 empleos de la rama *B1* en la zona *Z1*, es poco?" o "¿325 empleos de la rama *B2* en la zona *Z1*, es mucho?".

En una primera etapa, podemos examinar las distribuciones.

Distribución del empleo de las ramas entre zonas

Rama	*B1*	*B2*	*B3*	Total
Zona				
Z1	0.400	0.542	0.598	0.550
Z2	0.225	0.308	0.308	0.300
Z3	0.375	0.150	0.094	0.150
Total	1.000	1.000	1.000	1.000

¿48 empleos de la rama *B1* en la zona *Z1*, es poco? El examen de la distribución del empleo entre las zonas muestra que estos 48 empleos representan 40% del total de empleo de la rama *B1*; es en la zona *Z1* donde encontramos el más grande número de empleo en esta rama. Por el contrario, la zona *Z1* contiene 55% del empleo sin distinción de ramas. Considerando la talla de la zona *Z1*, 48 empleos no son muchos.

Distribución del empleo de las zonas entre ramas

Rama	*B1*	*B2*	*B3*	Total
Zona				
Z1	0.073	0.492	0.435	1.000
Z2	0.075	0.514	0.411	1.000
Z3	0.250	0.500	0.250	1.000
Total	0.100	0.500	0.400	1.000

¿325 empleos de la rama *B2* en la zona *Z1*, es mucho? El examen de la distribución del empleo entre las ramas muestra que estos 325 empleos representan cerca de la mitad (49%)

del empleo en la zona *Z1*. Pero, en la economía total, la rama *B2* representa la mitad; por lo tanto, 325 empleos es "normal".

Calcular el cociente de localización es una manera de formalizar este tipo de razonamiento.

En el primer caso (48 empleos de la rama *B1* en la zona *Z1*) se mide la importancia de esta zona con la fracción de esta zona en el total de empleos de la rama (48/120 = 0.4 o 40%).[9] Sin embargo, es necesario referirse al porcentaje correspondiente a las actividades totales (660/1200 = 0.55 o 55%) para poder interpretar este 40%. La medición relativa que usamos de manera implícita para apreciar la importancia de *Z1* para *B1* es la razón 0.40/0.55: esto es un cociente de localización. En el segundo caso (325 empleos de la rama *B2* en la zona *Z1*), procedimos de manera análoga. Se mide la importancia de la rama con la fracción de esta rama en el empleo total de la zona (325/660 = 0.492 o 49%) Sin embargo, es necesario referirse al porcentaje del total de las zonas (660/1200 = 0.5 o 50%) para interpretar este 49%. La medición relativa que usamos de manera implícita para apreciar la importancia de *B2* para *Z1* es la razón 0.49/0.50: esto es también un cociente de localización. De esta manera, el cociente de localización compara dos puntos correspondientes en dos distribuciones (dos puntos correspondientes y no dos distribuciones; las distribuciones son objetos multidimensionales. Veremos en el capítulo 1-5 cómo se puede comparar).

Tabla de contingencia: simbología e identidades fundamentales

Antes de llevar a cabo una presentación más formal de los cocientes de localización, establecemos una simbología apro-

[9] Podemos ya hablar de una medición relativa considerando que es el resultado de la razón entre dos números comparables.

piada y recordamos las identidades fundamentales que se verifican en una tabla de contingencia del tipo del empleo por zona y por rama.

Simbología

x_{ij}	Número de empleos de la rama *j* en la zona *i*
$x_{\bullet j} = \sum_i x_{ij}$	Número total de empleos de la rama *j*
$x_{i\bullet} = \sum_j x_{ij}$	Número total de empleos en la zona *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	Número total de empleos de todas las ramas en todas las zonas
$p_{ij} = x_{ij} / x_{\bullet\bullet}$	Fracción del empleo total global que pertenece a la rama *j* y situado en la zona *i*
$p_{\bullet j} = \sum_i p_{ij}$	Fracción del empleo total global que pertenece a la rama *j*
$p_{i\bullet} = \sum_j p_{ij}$	Fracción del empleo total global situado en la zona *i*
$p_{j/i\bullet} = p_{ij} / p_{i\bullet}$	Fracción del empleo total de la zona *i* que pertenece a la rama *j*
$p_{i/\bullet j} = p_{ij} / p_{\bullet j}$	Fracción del empleo total de la rama *j* situado en la zona *i*

Hemos de caer en la cuenta de que estos símbolos $p_{\bullet j}$ y $p_{i\bullet}$ son probabilidades marginales: $p_{\bullet j}$ es la probabilidad que un empleo tomado al azar entre los $x_{\bullet\bullet}$ empleos censados pertenezca a la rama *j* ; $p_{i\bullet}$ es la probabilidad de que un

empleo tomado al azar esté en la zona *i*. De la misma manera, hemos de caer en la cuenta de que $p_{j/i\bullet}$ y $p_{i/\bullet j}$ son probabilidades condicionales: $p_{j/i\bullet}$ es la probabilidad que un empleo tomado al azar en la zona i pertenezca a la rama *j*; $p_{i/\bullet j}$ es la probabilidad de que un empleo tomado al azar de la rama *j* esté en la zona *i*.

Se deducen naturalmente las identidades siguientes:

$$p_{\bullet j} = \sum_i p_{ij} = \sum_i {x_{ij}}/{x_{\bullet\bullet}} = {x_{\bullet j}}/{x_{\bullet\bullet}} \text{ y}$$

$$p_{i\bullet} = \sum_j p_{ij} = \sum_j {x_{ij}}/{x_{\bullet\bullet}} = {x_{i\bullet}}/{x_{\bullet\bullet}}$$

$$p_{j/i\bullet} = {p_{ij}}/{p_{i\bullet}} = {x_{ij}}/{x_{i\bullet}} \text{ y } p_{i/\bullet j} = {p_{ij}}/{p_{\bullet j}} = {x_{ij}}/{x_{\bullet j}}$$

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$$

$$\sum_j p_{j/i\bullet} = {\sum_j p_{ij}}/{p_{i\bullet}} = 1 \text{ y } \sum_i p_{i/\bullet j} = {\sum_i p_{ij}}/{p_{\bullet j}} = 1$$

El cociente de localización: formalización

Se define el cociente de localización tanto a partir de la distribución del empleo entre ramas como a partir de la distribución entre zonas. A partir de la distribución entre zonas, se define el cociente de localización de la actividad *j* en la zona *i* como sigue:

$$QL_{ij} = \frac{\text{Fracción del empleo total del ramo } j \text{ ubicado en la zona } i}{\text{Fracción del empleo total global ubicado en la zona } i}$$

$$QL_{ij} = \frac{p_{i/\bullet j}}{p_{i\bullet}} = \frac{x_{ij} \big/ x_{\bullet j}}{x_{i\bullet} \big/ x_{\bullet\bullet}}$$

Por ejemplo

$$QL_{21} = 0.225 / 0.300 = 0.750$$

De manera equivalente, a partir de la distribución entre ramas, se define el cociente de localización de la actividad *j* en la zona *i* como sigue:

$$QL_{ij} = \frac{\text{Fracción del empleo total de la zona } i \text{ que pertenece al ramo } j}{\text{Fracción del empleo total global que pertenece al ramo } j}$$

$$QL_{ij} = \frac{p_{j/i\bullet}}{p_{\bullet j}} = \frac{x_{ij} \big/ x_{i\bullet}}{x_{\bullet j} \big/ x_{\bullet\bullet}}$$

Por ejemplo:

$$QL_{21} = 0.075 / 0.100 = 0.750.$$

No es casual que los dos cálculos alojen el mismo resultado, ya que

$$\frac{x_{ij} \big/ x_{\bullet j}}{x_{i\bullet} \big/ x_{\bullet\bullet}} = \frac{x_{ij} \big/ x_{i\bullet}}{x_{\bullet j} \big/ x_{\bullet\bullet}} = \frac{x_{ij}\, x_{\bullet\bullet}}{x_{i\bullet}\, x_{\bullet j}}$$

En nuestro ejemplo:

$$QL_{21} = \frac{x_{21} \big/ x_{\bullet 1}}{x_{2\bullet} \big/ x_{\bullet\bullet}} = \frac{x_{21} \big/ x_{2\bullet}}{x_{\bullet 1} \big/ x_{\bullet\bullet}} = \frac{x_{21}\, x_{\bullet\bullet}}{x_{2\bullet}\, x_{\bullet 1}}$$

$$QL_{21} = \frac{27 \big/ 120}{360 \big/ 1200} = \frac{27 \big/ 360}{120 \big/ 1200} = \frac{27 \times 1200}{360 \times 120} = 0.75$$

Hay efectivamente equivalencia.

Cocientes de localización

Rama	*B1*	*B2*	*B3*
Zona			
Z1	0.727	0.985	1.087
Z2	0.750	1.028	1.028
Z3	2.500	1.000	0.625

Los cocientes de localización pueden tomar valores entre el 0 y el infinito.[10] Cuando $x_{ij} = 0$, el cociente de localización alcanza el valor mínimo: $QL_{ij} = 0$. Por otra parte, alcanza el valor más elevado posible cuando $x_{ij} = x_{\bullet j} = x_{i\bullet}$, o sea, cuando el total de los empleos de la actividad *j* se encuentra

[10] Algunos autores normalizan el cociente de localización con la transformación $\frac{QL_{ij} - 1}{QL_{ij} + 1}$. Esta razón varia de –1 a +1.

en la zona *i* y que no existe otro tipo de actividad en esta zona: en estas condiciones, $QL_{ij} = x_{\bullet\bullet} / x_{ij}$.

En la expresión anterior, $x_{ij} = x_{\bullet j} = x_{i\bullet} \geq 1$, de lo contrario, la rama no existiría. Por lo tanto, el valor máximo de QL_{ij} es $x_{\bullet\bullet}$: este valor no tiene límite teórico, razón por la cual se dice que el cociente de localización puede tomar valores hasta el infinito; sin embargo, en la práctica, son los valores observados que limitan el máximo.

El punto de referencia natural para interpretar el cociente de localización, es 1.0. Además, las fórmulas anteriores mostraron que se pueden hacer dos lecturas del cociente de localización.

- Con relación a la primera lectura, si $QL_{ij} > 1$, se dice que la actividad *j* es *relativamente concentrada*[11] en la zona *i*; decimos "relativamente" en comparación con otras actividades, y esto porque la fracción del empleo en la zona *i* es *más* importante para la actividad *j que para las otras actividades*; con más precisión, decimos que la zona *i* es una zona de concentración relativa para esta actividad porque puede haber otras zonas de concentración relativa de esta misma actividad.

  Por ejemplo en $QL_{23} = 1.028$ la actividad *B3* es relativamente concentrada para la zona *Z2*; sin embargo, no es la zona *Z2* que tenga el más empleos en esta rama sino más bien la zona *Z1*.

- Con relación a la segunda lectura, si $QL_{ij} > 1$, se dice también que la zona *i* es *relativamente* especializada en la actividad *j*; decimos "relativamente" en compara-

[11] De aquí la expresión adecuada, "índice de concentración relativo", para designar el cociente de localización.

ción a las otras zonas porque, en esta zona, la actividad *j* ocupa un lugar *más* importante *que en otras partes*.

Por ejemplo en $QL_{31} = 2.500$ la zona *Z3* es relativamente especializada en la actividad *B1*; sin embargo, no es en la rama *B1* donde encontramos el más grande número de empleos de la zona *Z3* sino más bien en la rama *B2*.

- Por el contrario, si $QL_{ij} < 1$, se dice que la actividad *j* es relativamente menos presente en la zona *Z1* que en otras partes, o sea que la actividad *j* no es relativamente concentrada en la zona *i* y que la zona *i* no es relativamente especializada en la actividad *j*.

Por ejemplo, $QL_{12} = 0.985$ la rama *B2* es relativamente menos presente en la zona *Z1* aunque sea la rama con el más grande número de empleos en esta zona, y aunque sea en *Z1* que esta rama tenga el más grande número de empleos (325 es el número más grande de su línea y de su columna).

Los ejemplos citados nos muestran la importancia del adverbio "relativamente" en los enunciados interpretativos anteriores. De manera más general, si la zona *i* es pequeña comparada con otras zonas ($p_{i\bullet}$ pequeño), lo mismo cuando $QL_{ij} > 1$, es posible que la fracción del empleo de la actividad *j* en la zona *i* ($p_{i/\bullet j}$) no sea importante. En efecto,

$$QL_{ij} = \frac{\text{Fracción del empleo total del ramo } j \text{ ubicado en la zona } i}{\text{Fracción del empleo total global ubicado en la zona } i}$$

$$QL_{ij} = \frac{p_{i/\bullet j}}{p_{i\bullet}} = \frac{x_{ij} \diagup x_{\bullet j}}{x_{i\bullet} \diagup x_{\bullet\bullet}}$$

de tal manera que si $p_{i\bullet}$ es pequeño, es posible que $QL_{ij} > 1$, mismo si $p_{i/\bullet j}$ es pequeño, y en cuanto $p_{i\bullet}$ sea todavía más pequeño. En tales condiciones, sería erróneo pretender que la actividad *j* es concentrada (en términos *absolutos*) en la zona *i*.

Igualmente, si la actividad *j* es de menor importancia en la economía ($p_{\bullet j}$ pequeño), mismo cuando $QL_{ij} > 1$, es posible que la fracción del empleo de la actividad *j* con relación al total de empleos en la zona i ($p_{j/i\bullet}$) no sea importante. En efecto,

$$QL_{ij} = \frac{\begin{array}{c}\text{Fracción del empleo total de la} \\ \text{zona } i \text{ que pertenece al ramo } j\end{array}}{\begin{array}{c}\text{Fracción del empleo total global} \\ \text{que pertenece al ramo } j\end{array}}$$

$$QL_{ij} = \frac{p_{j/i\bullet}}{p_{\bullet j}} = \frac{x_{ij} \diagup x_{i\bullet}}{x_{\bullet j} \diagup x_{\bullet\bullet}}$$

de tal manera que si $p_{\bullet j}$ es pequeño, es posible que $QL_{ij} > 1$o mismo si $p_{j/i\bullet}$ es pequeño, y en cuanto $p_{\bullet j}$ sea todavía más pequeño. En tales condiciones, sería erróneo pretender que la zona *i* es especializada (en términos absolutos) en la actividad *j*.

Con una interpretación correcta, los cocientes de localización pueden servir para el análisis descriptivo de datos de empleo (vea Lemelin y Polèse, 1993).

Nota: es matemáticamente imposible que $QL_{ik} > 1$ para todas las zonas *i* al mismo tiempo (o, de manera simétrica que $QL_{ik} < 1$ para todas las zonas *i* al mismo tiempo). En efecto, ya que $QL_{ik} = \frac{p_{i/\bullet k}}{p_{i\bullet}}$, esto implicaría que $p_{i/\bullet k} > p_{i\bullet}$ para cada *i*, de tal manera que tendríamos $\sum_i p_{i/\bullet k} > \sum_i p_{i\bullet}$, algo imposible dado que las dos sumas deben sumar 1.

De igual manera, es matemáticamente imposible que $QL_{kj} > 1$ para todas las actividades *j* al mismo tiempo. En efecto, ya que $QL_{kj} = \frac{p_{j/k\bullet}}{p_{\bullet j}}$, esto implicaría que $p_{j/k\bullet} > p_{\bullet j}$ para cada *j*, de tal manera que $\sum_j p_{j/k\bullet} > \sum_j p_{\bullet j}$, algo imposible dado que los dos términos de la comparación deben sumar 1.

Puede ser de gran utilidad recordar estas reglas: llegar a tales resultados implica que se cometieron errores en los cálculos.

1-2.1.2 Estimación del empleo exportador por medio del cociente de localización

Usamos también los cocientes de localización en la teoría de la base económica (Polese, 1994, p. 125-138) para estimar el empleo "exportador" (para ver un ejemplo, vea Polese y Stafford, 1982). Dada la escasez de datos sobre los intercambios interregionales, esta posibilidad es atrayente. Sin embargo, la estimación del empleo exportador por medio del cociente de

localización se basa en hipótesis más bien restrictivas (Isserman, 1990, p. 157):

1. La productividad del trabajo es igual entre ciudades y regiones.
2. La absorción (uso local) del producto por empleo en la economía local es igual entre ciudades y regiones.[12]
3. No hay importaciones o exportaciones netas en todo el país.
4. La demanda local se aprovisiona primero con los productores locales; esto implica que no haya flujos cruzados entre ciudades o entre regiones ("cross-hauling").

En estas condiciones, podemos interpretar el excedente del cociente de localización con relación a 1.0 como una medición del empleo exportador. Con más precisión, se puede estimar el empleo "exportador" de la rama *j*, EXP_{ij}, lo cual pertenece a la base económica de la región, con la fórmula:

$$EXP_{ij} = \begin{cases} x_{ij} \dfrac{\left(QL_{ij} - 1\right)}{QL_{ij}}, \text{si } QL_{ij} > 1 \\ 0 \text{ si no} \end{cases}$$

Por ejemplo, $EXP_{31} = 45 \times \dfrac{2.5 - 1}{2.5} = 27$ de los 45 empleos de la rama *B1* en la zona *Z3*

La fracción de x_{ij} que pertenece al empleo exportador es la fracción de QL_{ij} que excede 1. Cuando $QL_{ij} < 1$, no hay exportación de la actividad *j* desde la región *i* y, por lo tanto, el empleo exportador es nulo.

Para entender mejor la significación de este cálculo, sustituimos QL_{ij} y simplificamos para obtener

[12] Isserman (1999 y Norcliffe (1993) usan el término "consumo" para designar tanto la demanda final como la demanda intermediaria. El término "absorción" parece más exacto.

$$EXP_{ij} = x_{ij} - \left(x_{i\bullet} \big/ x_{\bullet\bullet} \right) x_{\bullet j} = x_{ij} - p_{i\bullet} x_{\bullet j}$$, si

$x_{ij} > p_{i\bullet} x_{\bullet j}$

Por ejemplo: $EXP_{31} = 45 - \left(180 \big/ 1200\right) 120 = 27$

Vemos, así, que el empleo exportador es la diferencia entre el valor observado x_{ij} y el valor hipotético que tomaría la cifra del empleo si la región *i* produjera solamente "su parte" de *j* (o sea $p_{i\bullet}$, que implicaría que el cociente de localización QL_{ij} fuera igual a 1).

Para ver cómo intervienen las hipótesis enunciadas anteriormente, volvamos a escribir la fórmula en la forma siguiente:

$$EXP_{ij} = \left[\left(x_{ij} \big/ x_{\bullet j} \right) - \left(x_{i\bullet} \big/ x_{\bullet\bullet} \right) \right] x_{\bullet j} = \left(p_{i/\bullet j} - p_{i\bullet} \right) x_{\bullet j}$$

si $p_{i/\bullet j} > p_{i\bullet}$

Por ejemplo: $EXP_{31} = \left[\left(45 \big/ 120\right) - \left(180 \big/ 1200\right) \right] 120 = 27$

La primera hipótesis se refiere a la razón $p_{i/\bullet j}$ o $x_{ij} \big/ x_{\bullet j}$; esta razón es la participación de la región *i* en el empleo de la actividad *j*; la primera hipótesis permite considerar esta razón como una aproximación de la parte de la región en la producción del bien *j*.

La segunda hipótesis se refiere a la razón $p_{i\bullet}$ o $x_{i\bullet} \big/ x_{\bullet\bullet}$; esta razón es la parte de la región *i* en el empleo total; la segunda hipótesis permite considerar esta razón como una aproximación de la parte de la región en el uso (absorción) del bien *i*.

Las otras dos hipótesis permiten interpretar la diferencia como la parte del empleo nacional de la rama *j* que pertenece

a la base económica de la región *i*. O sea que la tercera hipótesis nos dice que las importaciones internacionales son nulas, de ahí que la identidad

producción + Importaciones de las demás regiones + Importaciones internacionales	=	Absorción + Exportaciones a las demás regiones + Exportaciones internacionales

llega a ser

Producción + Importaciones de las demás regiones	=	Absorción + Exportaciones a las demás regiones

o sea

Producción – Absorción	=	Exportaciones netas a las demás regiones

cuando el excedente de la producción con relación a la absorción es positivo.

Finalmente, la cuarta hipótesis nos dice que si hay exportaciones hacia otras regiones, no hay importaciones que vengan de otras regiones, y viceversa. Por lo tanto, cuando las importaciones netas son positivas, son iguales a las exportaciones brutas.

Después de calcular el empleo exportador de cada rama, sólo resta hacer la suma para obtener una estimación de la "base" económica de la región *i* (conocido también como empleo *básico*):

$$\text{Base exportadora} = \sum_{j \text{ cuando } QL_{ij}>1} EXP_{ij}$$

En el modelo de la base económica, hacemos la hipótesis que la razón

$$\theta = \frac{\sum_j x_{ij}}{\sum_j EXP_{ij}}$$

es constante. El modelo predice que, para cada empleo exportador que se crea (o que desaparece), el empleo total aumenta (o disminuye) de θ empleos. θ se llama el *multiplicador de la base económica.*[13]

1-2.2 El análisis de descomposición aditiva y multiplicativa de las variaciones

1-2.2.1 Principio

El análisis "shift-share" es un caso particular de una técnica más general conocida como el análisis de descomposición de las variaciones.[14] En principio, el análisis de descomposición de las variaciones puede aplicarse a toda diferencia entre dos valores observados de una misma variable. Puede tratarse de dos observaciones de un mismo objeto en momentos diferentes u observaciones de dos objetos distintos.

El análisis de descomposición de las variaciones consiste en descomponer la diferencia entre dos valores de una medición en una suma de términos (descomposición aditiva) o en un producto de factores (descomposición multiplicativa). Tal descomposición es siempre una tautología del tipo

$$x - y = (x - a) + (a - b) + (b - c) + (c - y)$$

[13] Múltiples variantes del cociente de localización se propusieron para aligerar lo exigente de las hipótesis que son bases del método.

[14] Encontramos otro ejemplo en el famoso artículo en ciencias regionales de Williamson (1965), en el cual propone una descomposición de la evolución en el tiempo de la medición de desigualdad interregional.

o

$$x/y = (x/a)\ (a/b)\ (b/c)\ (c/y)$$

o sea

$$\log x - \log y = (\log x - \log a) + (\log a - \log b) + (\log b - \log c) + (\log c - \log y)$$

Por lo tanto, la utilidad de la descomposición depende de la interpretación que le podemos dar a los términos de una descomposición aditiva o a los factores de una descomposición multiplicativa. Esta interpretación se basa en un modelo a menudo implícito. El lenguaje usado ("este efecto", "este otro factor") sobreentiende, a veces, connotaciones de causalidad no siempre justificadas.

1-2.2.2 Aplicación al análisis "shift-share"[*]

El análisis "shift-share"[15] es un método de análisis de descomposición muy conocido entre los practicantes de las ciencias regionales. Consiste en descomponer la variación del empleo de una ciudad o de una región. Examinaremos, ahora, el método de descomposición de la variación del empleo de una actividad, y luego el método de descomposición de la variación del empleo de un grupo de actividades.

Descomposición de la variación del empleo de una actividad

Para ilustrar el método shift-share, usaremos el ejemplo numérico que sigue:

[*] Réferencias: Page-Patton, cap. 9; Coffey y Polèse (1988); Polèse, 1994, pp. 349-357.

[15] Jayet (1993, pp. 29-34) emplea la expresión "análisis estructural geográfico". Por mi parte, prefiero mejor la expresión de Bonnet (1995): "análisis estructural-residual".

Empleo por zona y por rama

	Año 1				Año 2			
Rama	*B1*	*B2*	*B3*	Total	*B1*	*B2*	*B3*	Total
Zona								
Z1	48	325	287	660	24	388	300	712
Z2	27	185	148	360	11	173	200	384
Z3	45	90	45	180	25	99	52	176
Total	120	600	480	1200	60	660	552	1272

Variación del empleo por zona y por rama
entre el año 1 y el año 2

	Diferencias				Tasa de variación (%)			
Rama	*B1*	*B2*	*B3*	Total	*B1*	*B2*	*B3*	Total
Zona								
Z1	–24	63	13	52	–50.00	19.38	4.53	7.88
Z2	–16	–12	52	24	–59.26	–6.49	35.14	6.67
Z3	–20	9	7	–4	–44.44	10.00	15.56	–2.22
Total	–60	60	72	72	–50.00	10.00	15.00	6.00

Examinemos la variación del empleo de la rama *B1* en la zona *Z2*.

El análisis shift-share se basa en la comparación de tres escenarios:

1. Cuál hubiera sido la variación si el empleo de *B1* en *Z2* hubiera evolucionado con la misma tasa que el empleo total (de todas las ramas y todas las zonas)?
   - Tasa = 6%.
   - Número = 6% de 27 = 1.62.
2. ¿ Cuál hubiera sido la variación si el empleo de *B1* en *Z2* hubiera evolucionado con la misma tasa que el empleo total de la rama *B1*?
   - Tasa = –50%.
   - Número = –50% de 27 = –13.50.
3. ¿ Cuál fue la variación observada del empleo de *B1* en *Z2*?
   - Tasa = –59.26%.

– Número = –59.26% de 27 = –16.

La comparación de estos tres escenarios infiere la descomposición aditiva siguiente:

1. Efecto nacional = escenario 1.
   – Tasa = 6%.
   – Número = 6% de 27 = 1.62.
2. Efecto proporcional (o sectorial) = diferencia entre escenario 2 y escenario 1.
   – Tasa = –50% –6% = –56%.
   – Número = –56% de 27 = –15.12 = –13.5 – 1.62.
3. Efecto residual (o regional) = diferencia entre escenario 3 y escenario
   – Tasa = –59.26% – (–50%) = –9.26%.
   – Número = –9.26% de 27 = –2.5 = –16 – (–13.5).

Podemos verificar que la suma de los tres "efectos" es efectivamente igual a la variación observada:

– Tasa = 6 % + (–56 %) + (–9.26 %) = –59.26 %.
– Número = 1.62 + (–15.12) + (–2.5) = –16.

Se puede formalizar este método de descomposición con los símbolos que siguen:

x_{ij}^t	El empleo de la rama *j* en la región *i* en el momento *t*
$x_{\bullet j}^t = \sum_i x_{ij}^t$	El empleo de la rama *j* en el total de las regiones en el momento *t*
$x_{\bullet\bullet}^t = \sum_i \sum_j x_{ij}^t$	El empleo de todas las ramas en el total de las regiones al momento *t*

Tenemos la identidad siguiente:

$$\frac{x_{ij}^t}{x_{ij}^0} = \left(\frac{x_{ij}^t}{x_{ij}^0} - \frac{x_{\bullet j}^t}{x_{\bullet j}^0} \right) + \left(\frac{x_{\bullet j}^t}{x_{\bullet j}^0} - \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} \right) + \frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0}$$

Como símbolo para las tasas de crecimiento usamos

$$r_{ij} = \frac{x_{ij}^t}{x_{ij}^0} - 1$$

Entonces, con las tasas de crecimiento, la identidad anterior vuelve a escribirse

$$\left(\frac{x_{ij}^t}{x_{ij}^0} - 1\right) = \left[\left(\frac{x_{ij}^t}{x_{ij}^0} - 1\right) - \left(\frac{x_{\bullet j}^t}{x_{\bullet j}^0} - 1\right)\right] - \left[\left(\frac{x_{\bullet j}^t}{x_{\bullet j}^0} - 1\right) - \left(\frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} - 1\right)\right] - \left(\frac{x_{\bullet\bullet}^t}{x_{\bullet\bullet}^0} - 1\right)$$

o sea

$$r_{ij} = \left(r_{ij} - r_{\bullet j}\right) + \left(r_{\bullet j} - r_{\bullet\bullet}\right) + r_{\bullet\bullet}$$

En esta descomposición,

- $r_{\bullet\bullet}\, x_{ij}^0$ es el efecto nacional ("national share effect"): es el crecimiento esperado si el empleo de la rama *j* en la región *i* hubiera aumentado con la misma tasa que el empleo total del país (escenario 1);
- $\left(r_{\bullet j} - r_{\bullet\bullet}\right) x_{ij}^0$ es el efecto sectorial o el efecto de desplazamiento proporcional ("proportional shift effect"): es el crecimiento suplementario (positivo o negativo) del empleo esperado si el empleo de la rama *j* en la región *i* hubiera aumentado con la misma tasa que el empleo de la rama *j* en todo el país (es, por lo tanto, la diferencia entre el escenario 2 y el escenario 1); el efecto sectorial nos permite saber si, comparada al resto de la economía, la rama *j* es dinámico o sea, si goza de un crecimiento acelerado;
- $\left(r_{ij} - r_{\bullet j}\right) x_{ij}^0$ es el efecto regional o el efecto de desplazamiento diferencial ("differential shift effect"): es

la diferencia residual entre el crecimiento observado y el crecimiento resultado de la aplicación del efecto de parte y el efecto de desplazamiento proporcional.

La suma del efecto de desplazamiento proporcional y del efecto de desplazamiento diferencial es el efecto de desplazamiento total o neto ("total shift" o "net shift") para una actividad *j* en una región *i*.

Se interpreta, a menudo, el desplazamiento diferencial como una medición de la competitividad de la rama *j* en la región *i* ("competitive effect"). Este uso es muy discutible. En efecto, suponiendo que el empleo de la rama *j* crece más rápido en la región *i* que en la región *k*, ¿esto significa que la producción de esta rama crece más rápido en la región *i* que en la región *k*? No es tan seguro, sobre todo si la proporción en la mano de obra y los demás factores de producción varía de una región a otra y en el tiempo (en reacción a los cambios de los precios relativos).[16] Veremos, en un momento, que no es la única razón de dudar de la validez del efecto regional como una medición de la competitividad.

Descomposición de la variación del empleo total de una región

Cuando hacemos la suma de todas las ramas y de cada uno de los tres términos de la descomposición, obtenemos:

$$\sum_j r_{ij}\, x_{ij}^0 = \sum_j \left(r_{ij} - r_{\bullet j}\right) x_{ij}^0 + \sum_j \left(r_{\bullet j} - r_{\bullet\bullet}\right) x_{ij}^0 + \sum_j r_{\bullet\bullet}\, x_{ij}^0$$

donde

$\sum_j r_{\bullet\bullet}\, x_{ij}^0 = r_{\bullet\bullet} \sum_j x_{ij}^0$ es el efecto nacional.

[16] Para demostrar de manera formal esta proposición tendríamos que desarrollar un modelo de equilibrio general entre dos economías.

$\sum_j \left(r_{\bullet j} - r_{\bullet\bullet}\right) x_{ij}^0$ es el efecto estructural.

$\sum_j \left(r_{ij} - r_{\bullet j}\right) x_{ij}^0$ es el efecto regional.

Las cuatro tablas que siguen complementan los cálculos para nuestro ejemplo.

Análisis shift-share por rama

Rama *B1*

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	2.88	–26.88	0.00	–24.00	6.00	–56.00	0.00	–50.00
Z2	1.62	–15.12	–2.50	–16.00	6.00	–56.00	–9.26	–59.26
Z3	2.70	–25.20	2.50	–20.00	6.00	–56.00	5.56	–44.44

Rama *B2*

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	19.50	13.00	30.50	63.00	6.00	4.00	9.38	19.38
Z2	11.10	7.40	–30.50	–12.00	6.00	4.00	–16.49	–6.49
Z3	5.40	3.60	0.00	9.00	6.00	4.00	0.00	10.00

Rama *B3*

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	17.22	25.83	–30.05	13.00	6.00	9.00	–10.47	4.53
Z2	8.88	13.32	29.80	52.00	6.00	9.00	20.14	35.14
Z3	2.70	4.05	0.25	7.00	6.00	9.00	0.56	15.56

Total de ramas (clasificación de tres ramas)

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	39.60	11.95	0.45	52.00	6.00	1.81	0.07	7.88
Z2	21.60	5.60	–3.20	24.00	6.00	1.56	–0.89	6.67
Z3	10.80	–17.55	2.75	–4.00	6.00	–9.75	1.53	–2.22

Se interpreta el efecto de estructura como el efecto de la estructura económica o industrial de la región, o sea de la composición de su producción industrial ("industry mix effect"). Se interpreta, a menudo, el efecto regional como una medición de la competitividad de la región *i*; como la anterior, esta interpretación es igualmente discutible. Sin embargo, existe algo aún más grave.

En efecto, cuando se descompone la variación del nivel global del empleo de una región, la importancia del efecto de estructura depende del nivel de agregación de la clasificación de las actividades. Y como se calcula el efecto regional de manera residual, este último depende también del nivel de agregación. Para una región dada, el paso de una clasificación a otra puede aumentar o disminuir el efecto regional, de tal manera que puede implicar intervensiones de rango entre regiones: una región que parecía más competitiva que otra en una clasificación dada puede parecer menos competitiva en otra clasificación, algo incoherente.

Esto le resta aún más a la validez de la interpretación del efecto regional como medición de la competitividad de una región.

Se ilustra este fenómeno con la agregación de las ramas *B1* y *B2*. Podemos constatar que los resultados de la descomposición para todas las ramas son diferentes de los resultados obtenidos anteriormente con una clasificación de tres ramas.

Ramas *B*1 y *B*2 agregadas

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	22.38	–22.38	39.00	39.00	6.00	–6.00	10.46	10.46
Z2	12.72	–12.72	–28.00	–28.00	6.00	–6.00	–13.21	–13.21
Z3	8.10	–8.10	–11.00	–11.00	6.00	–6.00	–8.15	–8.15

Todas las ramas (clasificación de dos ramas)

Zona	Número de empleos				Tasa de variación (%)			
	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total	Efecto nacional	Efecto sectorial	Efecto residual	Efecto total
Z1	39.60	3.45	8.95	52.00	6.00	0.52	1.36	7.88
Z2	21.60	0.60	1.80	24.00	6.00	0.17	0.50	6.67
Z3	10.80	–4.05	–10.75	–4.00	6.00	–2.25	–5.97	–2.22

De igual manera, este problema afecta el análisis de descomposición de la variación del nivel de una sola actividad, ya que se define ésta según una clasificación que no puede más que tener un cierto grado de agregación. Dicho de otra manera, en caso de una actividad, el efecto regional calculado contiene el efecto estructural que acompaña los desplazamientos entre las subramas que componen la actividad considerada. Es por lo tanto, abusivo, aunque se considere una sola actividad, interpretar el efecto residual como una medición de la competitividad.

1-2.3 La medición del crecimiento (el cálculo de la tasa de variación en el tiempo)

De cierto modo el análisis de una serie cronológica posee el problema de la multidimensionalidad, de la cual hablaremos después. En efecto, el concepto de “crecimiento” o de “varia-

ción en el tiempo" encierra múltiples dimensiones. De hecho tiene tantas dimensiones como hay observaciones del crecimiento, a menos que la tasa de variación sea constante (caso cuando el crecimiento es uniforme).

1-2.3.1 Tasa de crecimiento por periodo

Como ejemplo, veamos la evolución del índice de precios al consumidor (IPC) en Canadá de 1984 a 1992:[17]

1984	92.4
1985	96.0
1986	100.0
1987	104.4
1988	108.6
1989	114.0
1990	119.5
1991	126.2
1992	128.1

Entre cada periodo y el siguiente, podemos calcular una tasa de crecimiento por periodo. Así, entre 1984 y 1985, la tasa de crecimiento por periodo fue de $\frac{96.0 - 92.4}{92.4} = 0.039$ o sea 3.9%. Con nueve observaciones consecutivas (de 1984 a 1992), podemos calcular ocho tasas de crecimiento por periodo:

[17] Media anual sin ajusto estacional (Estadística Canadá, número de catálogo 62-210).

de...	a...	tasa
1984	1985	0.039
1985	1986	0.042
1986	1987	0.044
1987	1988	0.040
1988	1989	0.050
1989	1990	0.048
1990	1991	0.056
1991	1992	0.015

En general, con una serie de T+1 observaciones, del periodo 0 al periodo T, x_0, x_1, x_2, ..., x_t, ..., x_T podemos calcular T valores de la tasa de crecimiento r_t de un periodo con relación al anterior:

$$r_t = \frac{x_t - x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - \frac{x_{t-1}}{x_{t-1}} = \frac{x_t}{x_{t-1}} - 1$$

Por ejemplo para t = 1985 $r_{1985} = \frac{96.0}{92.4} - 1 = 0.039 = 3.9\ \%$ donde t varía de 1 a T. En este caso, se dice que, entre -1 y t, x creció de R_t por ciento, sabiendo que $R_t = 100 \times r_t$

Nota: si x_t es inferior a x_{t-1}, hubo decrecimiento: r_t es negativo. Hablamos entonces de crecimiento negativo.

La serie r_1, r_2, ..., r_t, ..., r_T describe la evolución en el tiempo de la variable x. De hecho, si se invierte la fórmula de cálculo de la tasa de crecimiento periódica, se obtiene $x_t = (1+r_t)\, x_{t-1}$.

Pero tenemos también $x_{t-1} = (1+r_{t-1})\, x_{t-2}$ de tal modo que $x_t = (1+r_t)\,(1+r_{t-1})\, x_{t-2}$ y, haciendo sustituciones sucesivas, tenemos $x_t = (1+r_t)\,(1+r_{t-1})\, ...\, (1+r_2)\,(1+r_1)\, x_0$.

O sea que si conocemos los r_t y x_0 podemos reconstituir la serie de los x_t. Por lo tanto, es cierto que la serie

$$r_1, r_2, ..., r_t, ..., r_T$$

describe la evolución en el tiempo de la variable *x*.[18] Sin embargo, ¿podríamos resumir esta evolución con una sola cifra?

De alguna manera, los *T* valores de r_t son tantas dimensiones de la evolución en el tiempo de la variable *x*. Es por eso que la medición del crecimiento se parece al problema de multidimensionalidad y de la construcción de los números índice.

1-2.3.2 Promedio de las tasas de crecimiento por periodo

Para resumir la evolución de la variable *x* en el tiempo, se puede tomar el promedio aritmético de las tasas de crecimiento por periodo:

$$\frac{1}{T}\sum_{t=1}^{T} r_t = \frac{r_1 + r_2 + \cdots + r_t + \cdots + r_T}{T}$$

En el caso del IPC entre 1984 y 1992, el promedio de las tasas de crecimiento por periodo es de 0.042 (o sea 4.2%).

Sin embargo, esta manera de resumir la evolución de la variable *x* en el tiempo tiene el inconveniente de no tomar en cuenta la variabilidad de las tasas de crecimiento. Ahora bien, para una misma tasa promedio, más las tasas son uniformes, más el crecimiento acumulado es fuerte.[19] No vamos a demos-

[18] En la práctica, se utilizan valores redondeados de las tasas de crecimiento, de tal manera que no se podría reconstituir exactamente la serie original.

[19] Si bien el crecimiento acumulado depende de la variabilidad de las tasas, por el contrario, no depende del orden cronológico entre las diferentes tasas de crecimiento. Se puede observar al constatar en la fórmula siguiente

trar esta propiedad; un ejemplo bastará para ilustrarla. Comparemos las dos series siguientes:

100, 110, 121

y

100, 100, 120

En ambos casos, el promedio de las tasas de crecimiento por periodo es igual a 0.1 (o sea 10%). Sin embargo, el crecimiento acumulado en los dos periodos es de 21% en el primer caso, pero de solamente 20% en el segundo. La pregunta que provoca esto es: ¿en qué medida el promedio de las tasas por periodo es representativo de la evolución de una serie cuando las tasas por periodo son variables?

1-2.3.3 Cálculo de una tasa de crecimiento exponencial

La tasa de crecimiento exponencial es otra manera de resumir la evolución de la variable *x* en el tiempo. Se puede definir de dos maneras equivalentes:

La primera definición de la tasa de crecimiento exponencial se refiere al promedio geométrico.[20] Es la tasa de crecimiento *r* obtenida a partir del promedio geométrico de los factores[21] de crecimiento por periodo:

$$1+r = \left[(1+r_T)(1+r_{T-1})\cdots(1+r_2)(1+r_1)\right]^{1/T}$$
$$= \sqrt[T]{(1+r_T)(1+r_{T-1})\cdots(1+r_2)(1+r_1)}$$

o sea, en forma logarítmica:

---

que el valor de x_t sigue sin cambio si cambiamos el orden de los factores del miembro de la derecha: $x_t = (1+r_t)\,(1+r_{t-1})\,\ldots\,(1+r_2)\,(1+r_1)\,x_0$.

[20] Con relación al promedio geométrico y a sus aplicaciones, vea Wonnacott y Wonnacott (1991, p. 755).

[21] Observe la distinción entre la tasa de crecimiento *r* y el factor de crecimiento (1 + *r*).

$$\log(1+r) = \frac{1}{T}\sum_{t=1}^{T}\log(1+r_t)$$

Por el contrario, la medición anterior usaba el promedio aritmético de las tasas de crecimiento por periodo. Se puede simplificar el cálculo de la tasa de crecimiento exponencial explotando la relación

$$x_T = (1+r_T)\,(1+r_{T-1})\,...\,(1+r_2)\,(1+r_1)\,x_0$$

Ahora bien, según la definición del promedio geométrico

$$(1+r)^T = (1+r_T)(1+r_{T-1})\cdots(1+r_2)(1+r_1)$$

de tal manera que

$$x_T = (1+r)^T x_0$$

donde x_T y x_0 son valores conocidos y r es la incógnita. Al desarrollar esta fórmula, llegamos a un método de cálculo de la tasa de crecimiento exponencial.

$$(1+r)^T = \frac{x_T}{x_0}$$

$$\log(1+r)^T = \log\left(\frac{x_T}{x_0}\right)$$

$$T\log(1+r) = \log(x_T) - \log(x_0)$$

$$\log(1+r) = \frac{\log(x_T) - \log(x_0)}{T}$$

$$1+r = \text{antilog}\left(\frac{\log(x_T) - \log(x_0)}{T}\right)$$

donde $z = e^z$ o 10^z, según se tome el logaritmo neperiano de base e (= 2.71828...) o el logaritmo común de base 10.

$$r = \text{antilog}\left(\frac{\log(x_T) - \log(x_0)}{T}\right) - 1$$

O sea, con los logaritmos comunes,

$$r = 10^{\frac{\log x_T - \log x_0}{T}} - 1$$

y con los logaritmos neperianos

$$r = e^{\frac{\ln x_T - \ln x_0}{T}} - 1 = \exp\left(\frac{\ln x_T - \ln x_0}{T}\right) - 1$$

Por ejemplo, la tasa de crecimiento exponencial del IPC de 1984 a 1992 se calcula de la manera que sigue:

$x_0 = 92.4$ y $\log_e x_0 = 4.526126979$

$x_T = 128.1$ y $\log_e x_T = 4.852811209$

$T = 8$

$$r = \exp\left(\frac{4.852811209 - 4.526126979}{8}\right) - 1 = 0.042\,,$$

o sea 4.2 %

Existe otra definición de la tasa de crecimiento exponencial que contiene su propia interpretación. En efecto, acabamos de ver que la tasa de crecimiento exponencial r es la solución de la ecuación

$$x_T = (1+r)^T x_0$$

Por lo tanto, se puede ver la tasa de crecimiento exponencial como una tasa hipotética: es la respuesta a la pregunta: "¿si la variable x hubiera evolucionado con una tasa por periodo constante, con que tasa hubiera tenido que crecer para que su valor terminal fuese igual al valor terminal observado?" Por definición, la tasa de crecimiento exponencial es, por lo tanto, la tasa de crecimiento por periodo uniforme que da el mismo crecimiento acumulado que la serie r_1, r_2, ..., r_t, ..., r_T.

Así que, como el promedio aritmético de las tasas por periodo, la tasa de crecimiento exponencial resume la evolución de la variable x en el tiempo. Sin embargo, la tasa de crecimiento exponencial toma solamente en cuenta el primer y el

último valor de la serie. Esto constituye un inconveniente si el primer valor de la serie x_0 o el último x_T es excepcional (fuera de tendencia); en este caso, la tasa de crecimiento exponencial podría ser engañosa.

Como índice, se puede considerar como un índice truncado ya que no usa toda la información disponible. Los índices truncados pueden, a veces, ser útiles; además, no podemos dudar que tienen adeptos justamente por la poca exigencia en información que requieren: falta pensar en el "Big Mac cost-of–living index" del periódico *The Economist.*

1-2.3.4 Entre dos males...

Como índices de la evolución cronológica de una variable, tanto el promedio de las tasas por periodo como la tasa de crecimiento exponencial representan inconvenientes. Necesitamos, pues, escoger el menos malo, pero ¿cuál? Claro está que depende del objetivo con que queramos usarlo. Por ejemplo, ¿qué tanto es importante que la tasa de crecimiento seleccionada permita "predecir" con exactitud (o sea de reproducir) el valor terminal a partir del valor inicial? Respecto a esto, es preferible usar la tasa de crecimiento exponencial. ¿O por lo contrario, es más importante que la tasa de crecimiento seleccionada sea representativa de la tendencia? En este caso, nada garantiza, a priori, que los valores iniciales y terminales no estén fuera de tendencia, o sea, que la tasa de crecimiento exponencial no sea engañosa.

Sin embargo, vimos que el promedio de las tasas por periodo tiene el inconveniente de no tomar en cuenta la variabilidad de las tasas por periodo. ¿Tiene importancia este inconveniente? Eso depende. Por ejemplo, en el caso del IPC de 1984 a 1992 en Canadá, si la tasa de crecimiento hubiera estado constante, o sea igual al promedio de las tasas de crecimiento por periodo, el valor del IPC en 1992 hubiera sido de

128.16 en lugar de 128.1. La diferencia es mínima (¡seis centésimas de uno por ciento!).

¿Sería más grande la diferencia tomando en cuenta un gran número de periodos, con una tasa de crecimiento más volátil? Por ejemplo, el valor de cierre del índice bursátil Standart & Poor's 500 era de 470.34 el 17 de febrero 1994, y de 656.37 el 9 de febrero 1996, 499 sesiones de mercado más tarde,[22] el promedio de las tasas de crecimiento por periodo (de una sesión de mercado a otra) fue de 0.0007 (0.07%), con una desviación estándar 0.0057 (0.057%), lo que representa una volatilidad importante (el coeficiente de variación es de 8.37). Sin embargo, ¿cuál hubiera sido el valor de clausura el 9 de febrero de 1996 si el índice hubiese crecido con una tasa constante igual al promedio de las de las tasas periódicas? Hubiera sido de 661.76... Nuevamente, la diferencia es mínima (0.82%); en este caso, otra vez, el promedio de las tasas de crecimiento por periodo es bastante representativo de la tendencia.

1-2.3.5 Ajuste de una curva de tendencia

De todas maneras, existe un método más preciso para resumir la evolución de una serie: ajustándole una curva de tendencia. Para este fin, escogemos un modelo de la evolución de la serie. Por ejemplo, se puede tomar el modelo lineal simple $x_t = a + bt$ o el modelo exponencial simple $x_t = a\ b^t$ que se transforma en un modelo lineal simple cuando aplicamos los logaritmos:

$$\log x_t = \log a + (\log b)\ t$$

Se puede considerar el modelo exponencial simple como una versión mejorada de la tasa de crecimiento exponencial. En efecto, se puede establecer un paralelo entre el parámetro

[22] Fuente: www.fortitude.com/data.htm.

a y el valor inicial x_0 y entre b y el factor de crecimiento exponencial $(1+r)$. La estimación del modelo exponencial consiste en buscar los valores de x_0^* $(= a^*)$ y de $(1+r^*)$ $(= b^*)$ para que los valores x_t^* "predichos" por la relación

$$x_t^* = x_0^* (1+r^*)^t = a^* (b^*)^t$$

"se peguen" lo mejor posible con los valores observados. De esta manera, habremos calculado una tasa de crecimiento exponencial que estará menos expuesta a la influencia de valores fuera de tendencia.[23]

De manera más general, después de seleccionar un modelo, escogemos los valores de los parámetros que más se apegan a la realidad observada, aplicando los métodos de estadística más apropiados. La regresión lineal es una técnica que permite encontrar los "mejores" valores para a^* y b^*. Regresaremos sobre este tema en la tercera parte de esta obra.

1-2.3.6 ¿Qué recordar?

Dentro de poco examinaremos el problema de la multidimensionalidad en la medición. Mientras, vimos cómo algo a primera vista sencillo como la medición del crecimiento se dificulta a causa de esta multidimensionalidad. Estudiamos dos maneras sencillas de resumir la evolución de una variable en el tiempo con una tasa única (el promedio de las tasas de crecimiento por periodo y la tasa de crecimiento exponencial); cada una de las dos presenta inconvenientes. Evocamos otra técnica, como el ajuste de una curva de tendencia que no parece tener los defectos de las otras dos. Sin embargo, el uso de esta técnica tiene su propio precio: es mucho más compleja, tiene menos transparencia y los cálculos son más pesados.

[23] Vea Wonnacott y Wonnacott (1992, pp. 513-523).

Aquí tiene, pues, una perfecta ilustración de cómo la medición, mucho más que una ciencia, es también un arte.

- No hay una medición perfecta.
- Las mediciones menos imperfectas son, por lo general, más complejas y más pesadas en su manejo.

CAPÍTULO 1-3
EL PROBLEMA DE LA MULTIDIMENSIONALIDAD: LOS NÚMEROS ÍNDICE

1-3.0 PROBLEMÁTICA DE LA MULTIDIMENSIONALIDAD

Ya citamos a Gilles (1984, p. 24), quien refiriéndose al esquema clásico de Lazarsfeld (1971), define la operacionalización como el hecho de "someter los conceptos, por medio del análisis, a un proceso que los transforma en dimensiones y luego en indicadores que permite observarlos, medirlos o calificarlos". La reflexión teórica que permite identificar las dimensiones de un concepto pertenece a la disciplina o al campo de estudio pertinente. Aquí, consideramos de manera implícita que la mayor parte de los conceptos tienen dimensiones múltiples y examinamos las implicaciones de este hecho al momento de la construcción de mediciones asociadas a estos conceptos.

No faltan ejemplos de conceptos con dimensiones múltiples y que a cada una de ellas se le pueda asociar una medición distinta.

1. Se puede descomponer el concepto político de "nivel de satisfacción con respecto al gobierno" en varias dimensiones, como "satisfacción en cuanto a la política económica", "satisfacción en cuanto a la política so-

cial", "satisfacción en cuanto a la política exterior", etcétera.

2. Se puede descomponer el concepto "costo de la vida" en "costo de la vivienda, "costo de la alimentación", etcétera.

Cuando un mismo concepto encierra varias dimensiones pero se busca analizar como un todo, es necesario encontrar una manera de combinar las mediciones asociadas a las diferentes dimensiones en una sola medición que las resuma todas. O sea que el problema es sumar plátanos y naranjas.

La manera más conocida de tratar este problema consiste en construir números índices. Un número índice es una regla (una fórmula) para combinar varias mediciones en una sola cifra. Las diferentes mediciones que componen el índice se refieren a las diferentes dimensiones de un concepto; se usa el mismo índice como medición global del concepto estudiado. En general, no existe índice que sea perfectamente fiable (o sea, cuyas variaciones reflejen variaciones reales). Es hasta difícil construir un índice válido (o sea que mida bien lo que se quiere medir). Es uno de los puntos importantes que se busca resaltar en este capítulo con la ayuda de dos ejemplos: los índices de precios (en particular el índice de precios al consumidor) y el indice de Desarrollo Humano del Programa de las Naciones Unidas para el Desarrollo (PNUD).

1-3.1 ILUSTRACIÓN #1: LOS ÍNDICES DE PRECIO[*]

El concepto "nivel general de los precios" encierra tantas dimensiones como existen precios diferentes, o sea, tantas dimensiones como existen bienes o servicios diferentes. Un índice de precios sirve para comparar los precios de un grupo

[*] Referencias: Wonnacott y Wonnacott (1991, cáp. 22); Statistique Canada (1996 y 1997).

de bienes y servicios en dos momentos o, de manera excepcional, en dos lugares diferentes.

El ejemplo que sigue ilustra el problema.[24] Se trata de precios de la alimentación para un país ficticio donde el régimen alimenticio se compone únicamente de tres alimentos: el bistec, la pimienta y el pan.

Bienes	Precios ($)		Índices de los precios de bienes individuales
	1980	1985	
	p_{0i}	p_{ti}	p_{ti} / p_{0i}
Bistec (kg)	4.85 $	6.60 $	1.36
Pimienta (g)	0.07 $	0.07 $	1.00
Pan (kg)	1.10 $	1.32 $	1.20

Los tres índices de precio individual constituyen las tres dimensiones del concepto "precio de la alimentación". ¿Cómo podemos combinar estos tres índices de precio individual en una medición única? Los dos índices de precio más usados son el índice de Laspeyres y el índice de Paasche.

1-3.1.1 El índice de Laspeyres

Un índice de Laspeyres mide las variaciones del nivel general de los precios comparando el costo de adquisición de una canasta representativa de los bienes y servicios en dos momentos diferentes en el tiempo. La mayor parte de los índices de los precios al consumo son índices de Laspeyres.

Simbología:

- p_{ti} precio del bien *i* en el periodo *t.*
- p_{0i} precio del bien *i* en el periodo 0.

[24] Fuente: adaptado de Wonnacott y Wonacott (1991, p. 753).

- q_{0i} cantidad del bien *i* comprado por un hogar típico en el periodo 0.

En el índice de Laspeyres, se define la canasta representativa por los q_{0i}, o sea, las cantidades compradas por un hogar típico durante el periodo 0 conocido como el periodo de referencia o año de base. Supongamos que tenemos *n* bienes y servicios en la canasta representativa. Si usamos el operador suma entonces el costo de adquisición de la canasta representativa puede escribirse:

Costo de la canasta de referencia con los precios del periodo 0

$$= p_{01}q_{01} + p_{02}q_{02} + \cdots + p_{0n}q_{0n} = \sum_{i=1}^{n} p_{0i}q_{0i}$$

Costo de la canasta de referencia con los precios del periodo *t*

$$= p_{t1}q_{01} + p_{t2}q_{02} + \cdots + p_{tn}q_{0n} = \sum_{i=1}^{n} p_{ti}q_{0i}$$

El índice de Laspeyres compara estos dos valores $\sum_{i=1}^{n} p_{0i}q_{0i}$ y $\sum_{i=1}^{n} p_{ti}q_{0i}$.

El valor del índice de Laspeyres para el periodo *t* se calcula con la razón

$$\mathrm{I}_t^L = \frac{\sum_{i=1}^{n} p_{ti}q_{0i}}{\sum_{i=1}^{n} p_{0i}q_{0i}}$$

Por lo tanto, el índice de Laspeyres es la razón del costo de la canasta representativa cuando los precios son los del periodo *t* entre el costo cuando los precios son los del periodo 0.

Podemos ilustrar el cálculo con la ayuda de los datos ficticios del ejemplo anterior.

Bienes	Datos				Cálculos	
	Precios ($)		Cantidades		Costo de la canasta	
	1980	1985	1980	1985	1980	1985
	p_{0i}	p_{ti}	q_{0i}	q_{ti}	$p_{0i}\,q_{0i}$	$p_{ti}\,q_{0i}$
Bistec (kg)	4.85$	6.60$	23	18	111.55$	151.80$
Pimienta (g)	0.07$	0.07$	57	85	3.99$	3.99$
Pan (kg)	1.10$	1.32$	36	45	39.60$	47.52$
					155.14$	203.31$

Se calcula el índice de Laspeyres de los precios de 1995, con base 1990, con la fórmula

$$\mathrm{I}_t^L = \frac{203.31}{155.14} = 1.31$$

Nota: los índices de precios publicados por agencias estadísticas oficiales son, por lo general, expresados en porcentaje, de tal manera que veríamos normalmente

$$\mathrm{I}_t^L = 100 \times \frac{203.31}{155.14} = 131$$

Con tal de aligerar las fórmulas, ignoramos aquí este convenio.

Vamos, ahora, a demostrar que el índice de Laspeyres es un promedio ponderado de los índices de precios de los bienes individuales. Desarrollemos la fórmula del índice de Laspeyres:

$$\mathrm{I}_t^L = \frac{\sum_{i=1}^{n} p_{ti} q_{0i}}{\sum_{i=1}^{n} p_{0i} q_{0i}} = \frac{\sum_{i=1}^{n} p_{ti} q_{0i}}{\sum_{k=1}^{n} p_{0k} q_{0k}}$$

$$I_t^L = \sum_{i=1}^{n} \left(\frac{p_{ti} q_{0i}}{\sum_{k=1}^{n} p_{0k} q_{0k}} \right) = \sum_{i=1}^{n} \left[\left(\frac{q_{0i}}{\sum_{k=1}^{n} p_{0k} q_{0k}} \right) p_{ti} \right]$$

$$I_t^L = \sum_{i=1}^{n} \left(\frac{p_{oi} q_{0i}}{\sum_{k=1}^{n} p_{0k} q_{0k}} \right) \left(\frac{p_{ti}}{p_{0i}} \right) = \text{promedio ponderado de los índices de precios de los bienes.}$$

Por lo tanto, se puede interpretar el índice de Laspeyres como un promedio ponderado de los índices de precios $\left(\frac{p_{ti}}{p_{0i}} \right)$ de los bienes individuales donde el peso del bien *i*

$$w_{0i} = \frac{p_{0i} q_{0i}}{\sum_{k=1}^{n} p_{0k} q_{0k}}$$

es la parte de este bien en el presupuesto del hogar típico en el periodo 0.

Esto se aplica a nuestro ejemplo de la manera siguiente:

Bienes	Datos				Índice de precios de Laspeyres		
	Precios			Cant.	Costo de la canasta	Peso	Cálculo del índice
	1980	1985	Razón	1980	1980	1980	1985
	p_{0i}	p_{ti}	$\left(\frac{p_{ti}}{p_{0i}}\right)$	q_{0i}	$p_{0i}\,q_{0i}$	w_{0i}	$w_{0i}\left(\frac{p_{ti}}{p_{0i}}\right)$
Bistec (kg)	4.85$	6.60$	1.36	23	111.55$	0.719	0.978
Pimienta(g)	0.07$	0.07$	1.00	57	3.99$	0.026	0.026
Pan (kg)	1.10$	1.32$	1.20	36	39.60$	0.255	0.306
					155.14$	1.000	1.310

En realidad, el cálculo del índice de los precios al consumo de *Statistique Canada* es más complicado por varias razones:

- Para mantener la representatividad del índice, la canasta de referencia se actualiza entre los cambios de año de base (= 100) del índice.
- Se calculan los pesos con las cantidades de cada canasta de referencia y los precios de otro periodo.
- Se efectúa el cálculo del índice de tal manera que sus valores se sigan sin ruptura al momento de pasar de una canasta de referencia a la siguiente.

Por ejemplo, en 2003, se calcula el índice de los precios al consumo con la canasta de referencia de 1996, la cual se evalúa a los precios de diciembre 1997 y su año de base es 1992 (1992 = 100). Para más detalles, vea el *Document de référence de l'indice des prix à la consommation*, No 62-553 en el catálogo de *Statistique Canada*.

1-3.1.2 El índice de Paasche

¿Cuál es la diferencia entre un índice de Laspeyres y un índice de Paasche? Es la selección de la canasta representativa: en el caso del índice de Paasche, son los gastos de un hogar típico en el periodo *t* y no en el periodo de base 0 que indican la canasta representativa. Por lo tanto, el valor del índice de Paasche en el periodo *t* es:

$$I_t^P = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}}$$

De nueva cuenta, usemos nuestro ejemplo para ilustrar el cálculo.

Bienes	Datos				Cálculos	
	Precios ($)		Cantidades		Costo de la canasta	
	1980	1985	1980	1985	1980	1985
	p_{0i}	p_{ti}	q_{0i}	q_{ti}	$p_{0i}\,q_{ti}$	$p_{ti}\,q_{ti}$
Bistec (kg)	4.85$	6.60$	23	18	87.30$	118.80$
Pimienta (g)	0.07$	0.07$	57	85	5.95$	5.95$
Pan (kg)	1.10$	1.32$	36	45	49.50$	59.40$
					142.75$	184.15$

El índice de Paasche de los precios de 1995, base 1980 es

$$I_t^P = \frac{184.15}{142.75} = 1.29$$

De igual manera, se puede interpretar el índice de Paasche como el promedio ponderado de los índices de precio $\left(\frac{p_{ti}}{p_{0i}}\right)$ de precios de los bienes individuales. En el caso del índice de Paasche, el peso del bien *i*

$$w_{ti} = \frac{p_{0i} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}}$$

es la parte de este bien en el presupuesto ficticio de un hogar que hubiera consumido las cantidades q_{ti} de la canasta de consumo de un hogar típico en el periodo t, con los precios p_{0i} del periodo de base. Enseguida se muestra el desarrollo que conduce a este resultado:

$$\mathrm{I}_t^P = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{i=1}^{n} p_{0i} q_{ti}} = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}} = \sum_{i=1}^{n} \left(\frac{p_{ti} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}} \right)$$

$$\mathrm{I}_t^P = \sum_{i=1}^{n} \left[\left(\frac{q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}} \right) p_{ti} \right]$$

$$\mathrm{I}_t^P = \sum_{i=1}^{n} \left(\frac{p_{oi} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}} \right) \left(\frac{p_{ti}}{p_{0i}} \right) \quad = \quad \text{promedio ponderado de los índices de precios de los bienes.}$$

Esto se aplica a nuestro ejemplo como sigue:

Bienes	Datos				Índ. de precios de Paasche		
	Precios			Cant.	Costo de canasta	Peso	Cálculo del índice
	1980	1985	Razón	1985	N.A. [25]	N.A.	1985
	p_{0i}	p_{ti}	$\left(\frac{p_{ti}}{p_{0i}}\right)$	q_{ti}	$p_{0i}\,q_{ti}$	w_{ti}	$w_{ti}\left(\frac{p_{ti}}{p_{0i}}\right)$
Bistec (kg)	4.85$	6.60$	1.36	18	87.30$	0.612	0.832
Pimienta (g)	0.07$	0.07$	1.00	85	5.95$	0.042	0.042
Pan (kg)	1.10$	1.32$	1.20	45	49.50$	0.347	0.416
					142.75$	1.000	1.290

Se usa menos el índice de Paasche que el índice de Laspeyres porque implica efectuar de vuelta en cada periodo la encuesta con los hogares para definir la canasta representativa de bienes y servicios (los q_{ti}).

1-3.1.3 Uso de los índices de precios

Los índices de precio son a menudo usados como desinfladores[26] en el análisis de las series temporales. En cuanto las cifras se expresan en unidades monetarias y se refieren a diferentes años, se vuelve difícil compararlos si los precios han cambiado. Para poder comparar las cifras, es necesario aplicarles una corrección que tome en cuenta la evolución de los precios.

Veamos por ejemplo la evolución de los gastos personales en el consumo en Canadá de 1991 a 1999:[27]

[25] No se aplica, o sea que las cifras de las columnas no se aplican a ningún año.

[26] Se justifica el uso de esta palabra ya que, en la historia económica reciente, los periodos de aumento de los precios (inflación) fueron más frecuentes que los periodos de disminución de los precios.

[27] Gastos personales en bienes y servicios de consumo (millones de $) según las cuentas nacionales (*Statistique Canada,* "L'observateur économi-

Año	Gastos personales
1991	398 314
1992	411 167
1993	428 219
1994	445 857
1995	460 906
1996	480 427
1997	510 695
1998	531 169
1999	560 954

Se muestran estas cifras en dólares corrientes, o sea que no se toma en cuenta la evolución de los precios. ¿Cómo podemos, en estas condiciones, comparar, por ejemplo, el 480 billones de 1996 con el 411 billones de 1992, cuando los precios aumentaron de manera apreciable entre estos dos años? Para esto, podemos usar un índice de precios como desinflador. Este índice de precios podría ser un índice de Laspeyres, un índice de Paasche, o cualquier otro índice, siempre y cuando sea apropiado al tipo de cifras que queremos desinflar. Por ejemplo, aplicar un índice de precios industriales a una serie de cifras sobre los gastos personales de consumo no sería apropiado. Aquí, el índice de precios más recomendable es el índice de precios de los gastos personales calculado por *Statistique Canada.*[28]

En cuanto obtuvimos un índice de precios apropiado, resta para obtener cifras comparables, dividir cada cifra por el índice de precios del año correspondiente: si x_t es la cifra en dólares corrientes, entonces

---

que canadien, Supplément statistique historique 2001/02", No 11-210-XPB del catálogo).

[28] *Statistique Canada*, "L'observateur économique canadien, Supplément statistique historique 2001/02", No 11-210-XPB. Este índice es específico para los gastos personales en el producto interior bruto. Es preferible entonces al índice de los precios al consumidor.

$$y_t = \frac{x_t}{\mathrm{I}_t}$$

es una cifra *desinflada* en *dólares constantes* del año de referencia del índice.[29] Se dice que y_t es un dato en valor "real" cuando, por lo contrario, x_t es un valor "nominal".

Para los gastos personales de consumo, el resultado de usar el índice de precios al consumo como desinflador sería el siguiente:[30]

	Índice de precios	Gastos personales
1991	91.0	398 314
1992	92.5	411 167
1993	94.6	428 219
1994	95.6	445 857
1995	96.8	460 906
1996	98.4	480 427
1997	100.0	510 695
1998	101.2	531 169
1999	102.9	560 954

En dólares constante de 1997, el monto de los gastos personales de consumo de 1999 es igual a:[31]

$$100 \times \frac{560\,954}{102{,}9} = \frac{560\,954}{1{,}029} = 545\,145$$

[29] En caso de un índice expresado en porcentaje, la fórmula se convierte en $y_t = 100\frac{x_t}{\mathrm{I}_t}$.

[30] *Statistique Canada*, "L'observateur économique canadien, Supplément statistique historique 2001/02", No 11-210-XPB.

[31] La cifra publicada por *Statistique Canada* para los gastos personales de 1999 en dólares constantes de 1997 es más bien 545 162 millones. La diferencia es debida a que nosotros usamos un valor redondeado del índice de precios.

Preguntas engañosas:

- ¿Cuál es valor de los gastos personales de 1992 en dólares constantes de 1997?
- ¿Cuál es el valor de los gastos personales de consumo de 1999 en dólares constantes de 1992?[32]

Otro uso conocido de los índices, muy parecido al anterior, es la indexación. El ajuste busca mantener, año con año, el valor de un pago cuando los precios cambian.

Ejemplos:

> En el contrato colectivo que celebran un empleador y el sindicato de sus empleados, es común que el salario se fije únicamente para el primer año; para los que siguen, de común acuerdo, se ajustan los salarios en función de la evolución de los precios al consumo con una fórmula de indexación convenida.
>
> Ciertas categorías de ciudadanos (funcionarios jubilados, personas de edad avanzada...) perciben del Estado pensiones acordadas como en el caso anterior: se fija el monto inicial para luego, cada año, calcularlo con una fórmula de indexación.

Existen varias fórmulas de indexación; entre las que se usan, la mayor parte son fórmulas de ajuste parcial y algunas de ellas son algo complicadas. Veamos aquí la más simple fórmula de indexación. Un monto m_0 acordado en el año cero es ajustado, año con año, por medio de la fórmula

$$m_t = \mathrm{I}_t \, m_0$$

donde m_t es el monto ajustado para el año t e I_t es el índice de precios apropiado, con base el año cero ($\mathrm{I}_0 = 1$).[33] Si el índice

[32] Al momento de expresar un monto en dólares constante de un año θ diferente del año de base, es necesario aplicar la fórmula más general $y_t = x_t \dfrac{\mathrm{I}_\theta}{\mathrm{I}_t}$.

de precio tiene como base un año que no sea el año 0 entonces la fórmula se convierte de manera muy simple en

$$m_t = m_0 \frac{\mathrm{I}_t}{\mathrm{I}_0}$$

Por ejemplo, un monto de $35,000 en dólares de 1997, ajustado para el año de 1999, es igual a $35,000 × 1.029 = $36,015

y un monto de $35,000 en dólares de 1994, ajustado para el año de 1999 es igual a $35,000 × $\frac{1.029}{0.956}$ = $37,673

1-3.1.4 Índices de precios y costo de la vida

¿Son los índices de precios mediciones fiables?

Recuerdo: una variable es fiable cuando las variaciones en la medición corresponden a variaciones reales.

En el caso particular de la indexación de un ingreso (salario, pensión, etc,), si el índice de precios usado es fiable, el ingreso ajustado

$$m_t = m_0 \frac{\mathrm{I}_t}{\mathrm{I}_0}$$

le permitirá a la persona que lo recibe vivir en el periodo t con un ingreso de m_t tan bien como vivía (o que hubiera vivido) en el periodo 0 con un ingreso m_0.

Para empezar, está claro que la indexación "en vivo" no es posible puesto que se conoce el valor del índice de precios solamente después de un cierto lapso de tiempo. Esto implica que sólo se puede indexar un ingreso con un retraso (es cierto que el ajuste podría ser retroactivo). Nos referimos, por lo

[33] En el caso de un índice expresado en porcentaje ($\mathrm{I}_0 = 100$) la fórmula se convierte en $m_t = \frac{\mathrm{I}_t m_0}{100}$.

tanto, a una situación teórica. Además, las preferencias y las elecciones de consumo dependen de cada uno. Sería realmente un milagro que la canasta de referencia, que refleja el comportamiento general, correspondiese al consumo de un individuo o de una familia en particular.

Pero dejemos al lado los aspectos prácticos y preguntemos nos si, teorícamente, son perfectamente fiables los índices de Laspeyres y de Paasche. ¿Miden con exactitud las variaciones del costo de la vida? Pues no. De hecho, cuando los precios aumentan, el índice de Laspeyres sobrevalora el crecimiento del costo de la vida cuando, al contrario, el índice de Paasche lo subvalora. En el caso inverso, o sea cuando los precios disminuyen, el índice de Laspeyres subvalora la importancia de la disminución del costo de la vida mientras que el índice de Paasche la sobrevalora. De ahí que, tanto en tiempo de inflación como en tiempo de deflación (disminución general de los precios), un ingreso ajustado con el índice de Laspeyres será un poco más elevado que lo necesario cuando un ingreso ajustado con un índice de Paasche no será del todo suficiente.

Se puede demostrar lo anterior con la teoría económica del consumo. Sin embargo, podemos entenderlo de manera intuitiva. Los índices de precios de Laspeyres y de Paasche no miden con exactitud la evolución del costo de vida porque no toman en cuenta la capacidad de adaptación del consumidor.

En efecto, cuando los precios cambian, no cambian todos en la misma proporción: algunos precios aumentan o disminuyen más que otros (de hecho, puede pasar que algunos precios evolucionen en sentido contrario, o sea que unos aumenten mientras otros disminuyen). De ahí que los precios

relativos (o sea los precios comparados los unos con relación a los otros de los bienes y servicios) cambien.[34]

Ejemplo:

> Supongamos que la taza de café cueste 20¢ y la taza de té cueste 10¢: el precio del café es el doble que el precio del té. Supongamos que el precio del café aumenta 35% y el precio del té 50%: el nuevo precio del café es de 27¢, lo que equivale a 1.8 veces el nuevo precio del té (15¢). Aunque los dos precios hayan aumentado, el precio del café disminuyó con relación al precio del té porque pasó del doble a 1.8 veces el precio del té.

¿Cómo reaccionan los consumidores cuando dos bienes son sustitutos y sus precios relativos cambian? Se adaptan y mueven su consumo hacia los bienes cuyos precios relativos hayan disminuido. Observemos este hecho con un ejemplo numérico. Distinguimos tres bienes: el té, el café y un bien compuesto que engloba el resto de los bienes

		Té	Café	Et cœtera	Total
Año 0	Cantidad	1250	800	10000	
Año 0					
	Precio	0.40	0.50	1.00	
	Gasto	500	400	10000	10900
Año t					
	Precio	4.00	0.50	1.00	
	Gasto	5000	400	10000	15400

Calculemos el índice de Laspeyres:

$$I_t^L = \frac{(4.00 \times 1250) + (0.50 \times 800) + (1.00 \times 10000)}{(0.40 \times 1250) + (0.50 \times 800) + (1.00 \times 10000)}$$

[34] La traducción francesa de Wonnacott y Wonnacott (1991) usa la expresión “precio relativo” en un sentido diferente que la misma usada en la tabla 22-1 (p. 752) para designar los índices de precios de los bienes individuales.

$$I_t^L = 1.41284$$

Ingreso ajustado por año $t = 1.41284 \times 10900 = 15400$

Este ingreso indexado permite al consumidor típico comprar para el año *t* los mismos bienes que compraba en el año 0. Por lo tanto, con un ingreso ajustado y los nuevos precios, un consumidor común podría vivir exactamente como antes.

Al ver el precio muy alto del té, la mayor parte de nosotros buscará gastar de otra manera su ingreso ajustado; por ejemplo, podría decidir comprar menos té, beber un poco más de café para suplirlo y comprar un poco más del *et coetera*. Tomaríamos esta decisión porque nos permitiría vivir mejor adaptándonos a los cambios de los precios relativos gracias a la sustitución. Sin embargo, vivir mejor con un ingreso ajustado significa también que este ingreso es más elevado que lo necesario (para acabar de convencerse, basta preguntarse en qué situación preferiría estar: con el ingreso y los precios del año 0 o con el ingreso ajustado y los precios del año t).

En términos técnicos, el índice de Laspeyres no es exacto porque su numerador, por no tomar en cuenta las posibilidades de sustitución, es demasiado grande.

Claro está que los cambios en los precios relativos son en pocas ocasiones tan drásticos como los cambios del ejemplo numérico. Sin embargo, el principio de sustitución es el mismo y es por él que podemos afirmar que un ingreso ajustado con un índice de Laspeyres es siempre más alto que lo necesario.

¿Este sesgo es importante? En Canadá, Crawford (1993) estima que este sesgo varía entre 0.1% y 0.2% (o sea que el IPC sobreevaluaría el aumento del costo de vida por una o dos décimas del punto de porcentaje por año). En los Estados Unidos, la evaluación reciente de la CPI Comisión es de aproximadamente un medio punto de porcentaje por año (Boskin *et al.*, 1998). No es del todo despreciable. Además, los estudios citados identifican otros sesgos de aún más im-

portancia como el sesgo que provoca cambios en la calidad de los bienes. Crawford evalúa el efecto del total de los sesgos a un medio punto de porcentaje y, por su lado, la CPI Comisión entre, aproximadamente, 0.8 y 1.6 puntos de porcentaje.

A diferencia de lo que pasa con el índice de Laspeyres, un ingreso ajustado con un índice de Paasche nunca es del todo suficiente. Se puede observar este hecho al escribir el índice con la fórmula siguiente:

$$\mathrm{I}_t^P = \frac{\sum_{i=1}^{n} p_{ti} q_{ti}}{\sum_{k=1}^{n} p_{0k} q_{tk}}$$

En este caso, es el denominador del índice el que es demasiado grande porque no toma en cuenta las posibilidades de sustitución: al ver los precios p_{0k}, un hogar típico escogería cantidades diferentes de los q_{tk}.

Con el objetivo de corregir las distorsiones que resultan del uso tanto uno como otro índice, se propuso el índice "ideal" de Fischer, el cual es el promedio geométrico del índice de Laspeyres y del índice de Paasche:

$$\mathrm{I}_t^F = \sqrt{\mathrm{I}_t^L \times \mathrm{I}_t^P}$$

1-3.1.5 Conclusión: índices y modelos

Constatamos que los índices de precios de Laspeyres y de Paasche no eran mediciones del todo fiables de la evolución del costo de la vida. Se pudo demostrar al destacar las debilidades del modelo de comportamiento subyacente que provoca el uso de estos índices. En efecto, este modelo no toma en cuenta la capacidad de adaptación de los consumidores.

El examen de los índices de Laspeyres y de Paasche nos enseñó lo siguiente: el uso de cualquier índice se apoya explícita o implícitamente en un modelo subyacente.[35] En cuanto los modelos subyacentes son un reflejo fiel de la realidad que se busca medir, entonces los índices son mediciones fiables.

Ahora bien, los índices de Laspeyres y de Paasche son unos entre muchos índices del mismo tipo que se usan en todas las disciplinas. De hecho, el promedio ponderado es una forma de índice muy usado tanto para medir la evolución de los precios de un grupo de bienes como también, todo tipo de concepto con dimensiones múltiples. El cálculo de tales índices es fácil y su interpretación accesible ya que los pesos de los componentes representan la importancia relativa entre cada componente. Sin embargo, por su simplicidad, los índices calculados como promedios ponderados, implican generalmente, como los índices de Laspeyres y de Paasche unas hipótesis algo restrictivas. Por el contrario, los índices que derivan de modelos más elaborados pueden ser muy complicados.[36]

[35] Se dice que un índice es "exacto" con relación a un modelo cuando el índice es perfectamente coherente con este modelo.

[36] Por ejemplo, el índice de Törnqvist asociado a la función translog ("transcendental logarítmica").

1-3.2 ILUSTRACIÓN #2: EL ÍNDICE DE DESARROLLO HUMANO (IDH) DEL PROGRAMA DE LAS NACIONES UNIDAS PARA EL DESARROLLO (PNUD)*

1-3.2.1 Dimensiones del concepto y variables

Se propone el Índice de Desarrollo Humano (IDH) del programa de las Naciones Unidades para el Desarrollo (PNUD) como indicador de desarrollo para reemplazar (los más moderados dicen "para complementar") el producto interior bruto (PIB) usado por el Fondo Monetario Internacional (FMI) y el Banco Mundial, pues se critica fuertemente el PIB como medición del desarrollo por no tomar en cuenta varias dimensiones del desarrollo humano. Es la razón por la cual el IDH contiene tres componentes (dimensiones del concepto de desarrollo humano):

- Longevidad
- Saber
- Nivel de vida

Operacionalizar estas tres dimensiones del concepto de desarrollo humano conduce a escoger las tres variables siguientes:

- Longevidad: esperanza de vida al nacimiento
- Saber: tasa de alfabetización de los adultos y tasa de escolarización (sin distinción de niveles),[37] esto es dos variables
- Nivel de vida: producto interno bruto (PIB) real por habitante, en dólares ajustados en función del costo de la vida (paridad del poder de compra)

---

* Referencias: Programa de las Naciones Unidas para el Desarrollo (PNUD), *Informe sobre desarrollo humano* (anual, a partir de 1990) disponible en el sitio web: http://hdr.undp.org/

[37] Hasta el año de 1984 era el número promedio de años de escolaridad. No obstante, se dejó esta variable por falta de disponibilidad de datos.

Para cada variable, se mide la progresión efectuada para alcanzar el nivel máximo del índice con relación a la distancia total entre el nivel máximo y el nivel mínimo.[38]

1-3.2.2 Máximos y mínimos:

Antes de 1984 para calcular el IDH se usaban como valores máximos y mínimos los niveles más altos y más bajos que se habían observado ese año entre los países. Esto imposibilitaba la comparación de año en año. La versión actual del IDH considera como mínimos los valores más pequeños observados durante los últimos treinta años[39] y como máximos, los valores más altos que se prevé para los próximos treinta años.

Variable	Máximo	Mínimo
Esperanza de vida	85 años	25 años
Tasa de alfabetización	100 %	0 %
Tasa de escolaridad	100 %	0 %
PIB real/habitante	40 000 $	100 $

1-3.2.3 Ajuste del PIB real por habitante

Tasa de cambio aplicada a las conversiones monetarias

Con el fin de hacer comparaciones entre los países, se tiene que expresar los datos del PIB real por habitante en una misma unidad monetaria. De modo que se convierte en dólares US todas las cifras expresadas en Yenes japoneses, en Marcos alemanes o en colones de Costa Rica. Sin embargo, para efectuar esta conversión, en lugar de la simple tasa de cam-

[38] Por consiguiente, se trata de mediciones relativas.

[39] La variable de ingreso es una excepción: su valor mínimo debería ser de $200, en cambio, para poder incluir en los análisis los valores inferiores de la variante sexo especifica (vea más abajo) de la variable de ingreso, el nivel mínimo del PIB real por habitante se redujo a $100.

bio, se usa una tasa de conversión que refleja el poder de compra relativo de las monedas.

Ejemplo:

> Hasta hace poco, cuando se convertía en dólares el ingreso promedio de los japoneses con la tasa de cambio del momento, este ingreso parecía muy elevado. No obstante, el costo de vida en Japón es mucho más elevado que en los Estados Unidos cuando se compara con la tasa de cambio del momento. Es necesario tomar en cuenta este factor para poder comparar los ingresos promedio de Japón y de los Estados Unidos.

De la misma manera, se dice que a US$ 0.76, el dólar canadiense es subvaluado y que una tasa de aproximadamente US$ 0.85 reflejaría mejor su poder de compra relativo[40].

Es la razón por la cual que, al momento de calcular el IDH, se mide el nivel de vida con el PIB real por habitante, el cual se expresa en "PPA en USD", o sea en "paridad del poder adquisitivo (o sea poder de compra) en dólares de Estados Unidos".

Corrección con relación al efecto decreciente de los crecimientos sucesivos de ingreso en el desarrollo humano

Además, "el IDH refleja más la suficiencia que la saciedad" (PNUD, 1984, p. 97). Por esto, además de corregir el PIB real por habitante para tomar en cuenta el costo de vida, se ajusta con el fin de reflejar la posibilidad que los crecimientos sucesivos del ingreso per capita contribuyen cada vez menos al florecimiento humano. Desde el *Informe* de 1999, se traduce la aplicación de este principio con una transformación logarítmica: el indicador de nivel de vida usado en el cálculo del

[40] Lafrance y Schembri (2002).

IDH es, por lo tanto, el logaritmo del PIB real por habitante expresado en "PPA en USD".[41]

1-3.2.4 Cálculo del IDH

Cálculo del IDH para el país *j* (se encuentra un ejemplo numérico del cálculo en PNUD, 2003, p. 351.)

3. Para cada una de las cuatro variables, se calcula un subíndice que es la razón de la progresión efectuada en el recorrido (la distancia total entre el nivel máximo y el nivel mínimo).

$$I_{ij} = \frac{x_{ij} - \min x_i}{\max x_i - \min x_i}$$

donde

x_{ij} es el valor del índice *i* en el país *j*;

$\max x_i$ es el valor máximo del índice *i*;

$\min x_i$ es el valor mínimo del índice *i*.

4. El índice previsto para el saber es un promedio ponderado de dos variables usadas (tasa de alfabetización y tasa bruta de escolarización); el peso acordado para la alfabetización es el doble que el peso acordado para la escolaridad:

$$I_{\text{saber},j} = 0.67 \times I_{\text{alfa},j} + 0.33 \times I_{\text{escolar},j}$$

5. El índice de desarrollo humano es un promedio aritmético de los índices asociados a los tres componentes:

$$I_j = \frac{1}{3}\sum_{i=1}^{3} I_{ij}$$

[41] Hasta el año de 1998, se ajustaba el PIB real PPA por habitante con la dudosa "fórmula de Atkinson", fuertemente criticable (vea PNUD, 1999, p. 159 y Lemelin, 1999).

1-3.2.5 Reflexión sobre el IDH

Fundamentos teóricos y éticos

El IDH se inspira de las ideas de Amartya Sen[42] sobre la justicia social (vea Sugden, 1993). Sen propone conceptos difíciles de entender y aún más difíciles de traducir de manera operacional. No es tampoco muy claro que el IDH sea una traducción empírica confiable de los conceptos propuestos por Sen. Es el motivo por el cual muchos reprochan al IDH no tener fundamentos teóricos. Los partidarios del IDH admiten que constituye una medición imperfecta, pero afirman también que esta medición es útil y que contribuye a renovar la reflexión sobre el desarrollo. Falta mucho para que todos reconozcan la utilidad del IDH. El debate sobre los fundamentos teóricos y la validez del IDH está en el ojo del huracán (Aturupane, Glewne y Isenman, 1994; Srinivasan, 1994; Streeten, 1994; Ravaillon, 1997).

Formas funcionales y ponderaciones arbitrarias

Cuando algunos afirman que el IDH no tiene fundamentos teóricos, piensan que le falta solidez al vínculo que existe entre el concepto teórico y su operacionalización.

En particular, el PNUD no exhibe ningún argumento que fundamente de manera teórica otorgar un peso igual a los tres componentes del IDH, ni tampoco para justificar la ponderación que se usa en el cálculo del índice con relación a la educación (dos tercios-un tercio).

Antes de 1999, el tipo de relación entre el PIB real y el PIB corregido ("fórmula de Atkinson") era una contradicción al principio que estipula que, rebasando un cierto límite, el cre-

[42] Amartya Sen recibió el premio Nobel de economía en 1998. Vea http://www.nobel.se/announcement-98/economics98.html

cimiento del ingreso contribuye de menos en menos al florecimiento humano (Lemelin, 1999). La corrección logarítmica que se aplicó a partir del Informe de 1999 constituye una gran mejora. Aun así, la selección de esta fórmula de corrección sigue, al parecer, arbitraria.

Tomar en cuenta las disparidades al interior del país

Además, el IDH es una medición promedio y puede, por consiguiente, esconder fuertes disparidades al interior del país; por ejemplo, entre las regiones, los sexos, los grupos raciales o las clases socioeconómicas. Con miras a obtener una imagen más confiable de la realidad, el enfoque más preciso es claramente calcular el valor del IDH de manera independiente para las diferentes regiones o para los diferentes grupos del país. Es justamente lo que efectúa el PNUD en ciertos estudios de casos (PNUD, 1994, "Descomposición del IDH", p 104-107).

Algunos investigadores calcularon el valor del IDH para las regiones o para grupos particulares. Por ejemplo, al interior del Canadá, el valor del IDH para el Quebec es menor que el valor de todo Canadá. De igual manera, en Ontario el IDH de los franco-ontarianos es inferior al IDH de la provincia en su totalidad (esencialmente por una tasa de alfabetización menor). Cabe añadir que en México el Conapo cálcula el IDH por entidad y por municipio.[43]

Por otro lado, hasta 1996 se tomaban en cuenta las desigualdades económicas en el cálculo de un IDH ajustado según la repartición del ingreso. Sin embargo, se obtiene esta variable del IDH por un proceso un tanto mecánico; en efecto, se multiplica el IDH global de un país por un "coeficiente de

[43] http://www.conapo.gob.mx/m_en_cifras/principal.html (*Población de México en cifras*); en el "Menú de sección", ver "Índices de desarrollo humano".

disparidad" que corresponde a la razón de la parte del ingreso obtenido por el 20% de la población que se encuentra abajo de la escala entre la del 20% de la población que se encuentra arriba de la escala. Desde 1997, el PNUD busca más bien tomar en cuenta las desigualdades económicas con la ayuda de indicadores complementarios que se conocen como los "índices de la pobreza humana" (IPH).[44] El siguiente año, el PNUD empezó a calcular dos indicadores diferentes: uno (IPH-1) para los países en desarrollo y el otro (IPH-2) para los países industrializados.

Con el fin de calcular el IDH de manera separada para los hombres y para las mujeres, se acostumbra tener datos por sexo de la esperanza de vida, la alfabetización y la escolarización en varios países, pero no obstante es raro tener datos sobre la repartición del PIB. Antes de 1995, para obtener el PIB per cápita para las mujeres, se multiplicaba el PIB per cápita ajustado por una "razón global ingreso femenino-ingreso masculino"; esta última razón se calculaba multiplicando las dos razonas que siguen:

- La razón del salario de las mujeres en la industria entre el salario de los hombres.
- La razón de la tasa de participación de las mujeres en la población activa fuera de la agricultura entre esta misma participación de los hombres.

Según el PNUD, esta razón ingreso femenino-ingreso masculino "subestima la importancia de la discriminación en la medida que la diferencia entre las mujeres y los hombres es, por lo general, más grande en la agricultura y en los servicios que en la industria" (PNUD, 1994, p. 103).

A partir del *Informe* de 1995, el PNUD buscó tomar en cuenta con más precisión las diferencias socioeconómicas entre los sexos. Aunque las fórmulas que se usan sean compli-

[44] Se encuentra la matemática subyacente en las "Notas técnicas" del Informe del 1997.

cadas, el principio atrás del Índice de desarrollo relativo al género (IDG, antes llamado como "IDH sexoespecífico") es simple. De hecho, cada una de las variables que se usa en el cálculo del IDH es implícitamente un promedio[45] entre el valor de esta variable para los hombres y el valor para las mujeres. En el IDG, se reemplaza el promedio aritmético por un promedio armónico[46] que representa un valor que desinfla el valor del promedio en función del grado de desigualdad entre hombres y mujeres.

El IDG busca mejorar el IDH pero requiere también de formas funcionales a priori y de ponderaciones arbitrarias (aunque pueda parecer más científico por su complejidad). El motivo fundamental de esta situación es que no existen respuestas únicas a preguntas del tipo "¿Cómo tendríamos que medir la desigualdad?" y "¿Cuál peso tendríamos que darle a la desigualdad?"

Dimensiones ignoradas

A pesar de los trabajos de exploración efectuados en este sentido, los investigadores del PNUD no alcanzaron a proponer de manera satisfactoria que se tomaran en cuenta los alcances de los países en materia de medio ambiente.

Por otra parte, el IDH busca medir el desarrollo de las "capacidades" de los seres humanos siguiendo en esto las ideas de Sen. Respecto a esto, se debería considerar, sin duda, la tasa de desempleo puesto que ha de ser diferente recibir un

[45] Con más precisión, un promedio ponderado donde los pesos son proporcionales al número de personas de cada sexo.

[46] La fórmula que se usa es la siguiente:
$X = (p_m X_m^{1-\epsilon} + p_h X_h^{1-\epsilon})^{1/(1-\epsilon)}$
donde p_m es la proporción de mujeres entre la población y p_h, la proporción de hombres. El parámetro ϵ tiene que ser no negativo y diferente de 1. El valor escogido por el PNUD es 2. Si $X_m = X_h$, entonces $X = X_m = X_h$; por el contrario, el valor de X se sitúa en alguna parte entre los dos.

salario que una prestación social como ingreso (como pudo demostrarse en el caso particular de la relación entre el desempleo y la salud). De por sí, se toma en cuenta la tasa de desempleo para calcular el índice de la pobreza humana IPH2, el mismo que se aplica a los países industrializados.

> En el *Informe* del 2002, entre los 29 países con un desarrollo humano alto, Canadá resultó tercero en este rubro, después de Noruega y Australia, se clasifica al 11° lugar en cuanto a la pobreza humana después de Suecia, Noruega, los Países Bajos, Finlandia, Dinamarca, Alemania, Luxemburgo, Francia, Japón, España e Italia).

1-3.2.6 ¿A qué conclusión podemos llegar?

El IDH es un proyecto que moviliza recursos para recopilar y organizar datos sobre dimensiones del desarrollo humano que no aparecen en datos estrictamente económicos. Al enfocar la atención en estas dimensiones del desarrollo mucho tiempo olvidadas, el IDH destaca las debilidades del enfoque estrictamente económico, conlleva a una comprensión más justa de la situación y contribuye a abrir el diálogo sobre el desarrollo.[47] Además, es un instrumento de movilización política.

Sin embargo, nunca encontraremos una solución definitiva a este problema de medición del desarrollo y del progreso social, ya que esta solución no existe. Respeto a esto, sería

[47] Vea con respecto a este tema las palabras del mismo Amartya Sen (PNUD, 1998, p. 23): Se puede leer en el informe Rapport 2000 lo siguiente: "la información y las estadísticas constituyen un poderoso instrumento para construir una cultura de responsabilidad y aplicar los derechos del hombre. Los militantes, los juristas, los estadísticos y los especialistas del desarrollo necesitan cooperar con el pueblo y las comunidades con el fin de emitir datos y pruebas que sirvan a borrar el sentimiento de incredulidad y, por otro lado, fomenten el cambio de políticas y comportamientos" (p. 10).

muy grave e ilusorio dejar que la complejidad de los cálculos nos convenza de un rigor científico alcanzado. No obstante, es importante que sigamos con el empeño de buscar medir lo que no se puede medir, siempre y cuando esta conducta conlleve a una mejor comprensión de los límites de la medición y a un mejoramiento de los métodos de medición de las realidades sociales.

1-3.3 PARA SABER MÁS

En ciencias sociales y en estudios urbanos en particular, el tema de la construcción de índices es de gran actualidad. En efecto, a medida que se pierde el recuerdo de los "Treinta Gloriosos", la necesidad de administrar en condiciones apretadas se impone. Para las políticas sociales, esto significa que es necesario identificar muy bien las necesidades para establecer con precisión las intervenciones y, luego, medir los resultados con el fin de evaluar el grado de éxito de los programas. Estas tareas implican la medición de realidades complejas como la pobreza, la calidad de vida, la accesibilidad a la vivienda, etc. En pocas palabras, el IDH del PNUD procrea rápido.

En este contexto, los científicos tienen la responsabilidad de someter a examen crítico los múltiples indicadores propuestos, los cuales buscan, al igual que el IDH del PNUD, medir de la mejor manera lo que no se puede realmente medir. Como sugerencia para el lector que le interese emprender una exploración de los escritos sobre el tema, puede consultar las referencias completas en la bibliografía de este libro.

1-3.3.1 Los indicadores urbanos

El resumen que publicó la OCDE (1997) constituye un excelente punto de partida del tema al igual que Collin, Séguin y Pelletier (1999). La revista *Real Estate Economics* editó un

número especial con motivo de la conferencia *Habitat II* en Estambul en 1997. Además, el artículo de Combes y Wong (1994), aunque menos reciente, adopta un punto de vista metodológico y presenta una especie de "*how to...*", si bien muy interesante, un tanto confuso, a mi parecer, en relación con algunos puntos del método.

1-3.3.2 Un índice de estatus socioeconómico (Renaud y Mayer)

Hace ya varios años que Jean Renaud y Francine Mayer[48] trabajan en el desarrollo de un índice de estatus socioeconómico de los barrios urbanos que se basa en los datos del censo quinquenal. Su trabajo se parece mucho a la construcción de indicadores urbanos.

El modelo teórico subyacente a la construcción de este índice de status socioeconómico es el modelo de cohabitación de la ecología social urbana (Renaud *et al*., 1996, cap. 1). En este modelo, las personas que se parecen buscan agruparse en los mismos barrios "cada oveja con su pareja", lo que crea en el medio urbano una diferenciación espacial donde cada barrio se caracteriza por el tipo de gente que lo habita. De manera recurrente, los resultados de estudios empíricos arrojaron tres dimensiones "clásicas" que caracterizan la repartición espacial de la población:

- El status socioeconómico (riqueza/pobreza).
- El status familiar (presencia o no de niños, edad).
- La pertenencia étnica o lingüística.

El índice de estatus socioeconómico representa mucho más la primera de estas tres dimensiones (Renaud *et al*., 1996, pp. 35-51 y Anexo C, pp. 133-138). El proceso que se llevó a cabo pasa por cuatro grandes etapas:

[48] Renaud, Mayer y Lebeau (1996), Mayer-Renaud y Renaud (1989), Mayer-Renaud (1986).

1. Análisis de ecología factorial[49] de los datos del censo para confirmar de manera empírica las dimensiones que caracterizan la repartición espacial de la población.[50]
2. Identificación del contenido del o de los factores del tipo socioeconómico; el examen de los factores socioeconómicos revela que las variables socioeconómicas que son las que más contribuyen a caracterizar la repartición espacial, son el ingreso, la escolaridad y la profesión. Por lo tanto, las variables seleccionadas para construir el índice son:
   - Ingresos de los hogares.
   - Escolaridad de los individuos.
3. Construcción del índice con base en el modelo de la cohabitación.

Así que, la relación entre el índice de estatus socioeconómico y el modelo subyacente no es matemática en el sentido de que, como en el caso de los índices de precios, se puede deducir, de manera matemática, a partir de un modelo.[51] La

[49] Renaud *et al*. (1996, cap. 2). A grandes rasgos, el análisis factorial es un método estadístico de análisis multivariado por medio del cual se resume la información al reducir el número de variables gracias a la creación de variables "compuestas" (los "factores" que son análogos a unos índices formados a partir de las sumas ponderadas de las variables originales). Se busca dar una interpretación a los factores al momento de examinarlos, o sea que se busca asociarles un concepto. Es, para así decirlo, el proceso inverso de la construcción de índices: el concepto emerge de la interpretación de la composición de los factores en lugar de ser el punto de partida de la construcción del índice.

[50] "Para la selección de las variables y de la metodología [el índice] se apoya en los resultados de la ecología factorial aunque sin usar el marcador factorial para evitar la contaminación de las variables que pertenecen a otras dimensiones" (Renaud *et al*., 1986, p. 38).

[51] Es cierto que los modelos subyacentes a los índices de Laspeyres o de Paasche se basan en hipótesis extremadamente restrictivas, pero no obstante, se deduce de manera matemática la fórmula de cálculo de los índices a partir de los modelos.

relación es más bien "asociativa", es decir que se basa en mediciones de asociación estadística entre variables.

Sin embargo, el índice de estatus socioeconómico no tiene el carácter un tanto arbitrario del IDH. En este caso, los conceptos emergen del análisis estadístico: procedemos primero en el análisis de los datos (análisis factorial) para luego interpretar los resultados de este análisis a la luz de una hipótesis teórica (el modelo de ecología factorial); es basado en estos fundamentos que se definieron las dimensiones del concepto de estatus socioeconómico. Por el contrario, en el proceso de elaboración del IDH, se definió a priori el concepto de desarrollo humano y de inicio se le adjuntaron sus dimensiones (longevidad, saber y nivel de vida).

Con todo esto, el índice de estatus socioeconómico intenta medir una realidad compleja que encierra, además, dimensiones ordinales (o sea que no se pueden medir con variables de intervalo o racionales). Para combinar las múltiples dimensiones en una sola (en este caso, dos variables en un índice), es necesario considerar las variables ordinales[52] como si fueran racionales. En resumidas cuentas, la construcción del índice se fundamenta excesivamente en el juicio de valor del investigador, quien debe atribuir valores numéricos a las categorías.

1-3.3.3 Y más...

Cabe mencionar también que, en México, el Consejo Nacional de Población (CONAPO) elabora un índice de marginación de los municipios y hasta de las comunidades.[53] Este índice de marginación es un tanto parecido al

[52] La escolaridad, claro está, pero también el ingreso, ya que este último se conoce solamente por tramo (vea capítulo 1-1 con relación a las escalas de medición y, de manera más especifica, con relación a las variables racionales de intérvalo combinadas en clases).

[53] http://www.conapo.gob.mx/m_en_cifras/principal.html (*Población de México en cifras*); en el "Menú de sección", ver "Marginación".

índice de estatus socioeconómico de Renaud y Mayer, pero su formulación es más estrechamente relacionada a los resultados del análisis factorial.

CAPÍTULO 1-4
MEDICIÓN DE LA DESIGUALDAD Y DE LA CONCENTRACIÓN[*]

En este capítulo nos proponemos examinar los diferentes valores de una misma variable en un grupo de observaciones. Una medición de desigualdad (conocida también como "de disparidad") nos indica el grado en que los valores difieren unos de otros. Tomemos por ejemplo, los ingresos de los habitantes de un país; una medición de desigualdad del ingreso sirve para cuantificar el grado de desigualdad en la distribución del ingreso entre los habitantes de un país con el propósito de poder compararlo con el grado de desigualdad de otro país. En este ejemplo, la variable que se examina es el ingreso y las observaciones corresponden a los habitantes del país.

Cuando las observaciones corresponden a varias categorías y la variable examinada es el número de individuos (objetos) de una población dada que se encuentra en cada categoría, entonces una medición de desigualdad es también una medición de concentración. Por ejemplo, considerando la distribución de la población humana entre las regiones de un

[*] Referencias: Arriaga, 1975, pp. 65-71; Taylor, 1977, 179-185; Mills y Hamilton, 1989, pp. 413-414; Kendall y Stuart (1991, p. 58); Jayet (1993, pp. 18-29); Valeyre (1993); MacLachlan y Sawada (1997).

país, una medición de desigualdad indica el grado de concentración de su población.

En ciencias sociales, la medición de desigualdad es de utilidad en varios contextos diferentes: desigualdad en la distribución, concentración de la participación en el mercado (medición inversa del grado de competencia), concentración espacial de las poblaciones o de las actividades económicas, etcétera.

La construcción de medidas de desigualdad causa problemas parecidos a los de la multidimensionalidad al momento de definir los números índice. Se trata de resumir con una sola cifra una característica que posee todos los valores que toma una variable. No se puede esperar que haya una solución única.

En general, una medición de desigualdad compara la distribución observada con una distribución de referencia que representa la igualdad perfecta. Esta distribución de referencia es a menudo implícita. De otra manera, es a veces necesario hacerla explícita. Por ejemplo, en el caso de la distribución espacial de una población entre regiones, ¿una concentración cero corresponde a la situación cuando el número de habitantes sea lo mismo en todas las regiones? ¿O quizás, la distribución de referencia corresponde a la situación cuando el número de habitantes es proporcional a la superficie de las regiones? ¿O aún más, a la superficie habitable?

¿Cuáles son las propiedades deseables de una medición de desigualdad? Valeyre (1993) propone las seis propiedades siguientes:

1. Una medición de desigualdad no puede tomar valores negativos ya que es una medición del alejamiento de la distribución observada en relación con la distribución de referencia.

2. Una medición de desigualdad debe tomar el valor 0 si, y solamente si, la distribución observada es idéntica a la distribución de referencia.
3. Se deben tratar a todas las observaciones de la misma manera.
4. Una medición de desigualdad debe ser independiente del valor promedio de la variable examinada; una medición de concentración debe ser independiente del tamaño de la población cuya distribución se estudia.
5. La agregación de observaciones que tienen el mismo grado de especificidad no debe cambiar el valor de la medición.[54]
6. Principio de transferencia de Pigou-Dalton: una medición de desigualdad debe disminuir si se modifica la distribución de tal manera que reduce sin duda alguna la desigualdad.[55]

Estos principios permiten evaluar la validez de las diferentes mediciones de desigualdad propuestas. Así, la desviación estándar o la varianza no poseen la propiedad 4. Por lo contrario, el coeficiente de variación posee las seis propiedades enunciadas; correctamente usado, constituye, por lo tanto, una buena medición de desigualdad.

Veamos, enseguida, otros ejemplos de mediciones de desigualdad o de concentración.

---

[54] La especificidad se refiere a la razón entre un valor observado de la variable estudiada y el valor correspondiente en la distribución de referencia. Por ejemplo, los cocientes de localización son indicadores de especificidad. Una medición de la concentración geográfica del empleo de un ramo de actividad dada no debería verse afectada en caso de agregar dos regiones cuyos cocientes de localización sean iguales.

[55] Técnicamente, esto se traduce con la condición siguiente: si el valor de la variable disminuye al momento de una observación *i* y aumenta con la misma intensidad al momento de una observación *j* y, además, si el grado de especificidad de la observación *i* es superior al grado de especificidad de la observación *j*, entonces la medición de desigualdad debe disminuir.

1-4.1 El coeficiente de concentración de la economía industrial

Esta medición es sobre todo usada en economía industrial, aunque se usó también para medir el grado de concentración de la distribución por tamaño de las ciudades. Es, sencillamente, la suma de las partes de las *n* más grandes entidades. Por ejemplo, Rosen y Resnik (1980) miden el grado de concentración de una jerarquía urbana con la fracción de la población urbana total que se encuentra en las tres ciudades más grandes. En economía industrial, se mide a menudo la concentración del mercado con la suma de las partes de las cuatro empresas más grandes.

Esta medición tiene la ventaja de no ser muy exigente en términos de datos, sin embargo le falta la mayor parte de las propiedades deseables: posee únicamente la primera y la cuarta.

1-4.2 El índice de concentración de Hirschman-Herfindahl

Este índice es simplemente la suma de los cuadrados de las partes. Por ejemplo, para medir el grado de concentración en un sistema urbano compuesto de *n* ciudades, se puede calcular

$$H = \sum_{i=1}^{n} s_i^2$$

donde s_i es la fracción de la población urbana total que se encuentra en la ciudad *i*. El índice *H* varia entre $1/n$ y 1: toma el valor $1/n$ cuando todas las ciudades tienen el mismo tamaño y el valor 1 cuando toda la población urbana se concentra en una sola ciudad. Se interpreta a veces el índice *H* en

términos de "números equivalentes", particularmente, en economía industrial: en un mercado de, suponiendo, 40 empresas, si el índice *H* tiene un valor de *x*, se dice que el grado de concentración "equivale" al grado de un mercado de $1/x$ empresas que tienen partes de mercado iguales.

El índice Hirschman-Herfindahl no posee las propiedades 2 y 5. Además, este índice depende del número de observaciones *n*. Finalmente, es de notar que el índice *H* tiene un gran parecido a la varianza de partes: en efecto, esta última es igual a

$$\frac{1}{n}\sum_{i=1}^{n}\left(s_i - \frac{1}{n}\right)^2 = \frac{H}{n} - \frac{1}{n^2}$$

1-4.3 La curva de Lorenz y el índice de concentración de Gini

1-4.3.1 La diferencia promedio de Gini

Se conoce con este nombre el índice de concentración de Gini, en honor al estadístico italiano Corrado Gini (1884-1965). Este índice mide la desigualdad por medio de la diferencia entre todas los pares de observaciones (y_j, y_k). La suma ponderada de las diferencias se llama la "diferencia promedio de Gini" y se calcula, con datos agrupados, con la fórmula que sigue:[56]

$$\Delta = \frac{1}{N^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\left|y_j - y_k\right| f_j\, f_k$$

[56] En esta fórmula se compara cada observación con cada una de las observaciones, incluyendo ella misma; es la diferencia promedio con repetición. Kendall y Stuart (1991, p. 58) exhiben también la fórmula sin repetición. Cuando *N* es grande, la diferencia es despreciable.

donde

n es el número de valores distintos observados;

f_j es la frecuencia del valor y_j en la distribución,

de tal manera que

$N = \sum_{j=1}^{n} f_j$ es el número de observaciones.

Por ejemplo, en el caso de medir la desigualdad de la distribución del ingreso en Quebec, f_j sería el número de personas que tienen un ingreso de y_j; N es el número de personas en la población.

Cuando se agrupan las observaciones en clases, el valor y_j es el valor promedio de la variable Y en la clase j (no es el punto medio del intervalo de ingreso de la clase j).

Escribamos

$v_j = \frac{f_j}{N}$, la fracción de la población que pertenece a la clase j.

Entonces, el valor promedio de la variable Y se escribe

$$\mu = \frac{1}{N} \sum_{j=1}^{n} f_j \, y_j = \sum_{j=1}^{n} v_j y_j$$

Teniendo

$M = \sum_{j=1}^{n} f_j y_j$, la suma de los valores de la variable Y

y

$w_j = \frac{f_j y_j}{\sum_{k=1}^{n} f_k y_k} = \frac{f_j y_j}{N\mu} = \frac{v_j y_j}{\mu}$, la fracción de la suma correspondiente a la clase j.

Luego, arreglemos las observaciones con el objetivo de construir una curva de Lorenz (vea más abajo), por orden creciente de las razones w_j / v_j . Completamos la simbología con

$$Cw_j = \sum_{k=1}^{j} w_k$$

Es la fracción acumulada correspondiente a las clases desde la *i* hasta la *j*.

Desarrollando la fórmula de cálculo de la diferencia promedio de Gini, podemos mostrar que:

$$\Delta = 2\mu \left(1 - \sum_{j=1}^{n} v_j \, Cw_j \; - \sum_{j=1}^{n} v_j \, Cw_{j-1} \right)$$

1-4.3.2 Cálculo del índice de concentración de Gini

El índice de concentración de Gini es simplemente la razón de la diferencia promedio de Gini entre dos veces el promedio:

$$G = \frac{\Delta}{2\mu} = 1 - \left(\sum_{j=1}^{n} v_j \, Cw_j \; + \sum_{j=1}^{n} v_j \, Cw_{j-1} \right)$$

$$= 1 - \sum_{j=1}^{n} v_j \left(Cw_j + Cw_{j-1} \right)$$

Arriaga (1975, pp. 65-71), así como varios geógrafos, definen el coeficiente de Gini como

$$G = \sum_{i=2}^{n} Cw_i \, Cv_{i-1} - \sum_{i=2}^{n} Cw_{i-1} \, Cv_i$$

donde $Cv_j = \sum_{k=1}^{j} v_k$

Esta fórmula puede deducirse de la precedente:

$$G = 1 - \sum_{j=1}^{n} v_j \left(Cw_j + Cw_{j-1} \right)$$

$$G = 1 - \sum_{j=1}^{n} \left(Cv_j - Cv_{j-1} \right) \left(Cw_j + Cw_{j-1} \right)$$

$$G = 1 - \sum_{j=1}^{n} Cv_j \, Cw_j - \sum_{j=1}^{n} Cv_j \, Cw_{j-1} + \sum_{j=1}^{n} Cv_{j-1} \, Cw_j + \sum_{j=1}^{n} Cv_{j-1} \, Cw_{j-1}$$

donde

$$\sum_{j=1}^{n} Cv_{j-1} \, Cw_{j-1} = \sum_{j=0}^{n-1} Cv_j \, Cw_j \quad \text{et } Cv_0 = Cw_0 = 0,$$

de suerte que

$$-\sum_{j=1}^{n} Cv_j \, Cw_j + \sum_{j=1}^{n} Cv_{j-1} \, Cw_{j-1} = -\sum_{j=1}^{n} Cv_j \, Cw_j + \sum_{j=1}^{n-1} Cv_j \, Cw_j = -Cv_n Cw_n = -1$$

y

$$G = 1 - 1 - \sum_{j=1}^{n} Cv_j \, Cw_{j-1} + \sum_{j=1}^{n} Cv_{j-1} \, Cw_j$$

como $Cv_0 = Cw_0 = 0$, esto equivale a

$$G = -\sum_{j=2}^{n} Cv_j \, Cw_{j-1} + \sum_{j=2}^{n} Cv_{j-1} \, Cw_j$$

$$G = \sum_{j=2}^{n} Cw_j \, Cv_{j-1} - \sum_{j=2}^{n} Cw_{j-1} \, Cv_j$$

Cuando las observaciones están agrupadas, para calcular el índice de concentración de Gini sólo basta conocer la distribución entre las clases de la población (los v_j) y la distribución entre las clases de la suma de los valores de la variable Y (los w_j).

Primero, se acomoda de manera frecuente los habitantes (o los hogares) en orden creciente de ingreso; luego, se definen categorías de tamaños idénticos: quartiles, quintiles, deciles, etc. Con esto, se podrá comentar: "el 20% de la población con los ingresos más elevados (el quintil superior) acapara *xx*% del ingreso global mientras que los 20% con los ingresos más bajos (el quintil inferior) reciben únicamente el *zz*% de los ingresos globales. Tales enunciados dan de esta manera una medición de la concentración, pero, y a diferencia del coeficiente de Gini, son mediciones parciales que toman solo en cuenta una parte de la distribución.

Veamos, por ejemplo, la repartición del ingreso (y) entre las familias de Canadá en 1995 (la población que se toma en cuenta es, por lo tanto, las familias y no los individuos). Recientemente, *Statistique Canada* difundió los datos siguientes del censo de la población de 1996.[57]

[57] *Le Quotidien*, 3 de marzo de 1999. Es importante notar que los datos del censo de 1996 sobre los ingresos anuales se aplican al año anterior.

Tabla: Límites superiores (en $ de 1995) de los deciles del ingreso familiar y repartición del ingreso global familiar por decil, 1995

Decil	Límite superior	Parte del ingreso global (%)
Primero	15158	1.45
Segundo	23184	3.55
Tercero	31097	4.96
Cuarto	38988	6.42
Quinto	46951	7.86
Sexto	55355	9.37
Séptimo	64997	10.91
Octavo	77501	13.11
Noveno	98253	15.85
Décimo		26.53

Se pueden representar de otra manera los datos de esta tabla, como sigue:

Clase de ingresos ($ de 1995)	Fracción del número de familias (%)	Parte del ingreso global (%)
0-15158	10.00	1.45
15159-23184	10.00	3.55
23185-31097	10.00	4.96
31098-38988	10.00	6.42
38989-46951	10.00	7.86
46952-55355	10.00	9.37
55356-64997	10.00	10.91
64998-77501	10.00	13.11
77502-98253	10.00	15.85
98254 y más	10.00	26.53

En la tabla anterior, los w_j son parte del ingreso global; los v_j son todos iguales a 10%. A partir del cuadro antecedente, los cálculos preliminares del índice de Gini se hacen así:

Clase de ingresos ($ de 1995)	Fracción del número de familias (%)	Parte del ingreso global (%)			
	v_j	w_j	Cw_j	$v_j\, Cw_j$	$v_j\, Cw_{j-1}$
0-15158	10.00	1.45	0.0145	0.0015	0.0000
15159-23184	10.00	3.55	0.0500	0.0050	0.0015
23185-31097	10.00	4.96	0.0996	0.0100	0.0050
31098-38988	10.00	6.42	0.1638	0.0164	0.0100
38989-46951	10.00	7.86	0.2424	0.0242	0.0164
46952-55355	10.00	9.37	0.3361	0.0336	0.0242
55356-64997	10.00	10.91	0.4452	0.0445	0.0336
64998-77501	10.00	13.11	0.5763	0.0576	0.0445
77502-98253	10.00	15.85	0.7348	0.0735	0.0576
98254 y más	10.00	26.53	1.0001	0.1000	0.0735
Total	100.00	100.00		0.3663	0.2663

De modo que el índice de concentración de Gini del ingreso familiar en Canadá por decil en 1995 es igual a:

$$G = 1 - (0.3663 + 0.2663) = 0.3675$$

Es importante hacer dos observaciones en este momento:

- Los datos que se usaron en este caso fueron acomodados naturalmente en un orden creciente de las razones w_j / v_j . No es siempre el caso; en general, antes de calcular el índice de Gini, es necesario acomodar previamente los datos en orden correcto (vea el ejemplo extraído de Taylor, 1997, más abajo).
- Con datos agrupados, el índice de concentración de Gini depende del agrupamiento o del tipo de clasificación que se usa. Si se hubiera agrupado las familias por quintiles o por centiles los resultados del cálculo hubieran sido diferentes. Regresaremos más tarde en este punto.

1-4.3.3 La curva de Lorenz

La curva de Lorenz es un instrumento de comparación gráfico entre dos distribuciones.

Recordemos que

$Cv_j = \sum_{k=1}^{j} v_k =$ fracción acumulada de X (por ejemplo, de las familias, antes mencionado).

$Cw_j = \sum_{k=1}^{j} w_k =$ fracción acumulada de Y (por ejemplo, de los ingresos, antes mencionado).

Tenemos naturalmente:

$Cv_n = Cw_n = 1$

Método de construcción de la curva de Lorenz (vea el ejemplo numérico más abajo, extraído de Taylor, 1997, p. 179).

1. Calcular las razones $\frac{w_i}{v_i}$ [58].
2. Reordenar las categorías en orden creciente de $\frac{w_i}{v_i}$:

$$\frac{w_1}{v_1} < \frac{w_2}{v_2} < \cdots < \frac{w_n}{v_n}$$

3. Calcular las razones cumulativas Cv_i y Cw_i
4. La curva de Lorenz es el conjunto de los (Cv_i, Cw_i), donde los Cv_i se sitúan sobre el eje horizontal.

La curva de Lorenz tiene las propiedades que siguen:

[58] Estas razones no son más que las especificidades asociadas a las observaciones.

1. $Cv_0 = Cw_0 = 0$ (por definición, de Cv_i y de Cw_i): la curva empieza desde el origen; $Cv_n = Cw_n = 1$ (por definición, de Cv_i y de Cw_i): la curva termina en el punto de coordenadas [1,1] (o [100%, 100%]).
2. Cuando las dos distribuciones son idénticas, tenemos, para todo *i*,
3. $Cv_i = Cw_i$ es decir que la curva de Lorenz coincide con la diagonal.
4. $Cv_i \geq Cw_i$ para cada *i* diferente de 0 y de *n* (por construcción dado el reordenamiento de las categorías): la curva de Lorenz se encuentra debajo de la diagonal o coincide con ella;
5. La pendiente de cada segmento de la curva de Lorenz es igual al valor del indicador de especificidad asociado a la observación correspondiente:

   pendiente del segmento $i = = \dfrac{Cw_i - Cw_{i-1}}{Cv_i - Cv_{i-1}} = \dfrac{w_i}{v_i}$
6. La curva de Lorenz es cóncava hacia arriba, es decir que cada segmento tiene una pendiente más abrupta que la anterior: esto se infiere de la propiedad 5, donde por construcción $\dfrac{w_i}{v_i} < \dfrac{w_{i+1}}{v_{i+1}}$.

Construcción de una curva de Lorenz
(ejemplo numérico extraído de Taylor)

Primera etapa : cálculo de los w_i / v_i

Zona	x_i Número de hogares de clase media	v_i Distrib. de x	y_i Número de votos del partido Republícano	w_i Distrib. de y	w_i/v_i
A	30	0.25	30	0.30	1.20
B	20	0.17	15	0.15	0.90
C	10	0.08	8	0.08	0.96
D	10	0.08	5	0.05	0.60
E	20	0.17	19	0.19	1.14
F	30	0.25	23	0.23	0.92
Total	120	1.00	100	1.00	

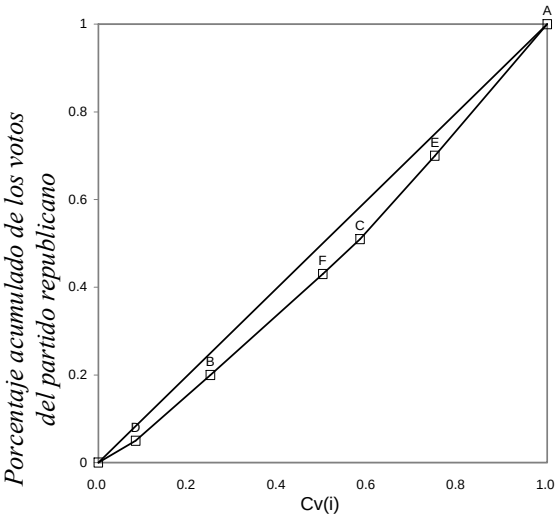
Segunda etapa: clasificación por orden creciente de los w_i/v_i.

Tercera etapa: cálculo de los abscisas y los ordenadas

Zona	x_i	v_i	y_i	w_i	w_i/v_i	Abscisas Cv_i	Ordenadas Cw_i	Diferencia $(Cv_i - Cw_i)$	Diferencia $\lvert v_i - w_i \rvert$
						0.00	0.00		
D	10	0.08	5	0.05	0.60	0.08	0.05	0.033	0.033
B	20	0.17	15	0.15	0.90	0.25	0.20	0.050	0.017
F	30	0.25	23	0.23	0.92	0.50	0.43	0.070	0.020
C	10	0.08	8	0.08	0.96	0.58	0.51	0.073	0.003
E	20	0.17	19	0.19	1.14	0.75	0.70	0.050	0.023
A	30	0.25	30	0.30	1.20	1.00	1.00	0.000	0.050
Total	120	1.00	100	1.00				0.147	

Nota: podemos constatar que la diferencia máxima entre la curva de Lorenz y la diagonal es igual a $\frac{1}{2}\sum_i |v_i - w_i|$.

Curva de Lorenz



Porcentaje acumulado de hogares de clase media

Cuarta etapa: cálculo del índice de concentración de Gini

						Abscisas	Ordenadas		
Zona	x_i	v_i	y_i	w_i	w_i/v_i	Cv_i	Cw_i	$v_i\ Cw_i$	$v_i\ Cw_{i-1}$
						0.00	0.00		
D	10	0.08	5	0.05	0.60	0.08	0.05	0.004	0.000
B	20	0.17	15	0.15	0.90	0.25	0.20	0.033	0.008
F	30	0.25	23	0.23	0.92	0.50	0.43	0.108	0.050
C	10	0.08	8	0.08	0.96	0.58	0.51	0.043	0.036
E	20	0.17	19	0.19	1.14	0.75	0.70	0.117	0.085
A	30	0.25	30	0.30	1.20	1.00	1.00	0.250	0.175
Total	120	1.00	100	1.00				0.554	0.354

$$G = 1 - (0.554 + 0.354) = 0.092$$

1-4.3.4 Cálculo geométrico del índice de Gini por medio de la curva de Lorenz

En realidad fue extraordinaria la hazaña de Corrado Gini al demostrar en 1914 que el índice de concentración que lleva su nombre es igual a la razón entre 1) la superficie contenida entre la diagonal y la curva de Lorenz, y 2) la superficie total debajo de la diagonal:

$$G = \frac{\text{Superficie contenida entre la diagonal y la curva de Lorenz}}{\text{Superficie total debajo de la diagonal}}$$

La superficie total del triángulo debajo de la diagonal se calcula de la manera siguiente:

$$\frac{Cw_n \times Cv_n}{2} = \frac{1 \times 1}{2} = \frac{1}{2}$$

La superficie contenida entre la diagonal y la curva de Lorenz se calcula como la diferencia entre la superficie total del triángulo debajo de la diagonal ($=\frac{1}{2}$) y la superficie debajo de la curva de Lorenz. La superficie debajo de la curva de Lorenz (vea ejemplo numérico anterior y la figura más abajo) es la suma de *n* trapecios que tienen, cada uno, una superficie igual a:

$$\frac{1}{2} v_i \left(Cw_i + Cw_{i-1} \right)$$

La superficie debajo de la curva de Lorenz es por consiguiente la suma de estas *n* superficies:

$$\frac{1}{2} \sum_{i=1}^{n} v_i \left(Cw_i + Cw_{i-1} \right)$$

Y el coeficiente Gini se calcula así:

$$G = \frac{\left(\frac{1}{2}\right) - \left(\frac{1}{2}\sum_{i=1}^{n} v_i \left(Cw_i + Cw_{i-1}\right)\right)}{\left(\frac{1}{2}\right)}$$

$$= 1 - \sum_{i=1}^{n} v_i \left(Cw_i + Cw_{i-1}\right) = \frac{\Delta}{2\mu}$$

lo que corresponde perfectamente a la fórmula enunciada con anterioridad.

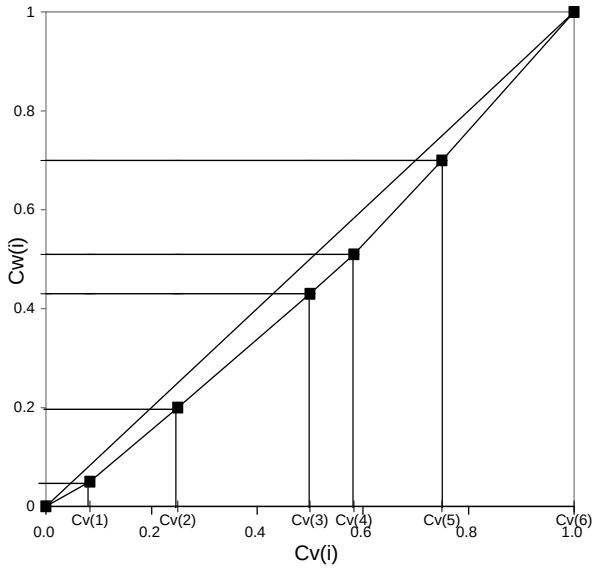
Para facilitar la interpretación de la curva de Lorenz y del índice de Gini asociado a esta curva, es importante recordar que es la distribución *V* que toma el papel de distribución de referencia (o sea de igualdad perfecta o de concentración cero). En la curva de Lorenz, los Cv_i se localizan en el eje horizontal y los Cw_i,, en el eje vertical.

Ejemplos:

Si *V* es una repartición del territorio entre las zonas y los *W*, la repartición de la población, el coeficiente de Gini es la medición de la concentración geográfica de la población.

Si *V* es una distribución de la población (o de los hogares) en *n* categorías y *W*, la distribución del ingreso anexada por categoría, entonces el coeficiente de Gini es la medición de la concentración del ingreso.

Cálculo geométrico del índice de concentración de Gini



1-4.3.5 Propiedades del índice de concentración de Gini

El índice de concentración de Gini posee las seis propiedades que debe tener una medición de desigualdad, tal y como fueron enunciadas al principio de este capítulo. Además, posee las propiedades siguientes:

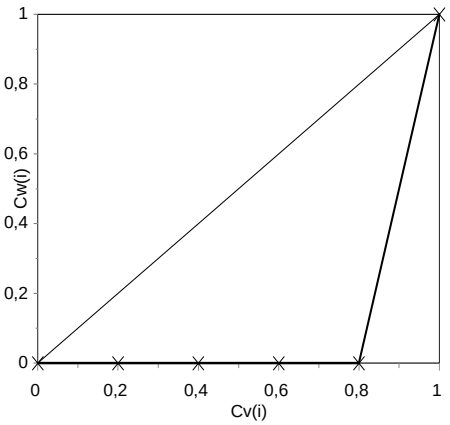
1. El índice de Gini varía entre 0 y 1.[59] El coeficiente de Gini toma el valor 0 cuando las dos distribuciones son idénticas. Toma su valor máximo teórico 1 cuando la curva de Lorenz sigue la base y el lado derecho de la "caja"; sin embargo, para alcanzarse el máximo teóri-

[59] O entre 0% y 100% cuando lo expresamos en porcentaje.

co, es necesario que el número de categorías tienda hacia el infinito de tal manera que v_n tienda hacia 0.[60]

2. Se puede demostrar que el índice de Gini es simétrico, es decir, que se puede intercambiar el papel de las dos distribuciones; en otras palabras, si invertimos los papeles, el valor del coeficiente de Gini no cambia.
3. Cuando los datos son agrupados, el índice de Gini es sensible a la definición y al número de categorías usadas (clases, zonas).
4. Cuando se usa como medición de concentración espacial, el índice de Gini no toma en cuenta, de ninguna manera, la proximidad en el espacio entre las diferentes zonas de fuerte densidad (se considera el espacio como un rompecabezas no hecho).

Guardando el valor máximo que puede lograr el coeficiente de Gini cuando el número de categorías no es infinito, precisamos que es igual a $1 - v_n$. Esta propiedad se ilustra en la siguiente figura.



[60] De otra manera, cuando $v_n > 0$, el valor máximo de G es igual a $1 - v_n$.

En este ejemplo, es fácil verificar, aplicando el método de cálculo geométrico, que $v_n = 0.2$ y $G = (1 - 0.2) = 0.8$.

La tercera propiedad merece que la examinemos con más énfasis. En particular, se manifiesta por lo siguiente: la agregación de dos o más categorías siempre implica una reducción del valor calculado del coeficiente de Gini (al menos que las dos categorías tengan la misma especificidad, y en este caso se aplica la propiedad 5 de las mediciones de la desigualdad). Este hecho se verifica fácilmente si pensamos en el cálculo efectuado con la ayuda de la curva de Lorenz: la agregación de dos categorías vecinas reduce el espacio contenido entre la curva de Lorenz y la diagonal. Viene a confirmar también la intuición de que la agregación de categorías implica borrar una parte de las diferencias.

Esta sensibilidad del Gini a la definición de las categorías puede seriamente comprometer su fiabilidad como medición de la concentración, y aún más cuando las categorías son de tamaños diferentes. Para ilustrar este fenómeno, imaginemos que queramos comparar la concentración de la población en dos momentos diferentes, en un territorio dividido en tres zonas con la misma superficie (acordamos que igual a 1):

	Superf.	Población		Densidad	
		al momento 0	al momento *t*	al momento 0	al momento *t*
Zona 1	1	10	80	10	80
Zona 2	1	80	10	80	10
Zona 3	1	10	10	10	10

Está claro en este ejemplo que la concentración quedó igual a la escala considerada ($G = 0.47$)), aunque el centro de gravedad de la población se haya desplazado hacia la zona 1.

Supongamos ahora que agregamos las zonas 2 y 3:

	Superf.	Población		Densidad	
		al momento 0	al momento *t*	al momento 0	al momento *t*
Zona 1	1	10	80	10	80
Zonas 2 y 3	2	90	20	45	10

Los datos agregados hacen creer que la concentración aumentó puesto que tenemos $G = 0.23$ al momento 0 y $G = 0.47$ al momento *t* (observe que el índice de Gini es más pequeño con los datos agregados al momento 0, pero es el mismo al momento *t* puesto que, en este último caso, las zonas agregadas son de misma densidad, es decir de la misma especificidad).

1-4.4 Conclusión con respecto a la medición de la desigualdad

Acabamos de citar solamente algunas de las múltiples mediciones de desigualdad que se proponen ahora. Entre las mediciones que no hemos mencionado, existen las mediciones de entropía, como la medición de Shanon, o la medición de la ganancia de información de Kullback-Leibter (también asociado al nombre de Theil). El lector interesado podrá consultar el resumen de Valeyre (1993).

Finalmente, recordemos que las mediciones de desigualdad son mediciones de alejamiento de una distribución observada con relación a una distribución de referencia. Por esta razón, se parecen mucho a las mediciones de disimilitud, las cuales comparan dos distribuciones cuyos papeles son simétricos (ninguna de las dos juega el papel de referencia).

CAPÍTULO 1-5
MEDICIÓN DE LA DISIMILITUD

1-5.1 MULTIDIMENSIONALIDAD, DISIMILITUD Y CONCENTRACIÓN

1-5.1.1 Problemática de la medición de la disimilitud

Vimos que una medición asociada a un concepto establece una correspondencia entre objetos y números. Lo que permite comparar objetos y determinar el valor de verdad de una o varias de las relaciones $=, \neq, <$ o $>$. Si un concepto contiene varias dimensiones, lo que es frecuente y queremos sin embargo tratarlo como un todo, vimos que al construir un índice se resuelve este problema.

No obstante, suele suceder que nos topemos con un concepto que no admite que se le asocie otra medición que no sea categórica; en estas condiciones, la construcción de un índice se vuelve imposible. Consideremos, por ejemplo, el concepto de estructura económica de una ciudad o de una región. Aunque definamos la estructura económica como la repartición del empleo entre las ramas de actividad, nunca podremos asociar a este concepto otra medición que no sea una clasificación (variable categórica): ciudad monoindustrial, ciudad de servicio, etc. Pero ¿cómo llegamos a construir

una clasificación que permita captar bien la realidad? Una manera de proceder consiste en comparar los objetos (en este caso, las estructuras económicas observadas) para constituir grupos de objetos lo suficientemente parecidos entre ellos, y claramente diferente de los objetos de los demás grupos. Una clasificación de este tipo puede, luego, servir de base para la elaboración de una tipología y para la definición de una variable categórica asociada al concepto.

Aprovechamos para mencionar que, aunque la construcción de un índice sea posible en principio, el proceso que acabamos de evocar puede ser de gran utilidad en caso de que no se pueda construir un índice que llene todas las expectativas del plan teórico. ¿Por ejemplo, al constituir una tipología de los países, podremos estudiar el desarrollo humano? Este tipo de tipología permitiría definir un índice apropiado para cada tipo de país con el objetivo de comparar únicamente países comparables con mediciones adaptadas a las características de estos países (es lo que ya efectúa el PNUD en relación con la medición de la pobreza: calcula dos "índices de la pobreza humana", uno para los países en vía de desarrollo y otro para los países desarrollados).

Formalizar el concepto de similitud y asociarle una medición no puede más que facilitar el proceso de clasificar objetos por tipos. De por sí existen procedimientos de clasificación automática basados en las mediciones de similitud[61]. Además, desearíamos, a veces, sólo tomar en cuenta un proceso heurístico con menos formalidades y examinar el grado de similitud entre los objetos sin tener que construir una tipología. Nuevamente, en este caso, la medición de la similitud puede convertirse en una herramienta de gran provecho. Vamos a hablar, por consiguiente, de la medición de similitud en esta parte.

[61] Dendrogramas, algoritmos de partición automáticos, etc. Vea Legendre y Legendre (1984 y 1998).

Para empezar, observemos que el concepto de similitud se aplica a un par de objetos. La similitud no es, por lo tanto, una propiedad de alguno de los dos objetos: es una propiedad del par.[62] Luego, el concepto de similitud es un concepto general que encierra en sí, una mirada de conceptos específicos; de hecho, examinar la similitud entre dos objetos es siempre *en relación con* un atributo en particular. Al momento de querer medir la similitud entre dos objetos, se define un concepto de similitud específico por el atributo que se selecciona para comparar estos objetos. En el caso de ciudades, por ejemplo, podemos considerar la similitud en relación con la estructura demográfica, con la tasa de criminalidad, con la calidad de vida, etcétera.

De entrada estamos de acuerdo para decir que la medición de similitud con relación a un atributo unidimensional es algo trivial puesto que no representa un gran problema para medir como, por ejemplo, la similitud entre dos países en cuanto al número de habitantes, a la tasa de criminalidad o al valor del IDH del PNUD.[63] Por el contrario, en caso de querer medir la similitud con relación a una propiedad multidimensional que no se resumió en un índice con anterioridad,[64] nos enfrentamos al mismo problema de construir un número índice. Por ejemplo:

> ¿Con relación a su estructura económica, cuál es el grado de similitud entre Quebec y Ontario?

[62] Se podría decir que el objeto al cual se aplica la similitud es un par de objetos.

[63] Este ejemplo es deliberadamente paradójico: si bien, el IDH es un indicador que permite medir una realidad multidimensional, la comparación de dos países en relación con el valor de este índice es, por su lado, unidimensional.

[64] O bien, y es exactamente lo mismo, medir al mismo tiempo la similitud bajo condiciones diferentes o, dicho de otra manera, medir la similitud entre dos objetos multidimensionales.

¿Con relación a su repartición en el territorio, cuál es el grado de similitud entre el cultivo de plátanos y la ganadería en Costa Rica?

Para medir la similitud en estos dos ejemplos, tenemos que tomar en cuenta más de una dimensión puesto que examinamos la similitud con relación a un concepto que contiene más de una dimensión.

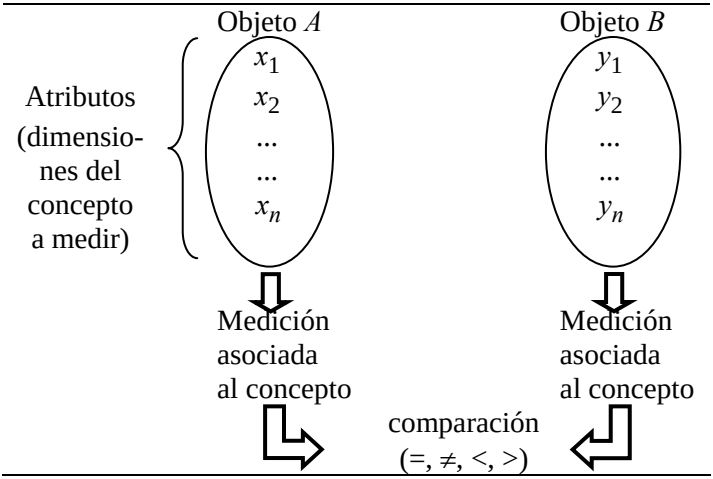
- En el caso de la similitud entre países en cuanto a su estructura económica, tenemos que considerar las diferentes ramas de producción.

En el caso de la similitud entre actividades en cuanto a la repartición espacial, tenemos que considerar las diferentes partes del territorio (zonas, distritos, provincias u otros en función del tipo de recorte geográfico que se usa).

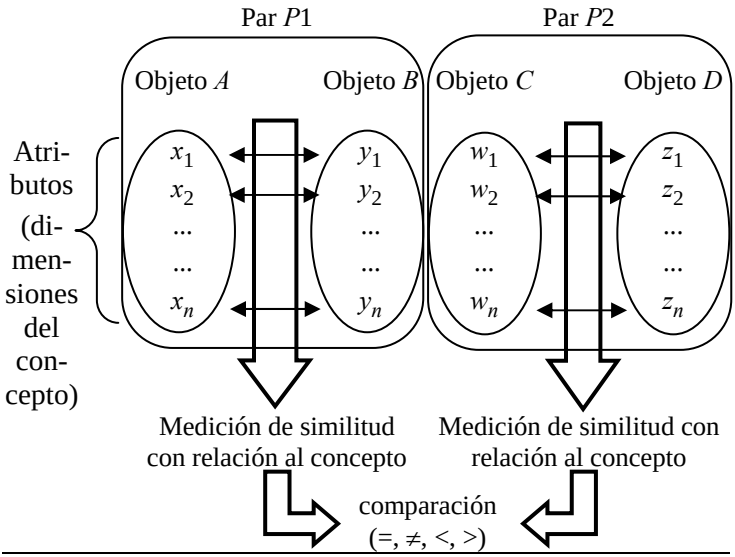
Así, las mediciones de la similitud entre objetos multidimensionales se parecen mucho a índices. A continuación, con el fin de resaltar las diferencias, vamos a centrar nuestro estudio en destacar las partes específicas de la medición de la similitud.

Como ya sabemos, un índice resume en una sola cifra los valores de los indicadores asociados a las múltiples dimensiones de un concepto. El índice es una medición porque permite comparar dos objetos en cuanto al grado que poseen de la propiedad definida por el concepto. Esto se resume con el diagrama siguiente.

Construcción de un número índice



En cambio, para medir la similitud, empezamos por comparar dos objetos detalle por detalle hasta obtener cuantas mediciones parciales de similitud como haya dimensiones que comparar. Es necesario, después, agregar todas estas mediciones parciales en una sola, lo que da por resultado una medición de la similitud. Esta medición permite comparar dos pares de objetos en cuanto a su similitud con relación a un atributo multidimensonal dado. Se resume esto en el diagrama que sigue, lo que al compararlo con el otro diagrama, permite entender las diferencias que existen entre la medición de la similitud entre objetos multidimensionales y la construcción de un índice.



De todas maneras, se trata efectivamente de una medición tal y como se definió en el capítulo 1-1. Recordemos que una medición asociada a un concepto establece una correspondencia entre los objetos y los números lo que permite comparar los objetos y determinar el valor de verdad de una o varias relaciones =, ≠, <, >. Por lo tanto, una medición de la similitud es una correspondencia que permite comparar dos pares de objetos cualesquiera en cuanto a su similitud con relación a un atributo dado. De manera formal, si convenimos que $f(A,B)$ es la medición de la similitud entre los objetos del par $[A,B]$ y que $f(C,D)$ es la medición de la similitud entre los objetos del par $[C,D]$, entonces, una medición de la similitud permite decidir el valor de verdad de una o varias de las relaciones siguientes:

$$f(A,B) = f(C,D)$$
$$f(A,B) \neq f(C,D)$$
$$f(A,B) < f(C,D)$$
$$f(A,B) > f(C,D)$$

Por ejemplo, si *A* es Nicaragua, *B* es Costa Rica, *C* es Costa Rica y *D* es Canadá,[65] una medición de la similitud permite contestar a la pregunta: "¿Con relación a la composición de su producción, Costa Rica se parece más a Nicaragua o a Canadá?" De igual manera, si *A* es el cultivo de plátano, *B* es la ganadería, *C* es el cultivo de plátano y *D* es el cultivo de cítricos, una medición de la similitud permite contestar a la pregunta: "¿Con relación a su repartición geográfica en Costa Rica, el cultivo de plátanos se parece más a la ganadería o al cultivo de cítricos?".

Notemos que nada de lo anterior mencionado implica que siempre tengamos que medir basándonos en una escala racional aunque las variables que se usan como medición de la similitud o de la disimilitud sean, a menudo, variables racionales. No obstante, el problema de la multidimensionalidad provoca que, por lo común, haya varias mediciones posibles sin que ninguna pueda, a priori, calificarse como la mejor. Es la razón por la cual, excepto en casos particulares, que es preferible interpretar las mediciones de similitud como mediciones ordinales y evitar interpretarlas de manera abusiva como mediciones de intervalo o racionales.

Además, y como lo veremos, las mediciones de similitud son, a menudo, mediciones inversas, es decir que son más bien mediciones de disimilitud. Debemos estar muy atentos a este aspecto que es causa de mucha confusión.

[65] Tal como nos lo muestra este ejemplo, puede suceder que $B = C$ (o $B = D$, o $A = C$, o $A = D$), sin embargo no sucede necesariamente.

1-5.1.2 La medición de la similitud entre distribuciones

Una distribución o una repartición es una propiedad (multidimensonal) de una población (en la aceptación general de una colección de personas u objetos) cuando se clasifica esta población en categorías: es el número de individuos o la fracción de la población que se encuentra en cada una de las categorías. En los ejemplos ya mencionados:

> Las personas empleadas en una economía constituyen una "población" que se puede clasificar en "categorías" como las ramas de actividad. Se puede describir la estructura económica de un país con una distribución que representa el número de personas empleadas por rama de actividad.
>
> Las hectáreas de un terreno que se dedican a una actividad dada (el cultivo de plátanos, por ejemplo) constituyen una "población" que se puede clasificar en las "categorías" como las subdivisiones (provincias u otras) de un territorio. Se pude describir la repartición espacial de las actividades con una distribución que representa el número de hectáreas que se dedican a la actividad en cada subdivisión del territorio.

Por consiguiente una distribución es un objeto multidimensional. No obstante, la comparación entre dos distribuciones no es tan complicada gracias a la existencia de una "regla de normalización" natural que marca lo siguiente: la medición asociada a cada una de las dimensiones de la distribución es simplemente la fracción de la población que pertenece a la categoría correspondiente. Ahora bien, en una distribución, la suma de las partes es necesariamente igual a 1. De entrada, esto elimina parte del problema de la multidimensionalidad que era, como lo vimos, el problema de los números índice y del peso que se le atribuye a cada una de las dimensiones.

Por el contrario, cuando intentamos comparar dos objetos que no sean distribuciones, la selección de la unidad de medición de cada dimensión de la comparación determina de manera implícita cuál será su peso en la medición de la disimilitud. Es entonces, en estas condiciones, que se intensifica el problema de la multidimensionalidad ya mencionado con relación a los números índice.

1-5.1.3 Disimilitud y desigualdad-concentración: ¿cuál es la diferencia?

En los ejemplos mencionados hasta el momento, sólo buscábamos examinar el grado de asociación entre dos fenómenos o a la inversa, el grado de segregación entre ellos. Sin embargo, existe otro uso de las mediciones de disimilitud entre dos distribuciones; este otro uso es la medición de la concentración o de la dispersión. Una medición de disimilitud se convierte en una medición de concentración cuando se compara la distribución estudiada con una distribución de referencia o teórica. Esta distribución teórica, que sirve de referencia, representa una concentración nula y sirve, de alguna manera, de patrón de medición (exhibiremos un ejemplo de esto más abajo).

Esto es del todo coherente con lo estudiado en el capítulo 1-4 puesto que, en general, una medición de la desigualdad compara la distribución observada con una distribución de referencia, la cual representa la igualdad perfecta, Por lo tanto, una medición de la desigualdad es una medición de disimilitud entre una distribución observada y una distribución de referencia.

De ahí que el índice de Gini resulte ser de igual manera adecuado como medición de disimilitud que como medición de desigualdad. De por sí, se mencionó con anterioridad que el índice de Gini, entre otras propiedades, era simétrico, es decir que los papeles de la distribución observada y de la dis-

tribución de referencia son intercambiables; en otras palabras, al intercambiar los papeles, el valor del coeficiente de Gini no cambia.

1-5.2 EL ÍNDICE DE DISIMILITUD

1-5.2.1 Un ejemplo numérico

Ahora, nos interesamos en una medición de disimilitud ampliamente usada, la cual se aplica a las distribuciones como, por ejemplo, la repartición geográfica del empleo. Mostramos un ejemplo numérico ficticio:

Empleo por zona y por rama

Ramo	*B1*	*B2*	*B3*	Total
Zona				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

Se trata de una medición de similitud entre las ramas de actividades en cuanto a su repartición geográfica. Por lo tanto, nos interesa conocer la fracción del empleo de cada rama en cada zona:

Distribución del empleo entre zonas

Ramo	*B1*	*B2*	*B3*	Total
Zona				
Z1	0.400	0.542	0.598	0.550
Z2	0.225	0.308	0.308	0.300
Z3	0.375	0.150	0.094	0.150
Total	1.000	1.000	1.000	1.000

La solución más simple que podemos imaginar para examinar la similitud entre dos distribuciones consiste en obser-

var las diferencias entre estas fracciones zona por zona. Hagamos esta comparación entre las ramas *B1* y *B2*:

Comparación de la repartición geográfica de las ramas B1 y B2

Rama	*B1*	*B2*	Diferencia
Zona			
Z1	0.400	0.542	0.142
Z2	0.225	0.308	0.083
Z3	0.375	0.150	–0.225
Total	1.000	1.000	0.000

Cada una de las diferencias calculadas constituye una de las dimensiones de la disimilitud entre las dos reparticiones geográficas. Para medir la disimilitud, es necesario combinar las diferencias en una cifra única. Una simple suma será siempre igual a 0 por razones obvias.[66] Es el motivo por el cual haremos una suma de los valores absolutos:

$$|0.142| + |0.083| + |-0.225| = 0.142 + 0.083 + 0.225$$

(y no – 0.225)

Por razones que más adelante nos parecerán evidentes, dividimos el resultado entre dos y obtenemos así:

$$\frac{|0.142| + |0.083| + |-0.225|}{2} = 0.225$$

1-5.2.2 Definición del índice de disimilitud

La medición de la disimilitud y una tabla de contingencia: recuerdo de la simbología

Con el fin de formalizar la presentación, usaremos nueva-

[66] Puesto que $\sum_i v_i = \sum_i w_i = 1$, entonces $\sum_i (v_i - w_i) = \sum_i v_i - \sum_i w_i = 0$.

mente y esta vez generalizándola, la simbología que se manejó en el apartado 1-2.1.[67] Analizamos una tabla de contingencia de dos dimensiones. Convenimos que la líneas corresponden a *n* grupos diferentes mientras que las columnas corresponden a *m* categorías diferentes (en nuestro ejemplo como en el ejemplo del apartado 1-2.1, los *n* grupos son las tres ramas de actividad mientras que las *m* "categorías" son las tres zonas).

x_{ij}	Número de empleos de la rama *j* en la zona *i*
$x_{\bullet j} = \sum_i x_{ij}$	Número total de empleos de la rama *j*
$x_{i\bullet} = \sum_j x_{ij}$	Número total de empleos en la zona *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	Número total de empleos de todas las ramas en todas las zonas
$p_{ij} = x_{ij} / x_{\bullet\bullet}$	Fracción del empleo total global que pertenece a la rama *j* y situado en la zona *i*
$p_{\bullet j} = \sum_i p_{ij}$	Fracción del empleo total global que pertenece a la rama *j*
$p_{i\bullet} = \sum_j p_{ij}$	Fracción del empleo total global situado en la zona *i*
$p_{j/i\bullet} = p_{ij} / p_{i\bullet}$	Fracción del empleo total de la zona *i* que pertenece a la rama *j*
$p_{i/\bullet j} = p_{ij} / p_{\bullet j}$	Fracción del empleo total de la rama *j* situado en la zona *i*

[67] Se invita al lector a referirse al apartado 1-2.1 para tener presente el enunciado de la identidades que se verifican en una tabla de contingencia.

En el ejemplo numérico arriba mencionado, aplicamos una medición de disimilitud entre dos divisiones geográficas, la primera de la rama B1 y la segunda de la B2. Según la simbología comúnmente usada, esto vuelve a aplicar una medición de disimilitud a las distribuciones que se representan por los vectores

$$Q_1 = \begin{bmatrix} p_{1/\bullet 1} \\ p_{2/\bullet 1} \\ \vdots \\ p_{m/\bullet 1} \end{bmatrix} \text{ y } Q_2 = \begin{bmatrix} p_{1/\bullet 2} \\ p_{2/\bullet 2} \\ \vdots \\ p_{m/\bullet 2} \end{bmatrix}$$

En un aspecto más general, comparamos las distribuciones

$$Q_h = \begin{bmatrix} p_{1/\bullet h} \\ p_{2/\bullet h} \\ \vdots \\ p_{m/\bullet h} \end{bmatrix} \text{ y } Q_k = \begin{bmatrix} p_{1/\bullet k} \\ p_{2/\bullet k} \\ \vdots \\ p_{m/\bullet k} \end{bmatrix}$$

o bien, las distribuciones

$$R_g = \begin{bmatrix} p_{1/g\bullet} & p_{2/g\bullet} & \cdots & p_{n/g\bullet} \end{bmatrix} \text{ y } R_i = \begin{bmatrix} p_{1/i\bullet} & p_{2/i\bullet} & \cdots & p_{n/i\bullet} \end{bmatrix}$$

Nota: Podemos trabajar tanto con fracciones como en la simbología arriba empleada como con porcentajes que obtenemos al multiplicar las fracciones por 100. En este momento, convenimos trabajar con fracciones para simplificar la escritura de las fórmulas. Sin embargo, en la práctica, con tal de simplificar las tablas al eliminar el punto de las decimales, se acostumbra presentar porcentajes.

Definición

A continuación, aplicamos el índice de disimilitud a una comparación de las distribuciones Q_h y Q_k. Se puede trasponer fácilmente todos los cálculos a una comparación entre las distribuciones R_g y R_i o, de hecho, a cualquier par de distribuciones que se pueda comparar de manera formal (es decir que tengan el mismo número de posibilidades).

Se define el índice de disimilitud como:

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

En el ejemplo numérico mencionado arriba,

$$Q_1 = \begin{bmatrix} 0.400 \\ 0.225 \\ 0.375 \end{bmatrix} \text{ y } Q_2 = \begin{bmatrix} 0.542 \\ 0.308 \\ 0.150 \end{bmatrix}$$

y

$$D = \frac{\left|0.400 - 0.542\right| + \left|0.225 - 0.308\right| + \left|0.375 - 0.150\right|}{2}$$

$$D = 0.225$$

Este índice de disimilitud y sus variantes cercanas se conocen con nombres diferentes según la disciplina que se trate. Por ejemplo, se conoce también el índice de disimilitud con las expresiones “índice de diferenciación” e “índice de disociación”.

Cuando una de las distribuciones es la repartición espacial de una actividad económica y la otra es el grupo de actividades, esta medición corresponde a lo que se conoce en ciencias regionales como el coeficiente de localización. No obstante, debemos hacer mención de que el coeficiente de localización no es exactamente un índice de disimilitud aunque se calculó con la misma fórmula; examinaremos el porqué más adelante.

En ciencias regionales, se usa también el coeficiente de especialización que compara la estructura económica de una zona (repartición del empleo entre las ramas de actividad) con la estructura del territorio completo que se estudia. Tampoco este índice es exactamente un índice de disimilitud.

En geografía, Taylor (1997, p. 180) nombra de varias maneras el índice de disimilitud; entre otros nombres, el de "coeficiente de asociación geográfica" nos parece particularmente inadecuado puesto que D es una medición de disimilitud o de disociación. Es hasta posible encontrar el término "índice de Gini" para designar el índice de disimilitud.

Los demógrafos y los sociólogos usan este mismo índice con el nombre "coeficiente de segregación residencial" o "índice de discriminación" para comparar las distribuciones espaciales residenciales de diferentes grupos étnicos o raciales (Mills y Hamilton, 1989, pp. 233-239; Waldorf, 1993).

¿Qué debemos entender de esta confusión de términos? Lo siguiente: al momento de enterarse de resultados de investigación que requieren el uso de índices de este tipo, verifique muy bien la fórmula matemática empleada.

Más allá de estas particularidades propias de cada disciplina, examinemos este índice de disimilitud como una medición de disimilitud de dos distribuciones.

1-5.2.3 El índice de disimilitud como medición de concentración o desigualdad

Hasta el momento hemos estudiado los diferentes usos del índice de disimilitud para medir la disimilitud entre dos distribuciones observadas. Pero podemos también usar el índice de disimilitud para medir la desigualdad o la concentración. De por sí y es importante recordarlo, las mediciones de desigualdad o de concentración son, en general, mediciones de disimilitud entre una distribución observada y una distribu-

ción de referencia. Para medir la desigualdad o la concentración, es necesario, por lo tanto, comparar una distribución observada con una distribución de referencia que representa la igualdad perfecta o una concentración nula (está claro que, en este caso, la tabla de datos no es una tabla de contingencia).

Ejemplo

Supongamos que queremos medir el grado de concentración geográfica de la población en un territorio dado que hubiéramos, con anterioridad, dividido en zonas (estados, provincias, distritos...). Una concentración nula corresponde a una situación donde la densidad de la población (habitantes / km^2) es, en todas partes, igual. De modo que podemos decir que la concentración es nula si la fracción de la población en cada zona es igual a la fracción del territorio contenida en esta zona.

Teniendo V, la distribución de la superficie del territorio y W, la distribución de la población,

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \text{ y } W = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$$

v_i es la fracción de la superficie total que contiene la zona i y w_i es la fracción de la población que se encuentra en la zona i.

La concentración es nula si $w_i = v_i$ para todo i.

En este caso, la distribución observada del territorio sirve de distribución de referencia para la población: es la distribución teórica o hipotética de una población con una concentración nula.[68] Podemos, entonces, usar el índice de disimilitud entre la distribución del territorio y la distribución de la po-

[68] En otras palabras, la distribución V es una distribución observada cuando se trata del territorio pero se convierte en una distribución hipotética cuando la aplicamos a la población.

blación como medición de la concentración geográfica de la población. Tenemos:

$$D = \frac{1}{2}\sum_i |w_i - v_i|$$

La tabla que presentamos a continuación ilustra esta situación del índice de disimilitud. En ella se mide el grado de concentración de la población de la ciudad de Montreal. Se extrajeron los datos de población del censo de 1991. El territorio es dividido según los 54 barrios de planificación de la ciudad ordenados de manera decreciente de densidad. Se obtiene D = 0.2361, es decir que, para obtener una densidad uniforme, se tendría que desplazar 23.61% de la población de un barrio a otro.

Medición de la concentración de la población
por medio del índice de disimilitud
Ciudad de Montreal (54 barros de planificación),
población del censo de 1991

	Datos			Reparticiones		
Barrio	Pob. 1991	Superf. (km^2)	Densidad (hab/km^2)	Pob.	Superf.	Diferencia absoluta
11	29469	1.65	17860	2.90%	0.88%	0.0201
8	10604	0.72	14728	1.04%	0.38%	0.0066
18	27022	2.03	13311	2.66%	1.08%	0.0157
34	24258	1.85	13112	2.38%	0.99%	0.0140
13	30314	2.39	12684	2.98%	1.28%	0.0170
35	14187	1.24	11441	1.39%	0.66%	0.0073
31	19652	1.73	11360	1.93%	0.92%	0.0101
33	15752	1.40	11251	1.55%	0.75%	0.0080
42	25495	2.32	10989	2.51%	1.24%	0.0127
15	19126	1.75	10929	1.88%	0.93%	0.0095
16	15030	1.38	10891	1.48%	0.74%	0.0074
29	15606	1.46	10689	1.53%	0.78%	0.0075
9	21348	2.02	10568	2.10%	1.08%	0.0102
32	14737	1.48	9957	1.45%	0.79%	0.0066
40	20350	2.15	9465	2.00%	1.15%	0.0085
14	15973	1.80	8874	1.57%	0.96%	0.0061
10	14165	1.65	8585	1.39%	0.88%	0.0051
27	11592	1.41	8221	1.14%	0.75%	0.0039
17	16167	2.00	8084	1.59%	1.07%	0.0052
30	29664	3.69	8039	2.91%	1.97%	0.0095

Medición de la concentración de la población
por medio del índice de disimilitud (continuación)

Barrio	Datos			Reparticiones		
	Pob. 1991	Superf. (km^2)	Densidad (hab/km^2)	Pob.	Superf.	Diferencia absoluta
45	24738	3.23	7659	2.43%	1.72%	0.0071
46	19880	2.60	7646	1.95%	1.39%	0.0057
39	34906	4.85	7197	3.43%	2.59%	0.0084
51	8452	1.20	7043	0.83%	0.64%	0.0019
23	18672	2.67	6993	1.83%	1.43%	0.0041
12	14980	2.21	6778	1.47%	1.18%	0.0029
6	16785	2.48	6768	1.65%	1.32%	0.0033
19	11499	1.75	6571	1.13%	0.93%	0.0020
4	23636	3.70	6388	2.32%	1.98%	0.0035
44	18699	2.96	6317	1.84%	1.58%	0.0026
24	13665	2.22	6155	1.34%	1.19%	0.0016
21	20564	3.62	5681	2.02%	1.93%	0.0009
48	17038	3.02	5642	1.67%	1.61%	0.0006
41	20092	3.59	5597	1.97%	1.92%	0.0006
5	18478	3.36	5499	1.82%	1.79%	0.0002
49	14687	2.73	5380	1.44%	1.46%	0.0001
20	27819	5.22	5329	2.73%	2.79%	0.0005
43	24957	4.84	5156	2.45%	2.58%	0.0013
3	18052	3.56	5071	1.77%	1.90%	0.0013
28	17764	3.56	4990	1.75%	1.90%	0.0015
2	25181	5.25	4796	2.47%	2.80%	0.0033
26	19073	4.01	4756	1.87%	2.14%	0.0027
22	9651	2.18	4427	0.95%	1.16%	0.0022
38	12512	3.16	3959	1.23%	1.69%	0.0046
7	22660	5.84	3880	2.23%	3.12%	0.0089
1	22613	5.85	3865	2.22%	3.12%	0.0090
52	35098	9.50	3695	3.45%	5.07%	0.0162
50	14403	4.07	3539	1.42%	2.17%	0.0076
47	13111	4.45	2946	1.29%	2.38%	0.0109
54	47534	19.04	2497	4.67%	10.16%	0.0549
37	3546	2.06	1721	0.35%	1.10%	0.0075
25	4009	4.28	937	0.39%	2.28%	0.0189
53	11970	13.92	860	1.18%	7.43%	0.0625
36	431	4.24	102	0.04%	2.26%	0.0222
Total	*1017666*	*187.34*	*5432*	*100.00%*	*100.00%*	*0.4720*

Índice de disimilitud: 0.2361

1-5.2.4 Propiedades del índice de disimilitud

El índice de disimilitud y las propiedades de una medición de desigualdad

Puesto que una medición de desigualdad es una medición de disimilitud entre una distribución observada y una distribución de referencia, las propiedades deseables para una medición de desigualdad son de igual manera deseables para una medición de disimilitud. ¿En qué resulta esto con el índice de disimilitud *D*?

Recordemos las propiedades deseables de una medición de desigualdad según Valeyre (1993):

1. Una medición de desigualdad no puede tomar valores negativos ya que es una medición del alejamiento de la distribución observada en relación con la distribución de referencia.
2. Una medición de desigualdad debe tomar el valor 0 si y solamente si la distribución observada es idéntica a la distribución de referencia.
3. Se deben tratar a todas las observaciones de la misma manera.
4. Una medición de desigualdad debe ser independiente del valor promedio de la variable examinada; una medición de concentración debe ser independiente del tamaño de la población cuya distribución se estudia.
5. La agregación de observaciones que tienen el mismo grado de especificidad, no debe cambiar el valor de la medicion.[69]
6. Principio de transferencia de Pigou-Dalton: una medición de desigualdad debe disminuir si se modifica la

[69] Se puede demostrar fácilmente esta característica usando la interpretación geométrica del índice de disimilitud como la distancia vertical máxima entre la curva de Lorenz y la diagonal. Vea más abajo.

distribución de tal manera que reduce sin duda alguna la desigualdad.

El índice de disimilitud posee las cinco primeras propiedades pero no posee la sexta: su valor no cambia después de una transferencia entre dos categorías cuyas especificidades son ambas superiores o ambas inferiores a 1.

Campo de variación

Si alguien le anunciase que obtuvo una calificación de 18 en un examen, ¿estaría usted contento? ¿Es esta calificación una buena o una mala calificación? Para saberlo es necesario, ante todo, conocer la calificación máxima.[70] Si el examen se califica sobre 20, un 18 es, de seguro, una buena calificación; por el contrario, si el examen se califica sobre 100, es muy probable que usted no esté del todo contento.

Ésta es la razón por la cual el campo de variación nos interesa. El campo de variación de una medición es el conjunto de los valores que esta medición puede tomar. En el caso de una medición continua, se define el campo de variación por el valor mínimo y el valor máximo de la medición. Para saber si una calificación dada es "alta" o no, es necesario, por lo menos, conocer su campo de variación y verificar si este valor se acerca más el máximo o el mínimo.

En el caso del índice de disimilitud, su valor mínimo es 0; este índice toma el valor 0 cuando $p_{i/\bullet h} = p_{i/\bullet k}$ para todo i, o sea cuando las distribuciones son idénticas.

¿Cuál es su valor máximo?

[70] No es la única consideración. La interpretación de la calificación depende también de la calificación que los demás obtuvieron y de los criterios que se usan para su interpretación (como la calificación de aprobación).

Cuando comparamos las distribuciones de dos poblaciones perfectamente distintas,[71] el valor máximo que puede tomar el índice es 1: esto se produce cuando $p_{i/\bullet h} = 0$ y $p_{i/\bullet k} > 0$ y viceversa, es decir cuando la separación entre las dos poblaciones es completa, lo que significa que nunca aparecen juntas en la misma categoría. En efecto, en esta situación, para la categoría *i*, tenemos:

Teniendo $p_{i/\bullet h} = 0$ entonces:

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \left| 0 - p_{i/\bullet k} \right| = p_{i/\bullet k}$$

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| 0 + p_{i/\bullet k} = p_{i/\bullet h} + p_{i/\bullet k}$$

Teniendo $p_{i/\bullet k} = 0$, entonces:

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \left| p_{i/\bullet h} - 0 \right| = p_{i/\bullet h}$$

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = p_{i/\bullet h} + 0 = p_{i/\bullet h} + p_{i/\bullet k}$$

Por lo tanto tenemos:

$$D^{\max} = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2} \sum_i \left(p_{i/\bullet h} + p_{i/\bullet k} \right)$$

$$D^{\max} = \frac{1}{2} \left(\sum_i p_{i/\bullet h} + \sum_i p_{i/\bullet k} \right) = \frac{1+1}{2} = 1$$

La división entre 2 en la fórmula del cálculo del índice de disimilitud permite, por consiguiente, normalizar su campo de variación en el intervalo [0, 1].

¿Es posible que el índice de disimilitud tome un valor superior a 1? No. Para convencerse, basta preguntarse, refiriéndose a la situación de separación completa que se describió anteriormente, cuál sería la consecuencia de desplazar un individuo de una categoría a otra (el efecto es nulo si el indivi-

[71] Esto significa que ningún individuo puede pertenecer a dos poblaciones al mismo tiempo.

duo se queda con los mismos de su especie y si no, el valor del indicador disminuye)

Índice de disimilitud: ejemplo de segregación total

Etnia	Números			Reparticiones			
	Marcianos	Terrícolas	Total	Martianos	Terrícolas	Total	Diferencia
	x_{i1}	x_{i2}	$x_{i1}+x_{i2}$	$p_{i/\bullet 1}$	$p_{i/\bullet 2}$	$p_{i\bullet}$	$\lvert p_{i/\bullet h} - p_{i/\bullet k} \rvert$
Planeta							
Tierra	0	6	6	0.00	0.75	0.40	0.75
Luna	0	2	2	0.00	0.25	0.13	0.25
Marte	3	0	3	0.43	0.00	0.20	0.43
Júpiter	4	0	4	0.57	0.00	0.27	0.57
Total	7	8	15	1.00	1.00	1.00	

Índice de disimilitud:

$$\frac{0.75 + 0.25 + 0.43 + 0.57}{2} = 1.00$$

Interpretación metafórica

Aunque conozcamos perfectamente el campo de variación de una medición, es a veces difícil intuir con claridad lo que es un valor "alto", motivo por el cual puede ser útil una interpretación metafórica. Como su nombre lo indica, una interpretación metafórica se basa en una comparación, una metáfora del tipo "es como si", Cuidado en no tomar esta metáfora al pie de la letra.

En el caso del índice de disimilitud éste compara la distribución de dos grupos perfectamente distintos,[72] digamos *h* y *k*. Se puede interpretar el índice como una fracción del grupo

[72] Esto significa que ningún individuo puede pertenecer a dos poblaciones al mismo tiempo.

h que tendríamos que desplazar de una categoría a otra para que su distribución quedara idéntica a la distribución del grupo *k*.

En el ejemplo numérico que mencionamos al principio de este apartado, el índice de disimilitud entre la repartición espacial de los empleos de la rama *B1* y los empleos de la rama *B2* es de 0.225. Esto significa que, con tal de que las reparticiones espaciales *B1* y *B2* sean idénticas, se tendría que desplazar 22.5% de los empleos de *B1*.

Se puede demostrar fácilmente este resultado. Para empezar, determinamos cuál es la fracción del grupo *h* que tendríamos que desplazar para pasar de la distribución representada por los $p_{i/\bullet h}$ a la distribución representada por los $p_{i/\bullet k}$. Para lograr esto, sólo basta sumar las fracciones de población que debe quitarse de las categorías (zonas, regiones,...) "excedentes" para redistribuirlas en las categorías "deficitarias". Designemos por A, el conjunto de las categorías "excedentes", o sea cuando $p_{i/\bullet h} > p_{i/\bullet k}$. Para cada una de las categorías que pertenece al conjunto *A*, la fracción "excedente" de la población *H* es igual a $p_{i/\bullet h} - p_{i/\bullet k}$. La suma total de la fracción de la población *h* que debe quitarse de las categorías "excedentes" es, por lo tanto:

$$\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right)$$

Es equivalente querer sumar las fracciones de población que debe añadirse en la categorías "deficitarias", es decir:

$$\sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right)$$

Está claro que

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)=\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

Puesto que

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)-\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

$$=\sum_{i\in A}p_{i/\bullet h}-\sum_{i\in A}p_{i/\bullet k}-\sum_{i\notin A}p_{i/\bullet k}+\sum_{i\notin A}p_{i/\bullet h}$$

$$=\sum_{i}p_{i/\bullet h}-\sum_{i}p_{i/\bullet k}=0$$

¿Cuál es la relación con el índice de disimilitud? Bueno, si sumamos las dos sumatorias del miembro a la izquierda de la ecuación anterior (en lugar de restar la segunda de la primera), obtenemos

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)+\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)=\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|$$

Además, puesto que los dos términos de miembro de derecha son iguales, tenemos:

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)=\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

$$=\frac{1}{2}\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|=D$$

Ahí está otra buena razón por dividir la suma entre 2.

Simetría

Es importante notar que el índice de disimilitud es simétrico con relación a los grupos *h* y *k*:

$$D=\frac{1}{2}\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|=\frac{1}{2}\sum_{i}\left|p_{i/\bullet k}-p_{i/\bullet h}\right|$$

Por consiguiente, se puede, de igual manera, interpretar el indicador como la fracción del grupo *k* que debería desplazarse para que su distribución fuese idéntica a la distribución del grupo *h*. Importando el grupo que querramos desplazar para que su distribución sea idéntica a la distribución de otro grupo, la fracción que debe desplazarse es la misma (de esta manera, podríamos decir que se necesita desplazar 22.5% del empleo del ramo B2 para que su distribución sea idéntica a la distribución de la rama B1). No obstante, en cuanto al número de individuos que se necesita desplazar, éste es obviamente igual a esta fracción multiplicada por el número de la población. En caso de que dos grupos tengan tamaños diferentes, el número de individuos que se deba desplazar (de manera hipotética) difiere dependiendo de la fracción de uno u otro grupo que se quiere desplazar.

De nueva cuenta, no debemos olvidar el aspecto metafórico de esta interpretación. Para empezar, la similaridad de las distribuciones no es siempre buena (recordemos la controversia que suscitó el *busing* que se realizó en Estados Unidos para lograr la integración escolar de los blancos y los negros). Luego, el desplazamiento (a fuerza) de poblaciones no resulta ser una acción aceptable cuando se trata de poblaciones humanas.

Otras propiedades

El índice de disimilitud, tal como todos los índices, tiene sus límites. Además de no respectar el principio de Pigou-Dalton, mencionamos:

- Cuando los datos son agrupados, el índice de disimilitud, tal como el de Gini, es sensible a la definición y al número de las categorías utilizadas (clases, zonas). Esa debilidad no es tan grave si se escoge una clasificación bastante fina –o sea si comprende un gran número de

categorías– pero las comparaciones entre clasificaciones diversas no tienen ninguna significación.[73]

- Cuando se utiliza como medición de concentración espacial, el índice de disimilitud, tal como el de Gini, no toma en cuenta la contigüidad o la proximidad de las unidades espaciales.
- El índice de disimilitud no admite datos negativos. Por ejemplo, no se puede utilizar el índice de disimilitud para medir la similitud entre dos ramas de actividad respecto a las *variaciones* del número de empleos por zona, porque esas variaciones pueden ser negativas.

El índice de disimilitud y la curva de Lorenz

Acabamos de ver que, como el índice de Gini, el índice de disimilitud puede servir para medir la concentración aunque no posea todas las propiedades deseables del índice de Gini (le falta el principio de transferencia de Pigou-Dalton). Vimos, también, que se puede calcular el índice de Gini de manera geométrica basándonos en la curva de Lorenz. ¿Existe, por lo tanto, una relación entre el índice de disimilitud y la curva de Lorenz? ¡Pues sí!

En efecto, el índice de disimilitud es justamente igual a la diferencia vertical entre la curva de Lorenz y la diagonal

$$D = \underset{k}{\text{MAX}}\left[Cv_k - Cw_k\right]$$

Demostración:

Puesto que

$$\sum_i \left(v_i - w_i\right) = \sum_i v_i - \sum_i w_i = 1 - 1 = 0\,,$$

[73] Eso tema es conocido en geografía como "MAUP", es decir "Modifiable Areal Unit Problem".

esta suma contiene términos positivos y términos negativos (al menos que todos los términos sean igual a 0). Sin embargo, al ordenar las observaciones en un orden creciente de las razones w_i / v_i, los términos $(v_i - w_i)$ que son positivos preceden los términos negativos. En estas condiciones, está claro que la diferencia vertical

$$Cv_k - Cw_k = \sum_{i-1}^{k} v_i - \sum_{i-1}^{k} w_i = \sum_{i-1}^{k} (v_i - w_i)$$

alcanza su valor máximo cuando se escoge k de tal manera que, únicamente los términos positivos se incluyen en la sumatoria, excluyendo así todos los términos negativos. Por lo tanto

$$\underset{k}{\text{MAX}}\left[Cv_k - Cw_k\right] = \sum_{\substack{i \text{ tal que} \\ v_i > w_i}} (v_i - w_i)$$

y, puesto que $\sum_i (v_i - w_i) = 0$

$$\sum_{\substack{i \text{ tal que} \\ v_i > w_i}} (v_i - w_i) = \sum_{\substack{i \text{ tal que} \\ v_i < w_i}} |v_i - w_i| = \frac{1}{2} \sum_i |v_i - w_i| = D$$

Se tiene, así, una interpretación geométrica para el índice de disimilitud D: es la distancia máxima entre la diagonal y la curva de Lorenz asociada a la distribución V (vea el ejemplo numérico extraído de Taylor, 1997, y analizado en 1-4.3).

Con la ayuda de esta interpretación, es fácil constatar que el índice de disimilitud es insensible a toda distribución que no reduce la diferencia vertical máxima pero que, sin embargo, acerca la curva de Lorenz a la diagonal. Es justamente esta insensibilidad la que viola el principio de transferencia de Pigou-Dalton.

Sumario de las propiedades del índice de disimilitud

1. Posee las 5 primeras propiedades deseables de una medición de desigualdad pero le falta la última (el principio de transferencia de Pigou-Dalton; Valeyre, 1993)
2. Campo de variación (valores máximo y mínimo)
   – $D = 0$ cuando $p_{i/\bullet h} = p_{i/\bullet k}$ para todo i (las dos distribuciones son idénticas)
   – $D = 1$ cuando hay segregación completa:
     teniendo $p_{i/\bullet k} > 0$, entonces $p_{i/\bullet h} = 0$
     teniendo $p_{i/\bullet h} > 0$, entonces $p_{i/\bullet k} = 0$
3. D es simétrico con relación a los grupos h y k:
   $$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet k} - p_{i/\bullet h} \right|$$
4. Interpretación metafórica (grupos perfectamente distintos)
   D = fracción del grupo h que convendría desplazar para que su distribución fuera idéntica a la distribución del grupo k o viceversa.
5. Cuando los datos son agrupados, tanto D como G son sensibles a la definición y al número de categorías (clases, zonas). En particular, esto implica que la agregación de una o varias categorías significa una disminución del valor del índice de disimilitud.
6. Como medición de concentración espacial y al igual que el índice de Gini, el índice de disimilitud no toma en cuenta la proximidad en el espacio de diferentes zonas con fuerte densidad.
7. No se aplica con datos negativos (ejemplo: comparación de variación del empleo).
8. D es igual a la máxima diferencia vertical entre la curva de Lorenz y la diagonal.

1-5.2.5 Aplicación de índice de disimilitud a una dicotomía

Equivalencia de la fórmula de Duncan-Duncan (1995)

Cuando distinguimos solamente dos grupos, decimos que tratamos con una dicotomía; en estas condiciones, se compara un grupo *h* con el resto de la población (que toma el papel del grupo *k*). Para el grupo *k* tenemos entonces:

$$p_{i/\bullet k} = \frac{x_{i\bullet} - x_{ih}}{x_{\bullet\bullet} - x_{\bullet h}} = \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}}$$

Así que se puede escribir el índice de disimilitud con la fórmula

$$D = \frac{\sum_i p_{i\bullet} \left| p_{h/i\bullet} - p_{\bullet h} \right|}{2\, p_{\bullet h} \left(1 - p_{\bullet h}\right)}$$

Esta segunda definición que aparece en el artículo clásico de Duncan-Duncan (1995) es equivalente a la definición que dimos anteriormente cuando se aplica a una dicotomía.

Se demuestra esta equivalencia entre dos definiciones en el caso de una dicotomía como sigue:

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2} \sum_i \left| \frac{p_{ih}}{p_{\bullet h}} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{ih}/p_{i\bullet}}{p_{\bullet h}} - \frac{1-\left(p_{ih}/p_{i\bullet}\right)}{1-p_{\bullet h}}\right|$$

$$D = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{h/i\bullet}}{p_{\bullet h}} - \frac{1-p_{h/i\bullet}}{1-p_{\bullet h}}\right|$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet}(1-p_{\bullet h}) - (1-p_{h/i\bullet})p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{h/i\bullet}\,p_{\bullet h} - p_{\bullet h} + p_{h/i\bullet}\,p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

Esta fórmula se presta a una interpretación interesante. En el numerador, aparece un promedio ponderado de las diferencias absolutas $\left|p_{h/i\bullet} - p_{\bullet h}\right|$ entre, por una parte, la fracción $p_{h/i\bullet}$ del grupo *h* en cada categoría *i* y, por otra parte, la fracción $p_{\bullet h}$ del grupo *h* en el total de la población, de modo que el peso $p_{i\bullet}$ de cada categoría es proporcional a su población, esto para cualquier grupo.

En cuanto a la expresión del denominador, ésta es igual a la diferencia absoluta promedio entre los individuos (y no entre las categorías) de la variable dicotómica de pertenencia al grupo *h*. Esta diferencia absoluta promedio es igual a dos veces la varianza de la misma variable.

En efecto, tenemos la variable dicotómica de pertenencia g_t

$$g_t \begin{cases} = 1 \text{ si el individuo } t \text{ pertenece al grupo } h \\ = 0 \text{ en otros casos} \end{cases}$$

donde el índice t se refiere a los individuos de los dos grupos: t varía entre 1 y $x_{\bullet\bullet}$.

La variable g_t tiene una distribución binómica cuyo promedio es

$$\mu_g = \frac{\sum_t g_t}{x_{\bullet\bullet}} = \frac{\sum_i x_{ih}}{x_{\bullet\bullet}} = \frac{x_{\bullet h}}{x_{\bullet\bullet}} = p_{\bullet h}$$

La diferencia absoluta promedio (desviación media) es entonces

$$d_g = \frac{\sum_t \left| g_t - \mu_g \right|}{x_{\bullet\bullet}} = \frac{\sum_t \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = \frac{\sum_{t \text{ tal que } g_t = 1} \left| g_t - p_{\bullet h} \right| + \sum_{t \text{ tal que } g_t = 0} \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = \frac{p_{\bullet h} x_{\bullet\bullet} \left| 1 - p_{\bullet h} \right| + (1 - p_{\bullet h}) x_{\bullet\bullet} \left| 0 - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = p_{\bullet h} \left| 1 - p_{\bullet h} \right| + (1 - p_{\bullet h}) \left| 0 - p_{\bullet h} \right|$$

$$d_g = 2 p_{\bullet h} (1 - p_{\bullet h})$$

La varianza, por su lado, se escribe con la fórmula

$$\sigma_g^2 = \frac{\sum_t (g_t - p_{\bullet h})^2}{x_{\bullet\bullet}} = \frac{\sum_t \left(g_t^2 - 2 s_t p_{\bullet h} + p_{\bullet h}^2 \right)}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{\sum_t g_t^2 - 2p_{\bullet h}\sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{\sum_t g_t - 2p_{\bullet h}\sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{p_{\bullet h}x_{\bullet\bullet} - 2p_{\bullet h}\left(p_{\bullet h}x_{\bullet\bullet}\right) + p_{\bullet h}^2 x_{\bullet\bullet}}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = p_{\bullet h} - p_{\bullet h}^2 = p_{\bullet h}\left(1 - p_{\bullet h}\right)$$

El coeficiente de localización y el índice de disimilitud: ¡cuidado, que no son lo mismo!

En ciencia regional es común el uso del coeficiente de localización[74] para medir el grado de especificidad de la repartición espacial de una actividad económica con relación a la economía total.

En una tabla de contingencia del empleo por zona y por rama, $p_{i/\bullet h}$ designa la fracción del empleo total de la rama *h* que se sitúa en la zona *i*; y $p_{i\bullet}$ designa la fracción del empleo total del total de las ramas que se sitúa en la zona *i*. Se define el coeficiente de localización como

$$CL = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right|$$

[74] Según Isard (1960, p. 251) fue P. Sargant Florence quien introdujo el coeficiente de localización como nueva herramienta de la ciencia regional; Duncan y Duncan (1995) citan a P. Sargant Florence, W.G. Fritz y R.C. Gilles, "measure of industrial distribution", cap. 5 en *National Resources Planning Board Industrial Location and National Resources*, Washington, Government Prinding Office, 1943.

A primera vista, es un índice de disimilitud. No obstante, no lo es. Más bien, la relación entre el coeficiente de localización CL y el índice de disimilitud D es la siguiente:

$$CL = (1 - p_{\bullet h})D$$

Demostración:
Sabiendo que *D* se aplica a una dicotomía, tenemos:

$$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h}(1 - p_{\bullet h}) - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet h}\, p_{\bullet h} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{ih} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right| = \frac{CL}{(1 - p_{\bullet h})}$$

Esta diferencia se debe a que el coeficiente de localización compara la distribución de un grupo (una rama de actividad) con la distribución del grupo completo al cual pertenece, cuando el índice de disimilitud compara la distribución de un grupo con la distribución del resto de la población (las demás actividades). Esto implica que no podemos dar al coeficiente de localización la misma interpretación metafórica que al índice de disimilitud, a saber la fracción del grupo a desplazar para obtener distribuciones idénticas. Además, el campo de variación de *CL*, de 0 a $(1 - p_{\bullet h})$, es más estrecho para las ramas más importantes lo que dificulta la comparación entre coeficientes de ramas de tamaños diferentes. Por lo contrario, si queremos medir cómo la reparti-

ción espacial de cada actividad económica depende de esta misma actividad, entonces el índice de disimilitud tiene el inconveniente de usar una distribución de referencia diferente para cada rama. Esta distribución de referencia corresponde a la distribución del conjunto de las demás ramas y, por consiguiente, es diferente para cada rama.

Se pueden ilustrar estas diferencias con el ejemplo utilizado al inicio del capítulo.

Empleo por zona y por rama y distribución del empleo entre las zonas

	Empleo					Distribución entre las zonas				
Rama	*B*1	*B*2	*B*3	*B*1+2	Total	*B*1	*B*2	*B*3	*B*1+2	Total
Zona										
Z1	48	325	287	373	660	0.400	0.542	0.598	0.518	0.550
Z2	27	185	148	212	360	0.225	0.308	0.308	0.294	0.300
Z3	45	90	45	135	180	0.375	0.150	0.094	0.188	0.150
Total	120	600	480	720	1200	1.000	1.000	1.000	1.000	1.000

Comparación de la distribución geográfica de la rama *B*3 con la del conjunto de las tres ramas, pues con la suma de *B*1 y *B*2

Rama	*B*3	Total	Dif.absol.	*B*1+2	Dif.absol.
Zona					
Z1	0.598	0.550	0.048	0.518	0.080
Z2	0.308	0.300	0.008	0.294	0.014
Z3	0.094	0.150	0.056	0.188	0.094
Total	1.000	1.000	0.113	1.000	0.188

Apliquemos la fórmula de cálculo del índice de disimilitud a cada una de ambas comparaciones. En el primer caso (*B*3 y total), se obtiene el coeficiente de localización:

$$CL = \frac{|0.048| + |0.008| + |-0.056|}{2} = 0.056$$

En el segundo caso ($B3$ y $B1+2$), se obtiene el índice de disimilitud:

$$D = \frac{|0.080| + |0.014| + |-0.094|}{2} = 0.094$$

Tal como se esperaba, los resultados son realmente diferentes. Sin embargo, son vinculados por la relación

$$CL = \left(1 - \frac{480}{1200}\right) D = 0.6 \times 0.094 = 0.056$$

donde el factor 0.6 es igual a la parte del empleo de las ramas *demás de B*3.

Cuando el grupo de la población que se considera no es más que una pequeña fracción de la población aparente, $p_{\bullet h}$ es pequeño y el valor del coeficiente de localización es cercano al valor del índice de disimilitud.

En el caso particular donde hay segregación total, el índice de disimilitud D es igual a 1 y el coeficiente de localización es igual a la parte del empleo de las ramas *de más de B*3. Se pueden ilustrar estas diferencias con el ejemplo de segregación total ya mencionado.

Coeficiente de localización: ejemplo de segregación total

Etnia	Números		Reparticiones		
	Marcianos	Total	Marcianos	Total	Diferencia
	x_i	y_i	v_i	w_i	$\lvert v_i - w_i \rvert$
Planeta					
Tierra	0	6	0.00	0.40	0.40
Luna	0	2	0.00	0.13	0.13
Marte	3	3	0.43	0.20	0.23
Júpiter	4	4	0.57	0.27	0.30
Total	7	15	1.00	1.00	

Coeficiente de localización:

$$\frac{0.40 + 0.13 + 0.23 + 0.30}{2} = 0.53 = 1 - \frac{7}{15}$$

= fracción de no-marcianos en la población = fracción de terrícolas.

De la misma manera, se calcula un coeficiente de localización para los terrícolas

$$0.47 = 1 - \frac{8}{15}$$

Post scriptum: el coeficiente de localización y los cocientes de localización

El parecido entre los dos nombres puede ser factor para confundir el coeficiente de localización y el cociente de localización. Sin embargo, cuando el coeficiente de localización compara dos distribuciones, el cociente de localización compara dos partes (vea más arriba), es decir dos puntos que corresponden en dos distribuciones. No obstante, existe una relación entre ambos que se puede ver al desarrollar la definición del coeficiente de localización:

$$CL = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right|$$

$$= \frac{1}{2}\sum_i p_{i\bullet} \left| \left(\frac{p_{i/\bullet h}}{p_{i\bullet}} \right) - 1 \right| = \frac{1}{2}\sum_i p_{i\bullet} \left| QL_{ih} - 1 \right|$$

El coeficiente de localización es un promedio ponderado de las diferencias absolutas entre los cocientes de localización y el valor 1 de referencia.

1-5.2.6 Un último vistazo crítico

Como cualquier otro índice, el índice de disimilitud tiene límites. Además de no respetar el principio de transferencia de Pigou-Dalton, debemos mencionar:

El índice de disimilitud no admite datos negativos. Por ejemplo, sería imposible usar el índice de disimilitud para medir la similitud entre dos ramas de actividad en cuanto a la variación del número de empleos por zonas porque estas variaciones pueden ser negativas.

Al igual que el índice de Gini, el índice de disimilitud es sensible a la definición y al número de las categorías (clases, zonas). Este defecto no es tan grave siempre y cuando la división que se seleccione sea lo suficiente fina, o sea que considere un buen número de categorías, pero también que las comparaciones entre divisiones no tengan ningún peso significativo.[75]

Cuando se aplica a datos espaciales, el índice de disimilitud, tanto como el índice de Gini, no toma en cuenta la contigüidad o la proximidad de las unidades espaciales.

Como en el caso de los índices de Laspeyres y de Paasche, cuando vimos que se podía construir índices de precios con fundamentos teóricos más satisfactorios pero, en cambio, con más alto grado de complejidad, se puede definir indicadores de disimilitud más refinados, de los cuales Waldorf (1993) nos da un ejemplo. Sin embargo, hemos de interrogarnos si, según el contexto, tales refinamientos son del todo pertinentes y concretos. Además, la presentación de Waldorf no deja por completo de caer en la trampa que consiste en pasar de la metáfora a la interpretación literal; de hecho, en el contexto de un estudio de la segregación racial en los Estados

[75] Se examina este problema en todos los escritos de geografía en la rúbrica MAUP, que significa “Modifiable Areal Unit Problem”.

Unidos, menciona el "esfuerzo requerido" para un desplazamiento de la población.

1-5.3 DISTANCIA Y DISIMILITUD

La medición de distancias geográficas tiene una gran importancia en los estudios urbanos y regionales. Se puede considerar la medición de la distancia como un caso particular de la medición de la disimilitud: la distancia es una medición de la disimilitud entre dos objetos con relación a su situación en el espacio, o sea entre dos lugares en el espacio.

Una superficie (como la superficie a condición de ignorar el relieve) es un espacio de dos dimensiones. La especificación de una situación en el espacio tiene, por lo tanto, dos dimensiones: longitud y latitud o coordenadas cartesianas (x,y). En consecuencia, la medición de la distancia geográfica tiene también dos dimensiones. De hecho, aunque en la vida cotidiana acostumbramos usar la distancia euclidiana de manera automática, existen otras maneras de medir la distancia.[76]

Una medición de distancia debe satisfacer algunas condiciones: la función $d(a,b)$ es una función de distancia si y solamente si, para todo conjunto de lugares a, b y c, esta función satisface las cuatro condiciones siguientes:

(c1) no negativo
$d(a,b) \geq 0$

(c2) identidad
$d(a,b) = 0$ si y solamente si, $a = b$

(c3) simetría
$d(a,b) = d(b,a)$

(c4) desigualdad triangular
$d(a,c) \leq d(a,b) + d(b,c)$

[76] Vea Huriot y Perreur (1990 y 1994).

La medición de distancia que más usamos es la distancia euclidiana. La distancia euclidiana entre el punto a, de coordenadas (x_a,y_a) y el punto b de coordenadas (x_b,y_b) es:

$$d_e(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

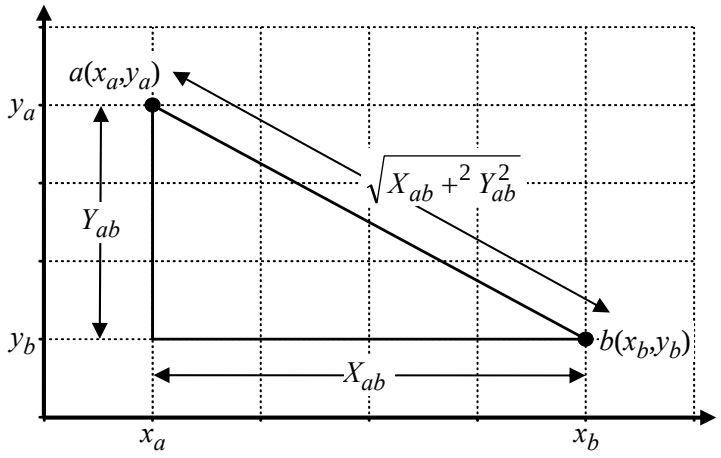
Entre las otras mediciones de distancia posibles, mencionemos la distancia rectilínea, conocida también como distancia según la métrica de Manhattan (vea Huriot y Perreur, 1994, p. 44):

$$d_r(a,b) = \left| x_a - x_b \right| + \left| y_a - y_b \right|$$

La distancia según la métrica de Manhattan es la distancia que se debe recorrer para ir del punto *a* al punto *b* siguiendo el trazo de las calles cuando estas forman una cuadra como en Manhattan.

Las dos métricas se enseñan en la figura que sigue teniendo

$X_{ab} = |x_a - x_b|$ et $Y_{ab} = |y_a - y_b|$



Puesto que se puede interpretar la distancia geográfica como una disimilitud, la recíproca es también verdadera: se

puede usar las mediciones de la distancia para medir las disimilitudes que no son distancias geográficas.

De esta manera, consideremos dos objetos que describimos con n variables que miden, cada una, una característica (dimensión) pertinente:

$x_{11}, x_{12}, ..., x_{1n}$ para el primer objeto y

$x_{21}, x_{22}, ..., x_{2n}$ para el segundo.

Ejemplo:

> Si los dos objetos fueran dos barrios de una ciudad, las características pertinentes podrían ser la densidad de la población, la proporción de la población menor de quince años, la proporción de la población que haya completado sus estudios de primaria, el ingreso promedio de los hogares, etcétera.

Para medir la disimilitud entre dos objetos multidimensionales, se usa a menudo la distancia euclidiana generalizada que tiene por fórmula

$$\sqrt{\sum_i (x_{1i} - x_{2i})^2}$$

Se usa también la distancia lineal generalizada o la distancia generalizada según la métrica de Manhattan, la cual se define con

$$\sum_i \left| x_{1i} - x_{2i} \right|$$

El lector perspicaz habrá observado la relación que existe entre el índice de disimilitud D y la distancia generalizada según la métrica de Manhattan. No obstante, en el presente contexto, los dos objetos que se comparan no son necesariamente distribuciones. Es de notar, en consecuencia, que no hay valor máximo inherente a la distancia recta generalizada (ni, tampoco, a la distancia euclidiana generalizada).

En general, el valor de una medición de distancia depende de las unidades de medición de las variables subyacentes. Por

este motivo, al momento de comparar dos objetos multidimensionales por medio de una distancia generalizada, no queda más que enfrentarnos a un problema parecido al problema encontrado cuando la construcción de un número índice. En efecto, la selección de la unidad de medición de cada variable determina de manera implícita cuál será el peso de la distancia-disimilitud en la medición. Sólo cuando los objetos que se comparan son distribuciones, el problema no se presenta.

1-5.4 La medición de la similitud en estadística

El problema de la medición de la similitud surge con frecuencia en estadística. Consideremos, por ejemplo, dos series de observaciones de dos variables:

$x_1, x_2, ..., x_n$ y $y_1, y_2, ..., y_n$

El coeficiente de correlación simple[77] es una medición de similitud entre dos series de datos.

Asimismo, para evaluar la exactitud de un modelo con relación a los datos que permitieron estimar sus parámetros, se mide la similitud entre los valores observados y los valores que la teoría predice. Una de las mediciones que más se emplea para este fin es el coeficiente de determinación múltiple R^2 (del cual hablaremos en la tercera parte de esta obra).

Finalmente, el Ji-cuadrado de Pearson[78] es una medición de la disimilitud entre los números observados y les números "teóricos" predichos por una hipótesis.

Todas estas mediciones pertenecen a la gran familia de las mediciones de similitud y disimilitud.

---

[77] Vea el anexo 2-A, "Recuerdo de algunas fórmulas usuales en estadísticas".

[78] Vea 4-1. Pero el Ji-cuadrado no es una medida simétrica : su valor varia si se intercambian lo papeles de los valores observados y teóricos.

En Webber (1984, pp. 41-45) se lleva a cabo una interesante discusión sobre el grado de pertinencia de diferentes mediciones de ajuste (en el contexto de la evaluación de la exactitud del modelo de repartición espacial de Lowry).

1-5.5 Otras mediciones de similitud y de disimilitud

Las mediciones de similitud y disimilitud existen en abundancia. Legendre y Legendre (1984, tomo 2, cap. 6, 1998) presentan y discuten numerosas mediciones que se usan en ecología numérica, las cuales se podrían emplear para el análisis espacial en ciencias sociales.

EN CONCLUSIÓN...

¿Qué podemos recordar de todas estas fórmulas, de todos estos números y de todas estas palabras que componen la primera parte de este curso? Puede ser que algunas ideas clave como...

- El método cuantitativo se basa en la medición y medir es comparar. Existen diferentes grados en la medición que dependen del tipo de comparaciones que se pueden efectuar (=, ≠, < o >). Es casi imposible encontrar una medición perfectamente válida y confiable. Por lo general, se puede asociar más de una medición a una sola dimensión de un concepto y hasta para resumir la evolución temporal de una serie, existe más de una posibilidad.
- Medir no tiene sentido si no se puede descubrir el significado de los números. Y, para interpretar magnitudes, es a menudo necesario recurrir a una "metacomparación" que permite poner un dato en perspectiva. El análisis de descomposición es, de igual manera, una técnica útil para examinar datos a condición de no confundir partes de la descomposición con causas, aún más cuando una de estas partes es un residuo...
- Los conceptos que encierran más de una dimensión causan un problema de medición que no tiene una so-

lución única. Con la construcción de índices, se busca resolver el problema de la multidimensionalidad lo mejor que se pueda. La validez de los índices depende, en gran parte, de la validez del modelo subyacente. En particular, los índices que son promedios ponderados se basan, a menudo, en modelos reductores (¡y a veces no tienen ninguna base!); y si esto fuera poco, las ponderaciones que se aplican son, a veces, arbitrarias, lo que despoja el índice de cualquier estatus de medición (puesto que, en estas condiciones, el orden que establece este índice entre las observaciones es igualmente arbitrario, y esto aunque el índice tenga supuestos fundamentos científicos).

- Tanto la medición de desigualdad o de la concentración como la medición de disimilitud están estrechamente emparentadas puesto que la mayoría de las mediciones de desigualdad son mediciones de disimilitud entre una distribución observada y una distribución de referencia. Existe toda una serie de mediciones de este tipo. No obstante, se prefiere usar algunas y no otras cuando poseen varias y hasta todas las propiedades deseables de tales mediciones. Por consiguiente, si se quiere usar con juicio estos diversos índices, será importante conocer, antes, sus propiedades.

ANEXO 1-A
HERRAMIENTAS MATEMÁTICAS DE BASE

1-A.1 EL OPERADOR SUMA*

1-A.1.1 Definición

El operador suma es sencillamente una manera compacta de escribir una suma cuando los términos sucesivos pueden escribirse bajo la forma de una expresión general que varía en función de un índice. Por ejemplo, la suma

$$x_1 + x_2 + x_3 + x_4 + x_5$$

puede escribirse

$$\sum_{i=1}^{5} x_i$$

En esta expresión, i es una variable que toma sucesivamente los valores 1, 2, 3, 4 y 5 : el "i=1" que se encuentra bajo el Σ indica que el valor inicial de la variable i es 1 ; el "5" que se encuentra encima del Σ indica que el valor terminal de la variable i es 5. La variable x_i es una función de la variable i, es decir que su valor depende del valor de i: cuando $i = 1$, $x_i = x_1$; cuando $i = 2$, $x_i = x_2$; y así sucesivamente. Finalmen-

* Lo que sigue fue sacado en gran parte del anexo I de Hohn (1964).

te, el signo Σ indica que hay que *sumar* x_1, x_2, x_3, x_4 y x_5, los valores sucesivos de x_i. Se lee esta expresión de la manera siguiente : “la suma de los x_i para i variando de 1 a 5”.

De manera más general, tendremos

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

Además, cuando no hay ambigüedad posible sobre los valores inicial y terminal del índice, puede escribirse de manera elíptica

$$\sum_{i} x_i = x_1 + x_2 + \cdots + x_n$$

Hay que observar que el índice i es un índice mudo (*dummy index*). La elección de la letra que sirve para representar el índice mudo es perfectamente arbitraria :

$$\sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j = \sum_{k=1}^{n} x_k = x_1 + x_2 + \cdots + x_n$$

Hay también un cierto grado de arbitrariedad en la elección de los valores inicial y terminal, como lo muestra el ejemplo siguiente :

$$\sum_{i=1}^{n} x_i = \sum_{i=0}^{n-1} x_{i+1} = \sum_{i=2}^{n+1} x_{i-1} = x_1 + x_2 + \cdots + x_n$$

En los desarrollos matemáticos, es a veces cómodo poder decalar así el índice mudo.

* * *

Para calcular el valor numérico de la expresión

$$\sum_{j} x_j = x_1 + x_2 + \cdots + x_n$$

hay que conocer los valores de x_1, x_2, ..., x_n. En ciertos casos, la notación permite conocer directamente el valor de cada uno de los términos de la suma. He aquí algunos ejemplos :

$$\sum_{t=1}^{n} t^2 = 1^2 + 2^2 + \cdots + n^2 \quad \text{(basta conocer } n\text{)}$$

$$\sum_{k=1}^{K} \left(\frac{1}{k}\right) = \left(\frac{1}{1}\right) + \left(\frac{1}{2}\right) + \left(\frac{1}{3}\right) + \cdots + \left(\frac{1}{K}\right)$$

(basta conocer K)

Se encuentran también expresiones como

$$\sum_{j=0}^{n} a_j x^j = a_0 x^0 + a_1 x^1 + a_2 x^2 + \cdots + a_n x^n$$

donde el índice mudo tiene a la vez un papel de índice propiamente dicho (en a_j) y un papel numérico (como exponente en x^j).

Se utiliza también el operador suma para tratar sumas infinitas como

$$\sum_{j=1}^{\infty} x_i = x_1 + x_2 + \cdots + x_n + \cdots$$

Se emplea entonces el símbolo ∞ para designar el valor terminal del índice.

1-A.1.2 Reglas de base (sumas finitas)

Las reglas de base para la utilización del operador suma son las siguientes :

1. $\sum_{i=1}^{n} c = nc$

2. $\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i = \sum_{i=1}^{n} x_i$

3. $\sum_{i=1}^{n} (c\, x_i) = c\left(\sum_{i=1}^{n} x_i \right)$

4. $\sum_{i=1}^{t} (x_i + y_i) = \sum_{i=1}^{t} x_i + \sum_{i=1}^{t} y_i$

Todas estas reglas, excepto la primera, pueden deducirse de la definición

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

La primera regla es más bien una convención, que se justifica de la manera siguiente. Supongamos que la variable x_j sea una constante :

$$x_1 = x_2 = ... = x_n = c$$

Entonces el valor de la suma se da por

$$\sum_{j=1}^{n} x_j = x_1 + x_2 + \cdots + x_n = c + c + \cdots + c = n\,c$$

La expresión "$\sum_{i=1}^{n} c$" se interpreta por lo tanto como "$\sum_{i=1}^{n} x_i$ donde $x_i = c$ para todo i". De ahí viene la primera regla. Así :

$$\sum_{i=1}^{5} 7 = 5 \times 7 = 35$$

1-A.1.3 Sumas dobles

Supongamos que haya que tratar un conjunto de $n \times m$ cantidades t_{ij}, con $i = 1, 2, ..., n$ y $j = 1, 2, ..., m$. Estas cantidades pueden disponerse en forma de cuadro :

$$\begin{matrix} t_{11} & t_{12} & \cdots & t_{1m} \\ t_{21} & t_{22} & \cdots & t_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1} & t_{n2} & \cdots & t_{nm} \end{matrix}$$

Para hacer la suma de todos los t_{ij}, puede primero hacerse el total de los términos de cada línea y luego sumar los totales de las líneas, lo que da

$$\sum_{j=1}^{m} t_{1j} + \sum_{j=1}^{m} t_{2j} + \cdots + \sum_{j=1}^{m} t_{nj}$$

La misma expresión puede escribirse de manera más compacta mediante un segundo operador suma :

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{m} t_{ij} \right)$$

Se hubiera obtenido el mismo resultado si se hubiera hecho primero el total de los términos de cada columna y luego se hubieran sumado los totales de las columnas :

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} t_{ij} \right)$$

Pues que ambos resultados son iguales, se tiene entonces

$$\sum_{i=1}^{n} \left(\sum_{j=1}^{m} t_{ij} \right) = \sum_{j=1}^{m} \left(\sum_{i=1}^{n} t_{ij} \right)$$

Por esta razón, se ometen generalmente las paréntesis. Se tiene entonces la regla

5. $$\sum_{i=1}^{n}\sum_{j=1}^{m} t_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} t_{ij}$$

(No se aplica siempre a las sumas infinitas)

La misma regla puede generalizarse en sumas triples, cuádruples, etc.

* * *

Hay que observar que en una doble suma, el índice de la suma exterior puede aparecer como valor inicial o terminal de la suma interior. Pero en tal caso, no se puede invertir las sumas como parece permitir la regla 5. Veamos por qué. Por ejemplo, supongamos que se desee hacer la suma de los valores del cuadro triangular siguiente :

$$\begin{array}{llllll} a_{11} & & & & \\ a_{21} & a_{22} & & & \\ a_{31} & a_{32} & a_{33} & & \\ \vdots & \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{array}$$

La suma de los totales de las líneas puede escribirse :

$$\sum_{i=1}^{n}\sum_{j=1}^{i} a_{ij}$$

y, de manera más elíptica,

$$\sum_{i}\sum_{j\leq i} a_{ij}$$

Asimismo, la suma de los totales de las columnas puede escribirse

$$\sum_{j=1}^{n}\sum_{i=j}^{n} a_{ij}$$

y, de manera más elíptica,

$$\sum_{j}\sum_{i\geq j} a_{ij}$$

No obstante, *no* se puede escribir $\sum_{i\geq j}\sum_{j} a_{ij}$: eso no tendría ningun sentido. Pues la expresión $\sum_{j}\sum_{i\geq j} a_{ij}$ significa $\sum_{j}\left(\sum_{i\geq j} a_{ij}\right)$: la segunda suma se hace adentro de la primera (se cálcula antes). Por tanto, el valor inicial de la primera suma (exterior) no puede depender del índice de la segunda suma (interior).

Supongamos que se desee excluir de la suma los términos de la diagonal del cuadro (a_{11}, a_{22}, ..., a_{nn}). Se escribe :

$$\sum_{i=1}^{n}\sum_{j=1}^{i} a_{ij} - \sum_{i=1}^{n} a_{ii} = \sum_{i}\sum_{j\leq i} a_{ij} - \sum_{i} a_{ii} = \sum_{i}\sum_{j<i} a_{ij}$$

(se nota la diferencia entre < y ≤)

o

$$\sum_{j=1}^{n}\sum_{i=j}^{n} a_{ij} - \sum_{i=1}^{n} a_{ii} = \sum_{j}\sum_{i\geq j} a_{ij} - \sum_{i} a_{ii} = \sum_{i}\sum_{j>i} a_{ij}$$

(se nota la diferencia entre > y ≥)

1-A.1.4 Nota : el operador producto

El operador producto es análogo al operador suma. Sirve para escribir productos de manera compacta :

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n$$

El uso del operador producto sigue las siguientes reglas :

1. $\prod_{j=1}^{n} c = c^n$
2. $\left(\prod_{j=1}^{k} x_j\right)\left(\prod_{j=k+1}^{n} x_j\right) = \left(\prod_{j=1}^{n} x_j\right)$
3. $\prod_{j=1}^{n} k\, x_j = k^n \left(\prod_{j=1}^{n} x_j\right)$
4. $\prod_{j=1}^{n} x_j y_j = \left(\prod_{j=1}^{n} x_j\right)\left(\prod_{j=1}^{n} y_j\right)$
5. $\prod_{i=1}^{m}\prod_{j=1}^{n} x_{ij} = \prod_{j=1}^{n}\prod_{i=1}^{m} x_{ij}$

1-A.1.5 Ejercicios sobre el operador suma

1. En los ejercicios 1.1 hasta 1.3, hay que calcular los valores de los x_i por medio de la ecuación:

   $x_i = 5 + 3\, i$

¿Cuál es el valor numérico de las expresiones siguientes?

1.1 $\sum_{k=1}^{4} x_k$

1.2 $\sum_{i=0}^{3} x_i$

1.3 $\sum_{i=0}^{n-1} x_{i+1}$, cuando $n = 4$

2. Calculen

2.1 $\sum_{x=2}^{3} x^3$

2.2 $\sum_{i=1}^{4} \left(\frac{1}{i}\right)$

2.3 $\sum_{j=1}^{10} a$,

cuando $a = 345$

3. Hagan la demostración de las siguientes reglas explicitando las expresiones a partir de la definición del operador suma:

3.1 $\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i = \sum_{i=1}^{n} x_i$

3.2 $\sum_{i=1}^{n} (c\, x_i) = c\left(\sum_{i=1}^{n} x_i\right)$

3.3 $\sum_{i=1}^{t} (x_i + y_i) = \sum_{i=1}^{t} x_i + \sum_{i=1}^{t} y_i$

4. En los ejercicios 4.1 hasta 4.8, se trata de datos en forma de un cuadro:

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} = \begin{bmatrix} 4 & 50 \\ 5 & 30 \\ 6 & 10 \end{bmatrix}$$

Cuál es el valor de las expresiones siguientes?

4.1 $\sum_i \sum_j a_{ij}$

4.2 $\sum_i a_{2i}$

4.3 $\sum_{j=1}^{2} \sum_{i=1}^{2} a_{ij}$

4.4 $\sum_{i=1}^{2} \sum_{j=1}^{2} a_{ij}$

4.5 $\sum_{j=1}^{2} \sum_{i=1}^{j} a_{ij}$

4.6 $\sum_{i=1}^{2} \sum_{j=1}^{i} a_{ij}$

4.7 $\sum_i \sum_{j>i} a_{ij}$

4.8 $\sum_i \sum_{j \leq i} a_{ij}$

Las soluciones se encuentran al fin del apéndice.

1-A.2 LOS LOGARITMOS Y LA FUNCIÓN EXPONENCIAL

1-A.2.1 Los exponentes

Para cualquier número real positivo *b* y cualquier entero positivo *n*, la expresión

b^n

significa, por su definición,

$b \times b \times \ldots \times b$

donde el número real positivo *b* aparece n veces. De esta definición, siguen

1. $b^m \times b^n = b^{m+n}$
2. $b^m \div b^n = b^{m-n}$ cuando $m > n$
3. $(b^n)^m = b^{m \times n}$

Según la definición dada inicialmente, la expresión b^n tiene un significado únicamente cuando *n* es un entero positivo. Sin embargo, las tres reglas anteriores conducen a las generalizaciones siguientes.

Cuando $m=n$,

$b^0 = b^{m-n} = b^m \div b^n = 1$

Tenemos también

$$b^{-n} = b^{0-n} = b^0 \div b^n = 1 \div b^n = \frac{1}{b^n}$$

Finalmente, sabemos que, si *a* es la n^{esima} raíz de *b*,

$a^n = b$

Por lo tanto

$$b^{\left(1/n\right)} = \left(a^n\right)^{\left(1/n\right)} = \left(a^{n \times \left(1/n\right)}\right) = a^{\left(n/n\right)} = a = \sqrt[n]{b}$$

Más generalmente

$$b^{\left(m/n\right)} = \sqrt[n]{b^m}$$

Con estas generalizaciones, la expresión b^r tiene significado para cualquier número real positivo *b* y para cualquier número racional $r=m/n$. En cuanto a números irracionales, ellos pueden identificarse como el límite de una sucesión convergente de números racionales ; eso permite de dar un significado a la expresión b^r, no solamente cuando *r* es un

número racional, sino también cuando *r* es cualquier número real, racional o irracional.

1-A.2.2 Los logaritmos

Los logarítmos, tal como el operador suma, son nada más un convenio de escritura. La expresión

$$x = \log_b y$$

simplemente significa

$$y = b^x$$

y se lee como "*x* es el logaritmo de *y* en la base *b*". Por ejemplo,

$\log_{10} 1 = 0$

$\log_{10} 10 = 1$

$\log_{10} 100 = 2$

$\log_{10} 1000 = 3$

etc.

El logaritmo en la base 10 de un número que no es un poder entero de 10 no es un número entero. Por ejemplo,[79]

$\log_{10} 2 = 0.30103$ significa $10^{0.30103} = 2$

$\log_{10} 12 = 1.07918$ significa $10^{1.07918} = 12$

Las bases más frecuentemente usadas para logaritmos son 10 (logaritmos "comunes") y el número irracional $e = 2.71828\ldots$ Logaritmos en la base *e* se llaman "naturales", o "neperianos"[80]. Usualmente, $\log_e y$ se denota ln *y*: naturalmente, "ln" significa "logaritmo natural".

[79] En las tablas de logarítmos que se usaban antes de los Lotus y otros Excel, el logaritmo estaba descompuesto en dos partes : la parte entera se llamaba la *característica*, y la parte fraccional, la *mantisa*. En log 12, la característica es de 1 y la mantisa de 07918.

[80] Del nombre de su inventor, el teólogo y matemático escoséz John Napier (1550-1617), cuyo nombre se escribe también "Neper".

La operación que consiste en encontrar un número a partir de su logaritmo se denota a veces por la palabra “antilog”; así tenemos la equivalencia

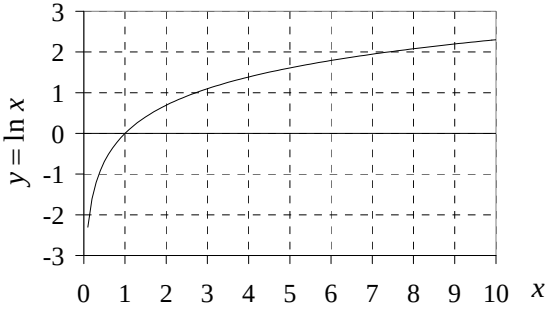
$$\text{antilog}_b\, x = b^x$$

Entonces, antilog $x = e^x$ o 10^x, según se considera x como un logaritmo neperiano o un logaritmo comun.

La figura siguiente ilustra la relación entre un número y su logaritmo.

Función logarítmica

$y = \ln x$



Se ve que la transformación logarítmica es una transformación *monótona creciente*: si $y_1 > y_2$, entonces $\log y_1 > \log y_2$, pues $y = b^{\log y}$.

Las reglas que se aplican a los exponentes se transponen a los logaritmos.

1. La regla $b^m \times b^n = b^{m+n}$ implica
   $$\log (y \times z) = \log y + \log z$$
2. La regla $b^m \div b^n = b^{m-n}$ implica
   $$\log\left(\frac{y}{z}\right) = \log y - \log z$$

3. La regla $(b^n)^m = b^{m \times n}$ implica

$$\log y^r = r \times \log y$$

Desde la última regla puede inferirse la regla para pasar de una base a otra. Supongamos que queremos pasar desde el logaritmo común en la base 10 al logaritmo neperiano en la base *e*. Pues $x = \log_{10} y$ significa $10^x = y$, tenemos

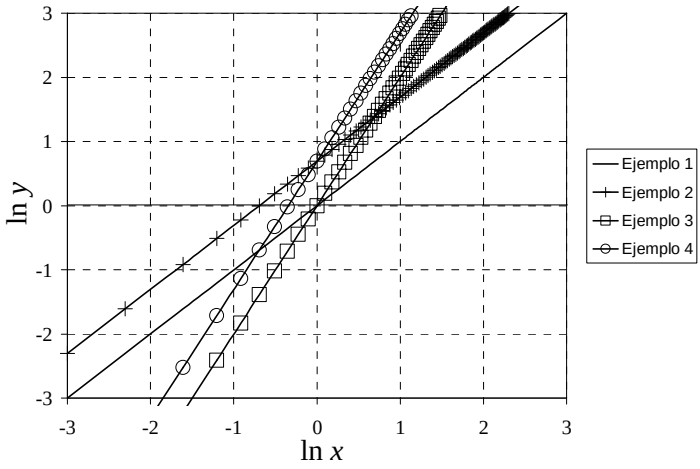
$$\log_e y = \log_e 10^x = x \log_e 10 = \log_{10} y \times \log_e 10$$

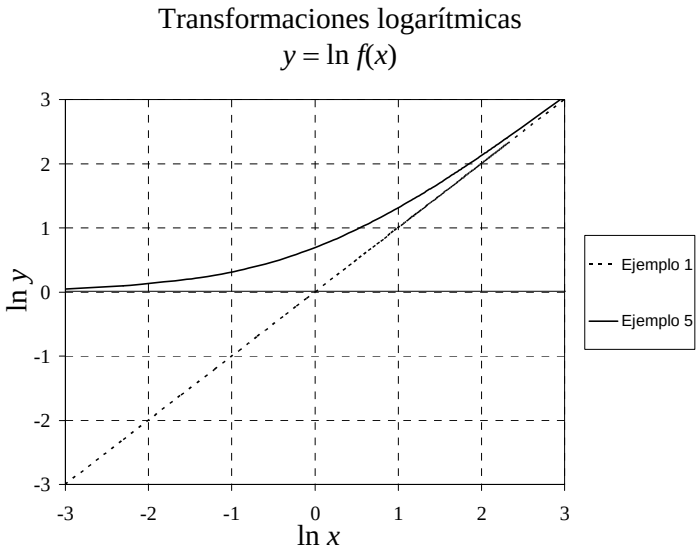
Las figuras que siguen muestran cómo cambia la forma de relaciones con la transformación logarítmica. Las relaciones que se representan son:

4. $y = x \quad \Rightarrow \quad \ln y = \ln x$
5. $y = 2x \quad \Rightarrow \quad \ln y = \ln 2 + \ln x$
6. $y = x^2 \quad \Rightarrow \quad \ln y = 2 \ln x$
7. $y = 2x^2 \quad \Rightarrow \quad \ln y = \ln 2 + 2 \ln x$
8. $y = x + 1 \quad \Rightarrow \quad \ln y = \ln (x + 1)$

Transformaciones logarítmicas

$y = \ln f(x)$





Se ve que después de una transformación logarítmica, hay relaciones no lineales que se vuelven lineales, y hay relaciones lineales que se vuelven no lineales.[81]

1-A.2.3 La función exponencial

La función

$$y = e^x$$

se llama la función exponencial. Se denota también como $y = \exp(x)$, o más raramente, $y = \text{antilog}_e\, x$.

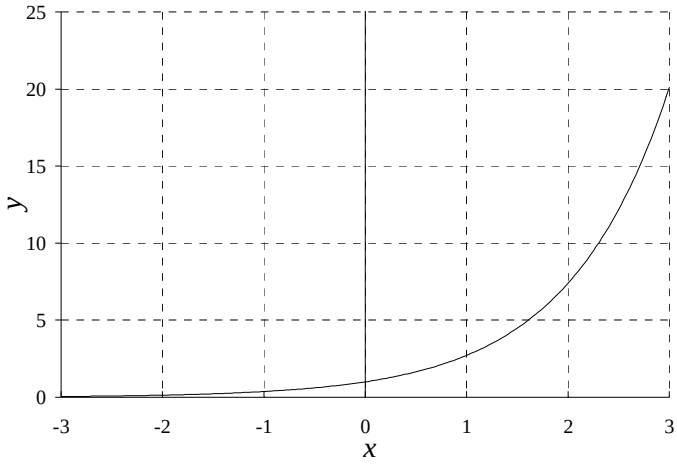
Se usa también la exponencial negativa, la cual tiene la forma

$$y = \exp(-x) = e^{-x} = \frac{1}{e^x}$$

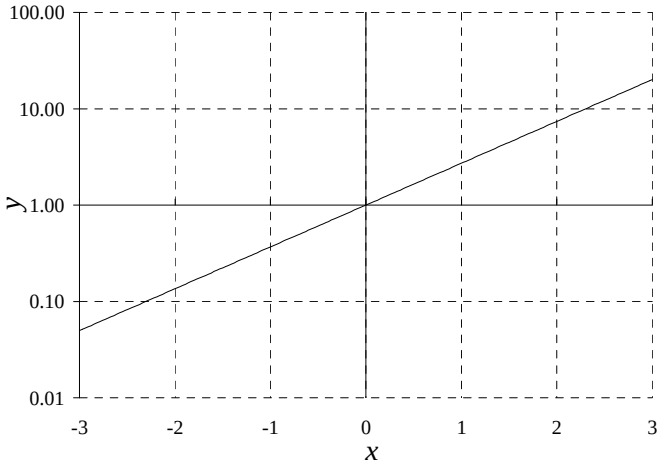
[81] Vea Wonnacott y Wonnacott (1992), p 513-523, "La non-linéarité résolue grâce aux logarithmes" (la no linealidad resuelta gracias a los logarítmos).

Las cuatro figuras siguientes ilustran la función exponencial.

Función exponencial
$y = \exp(x)$

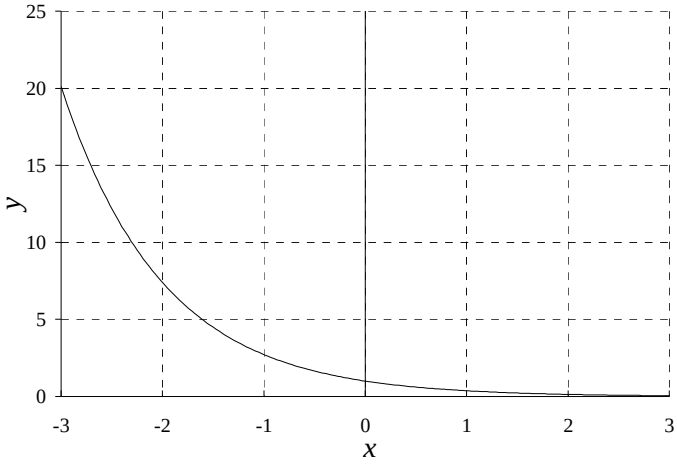


Función exponencial (escala logarítmica)
$y = \exp(x)$



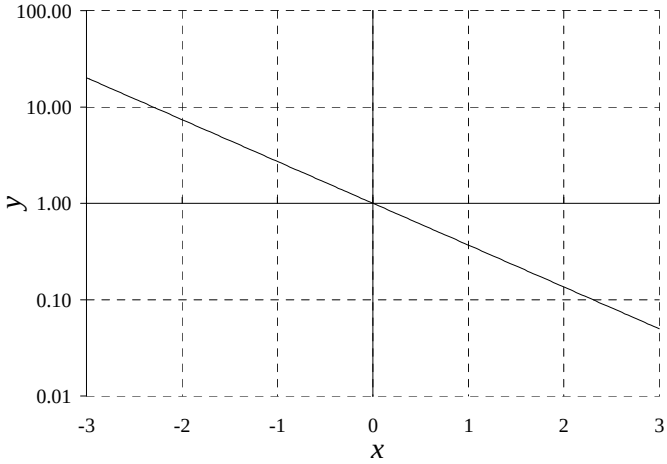
Función exponencial negativa

$y = \exp(-x)$



Función exponencial negativa (escala logarítmica)

$y = \exp(-x)$



1-A.2.4 ¿Por qué los logarítmos neperianos?

El número irracional *e* se define como el límite de una sucesión infinita :

$$e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$$

El interés para esta constante neperiana viene del análisis del crecimiento exponencial. Comenzando con un valor inicial q_0, una cantidad q se multiplica cada período por un factor $(1+r)$; entonces, después de t períodos, esta cantidad será igual a

$$q = q_0\,(1+r)^t$$

Esta es la fórmula para el crecimiento geométrico de una cantidad a la que se aplica un interés compuesto una vez cada período. Supongamos que la frecuencia en que el interés se compone sea multiplicada por *n*. Tenemos

$$q' = q_0\left(1+\frac{r}{n}\right)^{nt}$$

¿Qué sucede cuando *n* llega a ser muy grande (cuando *n* tiende a la infinidad y que el interés se compone continuamente)? Para verlo, escribamos la ecuación anterior en la forma siguiente

$$\lim_{n\to\infty}\left\{\left[1+\frac{1}{\left(n/r\right)}\right]^{n/r}\right\} = e$$

Cuando n tiende a la infinidad, *n*/*r* también tiende a la infinidad y la expresión entre los paréntesis rizados tiende a la constante *e*. Se obtiene :

$$\lim_{n\to\infty} q' = q_0\, e^{rt}$$

Ésta es la fórmula del crecimiento exponencial, que es la versión continua del crecimiento geométrico. Se nota que ambas fórmulas se vuelven lineales cuando se toma el logaritmo.

SOLUCIONES DE LOS EJERCICIOS SOBRE EL OPERADOR SUMA

1.1 $$\sum_{k=1}^{4} x_k = (5+3)+(5+6)+(5+9)+(5+12) = 50$$

1.2 $$\sum_{i=0}^{3} x_i = (5+0)+(5+3)+(5+6)+(5+9) = 38$$

1.3 Cuando $n = 4$, $$\sum_{i=0}^{n-1} x_{i+1} = \sum_{i=0}^{3} x_{i+1} = \sum_{i=1}^{4} x_i = 50$$

2.1 $$\sum_{x=2}^{3} x^3 = 2^3 + 3^3 = 8 + 27 = 35$$

2.2 $$\sum_{i=1}^{4} \frac{1}{i} = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} = \frac{25}{12}$$

2.3 Cuando $a = 345$, $$\sum_{j=1}^{10} a = 10\, a = 3450$$

3.1 $$\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i = \left(x_1 + x_2 + \cdots + x_k\right) + \left(x_{k+1} + x_{k+2} + \cdots + x_n\right)$$

$$\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i = x_1 + x_2 + \cdots + x_k + x_{k+1} + x_{k+2} + \cdots + x_n$$

$$\sum_{i=1}^{k} x_i + \sum_{i=k+1}^{n} x_i = \sum_{i=1}^{n} x_i$$

3.2 $$\sum_{i=1}^{n} (cx_i) = cx_1 + cx_2 + \cdots + cx_n$$

$$\sum_{i=1}^{n} (cx_i) = c(x_1 + x_2 + \cdots + x_n)$$

$$\sum_{i=1}^{n} (cx_i) = c\sum_{i=1}^{n} x_i$$

3.3 $$\sum_{i=1}^{t} (x_i + y_i) = (x_1 + y_1) + (x_2 + y_2) + \cdots + (x_t + y_t)$$

$$\sum_{i=1}^{t} (x_i + y_i) = (x_1 + x_2 + \cdots + x_t) + (y_1 + y_2 + \cdots + y_t)$$

$$\sum_{i=1}^{t} (x_i + y_i) = \sum_{i=1}^{t} x_i + \sum_{i=1}^{t} y_i$$

4.1 $$\sum_i \sum_j a_{ij} = 4 + 50 + 5 + 30 + 6 + 10 = 105$$

4.2 $$\sum_i a_{2i} = 5 + 30 = 35$$

4.3 $$\sum_{j=1}^{2} \sum_{i=1}^{2} a_{ij} = (4 + 5) + (50 + 30) = 89$$

4.4 $$\sum_{i=1}^{2} \sum_{j=1}^{2} a_{ij} = (4 + 50) + (5 + 30) = 89$$

4.5 $$\sum_{j=1}^{2}\sum_{i=1}^{j} a_{ij} = (4) + (50 + 30) = 84$$

4.6 $$\sum_{i=1}^{2}\sum_{j=1}^{i} a_{ij} = (4) + (5 + 30) = 39$$

4.7 $$\sum_{i}\sum_{j>i} a_{ij} = 50$$

4.8 $$\sum_{i}\sum_{j\leq i} a_{ij} = (4) + (5 + 30) + (6 + 10) = 55$$

ANEXO 1-B
TABLA DEL ALFABETO GRIEGO

Rango el el alfabeto griego	Letra mayuscula	Letra minuscula	Correspondencia en el teclado (fuenta símbolo)	Nombre de la letra griega	Fonética griega moderna
1	Α	α	a	alfa	a
2	Β	β	b	beta	v
3	Γ	γ	g	gama	g
4	Δ	δ	d	delta	dh
5	Ε	ε	e	epsilón	e
6	Ζ	ζ	z	zeta	z
7	Η	η	h	eta	i
8	Θ	θ	q	theta	th
9	Ι	ι	i	iota	i
10	Κ	κ	k	kappa	k
11	Λ	λ	l	lambda	l
12	Μ	μ	m	mu	m
13	Ν	ν	n	nu	n
14	Ξ	ξ	x	xi	x
15	Ο	ο	o	omicron	o
16	Π	π	p	pi	p
17	Ρ	ρ	r	rho	r
18	Σ	σ	s	sigma	s
19	Τ	τ	t	tau	t
20	Υ	υ	u	upsilón	i
21	Φ	ϕ	f	fi	f
22	Χ	χ	c	ji o khi	kh
23	Ψ	ψ	y	psi	“ps”
24	Ω	ω	w	omega	ô
	ϑ	φ	j	Formas arcaícas de theta y phi	
	ς	ϖ	v	ς es la forma de sigma como letra final	

INTRODUCCIÓN A LA SEGUNDA PARTE

Esta parte del curso tiene un objetivo ambicioso: iniciar en la epistemología de los métodos estadísticos a través del estudio de la lógica de las pruebas de hipótesis. Representa tanto una apuesta para el profesor como un desafío para los estudiantes. Es importante, por lo tanto, antes de empezar, medir y delimitar la dificultad.

Antes que nada, acabemos con un mito: esta dificultad no surge de las matemáticas. Deriva, más bien, de la complejidad de la lógica y la obligación que requiere de apartarse de la lógica de todos los días. En efecto, es indispensable buscar liberarse de sus hábitos mecánicos de pensamiento para poder entender la estructura de los argumentos de la inducción estadística, los cuales se basan en razonamientos hipotéticos y en una lógica probabilista.

Por lo tanto, aunque recurrimos a algunos enunciados matemáticos, el texto es comprensible para cualquier lector con buenos conocimientos de la lógica formal y capaz de un esfuerzo de concentración, lo cual se espera de una persona que cursa un nivel superior de estudio.

Además, aparte de la complejidad de la lógica, su nivel de abstracción puede contribuir a incrementar el nivel de dificultad de esta parte del curso. De hecho, se trata de abordar, como si fuera una mecánica para estudiarla, una argumentación general que se puede aplicar para una multitud de casos par-

ticulares. Sin embargo, una formulación del argumento es general solamente si es abstracta, lo que, no obstante, no impide ilustrar el tema con ejemplos sencillos. Antes bien, el propósito de este trabajo es permitir al estudiante trascender lo particular con el fin de ser capaz de aplicar la misma herramienta general en todos los casos particulares que sean pertinentes. Y si bien se espera un buen manejo de la lógica formal de la parte de un estudiante del ciclo superior, también se espera de él una alta capacidad de abstracción.

Ahora bien, se pueden uno preguntar el porqué de un esfuerzo de este tipo y qué beneficios puede traer. Y la respuesta es la siguiente: permite poseer un marco conceptual claro y robusto gracias al cual se pueden usar a conciencia algunos métodos estadísticos y emitir una crítica asertiva con relación a como otros usan estos métodos sin que sea necesario poseer todos los detalles técnicos de éstos, ni conocer todos los desarrollos matemáticas subyacentes. Me parece que, para lograr este resultado, vale la pena el esfuerzo.

Además, puesto que los paquetes estadísticos que existen ahora en nuestras computadoras efectúan los cálculos en estadística, es aún más importante que un joven investigador encare los métodos estadísticos con un riguroso marco conceptual. Del revólver Colt en el viejo oeste, el actor John Wayne comentaba que era "the great equalizer" porque hizo desaparecer la ventaja de tener poderosos músculos (o el dinero para comprárselos). De la misma manera, la informática establece cierta igualdad entre los buenos en matemáticas y los demás. Y, como los grandes o chicos que traen armas de fuego deben aprender la no-violencia, de igual manera los investigadores buenos o no en matemáticas deben aprender la no-violencia hacia los datos, o sea el respeto a la lógica de la inducción estadística.

CAPÍTULO 2-1
DESCRIPCIÓN E INDUCCIÓN ESTADÍSTICAS EN CIENCIAS SOCIALES[82]

La estadística es esencialmente una colección de métodos matemáticos para procesar datos con el fin de resumir o bien generalizar la información que nos proporcionan. La estadística descriptiva es la parte de la estadística que sirve para resumir la información. El proceso de generalizar la información se llama inducción estadística.

2.1.1 ESTADÍSTICA DESCRIPTIVA

Al enfrentarse a un conjunto de datos aunque pequeño, el cerebro humano se ve incapaz de captar de manera inmediata toda la información detallada que contiene.[83] La solución a este problema consiste en dejar de lado los detalles para concentrar su atención en los grandes rasgos. La estadística descriptiva permite procesar metódicamente los datos con el objetivo de condensar la información que contienen. Con cal-

[82] Blalock (1979); Wonnacott y Wonnacott (1992).

[83] Según Georges Ifrah (1994, tomo 1, p. 33-34, "Les limites de la perception directe des nombres"), el cerebro no puede aprehender de manera concreta, es decir, sin contar (lo que constituye una abstracción) más de cuatro objetos al mismo tiempo. Por ejemplo, sin contar, no podemos, con la pura mirada, diferenciar entre cinco o seis objetos.

cular porcentajes, promedios, desviaciones estándar o coeficientes de correlación, es posible tener una visión global de los datos. Sin embargo, debemos estar conscientes de que al resumir los datos de esta manera, perdemos parte de la información que contenían, y esto puede ser causa de errores al menos que seamos prudentes al momento de interpretar.

2.1.2 La inducción estadística

Es casi imposible, y particularmente en ciencias sociales, conseguir los datos pertinentes de la totalidad de un fenómeno en estudio. Es más probable que las observaciones que están a nuestra disposición se apoyen en una parte del fenómeno. Entendemos, entonces, la necesidad de generalizar a partir de una información incompleta. Sin embargo, para que se considere una generalización ciertamente científica, es necesario fundamentarla con principios epistemológicos.[84]

Los métodos de inducción estadística son, de hecho, una expresión matemática de principios epistemológicos gracias a los cuales se pueden inferir proposiciones de alcance más general a partir de la información que se obtiene de un conjunto de datos particulares. La inducción estadística es, por lo tanto, una manera científicamente válida de pasar de lo particular[85] a lo general.

Concretamente, el proceso inductivo estadístico busca sacar conclusiones con relación a las diversas características de una población a partir de hechos observados en una muestra obtenida de ésta. La estadística emplea la palabra *parámetros* para designar las características de la población y la palabra *estadísticas* para designar las características de la muestra.

---

[84] La epistemología es una parte de la filosofía que consiste en el estudio crítico de las ciencias con el fin de determinar, de cada una, su origen lógico, su valor y su alcance.

[85] Hay una broma en inglés que ilustra el carácter particular de los datos: " 'Data' is the plural of 'anecdote' " ('Datos' es el plural de 'anécdota').

Es importante siempre recordar que se consideran, en una situación normal, los parámetros como unos valores fijos con relación a una población y que, por lo general, son desconocidos (puesto que no se conoce, en su totalidad, la población misma). Por lo contrario, dado que se puede obtener más de una muestra de una población específica, las estadísticas son valores que pueden variar de una muestra a otra; por otro lado, se conocen o se pueden calcular los valores de las estadísticas de una muestra dada. Sin embargo, no sabemos hasta que grado una muestra es representativa de la población en general, ni en qué medida una estadística calculada a partir de esta muestra se parezca al parámetro correspondiente a la población desconocida.

Ejemplos de inducción estadística:

1. Basándose en las respuestas obtenidas por sondeo efectuado a una muestra de la población de una ciudad, estimar la proporción de los ciudadanos que son favorables a algún proyecto urbanístico.
2. Partiendo de la hipótesis (del modelo), aceptada a priori, que la relación macroeconómica entre el ingreso de los hogares y las inversiones en construcción de vivienda privada se puede describir con la ecuación

   $$I = a + bR$$

   (donde I es el monto de las inversiones y R el ingreso agregado), estimar el valor de los parámetros a y b a partir de los datos publicados por INEGI para el Estado de Puebla de 1974 a 1994.[86]

Es de hacer notar que las medidas que se usan en estadística descriptiva (promedio, desviación estándar...) se usan también en el contexto de la inducción estadística. No obstante, en estadística descriptiva no se distingue entre parámetros y estadísticas porque la estadística descriptiva no hace distinción entre población y muestra.

[86] Este ejemplo de modelo es, de toda evidencia, demasiado simplista.

2-1.3 LAS PROBABILIDADES Y LA INDUCCIÓN ESTADÍSTICA: LA RELACIÓN ALEATORIA ENTRE UNA MUESTRA Y LA POBLACIÓN

Con la inducción estadística dejamos el dominio de certidumbre. En efecto, la inducción estadística parte de una muestra entre tantas otras posibles que se pueden obtener de la población en estudio: si obtenemos de una población dada una muestra de un tamaño específico con un procedimiento determinado, en caso de repetir el proceso, la segunda muestra tiene mucha probabilidad de ser diferente a la primera. Para una población dada existe, por consiguiente, muchas muestras posibles. El conjunto de muestras posibles forma también una población en términos estadísticos: los "individuos" de esta población son las muestras.

Por ejemplo, el total de los abonados del teléfono en la cuidad de Montreal forma una población. Podríamos obtener de esta población una muestra de 1000 abonados seleccionados al azar en un directorio telefónico. Luego podríamos repetir el proceso y obtener una segunda muestra, luego una tercera, etc. (de hecho, podríamos repetir este proceso al infinito siempre y cuando se pueda seleccionar nuevamente los abonados que pertenecen a una muestra[87]). El conjunto de todas las muestras posibles de 1000 abonados seleccionados al azar en el directorio es la población de las muestras.

Entre las muestras posibles, algunas son representativas de la población estudiada mientras que otras lo son menos. Puesto que no conocemos la población sino es por la muestra, nunca podemos saber con certeza hasta que grado la muestra especifica que se obtuvo es representativa de la población en

[87] No se debe confundir este proceso con la muestra con reemplazo (sustitución) cuando se sortean los individuos que constituyen una muestra de manera secuencial y cuando un individuo que se sorteó es nuevamente elegible para el sorteo que sigue.

general. La relación entre la muestra y la población es, por lo tanto, esencialmente aleatoria (o sea, influenciada por el azar).

En resumen, está claro que no es posible saber con certeza en qué medida una estadística calculada a partir de datos de una muestra se parece al parámetro desconocido correspondiente en la población. Sin embargo, la teoría de las probabilidades nos facilita herramientas para evaluar la probabilidad de que la diferencia (el error de estimación) entre la estadística y el parámetro se sitúe en el interior de un cierto margen. Es por consiguiente con la teoría de las probabilidades que fundamentamos las reglas de la inducción estadística.

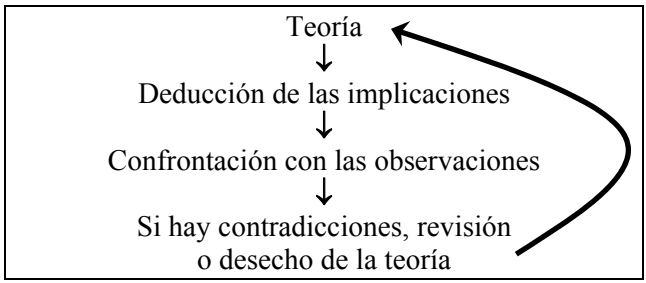
CAPÍTULO 2-2
LA INDUCCIÓN ESTADÍSTICA

2-2.1 LA INDUCCIÓN ESTADÍSTICA EN EL MÉTODO CIENTÍFICO: MODELOS TEÓRICOS Y MODELOS ALEATORIOS[88]

Estudiaremos después la lógica fundamental de las pruebas de hipótesis, la cual es el punto final de la inducción estadística. Mientras, antes, es importante situar la inducción estadística en el marco del método científico. Nuestro punto inicial será un esquema simplista del metodo cientifico hipotético-deductivo.[89]

[88] Malinvaud (1969), caps. 1 y 2 y Blalock (1972), caps. 2 y 8.

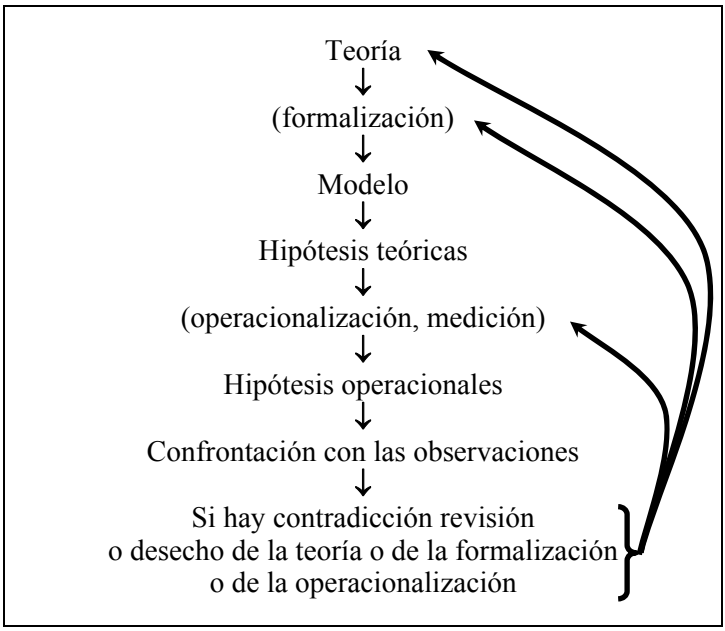
[89] Debemos reconocer que el esquema es truncado. Para tener acceso a una presentación más detallada del proceso de desarrollo del conocimiento, vea Robert (1993). Lo que se presenta ahora es más bien lo que Kuhn, en *La estructura de las revoluciones científicas* (1983), llama la ciencia "normal".



Tanto en ciencias sociales como en ciencias físicas, se formaliza, a menudo, las teorías como modelos: "Un modelo consiste en la representación formal de ideas o de conocimientos con relación a un fenómeno" (Malinvaud, 1969, p 45).

Cuando, como pasa muchas veces, se selecciona una representación formal del tipo matemático, tenemos un modelo matemático.

Una teoría es una construcción intelectual de carácter hipotético, es decir, una hipótesis global emitida con relación a un fenómeno. De igual manera, un modelo, y también las partes del modelo, son una representación formal del fenómeno. No obstante, tanto en el modelo como en la teoría, no se definen los conceptos en términos operacionales, esto es, no se pueden confrontar las hipótesis teóricas directamente con las observaciones. Para esto es necesario traducir las hipótesis teóricas en hipótesis operacionales, lo que implica definir mediciones (vea el capitulo 1-1). En estas condiciones, el esquema complejo es el siguiente:



En este esquema es fácil observar que una contradicción entre las observaciones y las predicciones de la teoría puede venir no solamente de la teoría misma o también de su formalización o de su operacionalización. En la realidad el proceso de la investigación no es tan fluido como lo enseña el primer esquema...

La estadística interviene al momento de la confrontación con las observaciones visto que, en la mayoría de los casos, se efectúan éstas sobre un muestreo cuando el modelo abarca una población. Ahora bien, la mayoría de los modelos matemáticos son deterministas, o sea que las variables que representan los conceptos se vinculan entre sí gracias a relaciones funcionales (funciones matemáticas) que no tienen ningún elemento aleatorio.

Sin embargo, la relación entre una muestra y la población de donde proviene, es esencialmente aleatoria puesto que cada muestra no es más que una de muchas muestras posibles.

Para confrontarlos con las observaciones, se necesita complementar los modelos deterministas para tomar en cuenta este elemento aleatorio. Al combinar un modelo determinista y un modelo de la relación aleatoria de la muestra con la población, obtenemos un modelo aleatorio (probability model).[90] Para traducir esta distinción, algunos autores designan como "modelo estructural" (structural model) el modelo teórico determinista y como "modelo muestral" (sampling model) el modelo de relación entre la muestra y la población.[91]

Hasta el momento mencionamos una sola fuente de lo aleatorio. De hecho, existen tres "puertas" por las cuales lo aleatorio se introduce en los modelos:[92]

1. Primero, existe el carácter aleatorio que ya se mencionó de la relación entre la muestra y la población de donde proviene.

[90] "[...] para todo conjunto de valores que se dan a las variables exógenas, un modelo aleatorio define la ley de probabilidad correspondiente a las variables endógenas" (Malinvaud, 1969, p. 59).

[91] Más precisamente, Upton y Fingleton (1985, p. 264) nombran "structural model" a la especificación de la relación funcional entre la variable dependiente y las variables independientes; nombran "sampling model" a la hipótesis con relación a la distribución de probabilidad de la variable dependiente (o, lo que equivale, del termino de error).

[92] Malinvaud escribe: "Sabemos que se justifica el empleo del cálculo de las probabilidades para el análisis de los datos estadísticos con una u otra de las dos consideraciones siguientes: o bien se asimila el fenómeno estudiado con un proceso que tenga una determinación aleatoria de algunas magnitudes; se consideran, entonces, estas magnitudes como aleatorias en el universo (NDLR: es decir, en la población) como en la muestra observada. O bien, la selección de los elementos observados resulta de un sorteo aleatorio; entonces, la composición de la muestra es aleatoria, y por lo tanto, los datos obtenidos lo son también aunque se refieran a variables no aleatorias." (Malinvaud, 1969, p. 62). Malinvaud prosigue diciendo que, en el contexto de la econometría, la primera consideración se adapta mejor.

2. Segundo, las variables operacionales son mediciones imperfectas de los conceptos, así que podemos considerar que el error de medición es aleatorio (es decir, determinado al azar). Por lo tanto, se puede representar con un modelo aleatorio, la influencia de los errores de medición que intervienen en el momento de la traducción de las hipótesis teóricas en hipótesis operacionales (los modelos de la "teoría de los errores" en ciencias físicas fueron de por sí unos de los primeros modelos aleatorios).
3. Finalmente, percibimos algunos fenómenos como aleatorios de por sí y no se pueden representar adecuadamente con modelos teóricos no aleatorios. El azar en estos modelos constituye un concepto que encierra unas veces una indeterminación fundamental (como en física de las partículas), otras, una multitud de factores que no se pueden observar (como pasa en la mayoría de los casos en ciencias sociales),[93] y cuyas manifestaciones aparecen como consecuencias de la aplicación de las leyes de probabilidad.

De todas maneras, un modelo aleatorio tiene un carácter hipotético puesto que se apoya en hipótesis sobre la estructura aleatoria, es decir sobre las leyes de probabilidad que rigen el azar. Durante la confrontación con las observaciones, estas hipótesis no se cuestionan (por lo menos no todas). Se consideran, por decirlo así, como el peaje que se exige para cruzar el puente de lo conocido a lo desconocido, dado que la inducción estadística va "más allá" de los datos observados. Sin embargo, aunque la inducción se basa en hipótesis, existe una ganancia epistemológica cuando las hipótesis que fun-

[93] En particular, pensamos en los modelos de utilidad aleatoria (random utility) subyacentes a los modelos de selecciones discretas (discrete choice) logit, probit, etc. Vamos a encontrar este tipo de modelos en el apartado 4-3.

damentan la inducción son menos restrictivas que los resultados obtenidos por medio de la inducción.

En resumen, lo mismo si la simplicidad del esquema que presentamos con anterioridad aparenta lo contrario, debemos reconocer que la confrontación de la teoría, de los modelos y de las hipótesis con las observaciones es, pocas veces, total. Cada ejercicio de confrontación se basa, de hecho, en un modelo más general que no se cuestiona. Esto es ciertamente el caso de la inducción estadística y de los tests de hipótesis, temas que trataremos más adelante. En efecto, los tests de hipótesis se aplican, casi siempre, a unas formas particulares de un modelo teórico general que no se cuestiona y se basan en un modelo aleatorio que no se cuestiona tampoco.[94]

2-2.2 ALGUNOS CONCEPTOS CLAVE DE LA TEORÍA DE LAS PROBABILIDADES*

Antes de estudiar propiamente la inducción estadística, es necesario recordar, aunque de manera resumida, las definiciones de algunos conceptos claves de la teoría de las probabilidades.

[94] Una "confrontación total" sería propia de una revolución científica a la Khun, Sin embargo, es dudoso que la inducción estadística tenga un papel predominante en el proceso de cambio de paradigma de una revolución científica. No obstante, es cierto que existen tests "de nivel superior", por así nombrarlos, que se aplican a ciertos aspectos del modelo aleatorio. Pero, estos mismos tests se basan en modelos aleatorios más generales los cuales, a este punto, no se cuestionan. Podemos imaginar un test del modelo aleatorio del test del modelo aleatorio... Antes bien, poco importa la "altura" del nivel al cual llegamos, siempre existirá un nivel superior donde el modelo de muestreo no se cuestiona.

* Referencias: Wonnacott y Wonnacott (1992, caps. 3 y 4, apartados 4.1-4.2)

2-2.2.1 Conceptos fundamentales

Azar (de la palabra árabe *az-zahr*, "el dado"). El Diccionario Enciclopédico Planeta lo define como "suceso que se presenta fortuitamente sin venir motivado por intención o plan alguno".

Evento aleatorio.[95] Evento cuya realización o no depende del azar. Por ejemplo, en el caso de las muestras que se pueden sortear de una población, cada posibilidad es un evento aleatorio. Cuando se sortea una muestra, solo uno de estos eventos se realiza mientras que los demás no se realizan.

Variable aleatoria. Es una variable cuyo valor es el resultado de eventos aleatorios.[96] Puesto que el resultado del sorteo de una muestra es un evento aleatorio, todas las mediciones que se pueden efectuar sobre una muestra son variables aleatorias. Esto se aplica tanto en los datos brutos como en las estadísticas calculadas a parir de estos datos.

Distinguimos las variables aleatorias discretas que no pueden más que tomar ciertos valores (números enteros en la mayoría de los casos), y las variables aleatorias continuas, cuyo valor puede ser cualquier número real en un intervalo dado (abierto o cerrado). Las variables aleatorias continuas forman un conjunto de posibilidades infinitas, cuando las va-

[95] "Alea" significa dado de jugar en latín. Recordemos el famoso "Alea Jacta est" (se jugaron los dados), significa la suerte está echada",de Julio Cesar al momento de cruzar el Rubicon.

[96] En la mayoría de los manuales de estadística, la distinción entre la variable aleatoria y sus valores posibles o observados se manifiesta a través de una simbolización donde X es la variable aleatoria y x sus valores posibles o observados. En nuestro contexto, escribiremos "variable aleatoria" textualmente cuando será necesario; de otra manera, usaremos una x para designar los dos.

riables aleatorias discretas pueden formar un conjunto de posibilidades finito cuando su campo de variación es finito.[97]

Probabilidad de un evento aleatorio. Todos tenemos una noción intuitiva de lo que es una probabilidad, sin embargo, no es dar de este concepto una definición rigurosa. Se puede tratar la noción de probabilidad de tres maneras.

Podemos concebir la probabilidad en el contexto de una serie de "experimentos" o de "pruebas", donde el resultado de cada intento es un "éxito" (el evento acontece) o de una "falla" (el evento no acontece); ésta es la definición "frecuentista" de la probabilidad en términos de frecuencias relativas de un evento aleatorio. Durante una serie de experimentos de este tipo (sortear cara o cruz o echar los dados), se define la probabilidad de un evento aleatorio (como obtener "sol" o un "seis") como la proporción de los experimentos cuando este evento se realiza en promedio.

Se puede definir la probabilidad de un evento como la suerte que pensamos tenga un evento de acontecer en una escala de 0 a 100% (definición subjetiva o bayesiana).

Finalmente, podemos considerar el concepto de probabilidad como primero y no definible para luego enunciar un sistema de axiomas al cual se debe conformar cualquier medición de probabilidad.

2-2.2.2 Distribuciones de probabilidad

Función de distribución de probabilidad o distribución de probabilidad. Es una correspondencia que asocia una probabilidad a cada evento de un conjunto exhaustivo de eventos que sean mutuamente exclusivos (posibilidades). Por ejem-

[97] Una variable aleatoria cuyo campo de variación es el conjunto de los enteros naturales es una variable discreta; sin embargo, el conjunto de sus valores posibles es infinito.

plo, cuando jugamos a cara o sol una vez con una moneda que no es trucada, la función de probabilidad es

$$\text{Prob}(\textit{cara}) = \text{Prob}(\textit{sol}) = 0.5$$

La distribución de probabilidad se parece a una distribución de frecuencia relativa, pero se distingue en el hecho de que la distribución de frecuencia específica frecuencias observadas cuando la distribución de probabilidad asigna a cada evento la frecuencia relativa que tendría en promedio en un contexto de una serie infinita de experimentos (vea más arriba, la definición "frecuentista" de la probabilidad).

Función de distribución acumulada de una variable aleatoria. La función de distribución acumulada de una variable aleatoria es una función $F(x)$ (una correspondencia) la cual, por cada valor posible de x de la variable aleatoria, da la probabilidad de que la variable aleatoria tome un valor inferior o igual a x.[98]

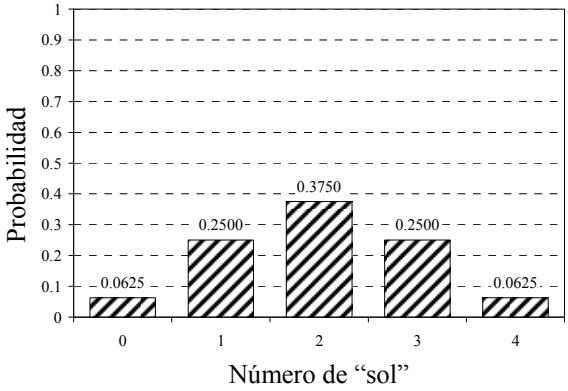
Por ejemplo, si sorteamos cara o sol cuatro veces, el número de veces que obtenemos cara es una variable aleatoria discreta cuya función de probabilidad y función de distribución acumulada se representan en la tabla que sigue (esta distribución se llama distribución binomial). Las figuras que acompañan la tabla ilustran las nociones de distribución de probabilidad y de función de distribución acumulada.

[98] En caso que los valores de la variable aleatoria no sean numéricos – como "cara" o "sol"– es necesario definir con anterioridad el orden en el cual se acomodan estos valores para que la relación "inferior o igual" tenga sentido.

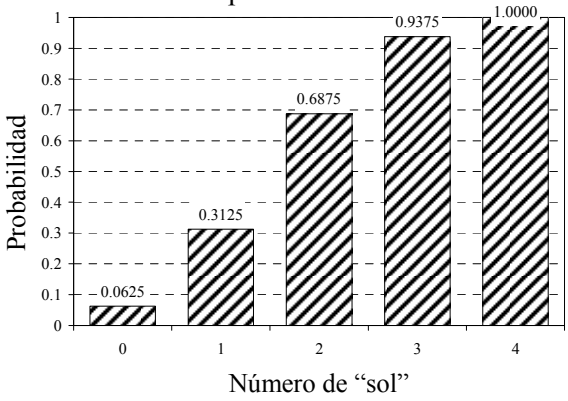
Función de distribución acumulada de una variable aleatoria

Número de "sol"	Probabilidad	Probabilidad acumulada	
x_i	$f(x_i)$	$F(x_{i-1})$	$F(x_i)$
0	1/16		1/16
1	4/16	+ 1/16	= 5/16
2	6/16	+ 5/16	= 11/16
3	4/16	+ 11/16	= 15/16
4	1/16	+ 15/16	= 16/16

Función de probabilidad



Función de probabilidad cumulativa



2-2.2.3 Distribución de muestreo[*]

El concepto de distribución de muestreo (sampling distribution) es primordial en inducción estadística. Es la forma operacional que toma el modelo de muestreo (el modelo de la relación entre una muestra y la población; vea 2-2.1).

En efecto, una distribución de muestreo es una distribución de probabilidades asociada a una estadística. Recordemos que una estadística es una característica de una muestra, mientras que un parámetro es una característica de una población.

Ahora bien, vimos que una muestra no es más que una de las muestras del mismo tamaño que se podría obtener de la población estudiada. Por consiguiente, según la muestra obtenida, la estadística podría tomar valores diferentes. Y, puesto que la muestra se obtiene al azar, el valor de la estadística es aleatorio y la estadística misma es una variable aleatoria. La distribución de muestreo de la estadística es su distribución de probabilidad en la población de las muestras de un tamaño específico que se pueden obtener al azar en una población estudiada.

En general, la distribución de muestreo de una estadística depende de los parámetros de la población estudiada. Es esta dependencia la que permite, a partir del valor observado de una estadística, formular enunciados probabilistas con relación con los parámetros. Explicitaremos este proceso cuando tratemos el tema de los tests de hipótesis.

Por ejemplo, imaginemos que queramos saber si una moneda que usamos para jugar a cara o sol es trucada. Si la moneda no es trucada, debería "en promedio" caer con la misma frecuencia sobre cara o sol. Sin embargo, para conocer el verdadero promedio, se tendría que lanzar la moneda un número infinito de veces porque, independientemente del nú-

[*] Referencias: Wonnacott y Wonnacott (1992, p. 224-226).

mero de lances, nunca estaremos seguros del resultado de los lances suplementarios que podríamos efectuar. La población estudiada es, por lo tanto, infinita. La única manera que queda para decidir si debemos considerar si la moneda es trucada o no es efectuar un cierto número de lances, calcular la proporción de cara y sol y aceptar la moneda como honesta si esta proporción (frecuencia relativa) es lo suficientemente cercana a 50%. La distribución de muestreo de esta proporción es la distribución de probabilidad de esta estadística. Esta distribución depende del número de lances y del verdadero valor de la probabilidad.

Diagrama 1

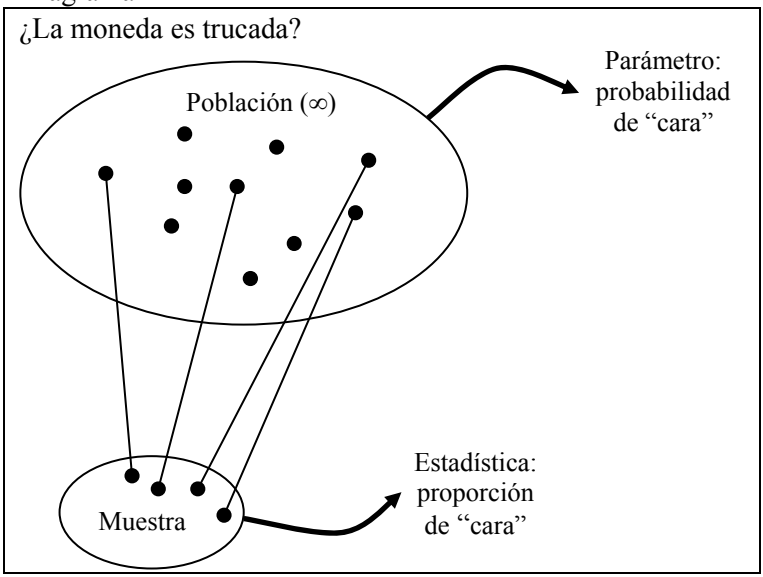
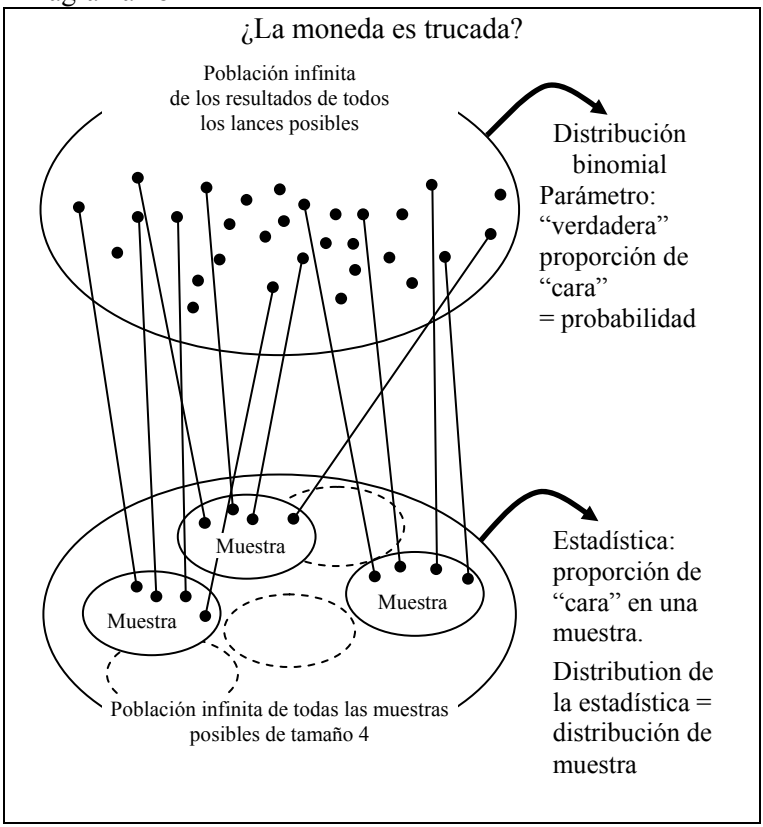


Diagrama 1b


¿La moneda es trucada?
Población infinita
de los resultados de todos
los lances posibles
Distribución
binomial
Parámetro:
"verdadera"
proporción de
"cara"
= probabilidad
Muestra
Muestra
Muestra
Estadística:
proporción de
"cara" en una
muestra.
Distribution de
la estadística =
distribución de
muestra
Población infinita de todas las muestras
posibles de tamaño 4

2-2.2.4 Variables aleatorias continuas: función de densidad de probabilidad y esperanza matemática

Función de densidad de probabilidad de una variable aleatoria continua[*]

Con el fin de entender correctamente la lógica de los tests estadísticos, es de suma importancia captar la diferencia entre una función de densidad de probabilidad y una función de distribución de probabilidad (aunque la expresión más usual, "distribución de probabilidad" se aplique indiferentemente para los dos).

Son variables aleatorias continuas. Ahora bien, con una variable aleatoria continua el número de valores posibles es infinito. En consecuencia, la probabilidad de que la variable aleatoria pueda tomar un valor específico es normalmente infinitamente pequeña; en otras palabras, no podemos asociar una probabilidad a cada valor posible de la variable aleatoria, así que el concepto de función de probabilidad como lo definimos anteriormente no se aplica.

Es la razón por la cual, cuando tratamos con variables continuas, es necesario recurrir a la noción de función de densidad de probabilidad. Se define la función de densidad de probabilidad a partir de la función de probabilidad acumulada puesto que, aunque la probabilidad que la variable aleatoria pueda tomar un valor x especifico es infinitamente pequeña, existe una probabilidad positiva[99] que su valor no rebase este valor x: tenemos por lo tanto una función

$$F(x) = \text{Prob (variable aleatoria} \leq x)$$

[*] Referencias: Wonnacott y Wonnacott (1992, p. 138-140) proponen otra presentación de la función de densidad de probabilidad.

[99] Es decir, no infinitamente pequeña.

Lo cual no es otra cosa que la distribución acumulada en función de los valores posibles de x.

Por ejemplo, se puede considerar el tiempo de vida de un foco eléctrico incandescente como una variable aleatoria continua. ¿Cuál es la probabilidad de que el foco dure exactamente 112 horas, 23 minutos, 14 segundos y tres centésimas? Es fácil entender que esta probabilidad es infinitamente pequeña. Sin embargo, la probabilidad que este foco dure 112 horas, 23 minutos, 14 segundos y tres centésimas o menos es, de seguro, positiva. Esta última probabilidad es la probabilidad acumulada

$F(x)$ = Prob (tiempo de vida del foco $\leq x$).

donde x = 112 hrs. 23 mn 14.03 s.

En resumen, con una variable aleatoria continua, el concepto de función de probabilidad tal como se definió más arriba no se aplica, pero la función de distribución acumulada existe por lo general. Y es a partir de la función de distribución acumulada $F(x)$ que se define la función de densidad $f(x)$; es una función que, por cada valor posible de la variable aleatoria, da la tasa (velocidad, densidad) a la cual aumenta la probabilidad acumulada en este punto de la función. Técnicamente, la función de densidad de probabilidad es una derivada (una pendiente) de la función de distribución acumulada de una variable continua:[100]

$$f(x) = \frac{d}{dx} F(x)$$

Con el fin de entender correctamente los tests estadísticos, es de suma importancia captar la diferencia entre una función

[100] Con relación en la función de distribución acumulada, la densidad juega el mismo papel que la velocidad con relación a la distancia recorrida: en un gráfico de la distancia recorrida en función del tiempo transcurrido, la pendiente de la curva da la velocidad en ese instante. La derivada es la velocidad instantánea, que es diferente de la velocidad promedio en un intervalo dado la cual corresponde en una gráfica de la distancia recorrida en la pendiente promedio en ese intervalo.

de densidad de probabilidad y una función de distribución de probabilidad, porque los tests son, de hecho, razonamientos sobre las probabilidades y se calculan estas probabilidades con la ayuda de funciones de densidad de probabilidad. Ahora bien, la ordenada de una función de densidad (su altura) no es una probabilidad (cuando la ordenada de una función de probabilidad sí es una probabilidad). Por lo contrario, la superficie debajo de una curva de una función de densidad es una probabilidad; técnicamente esto es cierto ya que si $f(x)$ es la derivada de $F(x)$ entonces se obtiene $F(x)$ con la integral de $f(x)$.

De esta manera,

$$\text{Prob}\left(a \leq \text{variable aleatoria} \leq b\right) = F(b) - F(a)$$

$$\text{Prob}\left(a \leq \text{var. aleatoria} \leq b\right) = \begin{pmatrix}\text{superficie debajo} \\ \text{de } f(x) \text{ entre } a \text{ y } b\end{pmatrix}$$

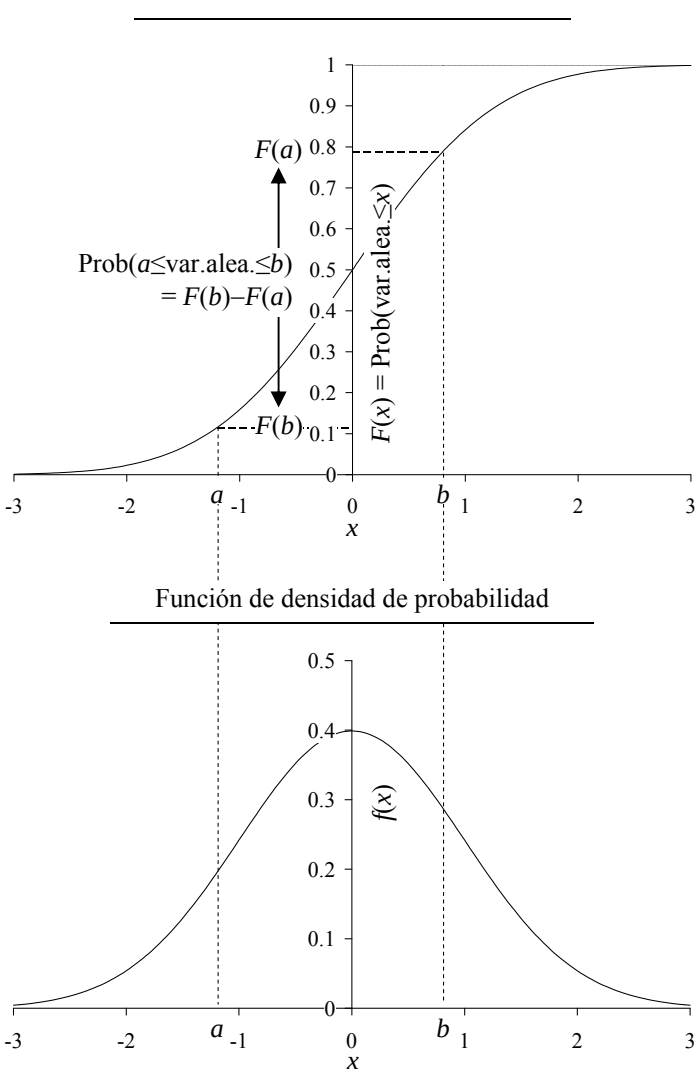
$$\text{Prob}\left(a \leq \text{variable aleatoria} \leq b\right) = \int_a^b f(x)\,dx$$

Naturalmente,

$$\int_{-\infty}^{+\infty} f(x)\,dx = 1 = F(+\infty)$$

Se ilustra en la figura 1 la relación entre la función de probabilidad acumulada y la función de densidad de probabilidad de una variable aleatoria continua.

Figura 1
Función de probabilidad acumulada y función de densidad



Esperanza matemática*

El promedio de una variable aleatoria continua en una población infinita no puede calcularse por medio de la famosa fórmula:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(donde x_i son los valores posibles de la variable aleatoria) simplemente porque el número de valores posibles n es infinito. El concepto de esperanza matemática es una generalización del promedio. Para una variable continua, la esperanza matemática es[101]

$$E(x) = \int_{-\infty}^{+\infty} f(x)\, x\, dx$$

Así, cuando hablamos del promedio de una variable aleatoria continua en una población, nos referimos a

$$\mu_x = E(x)$$

Y cuando hablamos de la varianza de una variable aleatoria continua en una población, nos referimos a:

$$\sigma_x^2 = E\left\{\left[x - E(x)\right]^2\right\} = \int_{-\infty}^{+\infty} f(x)\left[x - E(x)\right]^2\, dx$$

En el marco de este curso, sólo es necesario acordarse que las fórmulas de cálculo del promedio y de la varianza se pueden generalizar en el caso de una variable aleatoria continua. Por lo que resta, la intuición que se tenga del concepto del promedio y de la varianza a partir de las fórmulas de la estadística descriptiva es suficiente.

---

* Referencias: Wonnacott y Wonnacott (1992, pp. 154-155, 184-185).

[101] Si queremos comparar las dos fórmulas, se puede decir que $\int$ juega el papel de Σ y $f(x)$ el papel de $(1/n)$.

Ley normal*

La ley normal es un ejemplo de la distribución de probabilidad de una variable aleatoria continua. Es una distribución cuya función de densidad tiene la forma de una campana simétrica, como lo enseña la figura 2. Esta distribución es una buena aproximación de varias distribuciones de probabilidad observadas de manera empírica. Es, también, la distribución asintótica hacia la cual tienden muchas otras distribuciones (en el tema de distribución asintótica, vea 3.2).

Una de las características más importantes de las distribuciones normales es no tener más que dos parámetros: el promedio y la desviación estándar. Esto significa que, en caso de saber que una variable tiene una distribución normal y de conocer su promedio y su desviación estándar, se conoce entonces perfectamente su función de densidad de probabilidad.

Además, si la variable *x* tiene una distribución normal con un promedio μ_x y una desviación estándar σ_x, entonces la variable "estandarizada"

$$z = \frac{x - \mu_x}{\sigma_x}$$

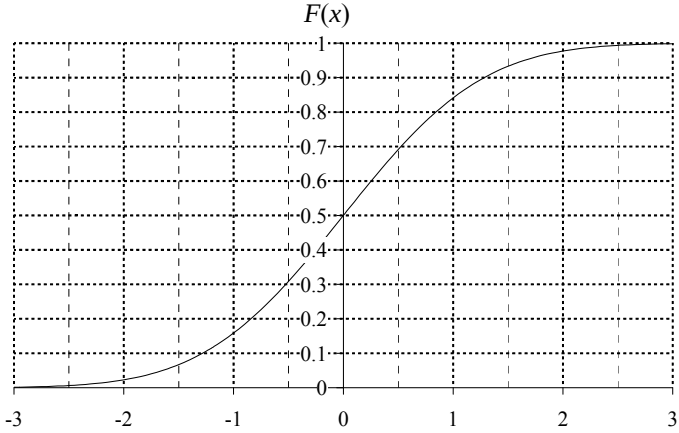
posee una distribución normal de promedio 0 y desviación estándar 1.[102]

---

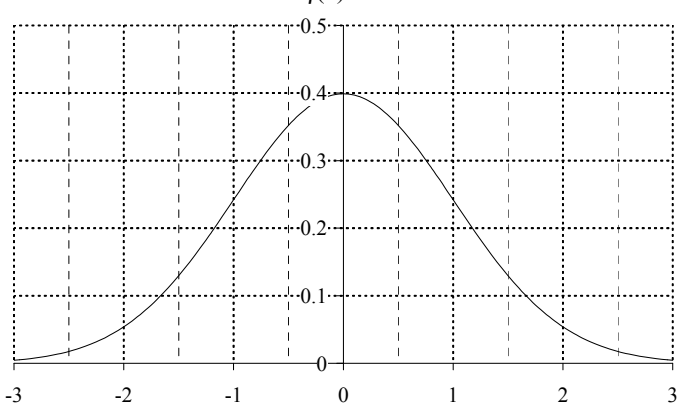
* Referencias: Wonnacott y Wonnacott (1992, pp. 142-148).

[102] Para diferentes valores de la variable normal estándar, las tablas estadísticas dan el valor de la densidad de probabilidad (ordenada de la función de densidad normal) y de la probabilidad acumulada (ordenada de la función de distribución acumulada normal). Estas tablas nos informan también de lo inverso, a saber, de el valor de la variable normal estándar que corresponde a diferentes probabilidades acumuladas. Se encuentra lo equivalente en el software Excel con las funciones *NORMSDIST* y *NORMSIMV*, las funciones que corresponden para las variables normales no estándares son *NORMLIST* y *NORMINV*.

Figura 2
Distribución normal
Función de distribución normal cumulativa



Función de densidad normal



2-2.3 MUESTREO, ESTIMACIÓN Y TESTS DE HIPÓTESIS

En grandes rasgos, el proceso de la inducción consiste en usar las estadísticas relativas a una muestra con el fin de aprender algo con relación a los parámetros de la población.

Este proceso conlleva tres tipos de preguntas:

- ¿Cuáles propiedades conviene que la muestra tenga? ¿Bajo cuáles condiciones es razonable considerar esta muestra como "representativa"?
- ¿Cuáles son las estadísticas de la muestra que se pueden usar para estimar el valor de los parámetros de la población? ¿Cuáles son las propiedades de estos estimadores?
- ¿Cómo se puede evaluar la fiabilidad de las estimaciones obtenidas? De manera más general, ¿qué se puede afirmar de la población y con qué grado de confianza?

Con relación a este tercer tipo de preguntas, los principios epistemológicos subyacentes a la inducción conducen a los tests de hipótesis.

Se puede, por lo tanto, dividir la inducción estadística en tres partes:

- La muestra.
- La estimación.
- Los tests de hipótesis.

2-2.3.1 Muestrario[*]

Un plan de muestreo debe contestar las preguntas siguientes:

1. ¿Cómo seleccionar la muestra de manera a respetar las condiciones que requiere el modelo de muestreo (*sampling model)* que corresponde al modelo de relación entre la muestra y la población? O, a la inversa, ¿cuál

[*] Referencias: Wonnacott y Wonnacott (1992, cap. 23).

modelo de muestreo refleja correctamente el modo de seleccionar la muestra?
2. ¿Cuál tamaño de muestra es necesario para obtener la precisión y el nivel de confianza deseados?

Antes de definir el plan de muestreo, hay que escoger la unidad de observación. Por ejemplo, en una encuesta a los consumidores, la unidad de observación puede ser la persona o el hogar. El conjunto de las unidades de observación constituye la población de la que se quiere sacar una muestra. Y la población también tiene que ser bien definida. Por ejemplo, en una encuesta a los hogares, por supuesto, hay que circunscribir el universo muestrario por límites geográficos u otros. Pero hay también que circunscribir el universo como concepto: por ejemplo, en una encuesta a los hogares, ¿se quiere tomar en cuenta los hogares de una persona sola? ¿Hogares "colectivos" (prisiones, conventos, cuarteles militares)?

Selección

Mencionemos los tres tipos más importantes de muestras.

La primera distinción que debemos operar en los métodos de selección es entre aquellas que conducen a una muestra aleatoria y las demás. Una muestra aleatoria es una muestra que se constituye por medio de un método de selección que permite conocer, para cada una de las muestras posibles, cuál es la probabilidad de que se seleccione. En la mayoría de los casos, esto equivale a conocer, para cada individuo, la probabilidad que se seleccione sabiendo que las probabilidades de los individuos son independientes entre sí (la probabilidad que se seleccione a un individuo no se ve afectada en caso de que se seleccione cualquier otro individuo).

Muestra aleatoria simple (sorteada): cada individuo de la población tiene igual probabilidad de ser seleccionado. Este método de muestreo exige, por lo general, que se haga un inventario previo de la población.

Un inventario incompleto puede crear un sesgo; por ejemplo, una selección aleatoria en el directorio telefónico descarta a priori los individuos que no tengan teléfonos o cuyo teléfono es confidencial. Se cometerá un error si los descartados son diferentes a los demás. Hay otras posibilidades de sesgo en la recolección de datos. Por ejemplo, los que la ciencia política suele llamar los "electores discretos" (que no contestan o "no saben") quizá tienen en promedio una opinión diferente de los que se expresan más espontáneamente.

Se hace la distinción entre el muestreo simple con o sin reemplazo: en el primer caso, se sortean sucesivamente los individuos miembros de la muestra y después de cada sorteo, el individuo seleccionado vuelve a ser elegible durante el sorteo que sigue (así, un individuo puede ser seleccionado más de una vez en la misma muestra); en el segundo caso, se quita el individuo que ha sido seleccionado de donde se sorteará los demás miembros de la muestra.

El muestreo sistemático es un método que se acerca del muestrario aleatorio simple.[103] Consiste en seleccionar un individuo a cada *n* vez (lo que supone que los casos ya sean ordenados: por ejemplo, el orden alfabético en el directorio telefónico o el orden de los números de dirección en las calles. Para sacar una muestra sistemática, se divide el tamaño de la población entre el tamaño de muestra que se quiere: eso es el intervalo muestral *n*. Después, basta sacar al azar el primer individuo y los demás siguen.

Muestra aleatoria estratificada: cuando la población se divide en varias subpoblaciones ("estratos") con parámetros que pueden ser diferentes, queremos que la muestra sea representativa no sólo de la población en general sino también de la subpoblación.

Esta representatividad no implica forzosamente que cada estrato de la muestra sea proporcional a la subpoblación que

[103] May (1993), p. 70.

representa. De hecho, si buscamos la misma precisión para todas los estratos de la población, es necesario que las subpoblaciones menos numerosas tengan una mayor representatividad. Eso porque la precisión en la estimación de los parámetros no es proporcional al tamaño de la muestra. Retomaremos este tema (vea capítulo 2-3, al final del apartado 2-3.5).

Mientras que la muestra aleatoria simple exige un inventario de la población, la muestra aleatoria estratificada necesita un inventario por estrato. Eso no es siempre disponible. A menudo se trata de aproximar la muestra aleatoria estratificada gracias al muestreo por cuotas.[104] Este método consiste en clasificar a los individuos al momento que son seleccionados, hasta que se cumpla el número esperado (cuota) en cada estrato. En una encuesta por cuestionario basada en el muestrario por cuotas, la primera parte del cuestionario sirve para clasificar a los individuos; no se completa el cuestionario con los individuos supernumerarios.

Muestra en racimos (cluster): este método consiste en dividir la población en grupos (racimos) para luego sortear un cierto número de éstos; los miembros de estos racimos seleccionados constituyen la muestra. Recurrimos a menudo a este método cuando no se tiene un inventario previo de la población.

Por ejemplo, para efectuar una encuesta en los hogares de una zona habitacional informal (donde los datos del censo no son fiables), se puede subdividir la zona en cuadras de casas, luego seleccionar una cierta proporción éstas y proceder a la entrevista en todos los hogares en el interior de las cuadras seleccionadas. La inducción estadística es más difícil con una muestra en racimos, porque las distribuciones muestrales de las estadísticas son más complejas.

[104] May (1993), p. 71.

Por supuesto, hay métodos de muestreos no aleatorios que, sin embargo, son adecuados en contextos no estadísticos. Así, encuestas de tipo cualitativo se basan a veces en el método "bola de nieve" o en el método de saturación. Pero esos métodos muestrarios no son pertinentes aquí.

Tamaño

Por lo general, más grande es la muestra, más probabilidad tiene de ser representativa y más alto es el grado de precisión de la estimación para un mismo grado de confianza. Sin embargo, el grado de precisión no es directamente proporcional al tamaño de la muestra (examinaremos este fenómeno con más precisión al momento de estudiar un test de hipótesis sobre el promedio; vea 2-3.5). Según los análisis que queremos efectuar, existen reglas que permiten determinar el tamaño de la muestra requerida para alcanzar la precisión y la confianza deseadas. Pero los costos de la recolección crecen con el tamaño.

2-2.3.2 Estimación

- ¿Cuáles son las estadísticas de la muestra que se pueden usar para estimar el valor de los parámetros de la población? ¿Cuáles son las propiedades de estos estimadores?

En este momento, hacemos una distinción entre un estimador que es una fórmula, un método de cálculo, y una estimación o valor estimado como el resultado de la aplicación de esta fórmula. Un estimador es una variable aleatoria puesto que el mismo estimador que se aplica a datos de muestras diferentes arroja, por lo general, valores estimados diferentes.

Métodos*

Existen tres enfoques:

1. Analógico.
2. Menores cuadrados.
3. Máxima verosimilitud (Theil, 1971, p. 89, Freund, 1962, p. 223).

1. Estimación según el enfoque analógico. El principio de estimación analógica (conocido también como método de los momentos) es simple: para estimar un parámetro, se aplica a la muestra la misma fórmula matemática que a la población. Ejemplo:

> Para estimar el valor promedio μ_x de una variable *x* en una población, se calcula el promedio de la misma variable en la muestra.
>
> $$m_x = \frac{1}{n}\sum_i x_i$$

Este procedimiento es totalmente mecánico. Sin embargo, en general un estimador es la expresión matemática de un principio de selección del "mejor" valor como estimación del parámetro. Diferentes principios de selección conducen a diferentes estimadores, sabiendo que los principios que más se emplean son el principio de mínimos cuadrados y el principio de máxima verosimilitud.

Se podría comparar la estimación con el hecho de sintonizar una radio: se prueba diferentes frecuencias hasta optimizar la recepción de la señal para que, al final, el valor seleccionado en el receptor sea un valor estimado del parámetro que se busca, o sea la frecuencia de emisión. La frecuencia seleccionada dependerá del criterio de selección usado (supongamos que se usa un solo criterio a la vez para

* Referencia: Wonnacott y Wonnacott (1992, cap. 18).

lograr la comparación): fuerza de la señal, ausencia de ruidos y de distorsión, ausencia de parásitos...[105]

2. Principio de los menores cuadrados. El principio de menores cuadrados puede aplicarse sin necesidad de modelo aleatorio. Consiste en "sintonizar" los valores estimados de los parámetros del modelo de tal manera que, cuando se aplica este modelo a la muestra, sus errores de predicción sean tan pequeños como se pueda. La expresión "menores cuadrados" se refiere a la medición de errores como la suma de los cuadrados de los errores de predicción, sabiendo que se mide cada error con la diferencia entre un valor observado y el valor predicho correspondiente. Esta medición de errores es, por lo tanto, el cuadrado de la distancia euclidiana generalizada entre la serie de las observaciones y la serie de las predicciones.

3. Principio del máximo de verosimilitud. La aplicación del principio del máximo de verosimilitud se refiere directamente al modelo aleatorio seleccionado para representar la relación aleatoria entre la muestra y la población o para representar el carácter aleatorio del fenómeno estudiado. Por consiguiente y contrario al principio de los menores cuadrados, el principio del máximo de verosimilitud no puede aplicarse sin modelo aleatorio. El principio del máximo de verosimilitud consiste en "sintonizar" los valores estimados de los parámetros del modelo de manera que, suponiendo que estos valores fueran los buenos, la muestra sea la más "verosímil" posible. Se mide la verosimilitud con la función de verosimilitud, la cual es

[105] Hay que tener cuidado con seguir demasiado lejos esta analogía puesto que cuando sintonizar un radio cuya frecuencia no se conoce se hace por lo general a tientas, la aplicación de uno u otro de los principios de estimación conduce, con frecuencia, a una fórmula que permite calcular directamente el valor estimado correspondiente.

la función de densidad de probabilidad de la muestra tomando los valores de los parámetros.

Cuando maximizamos la función de verosimilitud, los papeles de los valores observados de la muestra y de los parámetros en la función de densidad de probabilidad son invertidos: en lugar de considerar los valores observados como variables aleatorias cuya función de densidad de probabilidad depende del valor de los parámetros, son, al contrario, los valores observados que se consideran como fijos y se hace variar los valores estimados de los parámetros de tal manera que la verosimilitud alcance su máximo. Los valores seleccionados como valores estimados de los parámetros son, por lo tanto, los valores cuando la densidad de probabilidad de la muestra es la más grande (el modo de la distribución) y, por consiguiente, los intervalos alrededor de este punto tienen la probabilidad más alta.

En ciertas condiciones, los principios de los menores cuadrados y del máximo de verosimilitud conducen al mismo estimador. En algunos casos (como la estimación del promedio), este estimador es al mismo tiempo el estimador del proceso analógico.

Propiedades deseables[*]

Los estimadores son variables aleatorias de modo que sus propiedades son las propiedades de su distribución de muestreo.

1. Ausencia de sesgo. Entre las propiedades deseables de un estimador, la ausencia de sesgo es de suma importancia. Un estimador no sesgado es un estimador cuyo valor será en promedio igual al valor del parámetro estimado. La expresión

[*] Referencias: Wonnacott y Wonnacott (1992, pp. 262-266, 275-276); Freund (1962, p. 215-220).

"en promedio" implica, en este momento, examinar la distribución de muestreo del estimador.

Por ejemplo, si queremos estimar la varianza de una variable en la población por medio de los datos de la muestra, y si aplicamos la fórmula del método analógico

$$\frac{1}{n}\sum_i (x_i - m_x)^2$$

podemos demostrar que obtenemos un estimador sesgado: en caso de repetir el cálculo con muy grande número de muestras (un infinidad), el resultado sería, en promedio, diferente de la verdadera varianza. Es la razón por la cual usamos de preferencia un estimador corregido con tal de eliminar el sesgo; la fórmula de este estimador no segado es

$$s_x^2 = \frac{1}{n-1}\sum_i (x_i - m_x)^2$$

De la misma manera, el estimador de la covarianza mediante la fórmula del método analógico

$$\frac{1}{n}\sum_i (x_i - m_x)(y_i - m_y)$$

da un estimador sesgado de la covarianza entre x e y en la población, mientras que

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)$$

es un estimador no sesgado.

2. Eficacia relativa: los estimadores "best unbiased". En el universo de las muestras posibles, los resultados que cualquier estimador no sesgado arroja y apuntan en promedio hacia el objetivo que constituye el valor del parámetro que se pretende estimar. ¿Cómo escoger, en estas condiciones, entre dos estimadores no sesgados? Es obvio que se escogerá el estimador que apunta más al centro del objetivo y se evitará el estimador que arroja resultados más dispersos.

Es justamente la varianza que mide la dispersión de una variable aleatoria alrededor de su promedio. Llamamos "varianza de muestreo" la varianza de un estimador en la población de las muestras posibles (es la varianza de la distribución de muestreo); la raíz cuadrada de la varianza de muestreo constituye el "error de muestreo" (sampling error). Se dice que un estimador no sesgado es más eficaz que otro cuando su varianza de muestreo es inferior a la varianza del otro.

Llamamos "best unbiased" un estimador no sesgado cuya eficacia relativa es superior a la eficacia de cualquier otro estimador no sesgado. Esta misma apelación se usa de manera más restrictiva para una clase dada de estimadores. Por ejemplo, en la clase de estimadores cuyo valor es una función lineal de los datos, el estimador que detenta la mejor eficacia relativa es conocido como el "Best Linear Unbiased Estimate" o "BLUE".

3. Convergencia. Otra propiedad deseable de un estimador es que su precisión sea superior en cuanto la muestra sea de tamaño más grande o, dicho de otra manera, que su varianza de muestreo sea más pequeña cuando la muestra es más grande. Se dice que un estimador es convergente si su varianza de muestreo tiende hacia cero cuando el tamaño de la muestra tiende hacia el infinito (la distribución de muestreo tiende a concentrarse en un solo punto).

4. Suficiencia. Finalmente, un estimador es suficiente cuando incorpora toda la información contenida en la muestra con relación al parámetro que se pretende estimar; en cuanto se calculó el valor del estimador (a partir de los datos de la muestra), no se podrá aprender algo más sobre el valor del parámetro aunque se examinen nuevamente los datos de la muestra.

Técnicamente, esta propiedad se traduce de la manera siguiente: si un estimador es suficiente entonces la probabilidad de la muestra (su verosimilitud) dada el valor estimado es independiente del valor del parámetro.

2-2.3.3 La lógica fundamental de las pruebas de hipótesis[*]

Volvamos a examinar el esquema del método científico estudiado en el apartado 2-2.1. La lógica fundamental de este proceso es el siguiente:

- Si una teoría (o un modelo o una hipótesis) es verdadera, entonces sus implicaciones son también verdaderas.
- Por lo tanto, si las observaciones contradicen las implicaciones de una teoría, esta teoría no es verdadera; es falsa.

Con este razonamiento, se pretende aclarar algo de suma importancia: ¡si las observaciones no contradicen las implicaciones de una teoría, este hecho no nos da el derecho de concluir que esta teoría es verdadera! Con más precisión, para poder concluir que esta teoría es verdadera, es necesario que no exista otra teoría posible que sea compatible con las observaciones. En la práctica, esta condición es tan exigente que nunca se cumple.

En resumen, al momento de confrontar las implicaciones de una teoría con las observaciones, se rechaza la teoría cuando las observaciones contradicen las implicaciones; en el caso contrario, cuando las observaciones no contradicen las implicaciones, no se puede todavía confirmar la teoría, sólo queda en la categoría de "no rechazada".[106]

---

[*] Referencias: Blalock (1972), cap. 8, "The fallacy of affirming the consequent".

[106] Personalmente prefiero la expresión "no rechazada" en lugar de la palabra "aceptable" que usa Wonnacott y Wonnacott (1992) por el riesgo de pasar de la categoría "aceptable" a la categoría "aceptada", que no es lo mismo. Se reconocerá, aquí un parentesco con el falsificasionismo poppe-

Es esta misma lógica de rechazo/no rechazo que prevalece en los tests de hipótesis. Sin embargo, existe una diferencia capital: en los tests de hipótesis, la relación entre las hipótesis y la muestra observada es aleatoria, lo que implica que el razonamiento ya no puede ser determinista sino más bien probabilista.

En una lógica determinista, una observación es compatible con la hipótesis o no lo es, es decir, no existe un punto intermedio. En la lógica probabilista, una observación es más o menos compatible con la hipótesis: cuanto más improbable una observación mientras que se supone la hipótesis verdadera, menos compatible con la hipótesis.

Para ilustrar esto, enseñaremos un ejemplo un tanto caricatural:

> Consideremos la hipótesis de que el dromedario no es parte de la fauna salvaje del continente australiano. Supongamos que un viajero, con gran estupefacción, encuentre un dromedario sin amo en el desierto australiano. Esta observación contradice su hipótesis. Sin embargo, este dromedario pudiera haberse escapado de un circo o de un zoológico o bien pudiera ser un espejismo. La observación de un dromedario no es imposible, más bien es improbable: la observación de un solo dromedario o hasta de algunos no podría considerarse como incompatible con la hipótesis. Ahora bien, supongamos que el mismo viajero ubique dromedarios en diferentes momentos. Si la hipótesis fuese verdadera, estas repetidas observaciones serían extraordinariamente improbables. Al paso de tiempo, el observador acabará por concluir que estas observaciones no son compatibles con su hipótesis.[107]

---

riano; para Popper, una hipótesis que no es posible rechazar lógica o empíricamente no es "científica".

[107] Los dromedarios salvajes son parte de la fauna de los desiertos australianos desde que fueron abandonados por caravaneros afganos, quienes los

En este ejemplo, nuestro viajero se contentará, de seguro, con un enfoque intuitivo. En caso de los tests de hipótesis, está claro que se formaliza mucho más el proceso; en particular,

- Se debe cuantificar las probabilidades de las cuales trata el razonamiento. (Si bien se sabe que el dromedario no es parte de la fauna australiana, ¿cuál es la probabilidad exacta de, no obstante, encontrar un dromedario? ¿De encontrar dos? ¿Tres?);
- Se debe tomar la decisión de rechazar la hipótesis bajo criterios precisos, definido de antemano (¿cuál es la probabilidad arriba de la cual decidiremos que las observaciones son incompatibles con la hipótesis?)

Es la primera de estas dos exigencias la que, por mucho, causa grandes dificultades tanto conceptuales como prácticas. En cuanto a la segunda, veremos que es, ni más ni menos, una exigencia de transparencia.

importaron para asegurar las comunicaciones trans-continentales antes la construcción del ferrocarril.

CAPÍTULO 2-3
LAS PRUEBAS DE HIPÓTESIS*

2-3.1 INTRODUCCIÓN A LAS PRUEBAS DE HIPÓTESIS

Existen varios elementos que forman parte de la formulación de un test de hipótesis:

- La población (con más exactitud, los parámetros de la población), la cual corresponde a la realidad desconocida a propósito de la cual queremos probar una hipótesis...
- La hipótesis, la cual es un enunciado con relación a esta población (con más exactitud, con relación a uno o varios parámetros de esta población) y de la cual no se sabe si es verdadera o falsa.
- La muestra, es decir el conjunto de las observaciones obtenidas de la población a partir de las cuales buscaremos decidir[108] si consideramos la hipótesis como verdadera o falsa.

---

* Referencias: Wonnacott y Wonnacott (1992, cap. 9) presentan los tests de hipótesis una vez que presentaron la estimación por intervalos (intervalos de confianza). En estas condiciones, no es posible establecer un paralelismo perfecto entre el manual de estos autores y el presente documento.

[108] Observe que decimos claramente "decidir" y no "determinar". De hecho, para "determinar", necesitaríamos llegar a una certeza. Por lo con-

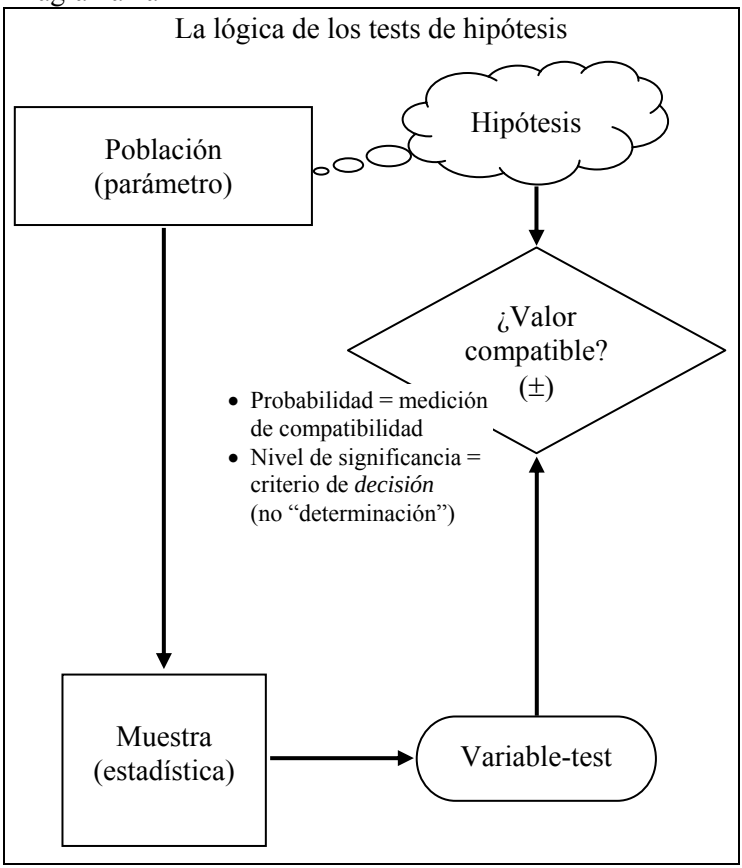
- La variable-test, la cual es una estadística de la muestra que se usará para decidir si consideramos la hipótesis como verdadera o falsa.
- La probabilidad, la cual, en este contexto, es una medición inversa[109] del grado de incompatibilidad del valor observado de la variable-test con la hipótesis.
- El nivel de significancia, el cual es el umbral de probabilidad crítica abajo del cual se decide que se juzgarán las observaciones (en sus formas resumidas en la variable-tests) lo suficientemente improbables como para ser incompatibles con la hipótesis.

Estos elementos así como las relaciones que los unen se representan en el diagrama 2a.

trario, podríamos, en dado caso, "decidir" rechazar una hipótesis sin estar seguro de no cometer un error.

[109] ¡Cuidado con la doble negación! Cuando más grande está la probabilidad, más el valor observado es compatible con la hipótesis, por consiguiente, cuando más pequeña está la probabilidad, más el valor observado tiende a ser incompatible con la hipótesis. La probabilidad es efectivamente una medición inversa de la incompatibilidad.

Diagrama 2a



En caso de pretender formalizar el ejemplo de los dromedarios australianos, podemos decir que:

- la población estudiada es la fauna australiana salvaje;
- la hipótesis para probar es que el número de dromedarios en esta fauna es nulo;
- la muestra se constituye del conjunto de animales observados hasta el momento del test;

- la estadística usada es el número de dromedarios observados.

Pero dejemos de un lado este ejemplo, pues desarrollarlo más no sería muy congruente puesto que lo que hace falta en este ejemplo es la posibilidad de medir el grado de incompatibilidad entre el valor observado de la variable-test y la hipótesis. De manera concreta, en el caso de los dromedarios, es imposible construir un enunciado del tipo "Ubiqué *X* dromedarios hasta el momento. Supongo que el dromedario no es parte de la fauna salvaje australiana (o sea que supongo que los que vi eran animales escapados de un zoológico o de un circo). Si mi suposición es cierta, la probabilidad de observar *X* dromedarios escapados son de *Y* en un millón"... En un test de hipótesis, es necesario poder cuantificar este *Y*.

Durante la formulación de un test de hipótesis, el meollo del problema es la selección de una variable-test. Es tan cierto que varias de las variables-test que se usan con frecuencia llevan el nombre de su inventor (Student,[110] Fisher, Durbin-Watson,...). Una variable-test debe poseer varias propiedades indispensables:

1. El valor de una variable-test depende, al mismo tiempo, de los datos de la muestra y de la hipótesis que se quiere probar.

   En efecto, la variable-test constituye el enlace entre el modelo (la hipótesis que se quiere probar) y las observaciones. Mide, en cierto modo, la distancia o la disimilitud entre las observaciones y las predicciones del modelo o de la hipótesis.[111] La variable-test no tendría ninguna utilidad si no fuera que incorpora la información contenida en la muestra, es decir si su valor fuera

[110] Student (estudiante) es seudónimo matemático W.S. Gossett (1876-1937).
http://www-gap.dcs.st-and.ac.uk/~history/Mathematicians/Gosset.html

[111] Es evidente con el Chi-cuadrado de Pearson que se utiliza para el test de independencia en los cuadros de contingencia. Vea 4-1.

independiente de las observaciones. No sería tampoco de gran utilidad si su valor fuera lo mismo no importando la hipótesis específica que se quisiera probar. Más bien, la variable-test debe permitir diferenciar entre las hipótesis que decidimos rechazar y las que decidimos no rechazar.

2. Se debe poder calcular el valor de la variable-test, es decir que no debe depender de valores desconocidos sino, más bien, únicamente de las observaciones y de los datos de la hipótesis.
3. La variable-test debe tener una distribución de muestreo cuya forma generalizada es conocida y cuya forma específica depende del propio contenido de la hipótesis que se quiere probar.

Veremos más tarde con un ejemplo el significado preciso y concreto de esta tercera propiedad. Mientras, examinemos un poco el aspecto aleatorio de la variable-test. De hecho, para tener una distribución de probabilidad (su distribución de muestreo), es necesario que la variable-test sea un variable aleatoria. En cambio, acabado el sorteo de la muestra, los valores observados ya no son aleatorios sino, más bien, fijos (así como el número de dromedarios detectados, una vez que hayan sido contados). La contradicción existe solamente en las apariencias en cuanto recordemos la distinción entre una variable aleatoria y los valores que puede tomar. En efecto, la muestra sorteada es sólo una de las muestras posibles. A cada una de ellas corresponde un valor de la variable-test (es poco probable que otros viajeros o el mismo viajero en otro momento hubieran visto el mismo número de dromedarios). Antes de sortear la muestra, existía por lo tanto una multitud (y en algunos casos, una infinidad) de valores posibles de la variable-test. En otras palabras, imaginando que nos encontramos justo antes del sorteo, entonces la variable-test es, por lo tanto, claramente una variable aleatoria a la cual se asocia

una distribución de probabilidad (la distribución de muestreo).

Este concepto no es tan exótico como parece y para demostrarlo, los lingüistas gustan citar este ejemplo de dos títulos de periódico:

Hombre mordido por un perro

y

Perro mordido por un hombre

En los dos casos se emplean las mismas palabras; sólo se modificó un tanto su orden. Entonces, ¿por qué el segundo título es digno de la portada de la sección de policía de *La Prensa*[112] y no el primero? Claro está que el segundo relata un evento sorprendente, sorprendente porque su probabilidad ex ante era muy pequeña.

De la misma manera, consideramos a una persona que acaba de ganar la lotería como una persona con mucha suerte sólo porque, ex ante, la probabilidad de que fuera ella era muy pequeña.

Resumiendo, la distinción entre el valor observado y su distribución ex ante es análoga a la distinción entre lo que efectivamente aconteció y lo que esperábamos. Siendo poetas, podríamos decir que el evento que se realiza no borra el recuerdo de lo que se esperaba de él sino todo lo contrario, la sorpresa nace del choque entre los dos.

Para efectuar un test de hipótesis es necesario, por lo tanto, poder medir la sorpresa, es decir, determinar, suponiendo que la hipótesis que queremos probar sea verdadera, cuál era la probabilidad de observar lo que observamos antes de observarlo (lo cual se resume con la variable-test). Para poder determinar esta probabilidad, se tiene que definir un modelo de muestreo, es decir un modelo de la relación entre la pobla-

[112] Diario de circulación nacional en México.

ción y la muestra. Esto implica que la selección de una variable-test y la especificación del modelo de muestreo van de la mano. El modelo de muestreo contiene, usualmente dos elementos:

- Una hipótesis en cuanto a la forma general de las leyes de probabilidad que rigen el fenómeno estudiado en la población.
- La especificación del proceso de muestreo.

Una vez que se determinó la probabilidad de lo que se observó, siempre y cuando se supone que la hipótesis es verdadera, sólo falta decidir si el resultado conduce o no al rechazo de la hipótesis. Para determinar esto se compara esta probabilidad con el umbral de probabilidad crítica, escogido previamente, abajo del cual pensamos que las observaciones son lo suficientemente improbables para ser incompatibles con la hipótesis. Este umbral crítico se llama nivel de significancia porque es el nivel de probabilidad debajo del cual se decide considerar el desacuerdo entre las observaciones y la hipótesis como estadísticamente significativo.

A fin de quedar con las ideas bien claras, se formaliza el argumento lógico del test de hipótesis clásico en el cuadro que sigue. En este enunciado, los términos en cursivas y entre corchetes son las "variables" del argumento. Para aplicar el argumento a un caso particular, se reemplaza estas variables con los datos pertinentes del caso particular. Por consiguiente, se presenta un poco el argumento como una fórmula matemática donde calculamos el resultado con reemplazar las variables por su valor. Una tabla subsiguiente presentará el "valor" que debe tomar cada "variable" para aplicar el argumento al test de una hipótesis simple sobre un promedio.

Argumento del test de hipótesis clásico

1. Modelo de muestreo, hipótesis y implicaciones (silogismo)

   Si es cierto {*modelo de muestreo*},
   entonces {*variable*}[113] tiene la distribución {*distribución de muestreo*}.

   Ahora bien

   si es cierta {*hipótesis*},
   entonces {*variable*} es igual a la estadística {*variable-test*}.

   Por lo tanto

   si son ciertos a la vez {*modelo de muestreo*} y {*hipótesis*},
   entonces {*variable-test*} tiene la distribución {*distribución de muestreo*}.

2. Regla de decisión: definición de la zona de rechazo

   Se rechazará {*hipótesis*} si el valor observado de {*variable-test*} pertenece a un conjunto de valores extremas cuya probabilidad es inferior o igual a {*nivel de significancia*}.[114]

   Teniendo

   - la distribución {*distribución de muestreo*},
   - la orientación del test (bilateral o unilateral, a la derecha o a la izquierda, dependiendo de la

[113] Esta variable, ni es una estadística, ni es un parámetro. No es un estadística porque su valor depende de parámetros pero, tampoco es un parámetro ya que su valor depende también de una estadística.

[114] Sería más sencillo hablar en términos de la probabilidad del valor observado. Sin embargo, se trata de una variable aleatoria continua lo que implica que la probabilidad de un valor específico es infinitamente pequeña. Ésta es la razón por la cual se razona en términos de un conjunto de valores extremos que se define con uno o dos valores críticos (dependiendo si se hace un test unilateral o bilateral).

hipótesis complementaria H_A),

el conjunto de valores extremos que tiene una probabilidad igual al {*nivel de significancia*} se define con {*zona de rechazo*}.

3. Decisión
   Ahora bien, el valor observado de {*variable-test*} {*forma* / *no forma*} parte del conjunto de valores extremos que se define con {*zona de rechazo*}.
   Por lo tanto, la regla de decisión seleccionada lleva a {*rechazar* / *no rechazar*} {*hipótesis*}.

2-3.2 CASO MODELO: UN TEST DE HIPÓTESIS SIMPLE SOBRE UN PROMEDIO

Ahora que expusimos la lógica fundamental de los tests de hipótesis, examinemos cuáles son las etapas a seguir para aplicar esta lógica. Lo haremos basándonos en un caso modelo: un test de hipótesis simple sobre un promedio.

Ejemplo:

queremos estudiar el tiempo que pasan los habitantes de la Isla de Montreal en escuchar la radio. El indicador posible (la variable *x*) podría ser el número de minutos durante los cuales un individuo escuchó la radio el miércoles 23 de septiembre 1998. El promedio desconocido μ_x podría ser el número promedio de minutos de audiencia radiofónica de los habitantes de la Isla de Montreal ese día.[115] En

[115] Observe que el promedio buscado podría ser, por ejemplo, el número promedio de audiencia radio durante un miércoles cualquiera del periodo del primero de septiembre al 30 de octubre de 1998. Este promedio diferente se refiere, también, a una población diferente. Sin embargo, si todas las observaciones se efectuaron el miércoles 23 de septiembre, la muestra

cuanto a la muestra, supondremos que encierra 25 observaciones y que el tiempo promedio de audiencia de la muestra es igual a 110 minutos, con una desviación estándar $s_x = 20$. La hipótesis que se quiere probar podría ser que, en promedio, los habitantes de Montreal escucharon la radio este día durante cien minutos:[116]

$H_0 : \mu_x = 100$

De manera más general, queremos estudiar en una población dada, una característica que se representa con la variable *x*. Nos interesa μ_x, el promedio de *x* en la población. Este promedio μ_x es desconocido. Sin embargo, disponemos de una muestra obtenida en la población y podimos calcular m_x, el promedio de *x* en la muestra. Se trata, ahora, de probar la hipótesis H_0 que el "verdadero" valor del promedio es igual a un valor especifico dado el cual se reconocerá con la letra griega gamma: γ.

$H_0 : \mu_x = \gamma$

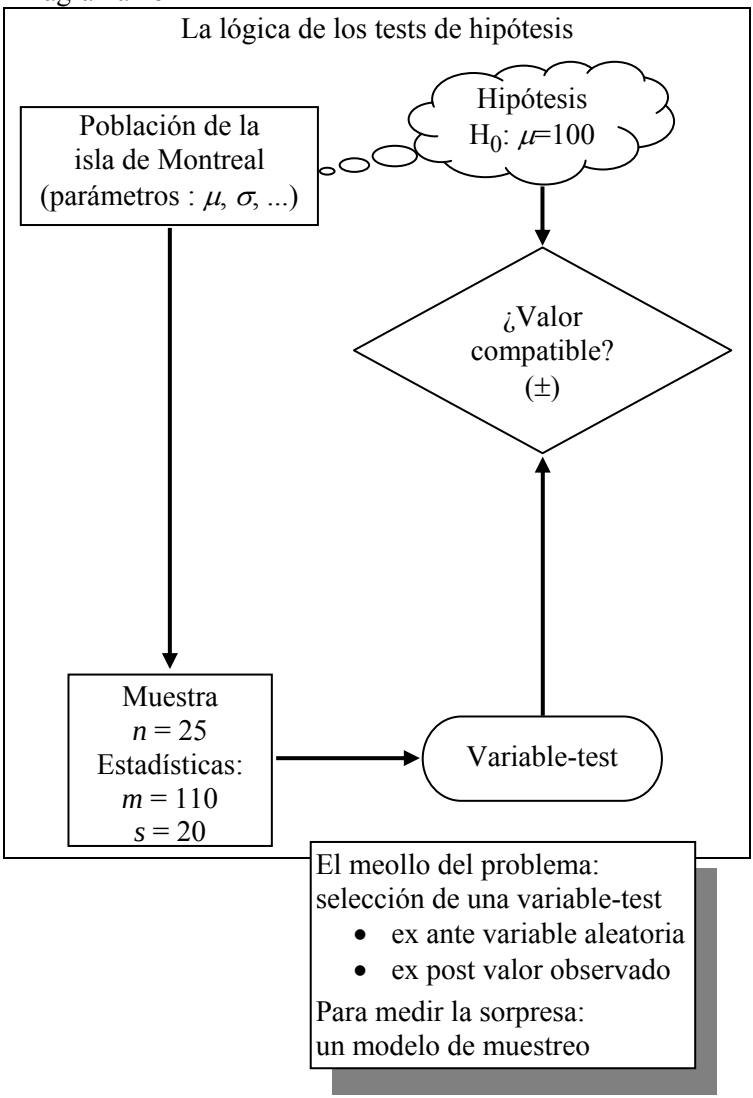
Para complementar la presentación, el diagrama 2b, como anexo, es una copia del diagrama 2a pero con los datos de nuestro ejemplo de test de hipótesis simple sobre un promedio.

---

corre el riesgo de no ser representativa de esta población más amplia, al menos que se crea que los comportamientos son similares durante todos los miércoles del periodo seleccionado (aunque hubo mal tiempo el miércoles, pero no así el 16...).

[116] Está claro que nadie nos impide testar la hipótesis que $\mu_x = m_x = 110$. Sin embargo, esta hipótesis específica no es más que una entre una infinidad de posibilidad.

Diagrama 2b


La lógica de los tests de hipótesis
Población de la isla de Montreal (parámetros : μ, σ, ...)
Hipótesis H_0: μ=100
¿Valor compatible? (±)
Muestra n = 25 Estadísticas: m = 110 s = 20
Variable-test
El meollo del problema: selección de una variable-test
• ex ante variable aleatoria
• ex post valor observado
Para medir la sorpresa: un modelo de muestreo

Las diferentes etapas a seguir para probar una hipótesis son las siguientes:

1. Seleccionar una variable-test;
2. Verificar que el modelo de muestreo asociado a esta variable es aceptable;
3. calcular el valor de la variable-test;
4. seleccionar un nivel de significancia;
5. detectar los valores críticos de la variable-test (zona de rechazo);
6. comparar el valor de la variable-test con los valores críticos y tomar la decisión de rechazar o no la hipótesis

Veamos ahora con más detalle en qué consiste cada una de estas etapas en nuestro ejemplo.

2-3.2.1 Primera etapa: selección de la variable-test

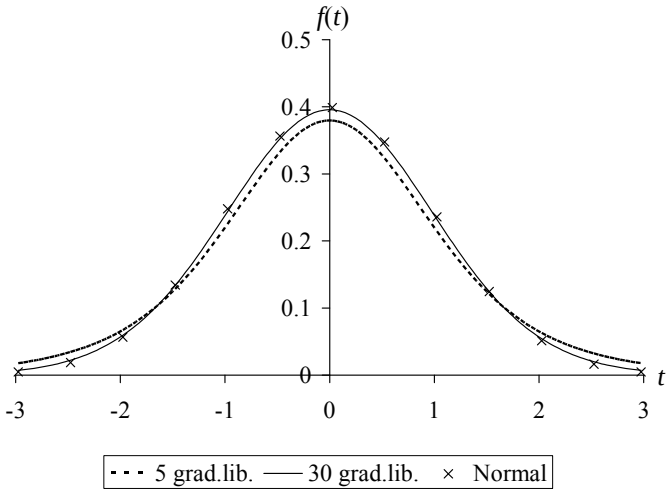
Por el tipo de usuarios que somos, no podemos pensar en inventar por completo una variable-test. Más bien, se trata de seleccionar una entre las que pone a nuestra disposición la estadística. En este caso particular, se aplicará el test de Student, el cual usa la variable-test que sigue:

$$t_{n-1} = \frac{m_x - \gamma}{\left(s_x / \sqrt{n} \right)}$$

La selección de esta variable se justifica puesto que, bajo ciertas condiciones (las examinaremos más tarde), la variable $\frac{m_x - \mu_x}{\left(s_x / \sqrt{n} \right)}$ posee una distribución conocida la cual designamos con el nombre de distribución de Student. La distribución de Student se parece a la distribución normal pero su forma cambia un poco con el valor de *n* tal como lo ilustra la

figura 3: Se dice que esta variable posee una distribución de Student con $n - 1$ grados de libertad.[117].

Figura 3
Comparación entre la distribución de Student y la normal según el número de grados de libertad
Función de densidad



Se calcula, por consiguiente, el valor de la variable-test con simplemente sustituir el valor γ con μ_x en la fórmula anterior. De esta manera se puede afirmar que *si la hipótesis* H_0 *es verdadera*, entonces $\gamma = \mu_x$ y la variable-test posee, una distribución de Student con $n - 1$ grados de libertad:

[117] No obstante, para valores de n superiores a 30, la distribución de Student se asemeja lo suficiente a la normal para considerar, con frecuencia, que la variable posee una distribución aproximadamente normal.

Si H_0 es verdadera, entonces $t_{n-1} = \dfrac{m_x - \gamma}{\left(s_x / \sqrt{n}\right)} = \dfrac{m_x - \mu_x}{\left(s_x / \sqrt{n}\right)}$

Es importante observar que, en fórmula de cálculo de la variable-test t_{n-1}, la desviación estándar que se emplea es efectivamente la desviación estándar de la muestra

$$s_x^2 = \frac{1}{n-1}\sum_i \left(x_i - m_x\right)^2$$

y no la desviación estándar de la población

$$\sigma_x^2 = \frac{1}{n}\sum_i \left(x_i - m_x\right)^2$$

Es posible verificar que esta variable-test posee las cualidades de requisito. Para empezar, su valor depende al mismo tiempo de los datos de la muestra (m_x, s_x, y n) y de la hipótesis que se quiere probar (γ). Luego, este valor no es una incógnita puesto que se puede calcular. Finalmente, esta variable-test posee una distribución de muestreo cuya forma general es conocida (distribución de Student) y cuya formar particular depende de que trata la hipótesis (el promedio μ_x).

2-3.2.2 Segunda etapa: ¿Es aceptable el modelo de muestreo?

Escogimos el test de Student porque, bajo ciertas condiciones, la variable

$$\frac{m_x - \mu_x}{\left(s_x / \sqrt{n}\right)}$$

posee una distribución de Student. ¿Cuáles son, pues, estas condiciones?

Las condiciones que siguen son suficientes[118]

- En la población, la variable *x* posee una distribución (aproximadamente[119]) normal con un promedio μ_x y una diferencia tipo σ_x desconocido.
- La población es de gran tamaño y en ella se sorteó una muestra aleatoria simple de tamaño *n*.

Estas condiciones constituyen un modelo de muestreo que especifica la forma general de la distribución de la probabilidad de *x* en la población y el tipo de muestreo. En cuanto a la distribución de probabilidad de *x*, puede ser un hecho declarado o una hipótesis dependiendo del contexto. En cuanto al tipo de muestreo, se tomó claramente la decisión al momento de la constitución de la muestra: en una muestra aleatoria simple, cada individuo tenía la misma probabilidad de formar parte de la muestra.

Es responsabilidad del investigador decidir si las condiciones que constituye el modelo de muestreo son aceptables. No se aplica el test sobre el modelo de muestreo, por lo tanto no se cuestionará más en el marco de este test.[120] El test se aplica únicamente sobre la hipótesis H_0.

---

[118] Estas condiciones son suficientes pero no necesarias. En caso que estas condiciones se realicen y que H_0 sea verdadera, entonces la variable-test t_{n-1} tendrá una distribución de Student. Sin embargo, existen otros grupos de condiciones bajo las cuales la variable-test t_{n-1} tendrá también una distribución de Student.

[119] Es imposible que la variable tenga una distribución exactamente normal puesto que no puede tomar valores negativos cuando una variable normal sí puede.

[120] Es cierto que existen tests de "nivel superior", para nombrarlos de alguna manera, que se aplican sobre algunos aspectos del modelo de muestreo. Sin embargo, estos mismos tests se basan en modelos aleatorios más generales, los cuales a este nivel, no se cuestionan. Es posible imaginar un test del modelo de muestreo del test del modelo de muestreo... No obstante, poco importa la "altura" del nivel al cual nos elevamos, siempre existirá en el nivel superior un modelo de muestreo que no se cuestiona.

2-3.2.3 Tercera etapa: cálculo del valor de la variable-test

Una vez seleccionada la variable-test, solo basta calcular su valor al reemplazar los símbolos por su valor numérico.

En nuestro ejemplo del tiempo de audiencia radiofónica, el tamaño de la muestra es de $n = 25$, el promedio de x en la muestra $m_x = 110$ y la desviación estándar $s_x = 20$; la hipótesis que se quiere probar es

$$H_0 : \mu_x = 100$$

$$\text{Entonces, } t_{24} = \frac{m_x - \gamma}{\left(s_x / \sqrt{n} \right)} = \frac{110 - 100}{\left(20 / 5 \right)} = 2.5$$

2-3.2.4 Cuarta etapa: selección del nivel de significancia

Debemos, ahora, seleccionar un umbral de probabilidad crítico abajo del cual juzgaremos que las observaciones son lo suficiente improbables como para ser incompatibles con la hipótesis. Los valores que más se emplean en ciencias sociales son 1%, 5% y 10%. Para nuestro ejemplo, se tomará 5%.

2-3.2.5 Quinta etapa: detectar los valores críticos de la variable-test (zona de rechazo)

Siguiendo, consultemos una tabla estadística (vea al final del capítulo la "Tabla de valores críticos del test de Student"). Con esta tabla, nos enteramos que, con $n - 1 = 24$ grados de libertad, existe una probabilidad de 0.05 (o sea, de 5%) que

$$t_{24} < -2.064 \text{ o que } t_{24} > 2.064$$

De manera general, la tabla estadística del t de Student nos entrega los valores críticos $\theta_{n-1}(\alpha)$ para los cuales, con $n - 1$ grados de libertad, existe una probabilidad de α que

$$t_{n-1} < -\theta_{n-1}(\alpha) \text{ o } t_{n-1} > +\theta_{n-1}(\alpha)$$

Nota: Algunos autores emplean la anotación $t_{\alpha,n-1}$ para designar los valores críticos de la distribución de Student;. En esta anotación, α es el nivel de significancia (aquí, 0.05) y $n - 1$ es el número de grados de libertad (aquí, 24). Para evitar cualquier confusión entre los valores críticos y la variable-test misma, evitaremos esta anotación y, más bien, designaremos los valores críticos con $\theta_{n-1}(\alpha)$: aquí en nuestro ejemplo.

$\theta_{24}(0.05) = 2.064$

2-3.2.6 Sexta etapa: comparación del valor de la variable-test con los valores críticos y toma de decisión

En este momento tenemos en nuestras manos todos los elementos necesarios para concluir. Calculamos la variable-test $t_{n-1} = 2.5$. Con la tabla estadística, nos enteramos que *si H_0 es verdadera*, este valor es bastante improbable, es decir que la probabilidad de observar un valor tan alejado de cero es de menos de 5%. Puesto que seleccionamos 5% como nivel de significancia, decidimos rechazar H_0. Esto significa como conclusión que el promedio de *x* en la población no es igual a 100 porque pensamos que nuestras observaciones son probablemente incompatibles con esta hipótesis.

De manera general, rechazamos la hipótesis, teniendo un nivel de significancia de α, si $t_{n-1} < -\theta_{n-1}(\alpha)$ o $t_{n-1} > +\theta_{n-1}(\alpha)$.

Está claro que si el valor de la variable-test no hubiera rebasado los valores críticos (lo que hubiera podido suceder con otra muestra), no hubiéramos rechazado la hipótesis.

Para resumir, seguimos las etapas siguientes:

1. Seleccionamos una variable-test que tuviera las propiedades de requisitos, es decir el *t* de Student.
2. Examinamos las condiciones bajo las cuales el test de Student se aplica (el modelo de muestreo) y decidimos que eran aceptables.
3. Calculamos el valor de esta variable-test ($t_{24} = 2.5$).
4. Seleccionamos un nivel de significancia ($\alpha = 5\ \%$).
5. Detectamos los valores críticos en la tabla estadística: si la hipótesis es verdadera, existe una probabilidad α de que la variable caiga al exterior del intervalo definido por los valores críticos $-\theta_{n-1}(\alpha)$ y $+\theta_{n-1}(\alpha)$ (en nuestro ejemplo, la probabilidad de que t_{24} sea inferior a –2.064 o superior a +2.064 es de 5%).
6. Comparamos el valor de la variable-test ($t_{24} = 2.5$) con los valores críticos vistos en la tabla. En nuestro ejemplo constatamos que si la hipótesis fuera verdadera, las observaciones tal como se resumieron en la variable-test hubieran sido improbables (probabilidad inferior a 5%). Y, puesto que esta probabilidad era inferior al nivel de significancia seleccionado, rechazamos la hipótesis.

La tabla que sigue da el "valor" que se necesita atribuir a cada "variable" en el argumento del test de hipótesis clásico con el objetivo de aplicar el argumento al test de una hipótesis simple sobre un promedio.

Aplicación del argumento al test
de una hipótesis simple sobre un promedio

Formulación general	Ejemplo: $n = 25$; $m_X = 110$; $s_X = 20$; $\alpha = 0{,}05$
{*hipótesis*}	
H_0: $\mu_X = \gamma$	H_0: $\mu_X = 100$
{*modelo de muestreo*} • En la población la variable x tiene una distribución (aproximadamente) normal, con un promedio μ_X y una desviación estándar σ_X desconocidos. • La población es de gran tamaño y en ella se sorteó un muestra aleatoria simple de tamaño…	
n	25
{*variable*}	
$\dfrac{m_x - \mu_x}{\left(s_x/\sqrt{n}\right)}$	$\dfrac{110 - \mu_x}{\left(20/\sqrt{25}\right)}$
{*distribución de muestreo*}: distribución de Student con...	
n–1 grados de libertad	24 grados de libertad
{*variable-test*}	
$t_{n-1} = \dfrac{m_x - \gamma}{\left(s_x/\sqrt{n}\right)}$	$t_{24} = \dfrac{110 - 100}{\left(20/\sqrt{25}\right)} = 2.5$
α ←{*Nivel de significancia*}→	0.05
Orientación del test ⇒ {*zona de rechazo*}: test bilateral	
H_A: $\mu_X \neq \gamma$ $\Rightarrow t_{n-1} < -\theta_{n-1}(\alpha)$ o $t_{n-1} > +\theta_{n-1}(\alpha)$	H_A: $\mu_X \neq 100$ $\Rightarrow t_{24} < -2.064$ o $t_{24} > 2.064$
O test unilateral a la derecha	
H_A: $\mu_X > \gamma \Rightarrow t_{n-1} > +\theta_{n-1}(2\alpha)$	H_A : $\mu_X > 100 \Rightarrow t_{24} > 1.711$
O test unilateral a la izquierda	
H_A : $\mu_X < \gamma \Rightarrow t_{n-1} < -\theta_{n-1}(2\alpha)$	H_A : $\mu_X < 100 \Rightarrow t_{24} < -1.711$

Examinemos nuevamente el criterio de selección del nivel de significancia. ¿Qué hubiera pasado si hubiésemos seleccionado un criterio diferente, 1% por ejemplo? La tabla, en anexo, nos informa que $\theta_{24}(0.01) = 2.797$ es decir que, con n –1 = 24 grados de libertad, la probabilidad que $t_{24} < -2.797$ o que $t_{24} > 2.797$ es de 0.01 (o sea 1%). Por consiguiente, si hubiésemos seleccionado un nivel de significancia de 1%, el valor de la variable-test (2.5) se hubiera encontrado en el interior del intervalo delimitado por los valores críticos –2.797 y +2.797. Así que, con este criterio más exigente, no podríamos rechazar la hipótesis.¡Sin embargo, esto no significa tampoco que aceptaríamos la hipótesis!

En términos generales, más pequeño es el nivel de significancia seleccionado, más grande es el valor crítico. En caso de comparar las decisiones que se tomarían con dos umbrales de significación diferentes, es evidente que existen hipótesis, las cuales corresponden a valores de la variable-test, que se encontrarán arriba del valor crítico para el nivel de significancia más elevado pero abajo del valor crítico para el nivel de significancia más exigente (el más pequeño). Tales hipótesis se rechazarían con el nivel de significancia más elevado (menos exigente) pero no con el nivel de significancia más pequeño (más exigente).

La relación entre la selección del nivel de significancia, los valores críticos y la zona de rechazo se ilustra en la figura 4.

Como una síntesis, el diagrama 2c es una copia de la estructura del diagrama 2a pero integra el conjunto de conceptos que acabamos de explicitar, los cuales intervienen en un test de hipótesis.

Figura 4
Test de Student (bilateral)

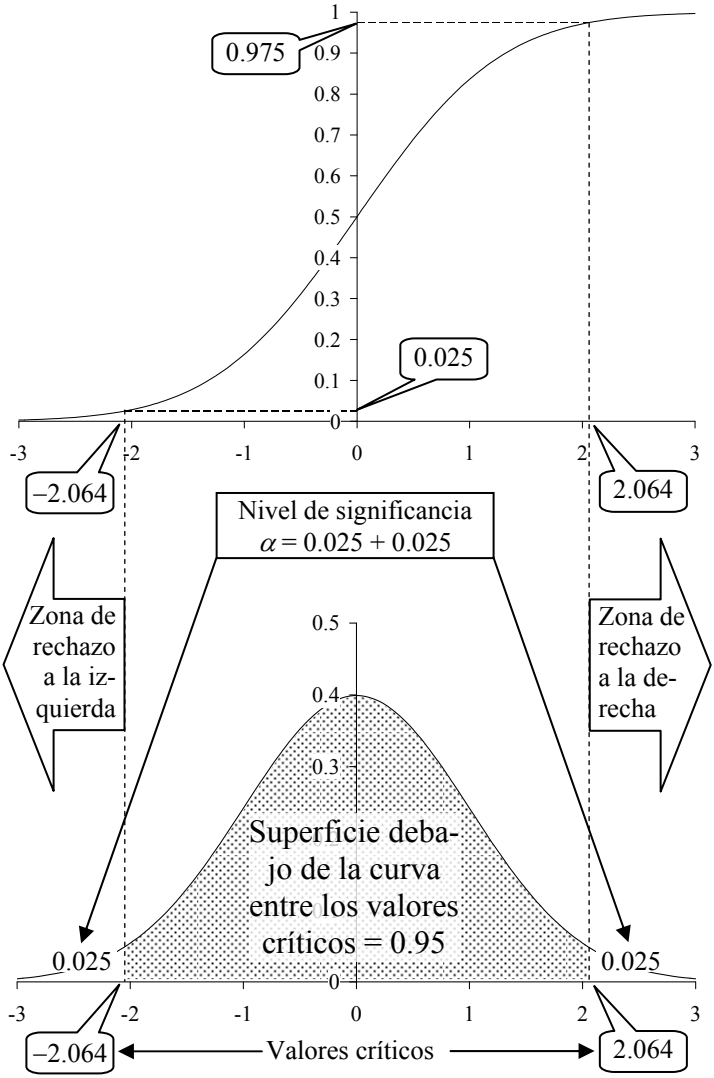
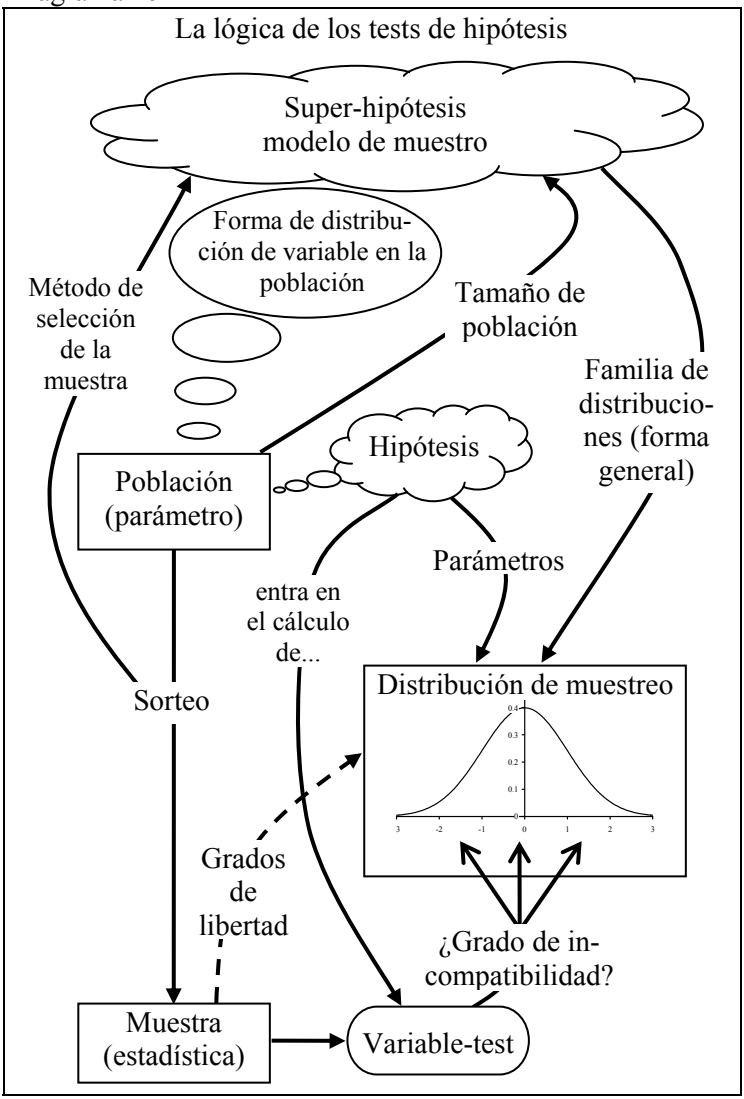


Diagrama 2c


La lógica de los tests de hipótesis
Super-hipótesis
modelo de muestro
Forma de distribu-
ción de variable en la
población
Método de
selección
de la
muestra
Tamaño de
población
Familia de
distribucio-
nes (forma
general)
Hipótesis
Población
(parámetro)
Parámetros
entra en
el cálculo
de...
Sorteo
Distribución de muestreo
Grados
de
libertad
¿Grado de in-
compatibilidad?
Muestra
(estadística)
Variable-test

2-3.3 UN POCO DE TERMINOLOGÍA EN RELACIÓN CON LOS TESTS DE HIPÓTESIS*

2-3.3.1 Hipótesis simple, hipótesis compuesta; hipótesis nula, hipótesis complementaria

Acabamos de exponer el proceso de todos los tests de hipótesis simples. Una hipótesis simple es una hipótesis que especifica en su totalidad la distribución de la variable-test: en la práctica, una hipótesis simple atribuye un valor único a un parámetro. Una hipótesis que albarca una serie de valores posibles es una hipótesis compuesta.. Por ejemplo:

- hipótesis simple: $\mu_x = 0$
- hipótesis compuesta: $\mu_x > 0$

En un test de hipótesis simple, la hipótesis que se quiere probar es con frecuencia llamada la hipótesis nula[121] y se designa con H_0. Cuando un test conlleva al rechazo de la hipótesis, implica, lógicamente, que aceptamos la hipótesis complementaria (*alternate hypothesis*) la cual se designa, a menudo, con H_A. La hipótesis complementaria de una hipótesis simple es, por lo general, una hipótesis compuesta. Por ejemplo:

- $H_0 : \mu_x = 0$
- $H_A : \mu_x \neq 0$

* Referencias: Freund (1962), cap. 11.

[121] Según Knapp (1996), esta expresión tiene varias explicaciones: (1) la hipótesis que se quiere testar es, a menudo que el valor del parámetro es nulo; (2) es la hipótesis neutral según la cual nada sale de lo ordinario; (3) con frecuencia, el investigador desea que los datos "anulan" esta hipótesis (por lo personal, considero esta ultima explicación muy ligera).

2-3.3.2 Nivel de significancia, zona de rechazo y errores del tipo I y II[*]

Una vez seleccionado el nivel de significancia, el conjunto de los valores de la variable-test cuya probabilidad de realización se encuentra abajo del nivel de significancia se llama la zona de rechazo (o región crítica o zona crítica), del test (vea figura 3).

Un test estadístico se basa en un razonamiento probabilista. Su conclusión no es, por consiguiente, cierta es más bien solamente probable. Siempre habrá ciertos riesgos al momento de tomar decisiones a la luz de un test estadístico.

La tabla que sigue resume los diferentes tipos de errores que se puede cometer:

		Situación (inobservable)	
		H_0 es verdadera	H_0 es falsa
Decisión	Rechazar H_0	Error de tipo I	Buena decisión
	No rechazar H_0	Buena decisión	Error de type II

En cada una de las situaciones posibles, las probabilidades asociadas a estas posibilidades son:

[*] Referencias: Wonnacott y Wonnacott (1992, p. 344-345, 349-350 y 354-357).

		Situación (inobservable)	
		H_0 es verdadera	H_0 es falsa
Decisión	Rechazar H_0	Nivel de significancia α	Potencia $(1 - \beta)$
	No rechazar H_0	$(1 - \alpha)$	β

En el habla de la estadística, el error del tipo I corresponde al error que se comete cuando se rechaza la hipótesis mientras que, de hecho, era verdadera. La probabilidad de cometer un error del tipo I es la probabilidad de que la variable caiga en la zona de rechazo aunque H_0 sea verdadera. ¿Y cuál es esta probabilidad? ¡Es, por definición, el nivel de significancia seleccionado! Si H_0 es verdadera, la probabilidad α de cometer un error del tipo I es el nivel de significancia seleccionado para el test. La selección del nivel de significancia es, por lo tanto, una selección del nivel que se acepta arriesgar de cometer un error del tipo I en caso que H_0 sea efectivamente verdadera.

Un error del tipo II consiste en aceptar (más bien no rechazar) una hipótesis cuando ésta es falsa. En caso que H_0 sea falsa, la probabilidad β de cometer un error del tipo II es difícil de evaluar dado que existen, en estas condiciones, varias distribuciones posibles para la variable-test. En caso de poder evaluar esta probabilidad, entonces la probabilidad de evitar un error del tipo II $(1-\beta)$ se llama la potencia del test.

En condiciones ideales, desearíamos un test cuyas probabilidades de los dos tipos de errores fueran muy pequeñas (α y β pequeños). Sin embargo, podemos lograr entender de manera intuitiva que, para un test dado, más α es pequeño y más β es grande. En efecto, cuando α es pequeño, la zona de rechazo es pequeña también, lo que aumenta la probabilidad

de no rechazar H_0 y, por consiguiente, esto aumenta β. En resumen, la decisión que se toma basándose en un test de hipótesis es una apuesta donde hacemos un compromiso entre dos riesgos de errores: el riesgo de error de tipo I o el riesgo de error de tipo II. Un buen test de estadística es, por lo tanto, un test que, para cualquier nivel dado de probabilidad de error de tipo I, posee la más pequeña probabilidad posible de error de tipo II; en otras palabras, el mejor test es el test más potente para cada nivel de significancia.

2-3.3.3 Distribuciones asintóticas[*]

El modelo de muestreo no siempre permite especificar por completo la distribución de muestreo de una variable-test. A menudo se puede lograr resolver este problema con la distribución asintótica de la variable-test. En efecto, se puede demostrar que varias distribuciones de muestreo tienden a aproximarse de una distribución conocida a medida que el tamaño de la muestra aumenta. Esta distribución conocida se llama una distribución asintótica. Cuando la muestra es lo "suficiente grande" se puede tomar la distribución asintótica como aproximación de la distribución de muestreo exacta.

Por ejemplo, la distribución asintótica de una distribución de Student es la distribución normal (vea figura 3). En este caso particular. No existen verdaderos problemas y se podrá especificar por completo la distribución de muestre de la variable-test para cada valor de número de grados de libertad. No obstante, al momento de rebasar un cierto tamaño de muestra (y de número de grados de libertad que esto implica), se considera que la distribución de Student es tan próxima de la normal que ya no vale la pena referirse a la distribución exacta. En la practica, cuando la distribución de Student tiene

[*] Referencias: Wonnacott y Wonnacott (1992, pp. 224-228).

más de 30 grados de libertad, se estima, con frecuencia, que la muestra es "lo suficiente grande".

2-3.4 TESTS UNILATERALES (ONE-SIDED TESTS)

El ejemplo que se presentó en el apartado 2-3.2 era un test bilateral (two-side test). En este test, se concede la misma importancia a las desviaciones tanto hacia arriba como hacia abajo con relación a la hipótesis nula.

$H_0 : \mu_x = \gamma$

Sin embargo, en algunas circunstancias, importa sólo una de las dos posibilidades. Por ejemplo, supongamos que un comprador quiera asegurarse que un producto respete una norma de calidad promedio. Digamos que se mide la calidad con un indicador x y que la norma que se debe respetar es que el valor promedio μ_x del indicador de calidad x sea, por lo menos, igual a γ. Para decidir aceptar el lote (la población), el comprador examina una muestra obtenida del lote y calcula la calidad promedio m_x de esta muestra. Está claro que el comprador no se decepcionará si la calidad promedio del producto rebasa la norma. En este caso, la hipótesis complementaria no es $H_A : \mu_x \neq \gamma$ sino, más bien $H_A : \mu_x > \gamma$

Dicho de otra manera, rechazar H_0, significa, para el comprador, aceptar el lote, es decir aceptar la hipótesis de que la calidad del lote rebasa la norma. En estas condiciones, la zona de rechazo se sitúa de un solo lado del cero, a la derecha. La lógica es simple: si m_x es lo suficiente grande para que se rechace la hipótesis $H_0 : \mu_x = \gamma$, entonces, con mucha más razón, se rechazará cualquier hipótesis $H_0' : \mu_x = \gamma$ para cualquier valor $\gamma' < \gamma$. Debemos notar que este razonamiento induce al comprador para aceptar únicamente el lote cuando la calidad promedia de la muestra rebasa la norma con un

margen suficiente.[122] En otras palabras, cuando un comprador no acepta el lote, no es porque rechaza la hipótesis que el lote respeta la norma, sino más bien es porque se siente estadísticamente hablando, incapaz de rechazar la hipótesis de que el lote no respeta la norma.

Por otra parte, debemos recordar que la aplicación de un test unilateral cambia la relación entre el nivel de significancia y los valores críticos que definen la zona de rechazo. Por ejemplo, si la zona de rechazo que se usa se define con $t_{24} > 2.064$ al lugar de $t_{24} < -2.064$ o $t_{24} > 2.064$ entonces, la probabilidad de rechazar H_0, mismo si esta hipótesis fuera verdadera, no es de 5% pero de 2.5%. El nivel de significancia de este test unilateral sería, por lo tanto, de 2.5%. Se ilustra esa situación en la figura 5a.

Ejemplo:[123]

Un comprador debe decidir si acepta un lote de 100 000 tubos catódicos. La norma de calidad exigida es que el tiempo de vida promedio de los tubos del lote sea por lo menos de 1200 horas. Se efectúan algunos tests en una muestra de 100 tubos que revelan un tiempo de vida promedio de los tubos de 1265 horas. Teniendo una diferencia tipo de 300 horas, la hipótesis que se quiere probar es

$$H_0 : \mu_x = 1200$$

Se define la estadística t con

$$t_{n-1} = \frac{m_x - \gamma}{\left(s_x / \sqrt{n} \right)} = \frac{1265 - 1200}{\left(300 / 10 \right)} = 2.17$$

[122] Para un test unilateral con un nivel de significancia igual a α, este margen es el margen de error bilateral asociado a un nivel de confianza de $(1-2\alpha)$. Vea 2-3.4.

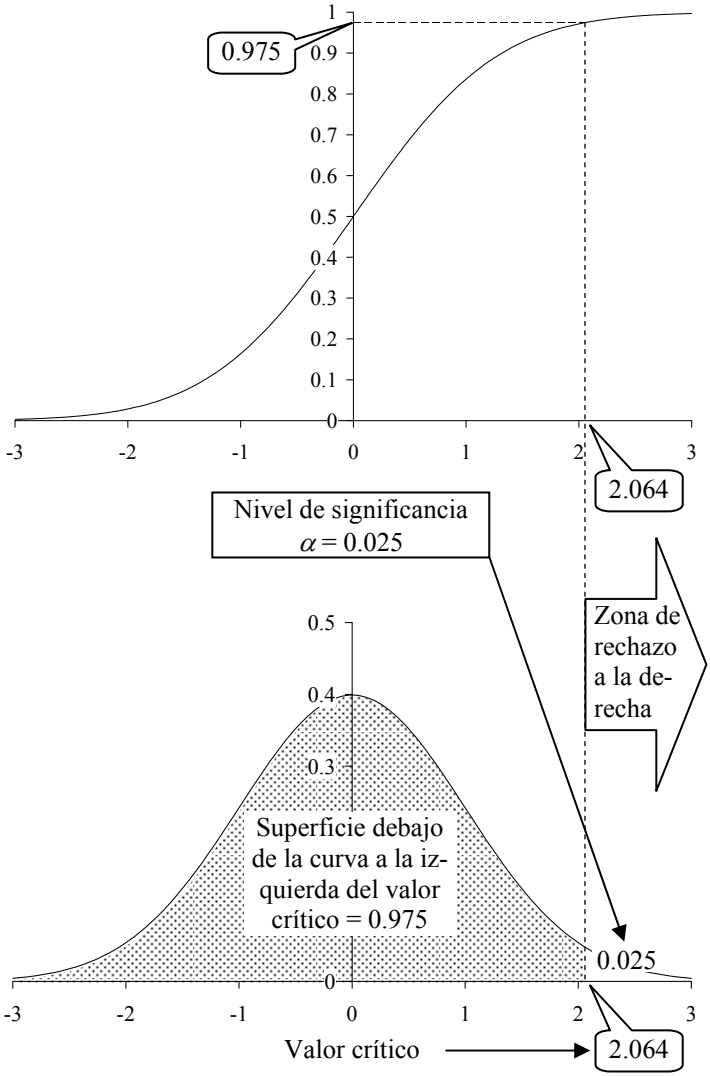
[123] Se toma este ejemplo de Wonnacott y Wonnacott (1991, pp. 333-334).

Si el umbral significación que se escogió es de 0.005, con 99 grados de libertad, el valor crítico de sitúa entre 2.626 y 2.632 (vea la tabla de los valores críticos del test de Student)[124]. No podemos rechazar H_0 y el comprador no aceptará el lote.

En este ejemplo, la zona de rechazo se encuentra a la derecha del cero. Existe, obviamente, circunstancias cuando la zona de rechazo estaría a la izquierda.

[124] Claro está que hubiéramos podido calcular el valor crítico exacto por medio de la función TINV del software Excel de la misma manera que se calculó los valores de la tabla.

Figura 5a
Test de Student (unilateral)



El argumento lógico del test de hipótesis que se formalizó en el cuadro al final del apartado 2-3.1 se aplica también a un test unilateral. Se puede aplicar a un test de hipótesis unilateral a la derecha efectuando en la tabla del apartado 2-3.2 que da el "valor" que es necesario atribuir a cada "variable", la sustitución que sigue:

Para un test unilateral a la derecha ($H_A : \mu_x > \gamma$) con 24 grados de libertad y $\alpha = 0.05$,

$\{zona\ de\ rechazo\} = t_{n-1} > +\theta_{n-1}(2\ \alpha) = t_{24} > 1.711$

al lugar de

$\{zona\ de\ rechazo\} = t_{n-1} < -\theta_{n-1}(\alpha)$ o

$t_{n-1} > +\theta_{n-1}(\alpha) = t_{24} < -2.064$ o $t_{24} > 2.064$

2-3.5 TEST DE PROBABILIDAD CRÍTICO SIN UMBRAL DE SIGNIFICADO PRE-DETERMINADO (P-VALUE TEST)*

Los tests estadísticos clásicos se efectúan comparando el valor calculado de una variable-test con los valores de referencias que se encuentran en las tablas. Sin embargo, para varias variables-tests que se emplean con frecuencia los paquetes de aplicación de estadística procuran hoy en día el nivel de significancia con el cual el valor de la estadística estaría exactamente al límite de la zona de rechazo.[125] Este nivel de significancia se llama la probabilidad crítica (p-value). En la presentación de resultados, se procura de más en más el valor de esta probabilidad en lugar de indicar si se rechaza o no la hipótesis con el nivel de significancia de 1%, 5% o 10%. Es ésta, una manera de entregar los resultados con un máximo de transparencia lo cual deja el lector libre de escoger el nivel

* Referencias: Wonacott y Wonnacott (1992, pp. 333-337).

[125] En caso de que se trate del *t* de Student, es posible encontrar su valor con la función del logicial Excel.

de significancia y decidir estar de acuerdo o no con el rechazo de la hipótesis.

En seguida se ilustra el test de probabilidad crítica en las figuras 5b y 5c; luego, se representa su argumento lógico en el cuadro siguiendo el modelo exhibido en el apartado 2-3.1.

Figura 5b
Test de de probabilidad crítica (Student bilateral)

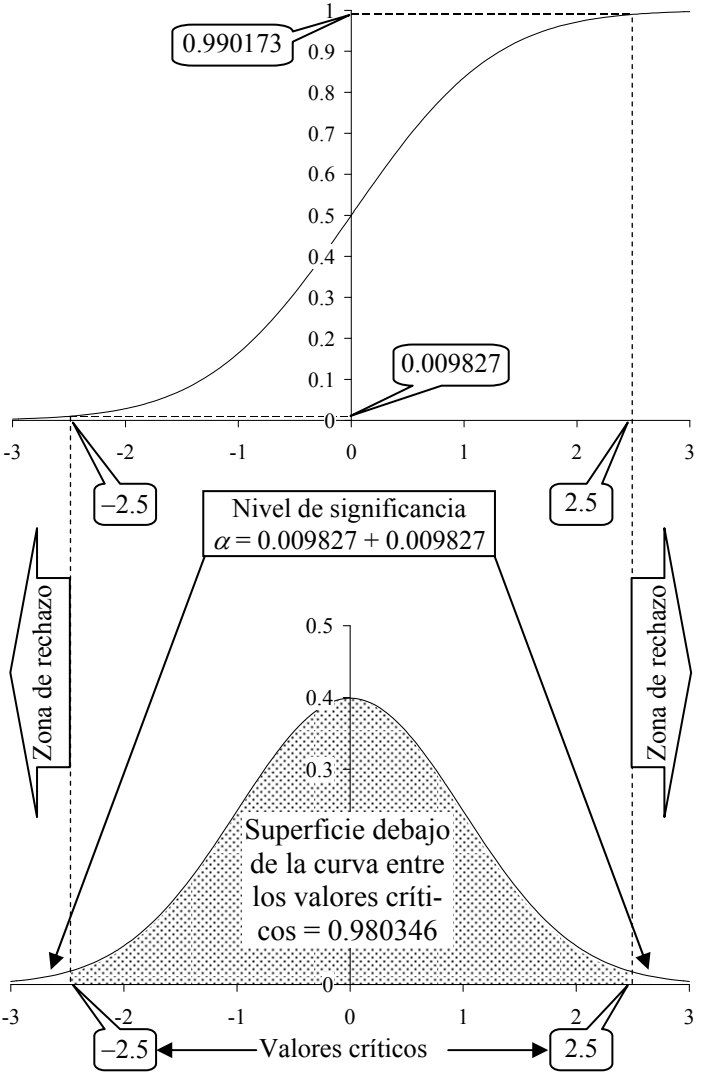
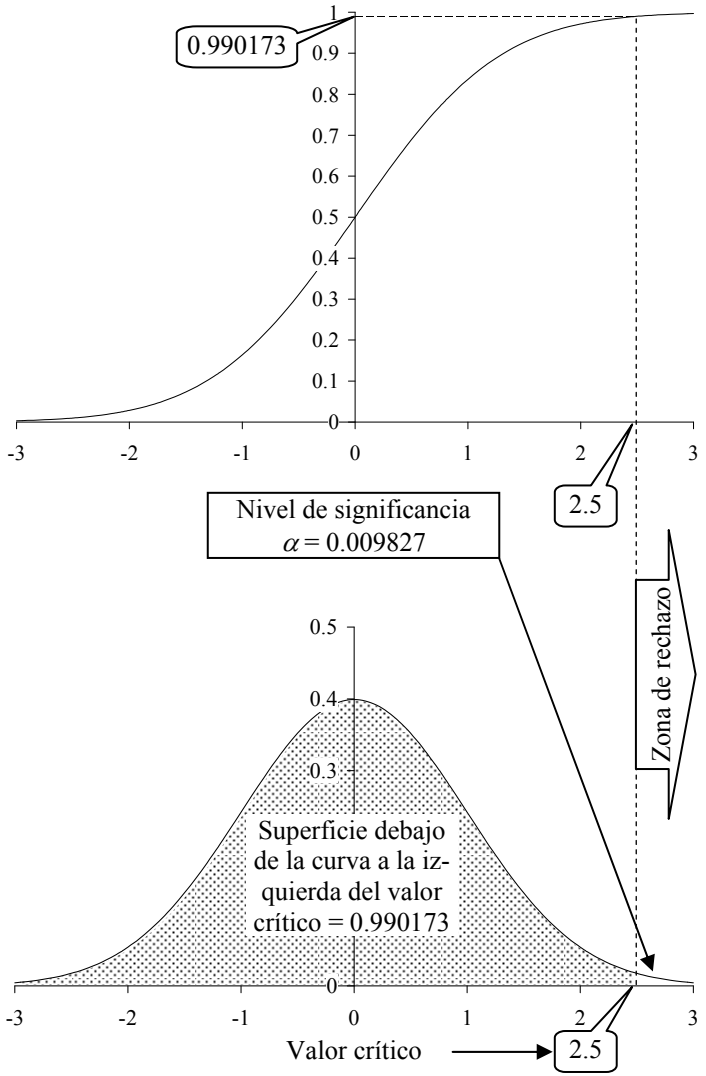


Figura 5c
Test de de probabilidad crítica (Student unilateral)



Argumento del test de probabilidad crítica
(test de hipótesis sin nivel de significancia predeterminado)

1. Modelo de muestreo, hipótesis y implicaciones (silogismo)

   Si es cierto {*modelo de muestreo*},
   entonces {*variable*}[126] tiene la distribución {*distribución de muestreo*}.

   Ahora bien

   si es cierta {*hipótesis*},
   entonces {*variable*} es igual a la estadística {*variable-test*}.

   Por lo tanto

   si son ciertos a la vez {*modelo de muestreo*} y {*hipótesis*},
   entonces {*variable-test*} tiene la distribución {*distribución de muestreo*}.

2. Evaluación de la credibilidad de la hipótesis

   Teniendo

   - la distribución {*distribución de muestreo*},
   - la orientación del test (bilateral o unilateral, a la derecha o a la izquierda, dependiendo de la hipótesis complementaria H_A),

   el conjunto de valores extremos cuyo límite es definido por el valor observado de {*variable-test*} tiene una probabilidad igual a {*probabilidad crítica*}.

   Por lo tanto, si es cierta {*hipótesis*}
   entonces, el valor observado de {*variable-test*} forma parte del conjunto de valores extremos con una probabilidad igual a {*probabilidad crítica*}.

[126] Esta variable, ni es una estadística, ni es un parámetro. No es un estadística porque su valor depende de parámetros pero, tampoco es un parámetro ya que su valor depende también de una estadística.

3. Conclusión:
   Se determina si la {*probabilidad crítica*} {*es* / *no es*} suficientemente pequeña para concluir que las observaciones son con toda probabilidad incompatibles con {*hipótesis*} para, luego, {*rechazar* / *no rechazar*} {*hipótesis*}.

2-3.6 INTERVALOS DE CONFIANZA Y MÁRGENES DE ERRORES (ESTIMACIÓN DEL PROMEDIO)*

En el ejemplo que se exhibió en el apartado 2-3.2, rechazamos la hipótesis que $\mu_x = 100$ con un nivel de significancia de 5%. Podríamos repetir el test con otras hipótesis, con $\mu_x = 101$, $\mu_x = 102$, ..., $\mu_x = 110$, etc. Al efectuar el test para todos los valores posibles, podríamos hacer un inventario de las hipótesis que no se rechazarían (es decir, que serían "aceptables") con un nivel de significancia de 5%. El conjunto de hipótesis que no se rechace con un nivel de significancia dado constituye un intervalo de confianza.[127]

Existe, sin embargo, una solución más directa para llegar al mismo resultado. Se sabe que para cualquier hipótesis posible del tipo $\mu_x = \gamma$, se tendrá la variable-test

$$t_{n-1} = \frac{m_x - \gamma}{\left(s_x \Big/ \sqrt{n} \right)}$$

es decir, en nuestro ejemplo,

$$t_{24} = \frac{110 - \gamma}{\left(20 \Big/ 5 \right)}$$

* Referencias: Wonnacott y Wonnacott (1992, pp. 286-296).

[127] La terminología estadística tradicional distinguía entre la estimación "puntual" y la estimación "por intervalo", esta ultima refiriéndose a los intervalos de confianza.

Se rechazará todas las hipótesis por las cuales

$$t_{n-1} < -\theta_{n-1}(\alpha) \text{ o } t_{n-1} > +\theta_{n-1}(\alpha)$$

o sea, en el caso de nuestro ejemplo,

$$t_{24} < -2.064 \text{ o } t_{24} > +2.064$$

El total de las hipótesis que NO se rechazaría con un nivel de significancia α se define, por lo tanto, de la manera siguiente:

$$-\theta_{n-1}(\alpha) < \tau_{n-1} < +\theta_{n-1}(\alpha)$$

o sea, en el caso de nuestro ejemplo,

$$-2.064 < t_{24} < +2.064$$

Al reemplazar t_{n-1}, obtenemos

$$-\theta_{n-1}(\alpha) < \frac{m_x - \gamma}{\left(s_x / \sqrt{n}\right)} < +\theta_{n-1}(\alpha)$$

o sea,

$$-\theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) < \left(m_x - \gamma\right) < +\theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$$

$$-m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) < -\gamma < -m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$$

$$m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) > \gamma > m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$$

$$m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) < \gamma < m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$$

o, en nuestro ejemplo,

$$110 - 2.064\left(20/5\right) < \gamma < 110 + 2.064\left(20/5\right)$$

$$101.744 < \gamma < 118.256$$

Así que, siempre y cuando γ no forme parte del intervalo

$$\left[m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right);\ m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)\right]$$

([101.744; 118.256] en nuestro ejemplo), la hipótesis no se rechazará con un nivel de significancia de α (5%). Esto implica naturalmente que se formule la hipótesis compuesta:

$$\mathrm{C}: \boxed{m_x - \theta_{n-1}(\alpha)\left(s_x/\sqrt{n}\right) < \mu_x < m_x + \theta_{n-1}(\alpha)\left(s_x/\sqrt{n}\right)}$$

es decir, en nuestro ejemplo, donde $\theta_{n-1}(\alpha) = 2.064$,

$$110 - 2.064\ (20/5) < \mu_x < 110 + 2.064\ (20/5)$$

¿Cuál es la probabilidad de que la condición C sea verdadera? De cierta manera, esta pregunta no tiene sentido puesto que la muestra misma nos da los valores de m_x, s_x y n mientras que μ_x es desconocida pero fija; por lo tanto nada es aleatorio en el enunciado de la condición C. Sin embargo, si imaginamos que estamos justo antes del momento del sorteo de la muestra,[128] sabemos que por cualquier valor fijo pero desconocido de μ_x, existe una probabilidad de α (5% o 0.05 en nuestro ejemplo), que los valores de m_x y de s_x extraídos de la muestra no respeten la condición C. Dicho de otro modo, antes de sortear la muestra, existe una probabilidad de $(1-\alpha)$ de que la condición C se respete (95% o 0.95).

El intervalo

$$\left[m_x - \theta_{n-1}(\alpha)\left(s_x/\sqrt{n}\right);\ m_x + \theta_{n-1}(\alpha)\left(s_x/\sqrt{n}\right)\right]$$

se llama un intervalo de confianza cuyo nivel de confianza se define con

$1 - \alpha = 1 -$ nivel de significancia

(en nuestro ejemplo, $0.95 = 1 - 0.05$).

El intervalo de confianza y el nivel de confianza son indisociables. Hablar de un intervalo de confianza sin mencionar su nivel de confianza, es como reportar el resultado "parcial" de un juego deportivo con sólo anunciar el número de goles

[128] Es decir, justo antes de conocer los valores de m_x y s_x.

que contó uno de los dos equipos sin mencionar el número de goles que contó el otro...

De manera paralela, se calcula el margen de error: si se considera m_x como valor estimado de μ_x, diremos que el margen de error es de $\pm\theta_{n-1}(\alpha)\left(s_x/\sqrt{n}\right)$ con un nivel de confianza de $1 - \alpha$ (en nuestro ejemplo ± 8.256 con un nivel de confianza de 95% o como se acostumbra mencionar en los reportes periodísticos de sondeo, "19 veces de 20"). Tal como el intervalo de confianza, el margen de error pierde todo significado al no ser acompañado de su nivel de confianza.

Puesto que μ_x es fijo, no es una variable aleatoria y su valor no depende de una distribución de probabilidad, no es del todo riguroso afirmar que el valor del parámetro se encuentre en el intervalo de confianza con una probabilidad de 95%. Es, de por si, la razón por la cual la estadística emplea una formulación diferente cuando se trata de "confianza" (probabilidad subjetiva). En cambio, es exacto concluir que, al momento de sortear muestras repetidas de la misma población, la diferencia entre el valor estimado del parámetro y su verdadero valor sería inferior al margen de error[129] en 95% de los casos; es lo que significa el famoso "19 veces de 20" a saber que, en promedio, de 20 muestras diferentes, habría 19 para las cuales no se rebasaría el margen de error.

Es importante notar aquí que el proceso de inducción estadística nos permite formular un enunciado afirmativo en lugar de un no rechazo. Sin embargo, esta afirmación, muy matizada de por sí, se infiere de una lógica de no rechazo: afirmamos que, en un conjunto dado de hipótesis, existe probablemente una que es verdadera y calificamos este "proba-

[129] Este margen es, sin embargo, diferente de una muestra a otra puesto su valor depende de la diferencia tipo de la muestra s_x.

blemente" con una evaluación de la confianza que se dicta de manera afirmativa.

El cuadro de la página siguiente resume el desarrollo que permite definir un intervalo de confianza o un margen de error.

Dos conclusiones se pueden sacar de lo anterior:

1. El ejemplo del promedio muestra con claridad que, mientras más el nivel de confianza seleccionado es alto, más el intervalo de confianza ha de ser amplio y más el margen de error es grande; es decir, cuanto más ganamos en confianza, menos precisión se tiene.
2. Este ejemplo ilustra, también, cómo la precisión de las estimaciones depende del tamaño de la muestra. Cuando se trata de estimar el promedio al momento de aumentar el tamaño de la muestra, el margen de error disminuye con la raíz cuadrada del tamaño de la muestra.[130] La ganancia de precisión es menos que proporcional al aumento del tamaño de la muestra: algo muy parecido a la ley de los rendimientos decrecientes de la economía, trasladada al campo de la estadística.

Presentamos las nociones de intervalos de confianza y de margen de error en el contexto de la estimación del promedio de una variable aproximadamente normal, con la ayuda de una muestra aleatoria simple obtenida de una población de muy gran tamaño. Está claro que estas nociones se pueden aplicar en otras situaciones donde las conclusiones que acabamos de sacar siguen válidas.[131]

---

[130] Hay, también, una ganancia de precisión cuando el número de grados de libertad, asociados al t de Student, aumenta; en la tabla, podemos ver cómo los valores críticos $\theta_{n-1}(\alpha)$ disminuyen cuando el número de grados de libertad aumenta. Sin embargo, a medida que nos aproximamos de 30 grados de libertad, las ganancias son cada vez menores.

[131] Aunque las conclusiones siguen siendo válidas en otras situaciones, es importante recordar que la forma particular de las fórmulas depende del modelo de muestreo que se definió en el apartado 2-3.2.

Intervalos de confianza y márgenes de error

Formulación general	Ejemplo: $n = 25$; $m_X = 110$; $s_X = 20$; $\alpha = 0{,}05$
{*variable-test*}	
$t_{n-1} = \dfrac{m_x - \gamma}{\left(s_x / \sqrt{n}\right)}$	$t_{24} = \dfrac{110 - 100}{\left(20 / \sqrt{25}\right)} = 2.5$
Hipótesis rechazadas con un nivel de significancia de α (5 %)	
$t_{n-1} < -\theta_{n-1}(\alpha)$ o $t_{n-1} > +\theta_{n-1}(\alpha)$	$t_{24} < -2.064$ o $t_{24} > 2.064$
Hipótesis no rechazadas con un nivel de significancia de α (5 %)	
$-\theta_{n-1}(\alpha) < t_{n-1} < +\theta_{n-1}(\alpha)$	$-2.064 < t_{24} < +2.064$
$m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) < \gamma < m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$	
	$110 - 2.064\left(20/5\right) < \gamma < 110 + 2.064\left(20/5\right)$
	$110 - 8.256 < \gamma < 110 + 8.256$
	$101.744 < \gamma < 118.256$
Antes de sortear la muestra e independientemente del valor de m_x, existe una probabilidad de $(1-\alpha)$ (95 %) que se respete la condición C : $m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right) < \mu_x < m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$	
El intervalo de confianza...	
$\left[m_x - \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right); m_x + \theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)\right]$	
	$\left[110 - 2.064\left(20/5\right); 110 + 2.064\left(20/5\right)\right]$
	[101.744 ; 118.256]
... y su nivel de confianza	
$1-\alpha$=1−nivel de significancia	0.95 = 1 − 0.05
Margen de error con un nivel de confianza de $(1-\alpha)$ (95 %)	
$\pm\,\theta_{n-1}(\alpha)\left(s_x / \sqrt{n}\right)$	± 8.256

2-3.7 DETERMINACIÓN DEL TAMAÑO REQUISITO DE UNA MUESTRA (ESTIMACIÓN DEL PROMEDIO)

¿Cómo determinar el tamaño de la muestra en función del nivel de precisión deseado? En el contexto de la estimación del promedio, observamos que el valor del margen de error depende de la desviación estándar de la muestra s_x, de tal manera que ninguna fórmula permite conocer el grado de precisión mientras no se efectue el sorteo de la muestra. A lo más, se puede determinar el tamaño necesario para que, en los peores de los casos, el margen de error esté aceptable. ¿Pero qué queremos decir con "en los peores de los casos"? Es evidente que el peor de los casos es el caso cuando, en la muestra sorteada, la desviación estándar es la más grande. Examinemos esto en detalle.

Vimos que el margen de error se define con

$$\varepsilon = \pm\theta_{n-1}(\alpha)\left(s_x \Big/ \sqrt{n}\right)$$

Buscamos determinar n, el tamaño de la muestra. Las tablas estadísticas nos procuran lo valores de $\theta_{n-1}(\alpha)$ y el valor de s_x es desconocido en cuanto no se haya sorteado la muestra. Así que éstos son los pasos que se debe seguir:

1. Decidir el margen de error aceptable ε.
2. Escoger el nivel de confianza deseado $(1 - \alpha)$.
3. Detectar en la tabla los valores de $\theta_{n-1}(\alpha)$ para los diferentes tamaños de la muestra n.
4. Formular con relación a s_x la hipótesis del peor, o sea del más grande valor de s_x, que se puede obtener en la muestra.
5. Resolver para n, la ecuación siguiente.

$$\varepsilon = \pm\theta_{n-1}(\alpha)\left(s_x \Big/ \sqrt{n}\right)$$

$$\sqrt{n} = \theta_{n-1}(\alpha)\left(s_x \big/ \varepsilon\right)$$

$$n = \left[\theta_{n-1}(\alpha)\left(s_x \big/ \varepsilon\right)\right]^2$$

Por ejemplo, supongamos que el margen de error aceptable sea de 10 ($\varepsilon = 10$), el nivel de confianza deseado de 90% ($\alpha = 0.10$) y la hipótesis del peor $s_x = 20$. Tendríamos, entonces:

$$n = \left[\theta_{n-1}(\alpha)\left(s_x \big/ \varepsilon\right)\right]^2 = \left[\theta_{n-1}(0.10)\left(20 \big/ 10\right)\right]^2$$

$$n = 4\left[\theta_{n-1}(0.10)\right]^2$$

Puesto que $\theta_{n-1}(\alpha)$ depende de n, se trata de una ecuación en forma implícita. Se puede resolver con aproximaciones sucesivas.

Solución con aproximaciones sucesivas

Podemos iniciar el proceso de aproximaciones suponiendo una muestra de gran tamaño ($n \to \infty$). por el cual la tabla nos da

$\theta_{\infty}(0,10) = 1.645$

Entonces

$$n_0 = \left[\theta_{n-1}(\alpha)\left(s_x \big/ \varepsilon\right)\right]^2 = \left[\theta_{n-1}(0.10)\right]^2$$

$$n_0 = 4\left[1.645\right]^2 = 10.8$$

lo que significa que la muestra podría ser más pequeña que el infinito al mismo tiempo que más grande que 10.8. En cambio, con $n = 11$ (11 es el primer número entero superior a 10.8), tenemos

$\theta_{11-1}(0.10) = 1.812$ y

$$n_1 = 4\left[\theta_{10}(0.10)\right]^2 = 4\left[1.812\right]^2 = 13.1 > 11$$

lo que significa que la muestra debe ser más grande que 11. Con $n = 13$ (13 es el primer número entero inferior a 13.1). tenemos

$\theta_{13-1}(0.10) = 1.782$ y

$$n_2 = 4\left[\theta_{12}(0.10)\right]^2 = 4\left[1.782\right]^2 = 12.7 < 13$$

Esto significa que la muestra podría ser más pequeña. Sin embargo con $n = 12$

$\theta_{12-1}(0.10) = 1.796$ y

$4[\theta_{11}(0.10)]^2 = 4 \times (1.796)^2 = 12.9 > 12$

Por lo tanto, la muestra debe ser más grande.

Conclusión: puesto que 12 no es suficiente y 13 es más que suficiente para obtener el margen de error deseado, se necesita una muestra de tamaño 13.

Se puede verificar el resultado calculando el margen de error:

$$\varepsilon = \pm\theta_{n-1}(\alpha)\left(s_x \big/ \sqrt{n}\right) = 1.782\left(20 \big/ \sqrt{13}\right) = 9.885 < 10$$

2-3.7.1 Caso en que el margen de error aceptable se fija en términos relativos

Por lo general, nos interesa mayormente el margen de error en términos relativos, o sea en fracción del promedio estimado (fracción que expresamos, con más frecuencia, como un porcentaje):

$$\frac{\varepsilon}{m_x} = \pm\theta_{n-1}(\alpha)\frac{\left(s_x \big/ m_x\right)}{\sqrt{n}}$$

Vemos que el método de determinación del tamaño de la muestra requerido es esencialmente lo mismo cuando se desea fijar el margen de error en porcentaje del promedio estimado. La única diferencia es el hecho de formular la

hipótesis del peor con relación al coeficiente de variación s_x/m_x, en lugar de con relación a la desviación estándar. La ventaja de este enfoque es de permitir la construcción de una tabla de uso general que da el tamaño de la muestra requerido en función del margen de error relativo aceptable y del coeficiente de variación.

2-3.7.2 Caso en que el promedio buscado es una proporción[*]

Puede ocurrir que no sea fácil formular la hipótesis del peor con relación a la desviación estándar o del coeficiente de varianza. No obstante, existe una clase de situaciones que no requieren de hipótesis: esto sucede al momento de querer estimar una proporción. Por ejemplo, no interesa saber cual proporción de una población es favorable a algún proyecto de planificación urbana. Se realiza un sondeo y se define una variable dicotómica que representa las respuestas a las preguntas sobre el proyecto de planificación.

$x_i = 1$ si el sondeado i es favorable

$x_i = 0$ si el sondeado i no es favorable

En estas condiciones, tenemos

$$m_x = \frac{\sum_i x_i}{n} = \frac{\text{Número de respuestas favorables}}{\text{Número total de respondientes}}$$

El promedio m_x es, por lo tanto, la proporción de personas favorables en la muestra; tal proporción se acostumbra representar con la letra p (por la p e proporción) mejor que con m_x. Se pretende estimar μ_x, la proporción de personas favorables en la población con un cierto margen de error.

Para determinar el tamaño de la muestra requerido, es necesario especificar lo que llamamos la hipótesis del peor.

[*] Referencias: Wonnacott y Wonnacott (1992, pp. 232-240 y 309-311).

Ahora bien, es posible demostrar que, para una variable dicotómica

$$s_x^2 = p\,(1 - p)$$

donde el valor de p se encuentra, forzosamente, entre cero y uno. Es posible demostrar también que, para todos los valores de p contenidos entre cero y uno, s_x alcanza su máximo cuando $p = 0.5$ lo cual implica que $s_x^2 = 0.25$ y $s_x = 0.5$. Se resuelve de esta manera el problema para especificar la más grande desviación estándar posible.

> Nota: Puesto que la variable estudiada es una variable dicotómica, no es posible pretender que tenga una distribución normal en la población. Siendo riguroso, esto implica que el test de Student y sus procedimientos respectivos no se aplican en el caso de una proporción. Sin embargo, si se trata con una población de muy grande tamaño y que, en ella, se sortea una muestra aleatoria simple, la estadística matemática nos indica que el test de Student es aproximadamente válido con la condición de que μ_x no sea muy alejado de 0.5.

Obviamente, existen tablas de estadísticas que procuran el tamaño de la muestra requerido en función del margen de error aceptable para diferentes hipótesis emitidas con el más grande valor posible de m_x (o sea de p).

Hay que tener un especial cuidado con no confundir el error relativo sobre un promedio que no es una proporción y el error absoluto sobre una proporción. Por ejemplo, se estima que en promedio el 23 de septiembre de 1998 los habitantes de la Isla de Montreal escucharon la radio durante 120 minutos con un margen de error de doce minutos (con un nivel de confianza de 95%), se calcula un margen de error relativo de 10% (12/120). Por otra parte, si se dice haber

estimado al 80%, la población de los habitantes de Montreal que escucharon la radio durante por lo menos diez minutos el 23 de septiembre de 1998, con un margen de error de más o menos 10% (con un nivel de confianza de 95%), hay ambigüedad: ¿Son 10% de 80% o 10% simplemente? Dicho de otra manera, ¿el intervalo de confianza con 95% de nivel de confianza se extiende de 72% a 88% o de 70% a 90%? Es, por lo general, la segunda interpretación correcta, porque hablar de un porcentaje de un porcentaje es un tanto trastornado (además si de hecho el intervalo de confianza se extendiera de 72% a 88%, la empresa de sondeo ganaría mucho con proclamar que su margen de error es de 8% en lugar de decir que es de 10% de 80%).

2-3.8 OTROS TESTS EMPLEADOS CON FRECUENCIA

Hasta el momento, hablamos del test de Student y de un solo uso que corresponde al test de una hipótesis simple sobre una media. Existen otras aplicaciones del test de Student. Por ejemplo, cuando se compara dos muestras, el test de Student sirve a testar la hipótesis que los dos promedio son iguales[132]. En caso de rechazar la hipótesis que los dos promedios son iguales, se rechaza automáticamente que los las dos muestras provienen de la misma población. Tenemos

$$H_0 : \mu_1 - \mu_2 = \delta$$

En el caso muy particular cuando las dos muestras tienen el mismo tamaño n, tenemos:

$$t_{2(n-1)} = \frac{m_1 - m_2 - \delta}{\left(s_1 + s_2 \Big/ \sqrt{n} \right)}$$

Por lo general, si las dos muestras son de tamaño n_1 y n_2 respectivamente:

[132] Wonnacott y Wonnacott (1992, pp. 299-307).

$$t_{2(n-1)} = \frac{m_1 - m_2 - \delta}{\left(\frac{\sqrt{\frac{(n_1 - 1)s_1 + (n_2 - 1)s_2}{n_1 + n_2 - 2}}}{\sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \right)}$$

Otro test empleado con frecuencia es el test de χ^2 (chi al cuadrado). Puede emplearse para, por ejemplo, probar una hipótesis simple sobre una varianza. En efecto, en una muestra aleatoria simple sorteada de una población normal de gran tamaño, la variable

$$\frac{s^2}{\left(\sigma^2 / n - 1 \right)}$$

posee la distribución del χ^2 con $n - 1$ grados de libertad.[133] El cuadro que sigue procura el "valor" que se debe atribuir a cada "variable" en el argumento del test de hipótesis clásica con el fin de aplicar el argumento al test de una hipótesis sobre la desviación estándar (vea el cuadro ubicado al final del apartado 2-3.1).

Mencionemos también el test *F* de Fisher, del cual trataremos al momento de estudiar el análisis de regresión. La distribución *F* de Fisher depende de dos parámetros: el número de grados de libertad del numerador y el número de grados de libertad del denominador (el significado de estas dos expresiones se aclarará un tanto en el contexto de los tests *F* sobre las regresiones). Puede emplearse para, por ejemplo, probar un coeficiente de correlación simple. En la hipótesis cuando el "verdadero" coeficiente de correlación $\rho = 0$, la variable-test

[133] Wonnacott y Wonnacott (1002, cap. 17).

$$\frac{r^2}{\left[\left(1-r^2\right)\Big/\left(n-2\right)\right]} = (n-2)\frac{r^2}{1-r^2}$$

posee la distribución *F* de Fisher con 1 grado de libertad en el numerador y $(n-2)$ grados de libertad en el denominador. En esta expresión, *r* es el coeficiente de correlación de la muestra:

$$r = \frac{s_{xy}}{s_x\ s_y}$$

Aplicación del argumento al test de una hipótesis simple sobre la diferencia tipo

Formulación general	Ejemplo: $n = 25$; $m_X = 110$; $s_X = 20$; $\alpha = 0{,}05$
H_0: $\sigma_x = \gamma$ ←{*hipótesis*}→	H_0: $\sigma_x = 70$
{*modelo de muestreo*} • En la población la variable x tiene una distribución (aproximadamente) normal, con un promedio μ_x y una desviación estándar σ_x desconocidos. • La población es de gran tamaño y en ella se sorteó un muestra aleatoria simple de tamaño…	
n	20
$\dfrac{s_x^2}{\left(\sigma_x^2 / n-1\right)}$ ←{*variable*}→	$\dfrac{65^2}{\left(\sigma_x^2 / 19\right)}$
{*distribución de muestreo*}: Distribución del χ^2 con...	
$n-1$ grados de libertad	19 grados de libertad
{*variable-test*}	
$\chi^2_{n-1} = \dfrac{s_x^2}{\left(\gamma^2 / n-1\right)}$	$\chi^2_{19} = \dfrac{65^2}{\left(70^2 / (20-1)\right)} = 16.38$
α ←{*Nivel de significancia*}→	0.05
Orientación del test ⇒ {*zona de rechazo*}: test unilateral a la derecha	
H_A: $\sigma_x > \gamma \Rightarrow \chi^2_{n-1} > \chi^2_{n-1}(\alpha)$	H_A : $\sigma_x > 70 \Rightarrow \chi^2_{19} > 30.144$
O test unilateral a la izquierda	
H_A: $\sigma_x < \gamma \Rightarrow \chi^2_{n-1} < \chi^2_{n-1}(\alpha)$	H_A : $\sigma_x < 70 \Rightarrow \chi^2_{19} < 30.144$
2α ←{*Nivel de significancia*}→	0.10
Test bilateral asimétrico	
H_A: $\sigma_x \neq \gamma$ $\Rightarrow \chi^2_{n-1} < \chi^2_{n-1}(1-\alpha)$ o $\chi^2_{n-1} > \chi^2_{n-1}(\alpha)$	H_A : $\sigma_x \neq 70$ $\Rightarrow \chi^2_{19} < 10.117$ o $\chi^2_{19} > 30.144$

Tabla de los valores críticos del test de student (test bilateral)

Grados de libertad	Probabilidad 0.10	0.05	0.01
1	6.314	12.706	63.656
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
40	1.684	2.021	2.704
50	1.676	2.009	2.678
60	1.671	2.000	2.660
70	1.667	1.994	2.648
80	1.664	1.990	2.639
90	1.662	1.987	2.632
100	1.660	1.984	2.626
∞	1.645	1.960	2.576

Fuente: Valores calculados con la ayuda de la función TINV del logicial Excel.

CONCLUSIÓN DE LA SEGUNDA PARTE

¿Qué debemos recordar de todo lo anterior?

Nuestro punto de partida fue el siguiente: los métodos de inducción estadística son una expresión matemática de principios epistemológicos gracias a los cuales se pueden inferir proposiciones de alcance más general (con relación a la población) a partir de la información que se obtiene de un conjunto de datos particulares (una muestra).

La idea clave del proceso es que toda muestra no es más que un individuo entre la población, usualmente infinita, de todas las muestras posibles. Entendemos así, la naturaleza aleatoria de la relación entre la muestra y la población, entre las estadísticas observadas y los parámetros desconocidos. Sólo la teoría de las probabilidades nos permite delimitar la irreducible incertidumbre con el fin de poder hablar con lucidez de lo que es la realidad de donde provienen nuestras observaciones fragmentarias.

Se formaliza el proceso inductivo con el test de hipótesis cuya lógica nos conduce a enfrentarnos con dos problemas mayores. Para empezar, esta lógica de los tests de hipótesis se fundamenta en el principio de no-contradicción, es decir que si las observaciones no son compatibles con la hipótesis examinada, entonces es necesario decidir no aceptar esta hipótesis, o sea rechazarla. En cambio, que las observaciones sean compatibles con la hipótesis no prueba nada puesto que

existen otras hipótesis que pudieran ser igual de compatibles con las observaciones. Para llegar a un enunciado afirmativo, es necesario delimitar el conjunto de hipótesis compatibles, lo que se logra al definir un intervalo de confianza y un margen de error.

La segunda dificultad mayor consiste en que la inducción estadística es probabilista. La "compatibilidad" no es más que una cuestión de grados, es decir que las observaciones no son más que más o menos compatibles o incompatibles con la hipótesis. Así que ¡Adiós certidumbre! La conclusión que se saca de los tests de hipótesis es una decisión, la cual no es impuesta de manera contundente por los hechos. No tomar decisiones a la ligera se convierte, en estas condiciones, en la responsabilidad social del investigador.

ANEXO 2-A
RECORDANDO ALGUNAS FÓRMULAS COMUNES EN ESTADÍSTICA

Para una presentación más detallada, el lector puede referirse a Gilles (1994), caps. 3 a 5, así como los apartados 6.1, 6.2 y 8.1. De igual manera, puede consultar el capítulo 2 de Wonnacott y Wonnacott (1992).

Nos contentamos aquí con reproducir algunas fórmulas de mediciones entre las más comúnmente empleadas en estadística. En el caso de algunas de ellas, daremos dos fórmulas, es decir la fórmula que se aplica a la población y la otra que se aplica a la muestra. Acabamos de estudiar en la inducción estadística las razones que motivan esta distinción (se emplea para una muestra una formula que produce un estimador no sesgado del parámetro correspondiente a la población). En cambio, puesto que la estadística descriptiva no distingue entre la población y la muestra, es, por lo general, la formula de la población que se emplea.

Simbología

n = número de observaciones

x_i = valor de la variable X en la i^a observación

y_i = valor de la variable Y en la i^a observación

2-A.1 MEDIDAS DE TENDENCIA CENTRAL

- Promedio

  Población:[134] $\mu_x = \left(\frac{1}{n}\right)\sum_i x_i$

  Muestra: $m_x = \left(\frac{1}{n}\right)\sum_i x_i$

- Mediana: es el valor $\breve{x}$ de la variable *X* cuando 50% de la población o de la muestra tienen valores inferiores a $\breve{x}$ mientras 50% tienen valores superiores.
- Moda: en una población o una muestra finita, es el valor la más frecuente de la variable *X* cuando se agrupan las observaciones por clases; es la clase con la frecuencia más alta; en una población infinita, es el valor con la más grande densidad de probabilidad correspondiente.

2-A.2 MEDIDAS DE DISPERSIÓN

- Varianza

  Población: $\sigma_x^2 = \frac{1}{n}\sum_i (x_i - \mu_x)^2$

  Muestra: $s_x^2 = \frac{1}{n-1}\sum_i (x_i - m_x)^2$

- Desviación estándar

  Población: $\sigma_x = \sqrt{\sigma_x^2}$

[134] Las fórmulas que se exhiben en este anexo se aplican a poblaciones finitas. Se pueden generalizar estas fórmulas para poblaciones infinitas por medio del concepto de esperanza matemática.

Muestra: $s_x = \sqrt{s_x^2}$

- Coeficiente de variación

Población: $C_x = \dfrac{\sigma_x}{\mu_x}$

Muestra: $C_x = \dfrac{s_x}{m_x}$

2-A.3 MEDIDAS DE ASOCIACIÓN

- Covarianza

Población: $\sigma_{xy} = \dfrac{1}{n}\sum_i (x_i - \mu_x)(y_i - \mu_y)$

Muestra: $s_{xy} = \dfrac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)$

- Coeficiente de correlación simple

Población: $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x\, \sigma_y}$, con $-1 < \rho_{xy} < +1$

Muestra: $r = \dfrac{s_{xy}}{s_x\, s_y}$, con $-1 < r_{xy} < +1$

INTRODUCCIÓN A LA TERCERA PARTE: UNA CLASIFICACIÓN DE LOS MÉTODOS DEL ANÁLISIS MULTIVARIADO

En un sentido amplio, el análisis multivariado designa un conjunto de métodos de análisis que tratan más de una variable al mismo tiempo. En particular, se requiere del análisis multivariado para:

- Medir el grado de asociación entre dos o varias variables.
- Estimar los parámetros de una relación entre dos o varias variables.
- Evaluar hasta qué punto las diferencias entre dos o varios grupos de observaciones son significativas.
- Intentar predecir a cuál grupo pertenece un individuo a partir de sus demás características.
- Buscar discernir una estructura en un conjunto de datos.

Varias técnicas de análisis multivariado necesitan distinguir entre las variables dependientes y las variables independientes. Las variables dependientes son aquellas cuyo valor se quiere predecir; se conocen las demás como las variables independientes.[135] Es posible clasificar los métodos de análi-

[135] Se encontra abajo una discusión de los términos "variable dependiente" y "variable independiente".

sis en función, por una parte, del número de variables dependientes e independientes y, por otra parte, de que sean, las unas o las otras, variables discretas o continuas.[136]

La tabla que sigue presenta una clasificación de algunos métodos de análisis multivariado.

<table>
<tr><th colspan="2">Variable dependiente</th><th>Variables independientes</th><th colspan="2">Método</th></tr>
<tr><td colspan="2" rowspan="2">Ninguna</td><td>2 variables categóricas</td><td rowspan="2">Análisis de tabla de contingencia</td><td>... con 2 dimensiones</td></tr>
<tr><td>Más de 2 variables categóricas</td><td>... con más de 2 dimensiones</td></tr>
<tr><td colspan="2" rowspan="2">Continua</td><td>Discretas (categóricas)</td><td colspan="2">Análisis de varianza o Regresión múltiple</td></tr>
<tr><td>Continuas y/o discretas</td><td colspan="2">Regresión múltiple</td></tr>
<tr><td rowspan="2">Categórica</td><td>2 categorías</td><td rowspan="2">Continuas y/o discretas</td><td rowspan="2">Logit o probit</td><td>... binomial</td></tr>
<tr><td>Más de 2 categorías</td><td>... multinomial</td></tr>
</table>

Esta parte de la obra trata del análisis de regresión. El análisis de regresión es un método de análisis de los datos que se aplica cuando nos basamos en un modelo teórico formalizado con una relación entre una variable dependiente continua y una o varias variables independientes continuas o discretas.

[136] Se infiere esto de la escala de mediciones de cada variable (ver cap. 1-1): las variables categóricas son discretas mientras que las variables racionales y las variables de intervalo se consideran, comúnmente, como continuas. En cuanto a las variables ordinales, pocos modelos son, de manera específica, idóneos para su tratamiento; en la práctica, se tratan como si fueran continuas pero, al momento de interpretar los resultados, es importante tomar en cuenta el hecho de que son variables ordinales.

La regresión es lineal siempre y cuando la forma funcional de la relación sea lineal.[137]

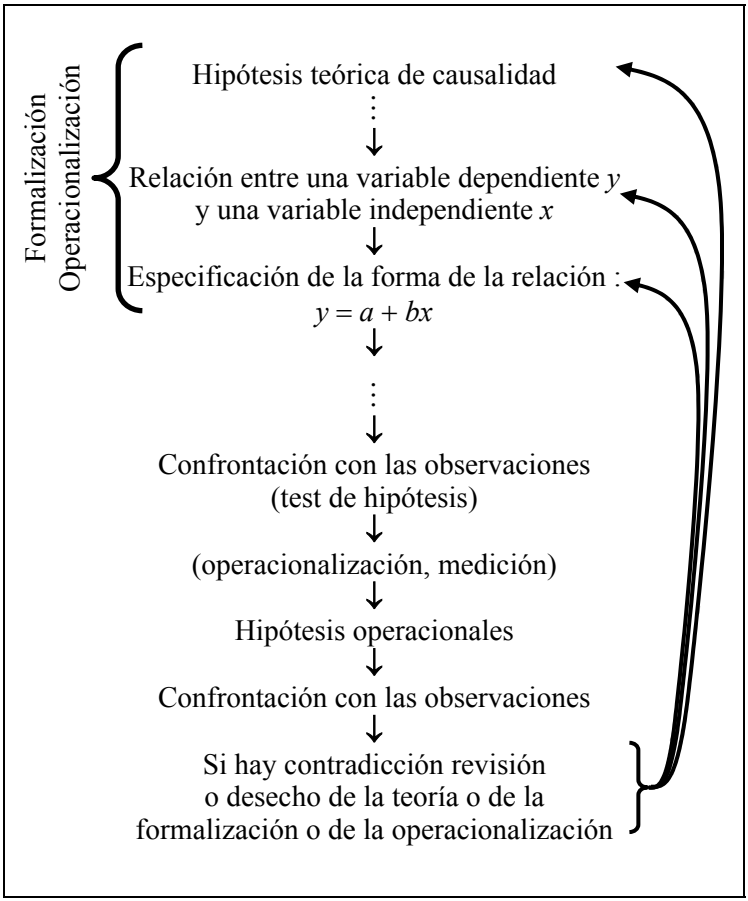
Antes de empezar de pleno el estudio del análisis de regresión, examinemos con más detenimiento la distinción entre las variables dependientes e independientes, particularmente desde el punto de vista del examen de las relaciones de causalidad. Los términos "variable dependiente" y "variable independiente" provienen de las ciencias experimentales cuando el investigador fija de manera "independiente" el valor de algunas variables (por ejemplo, la dosificación de un tratamiento) con el objetivo de observar luego el efecto sobre la variable "dependiente". En un modelo con una única ecuación, la variable "dependiente" es conocida también como "endógena", es decir que se determina en el interior del modelo, cuando las variables independientes son exógenas puesto que se determinan en el exterior del modelo. Estas mismas variables independientes son denominadas, también, "estímuli"; en este caso, las variables dependientes son "respuestas". En inglés, se emplean los términos *predictor/criterion*, *stimulus/response*, *task/performance*, *input/output.*

Esta multitud de términos es sintomática de los numerosos significados conceptuales o teóricos que puede representar la relación entre una variable dependiente y una variable independiente. Las variables independientes son, a veces, calificadas de "explicativas". Antes bien, esta expresión requiere emplearse con prudencia por la fuerte connotación de causalidad que implica. Es posible que la relación sea puramente estadística, de tal manera que la variable independiente pueda permitir "predecir" el valor de variable dependiente sin "explicar" este valor.

[137] Cuando no es posible traducir un modelo teórico con una relación lineal, hay que usar la regresión no lineal a la cual se aplica el método de estimación del máximo de verosimilitud.

Por ejemplo, sería posible observar una relación inversa entre la variable dependiente "variación de las ventas de menudeo de las tiendas de regalos con relación al mes anterior" y la variable "variación de la temperatura promedio con relación al mes anterior". Parece más que evidente que no es el frío el que incita a la compra de regalos sino, más bien, la llegada de la Navidad. En nuestro hemisferio, está la casualidad que este acontecimiento coincide con la llegada del invierno, pero en Australia, pasa lo contrario. Ahora bien, existe efectivamente una relación causal entre la tendencia a vestir ropa más caliente (variable dependiente) y la tendencia a la baja estacional de la temperatura (variable independiente). Lo que nos permite distinguir una relación causal (la segunda) de una relación no causal, es el modelo teórico que poseemos del fenómeno.

En particular, esto implica que los tests de hipótesis que podríamos aplicar a la relación entre variable dependiente y variable independiente no serían aptos para demostrar una relación de causalidad. Lo más que estos tests de hipótesis nos permitirían, sería constatar si la hipótesis de causalidad es o no rechazada por las observaciones. Ilustramos este punto con el diagrama que presentamos a continuación, el cual puede compararse con el diagrama del método hipotético-deductivo del capítulo 2-2.



Para resumir, la relación entre una variable dependiente y una variable independiente no es necesariamente una relación causal. La interpretación que hagamos de esta relación, el significado que le queramos dar dependen del modelo teórico que sirve de punto de partida. Además de la relación causal, es posible distinguir, en orden decreciente del contenido teórico:

- Modelos de simulación, algunas veces llamados modelos de previsión condicional, los cuales se apoyan en un modelo teórico que representa el funcionamiento del fenómeno[138] y que se conciben para contestar a las preguntas del tipo "¿qué pasaría (o qué hubiera pasado) si…?"
- Modelos de proyección que buscan contestar a la pregunta "¿qué va a pasar si la tendencia se mantiene?" o, a partir de un modelo más desarrollado del fenómeno, "¿qué va a pasar si los parámetros de las relaciones entre las variables siguen idénticas?"
- Modelos de previsión que, de manera puramente pragmática, buscan contestar a la pregunta "¿qué va a pasar?" y que, en función de este objetivo, pueden, con todo derecho, explotar relaciones de simple asociación estadística.

[138] El cómo más que el porqué de los modelos de relación causal.

CAPÍTULO 3-1
EL MODELO LINEAL GENERAL Y SU ESTIMACIÓN CON EL MÉTODO DE LOS MÍNIMOS CUADRADOS

3-1.1 EL MODELO LINEAL EN SU FORMA GENERAL

Para un modelo teórico determinista, la forma general del modelo lineal[139] se define con

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

donde el índice suscrito i designa un individuo en la población o una observación en la muestra, y_i es la variable dependiente y x_{i1}, x_{i2}, ..., x_{ik} son las variables independientes. Los coeficientes β_j son los parámetros desconocidos del modelo que se pretende estimar.

Por lo general, una de las variables independientes y a menudo la primera es una constante: $x_{i1} = 1$ para todo i

Es posible escribir entonces el modelo de la manera siguiente:

[139] Para ser precisos, tendríamos que referirnos a un modelo lineal general con variable dependiente única, puesto que el modelo lineal general puede contener varias variables dependientes.

$$y_i = \sum_{j=1}^{k} \beta_j \, x_{ij} = \beta_1 + \beta_2 \, x_{i2} + \cdots + \beta_k \, x_{ik}$$

Se usa la expresión "constante del modelo" para designar al mismo tiempo la variable independiente cuyo valor es constante (x_{i1}) y el parámetro que se le asocia ((β_1).

3-1.1.1 Ejemplo de un modelo lineal

Pretendemos estudiar la relación que existe entre el tamaño de la ciudad más grande (variable dependiente denotada *PLAR*) y la población total y el PIB per cápita del país (variables independientes denotadas *PTOT* y *GNPC* respectivamente). Uno de los modelos que podríamos considerar, sería:

$$\text{PLAR}_i = \beta_1 + \beta_2 \, \text{PTOT}_i + \beta_3 \, \text{GNPC}_i$$

Si fijamos los valores de los parámetros β_1, β_2 y β_3 y si se conoce la población total y el PIB per cápita de un país, es posible calcular lo que predice el modelo en cuanto a la población de su ciudad más grande. Por ejemplo, supongamos que fijamos[140]

$$\beta_1 = 3500$$
$$\beta_2 = 0.01$$
$$\beta_3 = 0.1$$

Presentamos los datos con relación a Brasil y Costa-Rica en 1990, extraídos de la tabla 1 de Lemelin y Polèse (1995):

[140] Estos valores son cercanos a los valores estimados con el método de los menores cuadrados ordinarios que se aplicó a los datos de 1990 que se presentaron en la tabla 1 de Lemelin y Polèse (1995). Los valores estimados exactos son $\beta_1 = 3431$, $\beta_2 = 0.01324$ y $\beta_3 = 0.09375$. El coeficiente de determinación múltiple de la regresión es de 0.26.

			PLAR (‘000)	*PTOT* (‘000)	*PURB* (‘000)	*GNPC* ($ US)
7	Brasil	Sao Paulo	17395	150368	112643	2680
13	Costa Rica	San José	1016	3015	1420	1900

A partir de estos datos, el modelo predice que, en 1980, la población de Sao Paulo era de, en miles:

$$3500 + (0.01 \times 150368) + (0.1 \times 2680) = 5272$$

y la población de San José,

$$3500 + (0.01 \times 3015) + (0.1 \times 1900) = 3720$$

Es fácil darse cuenta que estas predicciones son de muy mala calidad. La diferencia con los valores observados es de 12 123 en el primer caso y de –2704 en el segundo. Es todavía prematuro concluir que el modelo no sirve con sólo dos observaciones. Sin embargo, podemos sospechar que la relación lineal no es la más adecuada para el fenómeno que se estudia.

En el ejemplo anterior, el modelo cuenta con tres parámetros. Se asocia a cada parámetro una variable independiente. El parámetro β_1 es la constante del modelo, es decir que su variable independiente asociada es una constante. Así que si quisiéramos ser totalmente explícitos, tendríamos que presentan los datos de la forma siguiente:

			Constante	*PLAR* ('000)	*PTOT* ('000)	*PURB* ('000)	*GNPC* ($ US)
7	Brasil	Sao Paulo	1	17395	150368	112643	2680
13	Costa Rica	San José	1	1016	3015	1420	1900

Ahora, se escribe el cálculo de las “predicciones” como sigue:

$$(3500 \times 1) + (0.01 \times 150368) + (0.1 \times 2680) = 5272$$

$$(3500 \times 1) + (0.01 \times 3015) + (0.1 \times 1900) = 3720$$

Como los demás, se multiplica, entonces, el parámetro β_1 por el valor de la variable correspondiente. Es importante

siempre tener en claro que la constante es una de las variables del modelo, particularmente cuando se cuenta el número de variables independientes (en este caso, tres). Es de igual manera importante cuando el modelo cuenta con variables independientes dicotómicas (nombradas variables mudas) con el fin de no introducir redundancias en el modelo (vea capítulo 4-2).

Nota:

Algunos autores escriben

$$y_i = \sum_{j=0}^{h} \beta_j x_{ij} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_h x_{ih}$$

Es la razón por la cual al momento de dar el número de variables independientes, es necesario precisar si incluye o no la constante (hay $h + 1 = k$ variables contando la constante). Este detalle es importante al momento de contar el número los grados de libertad que se asocian a algunas variables-test. En este trabajo se incluye siempre la constante en el número de variables independientes (que indicamos con una k por lo general).

Cuando el modelo no tiene más que dos variables independientes incluyendo la constante, se trata de la regresión lineal simple:

$$y_i = \alpha + \beta x_i$$

Se estudiará aquí solamente el caso general de la regresión lineal múltiple cuyo caso particular es la regresión simple.

3-1.1.2 La representación de las relaciones no lineales en el modelo lineal

El modelo lineal general permite representar relaciones no lineales siempre y cuando sean lineales con relación a los parámetros o bien linealizables. Algunos ejemplos nos ayudarán a ilustrar lo que esto significa.

Ejemplo 1: la transformación logarítmica:

La relación exponencial

$$PLAR_i = K\, PURB_i^h$$

es lineal cuando tomamos los logaritmos:

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

Por lo tanto, en estas condiciones, las variables del modelo ya no son *PLAR* y *PURB* sino más bien ln *PLAR* y ln *PURB*.

Lemelin y Polèse (1995) estimaron los parámetros de esta relación; se efectuaron los cálculos con los logaritmos neperianos.[141] Se presentan los resultados en la tabla 2 del artículo: ln K = 2.067 y h = 0.636. A continuación, damos los valores redondeados para Brasil y Costa-Rica en 1990 que se calcularon a partir de la tabla 1 de Lemelin y Polèse (1995):

			ln *PLAR* (‘000)	ln *PURB* (‘000)
7	Brasil	Sao Paulo	9.76	11.63
13	Costa Rica	San José	6.92	7.26

[141] Como es posible observar al momento de aplicar los logaritmos neperianos en la relación exponencial. El valor estimado del parámetro h no tiene influencia de la selección de la base de los logaritmos (el número trascendental e para los logaritmos neperianos o 10 para los logaritmos comunes). La constante estimada, Log K o ln K, depende, sin embargo, de la selección de la base.

A partir de estos datos es posible calcular que el modelo "predice" que, en 1990, la población de Sao Paolo era de, en miles:

$$\text{EXP}[2.067 + (0.636 \times 11.63)] = \text{EXP}(9.46) = 12883$$

y aquella de San José,

$$\text{EXP}[2.067 + (0.636 \times 7.26)] = \text{EXP}(6.68) = 800$$

La linealización de un modelo con su transformación logarítmica es un procedimiento frecuente. Se examinó anteriormente al momento del ajuste de una curva de tendencia (vea 1-2.3):

$$y_t = y_0\,(1+r)^t$$

llega a ser

$$\log y_t = \log y_0 + t \log(1+r)$$

No obstante, en este caso el exponente t es una de las dos variables independientes del modelo (la otra es la constante) y log yt es la variable dependiente mientras que y0 y log(1+ r) son los parámetros que se pretende estimar.

En economía se aplica la transformación logarítmica a la función de producción Cobb-Douglas, que se define con:

$$Y_i = A\,K_i^B\,T_i^C$$

donde:

Y es la cantidad producida;

K es la cantidad de capital empleado;

T es la cantidad de mano de obra empleada.

A, B y C son los parámetros.

Al momento de aplicar la transformación logarítmica, el modelo llega a ser lineal:

$$\log Y_i = \log A + B \log K_i + C \log T_i$$

Ejemplo 2: el añadido de variables independientes:

La relación

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i + d (\ln GNPC_i)^2$$

encierra tres variables independientes (incluyendo la constante) pero una de ellas aparece, al mismo tiempo, en forma lineal y en forma cuadrática. No obstante, es posible tratar esta relación como si fuera una relación lineal. Para esto, sólo basta considerar que GNPC y (ln GNPC)2 son dos variables diferentes. Incluyendo la constante, el modelo cuenta, entonces, con cuatro variables independientes.[142]

Obviamente, se puede generalizar este procedimiento. Así, la relación cúbica

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3$$

llega a ser lineal cuando se define

$$z_{i1} = 1 \text{ (constante)}, z_{i2} = x_i, z_{i3} = x_i^2 \text{ et } z_{i4} = x_i^3$$

Se puede escribir, entonces, el modelo en la forma de una relación lineal:

$$y_i = \beta_1 + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} = \sum_{j=1}^{4} \beta_j z_{ij}$$

Este procedimiento permite, también, linealizar polinomios de cualquier grado, particularmente útiles para estimar superficies de tendencias. La estimación de superficie de tendencia puede servir para describir las variaciones en el espacio de los valores de una variable como, por ejemplo, el precio de las casas. Es efectivamente un modelo descriptivo,[143] puesto que la relación no se basa en ninguna teoría. Los

[142] Se puede comparar este procedimiento con el uso, en el teatro o en el cine, de más de un actor para representar un mismo personaje en edades diferentes.

[143] Del mismo modo que el ajuste de una curva de tendencia temporal es un modelo descriptivo. A menudo se usa un modelo descriptivo como complemento de un modelo teórico (vea, en particular, los trabajos de Francois Desrosiers y Marius Thériault de la Universidad Laval sobre los precios inmobiliarios en la región de Quebec).

datos que se requieren son los precios de venta de los inmuebles y su localización, en coordenadas XY[144]. Un polinomio de segundo grado se define como

$$Z_i = \beta_0 + X_i\,\beta_1 + Y_i\,\beta_2 + X_i^2\,\beta_3 + Y_i^2\,\beta_4 + X_i\,Y_i\,\beta_5$$

Este modelo cuenta con seis variables independientes: la constante, X_i, Y_i, X_i^2, Y_i^2, y X_iY_i.. En un polinomio de tercer grado, tendríamos las cuatro variables suplementarias siguientes: X_i^3, Y_i^3, $X_i^2Y_i$ y $X_iY_i^2$. Cuanto más alto es el grado del polinomio, más compleja puede ser la superficie que describe. Pero, por otro lado, más alto está el número de variables independientes. Veremos cómo el número de parámetros que se puede estimar es limitado por el número de observaciones.

3-1.2 ¿CUÁNDO INTERVIENE LO ALEATORIO?

En el análisis de regresión se busca conocer los parámetros de la relación entre los y_i y los x_{ij}. Ahora bien, si el modelo teórico determinista fuera verdadero, entonces cada observación se conformaría con exactitud al modelo; en estas condiciones, para conocer sus parámetros β_j, sólo bastaría recolectar, con relación a y_i y los x_{ij}, tantas observaciones como haya parámetros y resolver un sistema de k ecuaciones (una para cada observación i) con k desconocidas, o sea los β_j).[145]

---

[144] En un sistema de información geográfica (SIG) se registra la situación de los objetos en el espacio en la forma de coordenadas como la posición de un punto en el plano cartesiano; estas coordenadas son, a veces, la latitud y la longitud geográficas de la posición pero no necesariamente.

[145] Se emplean métodos similares en ciertas circunstancias. Se conoce, entonces, más bien como una "calibración" del modelo en lugar de una "estimación".

Así, en física, la velocidad v de un objeto en caída libre es igual al tiempo transcurrido t multiplicado por la constante de aceleración a[146].

$$v = at$$

De esta relación, se infiere que la distancia recorrida d es proporcional al cuadrado del tiempo de caída:

$$d = \frac{1}{2} a t^2$$

Puesto que la ley de aceleración de la gravedad es determinista, sólo basta una sola observación precisa del cuerpo en caída para conocer el valor de la constante *a*. Al medir *d* y *t*, es fácil calcular el valor de *a*.

De manera idéntica, si el modelo

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

fuera exacto, los valores observados para Brasil y Costa-Rica permitirían definir un sistema de dos ecuaciones lineales de dos desconocidas:

$$9.76 = \ln K + 11.63\ h \text{ (Brasil)}$$

$$6.92 = \ln K + 7.26\ h \text{ (Costa Rica)}$$

La solución de este sistema es

$$\ln k = 2.20$$

$$h = 0.65$$

¡Estaríamos en la gloria! Sin embargo, estamos concientes que los modelos, y particularmente en ciencias sociales, son demasiados simples, aún más si son lineales, para representar toda la complejidad de lo real. Nuestros modelos teóricos no son más que aproximaciones y, aunque sean buenos, solo de manera aproximativa, las observaciones se conforman a ellos. De modo que, si estimáramos los k parámetros con la ayuda de k observaciones, sería muy probable que una nueva observación (k + 1) fuera incompatible con el modelo (desde un

[146] La constante de aceleración de la gravedad es igual a 980.621 cm/s^2, o sea 32.1725 pies/s^2, al nivel del mar.

punto de vista determinista); dicho de otra manera, la $(k + 1)^{iesima}$ ecuación sería contradictoria con las demás.[147]

Por ejemplo, añadamos a las observaciones efectuadas sobre Sao Paolo y San José los datos relativos a Toronto (calculados a partir de la tabla 1 de Lemelin y Polèse, 1998):

			ln PLAR ('000)	ln PURB ('000)
7	Brasil	Sao Paulo	9.76	11.63
9	Canada	Toronto	8.15	9.93
13	Costa Rica	San Jose	6.92	7.26

Si aplicamos a Toronto los coeficientes que se calcularon anteriormente, obtenemos:

$$\ln PLAR = 2.20 + 0.65 \ln PURB$$

$$\ln PLAR = 2.20 + (0.65 \times 9.93) = 8.65 \neq 8.15$$

No se verifica la ecuación en el caso de Toronto. De hecho, sabemos que no existe solución para el sistema siguiente de tres ecuaciones y dos desconocidas:

$$9.76 = \ln K + 11.63\ h \text{ (Brésil)}$$

$$8.15 = \ln K + 9.93\ h \text{ (Can.)}$$

$$6.92 = \ln K + 7.26\ h \text{ (Costa Rica)}$$

Generalizando, con una muestra de n observaciones y k parámetros para estimar (uno para cada variable independiente), es posible construir un sistema de n ecuaciones con k desconocidas. En caso que el número de observaciones sea superior al número de parámetros para estimar, el número de ecuaciones es, entonces, superior al número de desconocidas.

[147] En las ciencias llamadas exactas como la física, es común enfrentarse con un problema muy similar. Los errores de medición introducen un elemento de inexactitud en las observaciones con que, lo mismo cuando los modelos son leyes "deterministas", subsiste un cierto grado de imprecisión con relación a los valores de los parámetros (como en el caso de la constante de aceleración de la gravedad).

Ahora bien, usualmente, un sistema de este tipo no tiene solución porque las ecuaciones son incompatibles entre sí.

Por lo tanto, aunque un modelo sea una buena aproximación de la realidad, subsiste, no obstante, una diferencia entre las predicciones del modelo y las observaciones. La ausencia entre las variables independientes de numerosos factores secundarios cuya influencia es pequeña (modelo incompleto y demasiado simple), es parte de la explicación de esta diferencia. Esta situación se traduce por un "error" que no parece ser sistemático sino, más bien, fruto del azar. Con el fin de tomar en cuenta este error, se añade una variable aleatoria en el modelo teórico:

$$y_i = \sum_{j=1}^{k} \beta_j \, x_{ij} + u_i$$

El término aleatorio u_i es, también, conocido como término de error, error estocástico o simplemente error o bien perturbación (disturbance term). Es importante entender que los valores que tome el término aleatorio son igual de inobservables que los parámetros de la relación, todo lo que podemos observar son los valores de la variable dependiente y las variables independientes.

Por ejemplo, el modelo

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

es un modelo teórico determinista. Sin embargo, el modelo en el cual se basa la estimación de los parámetros y los tests de hipótesis que aparecen en Lemelin y Polèse (1995) es, en realidad,

$$\ln PLAR_i = \ln K + h \ln PURB_i + u_i \,,$$

donde u_i es un término aleatorio.

Así, acabamos de descubrir la tercera "puerta" por la cual se introduce lo aleatorio en el análisis de regresión. Describimos estas tres puertas como sigue:

[...] existen tres 'puertas' por las cuales se introduce lo aleatorio en los modelos:

1. Para empezar, existe la naturaleza aleatoria que ya mencionamos del vínculo entre una muestra y la población de donde se obtuvo.
2. Las variables operacionales son medidas imperfectas de los conceptos y se puede considerar que el error de medición es aleatorio (o sea que se determina al azar). Es posible por lo tanto, representar con un modelo aleatorio la influencia de los errores de medición que intervienen al momento de traducir las hipótesis teóricas en hipótesis operacionales (entre los primeros modelos aleatorios, justamente hay que mencionar los modelos de la "teoría de los errores" en ciencias físicas).
3. Finalmente, algunos fenómenos son, por naturaleza, aleatorios y no pueden representarse adecuadamente con modelos teóricos no aleatorios. En estos modelos, el azar es el reflejo de, por un lado, una indeterminación fundamental (como en física de las partículas) y, por el otro, una multitud de factores inobservables (como suele suceder en ciencias sociales[148]) cuyas manifestaciones aparecen como reglas gracias a leyes de probabilidades". (Capítulo 2.2).[149]

[148] Pensemos, en particular, en los modelos de utilidad aleatoria (random utility) subyacentes a los modelos de selecciones discretas (discrete choice) logit, probit, etc. Vamos a encontrar este tipo de modelos en el apartado 4-3.

[149] Este pasaje es inspirado de Malinvaud, quien escribe: "Se sabe que se justifica el uso del cálculo de las probabilidades para el análisis de los datos de estadística con una u otra de las dos consideraciones siguiente. O bien, se asimila el fenómeno estudiado como un proceso que encierra una determinación aleatoria de algunas magnitudes; en este caso, se consideran, entonces, las magnitudes como aleatorias en el universo (NDLA: o sea en la población) así como en la muestra observada. O bien, la selección de las unidades observadas es el resultado de un sorteo aleatorio; la composición de la muestra es, entonces aleatoria y, por consiguiente, los datos obtenidos también aunque sean datos sobre magnitudes no aleatorias" (Malinvaud,

Según esta concepción, aunque los datos englobaran la totalidad de la población estudiada, el elemento aleatorio no desaparecería puesto que el aspecto aleatorio se debe no tanto por la relación entre la población y la muestra sino más bien, por la relación entre el modelo determinista (la ley matemática), cuyos parámetros son desconocidos, y las observaciones las cuales se alejan del modelo de manera aleatoria:[150] así, las observaciones dejan de ser incompatibles con el modelo para ser simplemente, desde el enfoque de la probabilidad, más o menos compatibles con el modelo. Agreguemos, sin embargo, que, en este contexto, la distinción entre población y muestra sigue subsistiendo pero esta distinción vale, primeramente, por el hecho que los valores que toman los términos aleatorios inobservables se sortean de la población infinita de los valores que el proceso aleatorio subyacente a cada uno de los términos aleatorios podría generar. Esto último nos permite entender que es posible que se engendren los valores de los términos aleatorios asociados a diferentes observaciones gracias a procesos aleatorios diferentes. Es para ilustrar esto que, en algunos contextos, se mencionan los términos aleatorios en plural.

Por lo tanto, en un primer nivel, la combinación de un término aleatorio y un modelo determinista permite acomodarse con el carácter aproximativo del acuerdo entre el modelo y las observaciones. Para ir más allá, es necesario caracterizar las distribuciones de probabilidad de los términos aleatorios u_i. Tendremos, entonces, un modelo aleatorio y se-

---

1969, p. 62). Malinvaud prosigue diciendo que el primer tipo de justificación le parece más apropiado en el contexto de la econométría.

[150] Hay algo de la caverna de Platón en esta concepción (ver en el anexo el texto de la alegoría). Tratamos con la realidad observable como si fuera el reflejo imperfecto (la sombra proyectada) del modelo teórico determinista (lo ideal). El término aleatorio del modelo representa las imperfecciones de la realidad observable. La inducción estadística busca discernir lo "ideal" (en el sentido que le daba Platón a esta palabra) a través de su reflejo.

remos capaces de aplicar los métodos de inducción estadística con el fin, en particular:

- de estimar los parámetros de las distribuciones de probabilidad de los términos aleatorios;
- de estimar los parámetros de las distribuciones de muestreo de los estimadores;
- de efectuar unos tests de hipótesis.

3-1.3 EL ESTIMADOR DE LOS MINIMOS CUADRADOS ORDINARIOS

Para complementar la simbología empleada, se conviene lo que sigue:

b_i: valor estimado del parámetro β_i.

$\hat{y}_i$: valor de y_i "predicho" o calculado por el modelo tal como se estimó.

Tenemos por definición:

$$\hat{y}_i \equiv \sum_j b_j x_{ij}$$

e_i: residuo calculado (o "error") de la regresión para la i^{esima} observacion

Tenemos por definición:

$$e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$$

NB: No se debe confundir e_i, el residuo calculado (observable), con el término aleatorio correspondiente u_i inobservable.

3-1.3.1 Definición

Aunque no se haya complementado la especificación del modelo aleatorio, es posible aplicar el método de los mínimos

cuadrados (vea el enunciado de este principio en el apartado 2-2.3). Sólo es necesario reconocer que el modelo no es más que una aproximación y que las observaciones no se conforman más que aproximadamente a él.

El principio de los mínimos cuadrados consiste en escoger los valores estimados b_j que minimizan la suma de los cuadrados de los residuos (o "errores"). Esto significa minimizar la suma de los cuadrados de las diferencias entre los valores observados de y_i y los valores "predichos" $\hat{y}_i$:

$$\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i \left(y_i - \sum_j b_j \, x_{ij} \right)^2 = \sum_i e_i^2$$

La expresión $\sum_i \left(y_i - \hat{y}_i\right)^2$ es, por lo tanto, el cuadrado de la distancia euclidiana generalizada entre los valores observados de la variable dependiente y los valores predichos. Se define la solución de este problema de minimización con el estimador de los mínimos cuadrados ordinarios.

Se presentan, con frecuencia, los resultados de la estimación en unas tablas: vea las tablas 2 y 4 de Lemelin y Polèse (1995), la tabla 1 de Heikkila et al. (1989) o la tabla 1 de Richardson et al. (1990).

3-1.3.2 Algunas propiedades del estimador de los mínimos cuadrados ordinarios

Estimador lineal

Este estimador es lineal, es decir que se calcula cada b_j como una función lineal de los y_i, o con más exactitud, como una suma ponderada de los y_i:

$$b_j = \sum_i w_{ji}\, y_i$$

donde cada uno de los coeficientes w_{ji} depende del conjunto de los xgh.[151]

Suma de los residuos nula

Cuando el modelo de regresión tiene una constante, como en la mayoría de los casos, la suma de los residuos la regresión es nula:

$$\sum_i e_i = 0$$

No se exhibe aquí la demostración porque se necesita, para el efecto, la escritura matricial.

Relación entre los promedios

Cuando el modelo tiene una constante, el promedio de los valores predichos es igual al valor predicho a partir de los valores promedios de las variables independientes y estos dos valores son iguales al valor promedio observado de la variable dependiente:

$$m_y = m_{\hat{y}} = \sum_j b_j\, m_{x_j}$$

Se deduce esta propiedad de la previa.

Demostración:

Sabemos que $\sum_i e_i = 0$

[151] Para ser más preciso, w_{ji} es el elemento j,i de la matriz $(X'X)^{-1}X'$. Observe que el hecho que el estimador sea lineal no es una consecuencia de que el modelo sea lineal.

Ahora bien $e_i \equiv y_i - \hat{y}_i \equiv y_i - \sum_j b_j x_{ij}$

Esto implica, en particular, que la suma de los valores "predichos" es igual a la suma de los valores observados:

$$\sum_i y_i - \sum_i \hat{y}_i \equiv \sum_i (y_i - \hat{y}_i) \equiv \sum_i e_i \equiv 0$$

$$\sum_i y_i = \sum_i \hat{y}_i$$

Pero, puesto que

$$\sum_i \hat{y}_i \equiv \sum_i \left(\sum_j b_j x_{ij} \right) \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

entonces

$$\sum_i y_i = \sum_i \hat{y}_i$$

implica que

$$\sum_i y_i \equiv \sum_i \hat{y}_i \equiv \sum_j b_j \left(\sum_i x_{ij} \right)$$

$$\left(\frac{1}{n}\right) \sum_i y_i \equiv \left(\frac{1}{n}\right) \sum_i \hat{y}_i \equiv \left(\frac{1}{n}\right) \sum_j b_j \left(\sum_i x_{ij} \right)$$

$$\left(\frac{1}{n}\right) \sum_i y_i \equiv \sum_j b_j \left[\left(\frac{1}{n}\right) \sum_i x_{ij} \right]$$

Ahora bien

$$m_y = \left(\frac{1}{n}\right) \sum_i y_i$$

$$m_{\hat{y}} = \left(\frac{1}{n}\right) \sum_i \hat{y}_i$$

$$m_{x_j} = \left(\frac{1}{n}\right) \sum_i x_{ij}$$

Por lo tanto, tenemos

$$m_y = m_{\hat{y}} = \sum_j b_j \, m_{x_j}$$

3-1.4 EL COEFICIENTE DE DETERMINACIÓN MÚLTIPLE Y EL ANÁLISIS DE LA VARIANZA

3-1.4.1 Construcción del coeficiente de determinación múltiple

El coeficiente de determinación múltiple es una medición de asociación que pertenece a la familia de las mediciones de similitud; con más precisión, es una medición del grado de acuerdo entre el modelo y las observaciones. En estadística, una medición de este tipo se llama "medición de ajuste" (goodness of fit measure).

El coeficiente de determinación múltiple se basa en un análisis de descomposición[152] de la variabilidad de la variable dependiente donde se calcula esta variabilidad con la suma de los cuadrados de las desviaciones con relación a la media:

$$\sum_i \left(y_i - m_y\right)^2 = (n-1)\, s_y^2$$

En estadística, este tipo de análisis de descomposición se llama una "análisis de varianza".[153]

---

[152] Sobre el análisis de descomposición, vea 1-2.2.

[153] Hay una forma más especializada de análisis de varianza que permite examinar la relación entre una variable dependiente y varias variables independientes categóricas, descomponiendo la varianza de la variable dependiente entre la varianza dentro los grupos (definididos por combinaciones de categorías de las variables independientes) y la varianza entre los grupos. Abordaremos este tema en el apartado 4-2.

Primera etapa: descomposición de la variabilidad

Cuando el modelo tiene una constante, es posible descomponer la variabilidad en dos componentes:[154]

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2$$

Principio de demostración:

$$\sum_i \left(y_i - m_y\right)^2 = (n-1)\, s_y^2$$

Si desarrollamos el miembro de la izquierda de esta expresión, obtenemos:

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left[\left(y_i - \hat{y}_i\right) + \left(\hat{y}_i - m_y\right)\right]^2$$

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2 + 2\sum_i \left(y_i - \hat{y}_i\right)\left(\hat{y}_i - m_y\right)$$

Se puede mostrar que, si el modelo tiene una constante, el último término es nulo[155], de tal manera que:

$$\sum_i \left(y_i - m_y\right)^2 = \sum_i \left(y_i - \hat{y}_i\right)^2 + \sum_i \left(\hat{y}_i - m_y\right)^2$$

[154] En caso de que no haya constante, la descomposición ya no es válida. Puede hasta ocurrir que R^2 sea negativo.

[155] Se requiere de la escritura matricial para esta demostración.

Segunda etapa: interpretación de los elementos de la descomposición

En la expresión $\sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - m_y)^2$, el segundo término es una medición de la variabilidad de los valores predichos por el modelo: es la parte "explicada" de la variabilidad. El primer término es una medición de la variabilidad de los residuos: es la parte de la variabilidad que el modelo no prevé. Por lo tanto, tenemos:

Variabilidad total	=	Variabilidad residual	+	Variabilidad "explicada"

De esta interpretación, se infiere la simbología siguiente la cual se usa frecuentemente en las salidas de los paquetes de aplicaciones estadísticas:[156]

- *SST*: Suma de los Cuadrados Totales (*Sum of Squares Total*)
  $$= \sum_i (y_i - m_y)^2 = (n-1)\, s_y^2$$
- *SSM*: Suma de los Cuadrados del Modelo (*Sum of Squares Model*),
  ya que, cuando existe una constante, $m_y = m_{\hat{y}}$,
  $$= \sum_i (\hat{y}_i - m_{\hat{y}})^2 = \sum_i (\hat{y}_i - m_y)^2 = (n-1)\, s_{\hat{y}}^2$$

[156] No obstante, tenga cuidado porque es posible encontrar la simbología siguiente: *SSR* por *Sum of Squares Regression* en lugar de *SSM* y *SSE* por *Sum of Squares Errors* en lugar de *SSR*.

- *SSR*: Suma de los Cuadrados de los Residuos (*Sum of Squares Residuals*)

$$= \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$$

En resumen:

Variabilidad total		Variabilidad residual		Variabilidad "explicada"
SST		*SSR*		*SSM*
Sum of Squares, Total		Sum of Squares, Residuals		Sum of Squares, Model
$\sum_i \left(y_i - m_y\right)^2 = (n-1)s_y^2$	=	$\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$	+	$\sum_i \left(\hat{y}_i - m_{\hat{y}}\right)^2 = \sum_i \left(\hat{y}_i - m_y\right)^2 = (n-1)s_{\hat{y}}^2$

Tercera etapa: construcción de una medición de ajuste ("goodness of fit")

El coeficiente de determinación múltiple es la parte de la variabilidad "explicada" en la variabilidad total:

$$R^2 = \frac{\text{Variabilidad «explicada»}}{\text{Variabilidad total}} = \frac{SSM}{SST}$$

O bien, puesto que $SST = SSR + SSM$, tenemos $SSM = SST - SSR$ y

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum_i e_i^2}{(n-1)s_y^2}$$

Recordemos que el método de los mínimos cuadrados consiste en tomar, como valores estimados de los parámetros, los valores de los coeficientes que minimizan $\sum_i e_i^2$. A la

luz de la formula enunciada arriba, entendemos que usar el método de los mínimos cuadrados es lo mismo que escoger los valores de los coeficientes que maximizan R^2 bajo la especificación, es decir en el marco del modelo seleccionado.[157]

> El valor del coeficiente de determinación múltiple es, por lo general, presentado en las tablas de resultados: vea las tablas 2 y 4 de Lemelin y Polèse (1995).

3-1.4.2 Campo de variación del coeficiente de determinación múltiple (valores extremos)

El coeficiente de determinación varía entre cero y uno. En efecto, matemáticamente hablando, *SST*, *SSM* y *SSR* son sumas de cuadrados y, por consiguiente, su valor no puede ser negativo. Además, *SST* = *SSR* + *SSM*, lo que nos permite deducir que ni *SSR*, ni *SSM* pueden exceder el valor de *SST*. Finalmente, puesto que $R^2 = \frac{SSM}{SST}$, se deduce de lo anterior que el coeficiente de determinación no puede ser inferior a cero o superior a uno. Examinemos ahora en que circunstancias R^2 podría alcanzar estos valores extremos.

El coeficiente de determinación es igual a uno cuando SSR = 0, es decir, cuando el modelo reproduce perfectamente las observaciones que sirvieron para estimar los parámetros de este mismo modelo. Es igual a cero cuando SSR = SST, es decir, cuando SSM = 0. Pero, ¿en qué circunstancias es posible tener SSM = 0? Bueno, para empezar, SSM es una suma de cuadrados:

---

[157] Como lo veremos más tarde, esto no es lo mismo que comparar los R^2 de diferentes modelos después de haber estimado los parámetros de cada uno de ellos con el propósito de obtener para cada modelo el R^2 más elevado posible con el método de los menores cuadrados.

$$SSM = \sum_i \left(\hat{y}_i - m_y\right)^2$$

Por lo tanto, sólo es posible tener $SSM = 0$ cuando todos los términos de la suma son nulos, o sea, si $\hat{y}_i = m_y$ para cada observación *i*. ¿Cómo puede ocurrir esto? Es posible mostrar que ocurre esta situación cuando, a partir de un modelo general

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=1}^{k} \beta_j x_{ij}$$

los coeficientes estimados con el método de los mínimos cuadrados son todos nulos con excepción de la constante. En estas condiciones,

$$b_2 = b_3 = ... = b_k = 0$$

y tenemos

$$\hat{y}_i = b_1 = m_y$$

donde m_y es un valor de b1 como se estimó con el método de los mínimos cuadrados ordinarios en este caso.

En resumen, que el coeficiente de determinación sea nulo muestra que no es posible detectar una relación entre la variable dependiente y las variables independientes; de hecho, al momento de estimar los parámetros del modelo, todas las variables independientes desaparecen con excepción de la constante, porque se multiplican por un coeficiente cuyo valor se estima en cero.

¿Es realmente el coeficiente de determinación múltiple una medición de similitud, como lo afirmamos al principio? Para convencerse, solo basta ver que SSR es el cuadrado de la distancia euclidiana generalizada entre el conjunto de los valores observados y el conjunto de valores predichos por el modelo. Es, por lo tanto, una medición de disimilitud. La razón $\frac{SSR}{SST}$ es, por consiguiente, una medición de disimilitud

normada cuyo campo de variación se extiende de cero a uno. Así que $R^2 = 1 - \frac{SSR}{SST} = \frac{SSM}{SST}$ es una medición de similitud cuyo campo de variación se extiende también de cero de uno.

3-1.4.3 Relación entre R^2 y el coeficiente de correlación simple

Es posible mostrar que el coeficiente de determinación, R^2, es igual al cuadrado del coeficiente de correlación simple entre los valores observados y_i y los valores predichos $\hat{y}_i$.

$$r_{\hat{y}y}^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y} \right)^2 = R^2$$

3-1.4.4 Coeficiente de determinación ajustado

Cuando se respetan las hipótesis clásicas (se definen estas hipótesis más abajo), el coeficiente de determinación ajustado

$$\overline{R}^2 = 1 - \frac{n-1}{n-k}\left(1 - R^2\right) = 1 - \frac{SSR \big/ (n-k)}{SST \big/ (n-1)}$$

es un estimador no sesgado del "verdadero" coeficiente de determinación.

El coeficiente de determinación ajustado puede interpretarse como un modo de tomar en cuenta el número de variables independientes en la evaluación del desempeño de un modelo. En efecto, es posible, por lo general, aumentar el coeficiente de determinación R^2 con solo añadir variables independientes en el modelo aunque la presencia de variables suplementarias no se base en una hipótesis teórica.

Podemos observar en la formula del coeficiente de determinación ajustado $\overline{R}^2$ que, al momento de añadir variables, $\overline{R}^2$ puede disminuir a condición que R^2 no aumente lo suficiente para compensar el incremento de k. Existen, sin embargo, otros procedimientos aun más fiables para decidir hasta que punto es oportuno añadir o quitar tal o tal variable: estos otros procedimientos son los tests de hipótesis.

Se presenta, por lo general, el valor del coeficiente de determinación ajustado en las tablas de resultados vea las tablas 2 y 4 de Lemelin y Polèse (1995), la tabla 1 de Heikkila et al. (1989) o la tabla 1 de Richardson et al. (1990).

CAPÍTULO 3-2
LA INDUCCIÓN ESTADÍSTICA APLICADA A LA REGRESIÓN MÚLTIPLE

Hasta el momento vimos cómo estimar los parámetros de un modelo teórico formalizado con una relación lineal. El método de estimación que decidimos usar, a saber el método de los menores cuadrados, consiste en escoger el valor de los parámetros para minimizar los errores de predicción que se cometen al momento de aplicar el modelo a las observaciones, mismas que sirvieron para la estimación.

Si no se fuera más allá, el análisis de regresión no sería más que un modo de resumir las relaciones que se observan en los datos entre las variables, es decir, el análisis de regresión sería entonces una técnica de estadística descriptiva (es, de por sí, uno de los usos legítimos del análisis de regresión). Sin embargo, por lo general, el análisis de los datos tiene como fin descubrir la relación subyacente cuyos parámetros son desconocidos pero que sirve para relacionar las variables entre sí.

Recordemos el contexto en el cual situamos el análisis de regresión. De principio, admitimos que el modelo (determinista) del cual pretendemos estimar los parámetros no es más que una aproximación de la realidad; subsiste una diferencia entre las predicciones del modelo y las observaciones. Se representa este "error" no sistemático con el término aleatorio

de la relación. El objetivo que se pretende es reconocer el valor de los parámetros (coeficientes) de la relación. Pero no es posible observar directamente estos parámetros. Se revelan indirectamente a través de los valores observados de las variables que el modelo relaciona. El problema es que, en las observaciones, la relación se ve afectada a causa del término aleatorio cuyo valor no es más observable que el valor de los parámetros de la relación.

En estas condiciones, sólo los métodos de la inducción estadística pueden permitirnos limitar la incertidumbre que afecta la estimación de los parámetros. Asimismo, la aplicación de estos métodos exige que completemos el modelo aleatorio con asociar al modelo determinista un modelo de muestreo, el cual será un modelo de relación aleatoria entre la muestra y la población. ¿Qué constituye la población y qué constituye la muestra en la regresión múltiple? En este contexto, se considera que el valor (inobservable) del término aleatorio asociado a cada observación es una muestra de tamaño 1 que se sorteó del conjunto infinito de valores posibles que podría tomar el término aleatorio en este caso; este conjunto infinito de valores posibles constituye una población. Por lo tanto, con n observaciones, existen n muestras de tamaño 1 que se sortea de n poblaciones. No se excluye, al principio, que estas n poblaciones sean diferentes entre sí; no se excluye tampoco que sean idénticas (lo que equivale a decir que los valores de los términos aleatorios se sortean dentro de una misma población). ¿Suponemos que las n poblaciones son idénticas? La respuesta forma parte del modelo de muestreo. Sin embargo, el modelo de muestreo se constituye sobre la base de hipótesis sobre las distribuciones de probabilidad de los términos aleatorios. Emitir la hipótesis que estas distribuciones de probabilidad son idénticas equivale a emitir la hipótesis de que las poblaciones son idénticas.

Un poco más tarde examinaremos nuevamente las hipótesis que constituyen el modelo de muestreo de la regresión li-

neal clásica. Para estudiar los procedimientos de inducción estadística, sólo basta, de manera provisoria, saber que en el marco de este modelo de muestreo es posible estimar los parámetros de las distribuciones de muestreo de los estimadores de los parámetros.

1. El estimador b_j del parámetro β_j es no sesgado. En otras palabras, el estimador de los menores cuadrados posee una distribución de muestreo cuyo promedio es igual al valor del parámetro.
2. Existe también un estimador no sesgado de la varianza de muestro $\sigma_{b_j}^2$ de cada uno de los coeficientes estimados b_j, y de la covarianza de muestra de cada par de coeficientes estimados, $\sigma_{b_j b_h}$.

Simbología:

$s_{b_j}^2$ = valor estimado de la varianza de muestro $\sigma_{b_j}^2$.

Esos valores estimados se entregan en los paquetes de estadística.

3-2.1 UNOS EJEMPLOS DE PRUEBAS DE HIPÓTESIS

3-2.1.1 Test bilateral de una hipótesis simple sobre el valor de un coeficiente (test de Student)

Queremos probar una hipótesis simple del tipo:

$$\mathrm{H}_0 : \beta_j = c$$

Por ejemplo, Lemelin y Polèse (1995) estimaron los parámetros del modelo siguiente:

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Este modelo equivale a

$$\ln\left(\frac{PURB}{PTOT}\right) = \ln PURB - \ln PTOT$$

$$\ln\left(\frac{PURB}{PTOT}\right) = \beta_1 + (\beta_2 - 1)\ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Queremos probar la hipótesis de que el coeficiente β_2 es igual a uno:

$$H_0 : \beta_2 = 1$$

¿Porqué esta hipótesis? Porque, en caso que esta hipótesis sea verdadera, esto significa que el grado de urbanización $\frac{PURB}{PTOT}$ es independiente de la población total; en otras palabras, poco importa el tamaño de la población de un país, se determina la fracción de la población que vive en zona urbana con el PIB por cápita. Con un ejemplo concreto, si $\beta 2 = 1$, el modelo predice que el grado de urbanización de China con 1.1 billones de habitantes en 1990 es el mismo que el grado de urbanización de Kenia, que cuenta con 24 millones de habitantes, porque los dos países poseen un PIB por capita de US$370.00 (vea la tabla 1 de Lemelin y Polèse, 1995). ¿Podemos rechazar esta hipótesis?

Recordemos los pasos a seguir para efectuar un test de probabilidad crítico (sin umbral de significación predeterminado – p-value test):

1. Escoger una variable test.
2. Verificar que el modelo de muestreo asociado a esta variable test sea aceptable.
3. Calcular el valor de la variable test.
4. Determinar el valor de la probabilidad crítica correspondiente;

5. Tomar la decisión de rechazar o no la hipótesis según juzguemos que esta probabilidad crítica es lo suficiente pequeña o no (cuanto más la probabilidad crítica es pequeña, menos las observaciones son compatibles con la hipótesis y más el rechazo pude ser categórico).

Vamos a aplicar el test de Student que, en caso de un test de hipótesis simple sobre un coeficiente de regresión, usa la variable test siguiente:

$$t_{n-k} = \frac{b_j - c}{s_{b_j}}$$

donde b_j es el valor estimado del parámetro β_j y s_{b_j} es el valor estimado de la diferencia type de muestreo de b_j. Se puede observar una analogía evidente entre esta variable test y la variable test que se usa para el test de una hipótesis simple sobre un promedio:

$$t_{n-1} = \frac{m_x - \gamma}{\left(s_x \Big/ \sqrt{n} \right)}$$

donde el denominador $\left(s_x \Big/ \sqrt{n} \right)$ es la desviación estándar de muestreo del promedio.

La selección de esta variable test se justifica porque, en las condiciones del modelo clásico de la regresión lineal normal, la variable $\frac{b_j - \beta_j}{s_{b_j}}$ posee una distribución de Student con $n - k$ grados de libertad, cuando n es el número de observaciones y k, el número de variables independientes del modelo (incluyendo la constante).

En el caso que nos interesa, el valor de la variable test se define con:[158]

$$t_{n-k} = \frac{0.971663 - 1}{0.0279321} = -1.0145$$

El número de observaciones n es igual a 64 (Lemelin y Polèse, 1995, tabla 2); el número de variables independientes *k* es igual a 4. Tenemos, por lo tanto, 60 grados de libertad. La probabilidad crítica asociada a este valor para el test bilateral es de 0.314 o 31.4% (esta probabilidad crítica se calculó con la función TDIST del programa Excel;[159] Lemelin y Polèse (1995, p. 322) presentan los resultados de un test equivalente).

Al menos de escoger un umbral de significación muy alto (superior a 0.314), no se puede rechazar la hipótesis que $\beta_2 = 1$. Por tanto, la hipótesis no rechazada tampoco es probada. Sin embargo, es legítimo mantenerla.

3-2.1.2 Test de hipótesis de un coeficiente nulo

Sucede, a menudo, que queramos probar la hipótesis

H_0: $\beta_j = 0$

Por ejemplo, en el modelo

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

el valor estimado del parámetro β_4 que presentan Lemelin y Polèse (1995, tabla 2) es de –0.045. Este valor parece peque-

[158] Los valores que se emplean para el cálculo que sigue provienen directamente de los resultados de computadoras, los cuales se reproducen en el anexo 3-A. Sin embargo, no se encuentran en su totalidad en Lemelin y Polèse (1995).

[159] Note que el valor de la estadística *t* sirve de argumento en la función *TDIST* no puede ser negativo. El usuario debe, por lo tanto, tomar en cuenta que la distribución de Student es simétrica.

ño sin embargo ¿es, "de manera significativa, diferente de cero"? En otras palabras, ¿es posible rechazar la hipótesis de que este coeficiente sea nulo y se pudiera entonces quitar la variable correspondiente?

Para probar este tipo de hipótesis, sólo basta aplicar el test que describimos con anterioridad, pero con $c = 0$. La variable-test toma, entonces, la forma siguiente:

$$t_{n-k} = \frac{b_j - c}{s_{b_j}} = \frac{b_j}{s_{b_j}} \text{ cuando } c = 0$$

En nuestro ejemplo,[160]

$$t_{64-4} = \frac{-0.0453}{0.01345} = -3.368$$

La probabilidad crítica asociada a este valor en un test bilateral con 60 grados de libertad es de 0.0013 o 0.13% (esta probabilidad crítica se calculó con la función TDIST del programa Excel). Con una probabilidad crítica tan pequeña, es casi imposible no rechazar la hipótesis. Diremos que el coeficiente es, de manera significativa, diferente de cero con un umbral de significación de menos de 1%, con más precisión de 0.0013, lo que representa un tanto más que 0.1%.

Es tan común probar este tipo de hipótesis que los paquetes de aplicación lo efectúan automáticamente: es el "*t*" que reportan los logiciales de aplicación estadística. Los logiciales procuran, también, el valor crítico correspondiente.

Las tablas de resultados de los artículos científicos presentan también, una evaluación del grado de significación de cada coeficiente. Lemelin y Polèse (1995) reportan la pro-

[160] Los valores que se emplean para el cálculo que sigue provienen directamente de los resultados de computadoras, los cuales se reproducen en el anexo 3-A. Sin embargo, no se encuentran en su totalidad en Lemelin y Polèse (1995).

babilidad crítica. Richadson et al. (1990) dan el valor del t de Student y Heikkika et al. (1989) presentan los dos.
Algunos autores dan la desviación estándar de cada coeficiente estimado (de por sí, al conocer el valor estimado del coeficiente, es posible calcular su desviación estándar a partir de su estadística t y viceversa, puesto que $t_{n-k} = \frac{b_j}{s_{b_j}}$). Cuando se presenta únicamente la desviación estándar o el valor t de Student, se identifica, por lo general, por medio de llamadas, los coeficientes que son, de manera significativa, diferentes de cero con un umbral de significación de 1%, de 5% o de 10%.

3-2.1.3 Test unilateral de una hipótesis simple sobre el valor de un coeficiente (test de Student)

En algunas ocasiones es pertinente aplicar un test unilateral (vea capítulo 2-3). Por ejemplo, el modelo

$$PLAR_i = K\, PURB_i^h$$

predice que la ciudad más grande de un país crece más rápido o menos rápido que el resto de la población urbana dependiendo de si el valor del exponente, el parámetro h, es superior o inferior a 1 (con la intención de facilitar las referencias al articulo, guardamos, en este momento, la misma simbología). En particular, si $h < 1$, el modelo predice que el peso relativo de la ciudad más grande disminuye a medida que crece la población urbana. ¿Es posible rechazar esta hipótesis de que $h \geq 1$?

Lemelin y Polèse (1995, tabla 2) estimaron los parámetros del modelo una vez después de haber sufrido una transformación lineal

$$\ln PLAR_i = \ln K + h \ln PURB_i$$

El valor estimado de h es 0.636. Se efectúa el test unilateral de la hipótesis H_0: $h = 1$ con, como hipótesis complementaria, H_A: $h < 1$.

La zona de rechazo se sitúa, por lo tanto, a la izquierda, es decir que, para rechazar H_0 y aceptar $1 < h$, es necesario que la diferencia $1 - h$ sea lo suficiente grande como para juzgar muy improbable que $h \geq 1$. La variable test es, nuevamente, el t de Student:[161]

$$t_{64-2} = \frac{h-1}{s_h} = \frac{0.636-1}{0.0426} = -8.54$$

Con $n - k = 62$, la probabilidad crítica unilateral asociada a un valor absoluto tan grande del t de Student es menor que 0.0001.[162] Podemos, pues, decidir con toda confianza rechazar H_0 y aceptar la hipótesis de que la importancia relativa de la ciudad más grande disminuye a medida que la población urbana crece (lo que sorprenderá a más de uno...).

3-2.1.4 Intervalos de confianza y márgenes de error

Es obvio que, como en el caso del promedio, la variable test t_{n-k} puede emplearse también para definir intervalos de confianza del tipo

$$b_j - s_{bj}\ \theta_{n-k}(\alpha) < \beta_j < b_j + s_{bj}\ \theta_{n-k}(\alpha)$$

---

[161] Los valores que se emplean para el cálculo que sigue provienen directamente de las salidas de computadoras, las cuales se reproducen en el anexo 3-A. Sin embargo, no se encuentran en su totalidad en Lemelin y Polèse (1995).

[162] Para $t = 4$, la función *TDIST* del logicial Excel da una probabilidad crítica unilateral de 0.0000857. Abajo de esta probabilidad, la función *TINV* empieza a arrojar resultados aberrantes. Nada nos garantiza que la función *TDIST* nos dé resultados válidos para valores de t superiores a 4. Por lo tanto, es preferible y además suficiente, cerciorarse de que la probabilidad crítica asociada a $t = 8.54$ sea inferior a 0.0001.

con un nivel de confianza de $(1-\alpha)$. El margen de error correspondiente, con el mismo nivel de confianza, es igual a

$$\pm s_{bj}\, \theta_{n-k}(\alpha)$$

Por ejemplo, calculemos un intervalo de confianza del parámetro β_4 en el modelo

$$\ln PURB = \beta_1 + \beta_2 \ln PTOT + \beta_3 \ln GNPC + \beta_4 (\ln GNPC)^2$$

Con $n - k$ grados de libertad y un nivel de confianza de 0.99 (99%), los valores críticos del t de Student son –2.66 y +2.66 [$\theta_{64}(0.01) = 2.66$; valores que se calcularon con la función *TINV* del logicial Excel. Dado que $b_j = -0.0453$ y $s_{b_j} = 0.01345$ el intervalo de confianza de β_4 con 99% se define con:

$$-0{,}0453-(0{,}01345\times 2{,}66) < \beta_4 < -0{,}0453+(0{,}01345\times 2{,}66)$$

$$-0{,}0811 < \beta_4 < -0{,}0095$$

y el margen de error con un nivel de confianza de 99% es igual a

$$\pm\, 0{,}01345 \times 2{,}66 = \pm\, 0{,}358$$

En este momento, se deducen los intervalos de confianza igual como en el caso de un test de hipótesis simple sobre un promedio; el conjunto de hipótesis que no se rechazarían con un nivel de significación de α se define con

$$-\theta_{n-k}(\alpha) < \frac{b_j - c}{s_{b_j}} < +\theta_{n-k}(\alpha)$$

$$-\theta_{n-k}(\alpha)\, s_{b_j} < (b_j - c) < +\theta_{n-k}(\alpha)\, s_{b_j}$$

$$-b_j - \theta_{n-k}(\alpha)\, s_{b_j} < -c < -b_j + \theta_{n-k}(\alpha)\, s_{b_j}$$

$$+b_j + \theta_{n-k}(\alpha)\, s_{b_j} > +c > +b_j - \theta_{n-k}(\alpha)\, s_{b_j}$$

$$b_j - \theta_{n-k}(\alpha)\, s_{b_j} < c < b_j + \theta_{n-k}(\alpha)\, s_{b_j}$$

3-2.1.5 Test de una o varias relaciones lineales entre coeficientes (Test F de Fisher)

El test de Student permite examinar una sola hipótesis al mismo tiempo; además esta hipótesis se emite únicamente con relación a un solo coeficiente. El test de Fisher es mucho más polivalente; de hecho, permite examinar varias hipótesis al mismo tiempo y permite examinar hipótesis que asocian más de un coeficiente. Como en el caso del test de Student, el test de Fisher exige que se respeten las condiciones del modelo clásico de la regresión lineal normal. No detallaremos, en este momento, la mecánica del test de Fisher, sin embargo daremos a continuación algunos ejemplos de su uso.

No obstante, recordemos que la variable test de Fisher se calcula por medio de la formula siguiente:

$$F_{p,n-k} = \frac{\left(\frac{SSR_H}{p}\right)}{\left(\frac{SSR}{n-k}\right)}$$

donde p es el número de las restricciones lineales simultáneas que constituyen la hipótesis y $SSRH$ es la suma de los cuadrados de los residuos que se obtuvieron bajo las restricciones (es decir, cuando estimamos los parámetros del modelo forzándolo a respetar la hipótesis).

De esta manera, Lemelin y Polèse (1995) estimaron los parámetros de los modelos siguientes (para facilitar las referencias al artículo, usamos en este momento la misma simbología):

$$\ln PLAR = p' + q' \ln PTOT + r' \ln GNPC + t' (\ln GNPC)^2 + s \ln PURB$$

$$\ln PLAR = \ln K + h \ln PURB$$

La hipótesis que queremos probar que es aceptable (o sea no rechazado) dejar de lado las variables que están ausentes en la segunda ecuación. De hecho, esta hipótesis la constituyen tres hipótesis simples:

H_1: $q' = 0$

H_2: $r' = 0$

H_3: $t' = 0$

Las probabilidades críticas que podemos encontrar en la tabla 2 de Lemelin y Polèse (1995) permiten concluir que, consideradas una por una, ninguna de estas hipótesis se puede rechazar: para H_1, la probabilidad crítica es de 0.484; para H_2, es de 0.189; para H_3, es de 0.173. Pero, cada uno de estos tres test de hipótesis se basa en un modelo de muestreo donde aparecen las otras dos variables: en el primer caso, por ejemplo, dado que el modelo contiene las variables ln *GNPC* y $(\ln GNPC)^2$, sólo es posible rechazar la hipótesis que $q' = 0$. ¿Qué sucede, entonces, con la hipótesis que los tres coeficientes sean nulos al mismo tiempo? Es justamente este tipo de hipótesis de que el test de Fisher permite examinar. Lemelin y Polèse (1995, p. 323) reportan que la aplicación de este test arroja una probabilidad crítica de 0.53 para la hipótesis que q', r', y t' y son nulos de manera simultánea, por lo tanto no es posible rechazar esta hipótesis.

Consideremos, ahora, la función de producción Cobb-Douglas:

$$Y = A\, K^B\, T^C$$

donde

Y es la cantidad producida;

K es la cantidad de capital empleada;

T es la cantidad de mano de obra empleada;

A, *B* y *C* son los parámetros;

Al momento de aplicar la transformación logarítmica, el modelo se vuelve lineal:

$$\log Y = \log A + B \log K + C \log T$$

Una de las hipótesis que queremos examinar en este modelo es la siguiente:

H_0: $B + C = 1$

Esta hipótesis tiene un especial interés porque, si $B + C = 1$, se trata de una función de producción con rendimientos constantes a la escala. Esto significa que, si aumentamos (o disminuimos) todos los factores de producción de manera proporcional, entonces la producción aumenta (o disminuye) en la misma proporción.

> No es tan complicado demostrar esta propiedad. Teniendo K_0, T_0 y Y_0, los valores iniciales de K, T y Y, y teniendo λ la proporción según la cual aumentamos las cantidades de los factores, entonces
>
> $$Y = A\,(\lambda K_0)^B\,(\lambda T_0)^C = \lambda^{B+C} A K_0^B T_0^C = \lambda^{B+C} Y_0$$
>
> y si $B + C = 1$,
>
> $$Y = \lambda Y_0$$

No es posible aplicar el test de Student a la hipótesis "H_0: $B + C = 1$", porque no se trata de una hipótesis con un parámetro sino, más bien, de una hipótesis sobre una relación entre dos parámetros. Sin embargo, el test de Fisher permite examinar este tipo de hipótesis.

Generalizando y de manera más formal, el test de Fisher permite probar cualquier hipótesis que podamos expresar con una o varias restricciones lineales con relación a los coeficientes. Una restricción lineal con relación a los coeficientes β_1, β_2, β_3, etc. se escribe como sigue:

$$\sum_{j=1}^{k} w_j \, \beta_j = w_1 \, \beta_1 + w_2 \, \beta_2 + \cdots + w_k \, \beta_k = c$$

donde c y los w_j son constantes que el usuario definió en función de la restricción que desea representar.

La hipótesis que más se acostumbra probar con el test de Fisher (pero no forzosamente la más interesante) es:

H_0: $\beta_2 = \beta_3 = ... \beta_k = 0$

Es la hipótesis de que todos los coeficientes de la regresión, menos la constante β_1, sean nulos; se constituye, por lo tanto, esta hipótesis con (k–1) hipótesis simples. Dicho de otra manera, es la hipótesis de que el "verdadero" valor del coeficiente de determinación múltiple R^2 es cero y que el R^2 calculado no es más que la correlación fortuita que los términos aleatorios causaron. Los logiciales de aplicación estadística proporcionan de manera automática el valor de la variable-test asociada a esta hipótesis.

3-2.2 ESPECIFICACIÓN DE UN MODELO DE MUESTREO: LAS CONDICIONES DEL MODELO CLÁSICO DE REGRESIÓN LINEAL NORMAL

Hicimos hincapié de que la validez de los test de hipótesis que acabamos de describir depende de la validez del modelo de muestreo sobre el cual se fundamentan. Por lo tanto les vamos a echar un ojo a continuación.

3-2.2.1 El modelo clásico de la regresión lineal[*]

El modelo de muestreo clásico se constituye de cuatro hipótesis. Estas hipótesis son bastante generales, en el sentido de que imponen pocas restricciones para la forma general de la distribución de probabilidad del término aleatorio.

Las dos primeras hipótesis tratan con los parámetros de las distribuciones de probabilidad de los términos aleatorios:

H1: Para cada observación, el valor del término aleatorio es sorteado de una población teórica de promedio nulo; por consiguiente: $E(u_i) = 0$ para todo i.

H2a): Para todas las observaciones, las poblaciones teóricas de donde se sortean los valores de los términos aleatorios tienen la misma varianza:[163]
$\sigma_i^2 = \sigma^2$ para todo i.

H2b): Para cada observación, el valor del término aleatorio es estadísticamente independiente de los valores de los términos aleatorios de las demás observaciones:[164]
$\sigma_{ij} = 0$ para todas las combinaciones i,j cuando $i \neq j$.

Veremos, con más detalle, lo que significan estas condiciones al momento de examinar lo que sucede cuando no se respetan. La tercera hipótesis circunscribe el papel de lo aleatorio en el modelo:

H3: Las variables independientes x_{ij} son no aleatorias.

---

[*] Referencias: Kennedy (1992, pp. 43-45).

[163] Esta propiedad se llama "homoescedasticidad" (de la palabra griega ομος, igual y σκεδασις, dispersión). Lo contrario es "heteroescedasticidad" (de la palabra griega ετερος, otro).

[164] Esta propiedad se llama la ausencia de autocorrelación.

La hipótesis H3 exige, en particular, que se midan los valores de las variables independientes sin error. Entre las otras situaciones que no son compatibles con esta condición, mencionemos la presencia de los valores atrasados de la variable dependiente del lado de las variables independientes, como sucede, por ejemplo, en un modelo donde la tasa de desempleo de cada mes depende, entre otras cosas, de la tasa de desempleo del mes anterior (un modelo del tipo $C_t = a + bC_{t-1} + \cdots$).

Mencionemos también los sistemas de ecuaciones simultáneas donde la variable dependiente de una ecuación aparece entre las variables independientes de otra (por ejemplo, un modelo donde el PIB depende del consumo y los demás componentes de la demanda mientras que el consumo depende del PIB: $Y = C + X$ y $C = a + bY$). En estas circunstancias, se deben adaptar los métodos con el fin de tomar en cuenta el no respeto de H3.

Finalmente, la cuarta hipótesis se refiere a las relaciones entre las variables independientes y al número de observaciones.

H4: Existe menos parámetros para estimar que observaciones y no hay redundancias entre las variables independientes.[165]

"No hay redundancias" significa, aquí, que ninguna variable independiente puede representarse como una combinación de las demás; es decir, las variables independientes son linealmente independientes entre sí. H4 no es tanto una hipótesis como una condición de aplicación, puesto que se puede determinar con un análisis de datos si esta condición se respeta.

[165] Técnicamente, esto se traduce por la condición que el rango de la matriz X de orden $n \times k$, sea igual a $k < n$.

Al momento de combinar las hipótesis H1 hasta H4 con el modelo lineal general que se expuso en el apartado 1, definen un modelo aleatorio que acostumbramos designar como el "modelo clásico de la regresión lineal". La especificación de este modelo de muestreo es, no obstante, incompleta puesto que el tipo de la distribución de probabilidad de los términos aleatorios no se define. Sin embargo, cuando las hipótesis H1 hasta H4 se respetan, el estimador de los menores cuadrados ordinarios posee varias propiedades deseables. Se demuestran estas propiedades en el teorema de Gauss-Markov.

3-2.2.2 Propiedades del estimador de los menores cuadrados bajo el modelo clásico de la regresión lineal: el teorema de Gauss-Markov

En este momento, nos contentaremos de enunciar, sin demostrarlas, las principales conclusiones del teorema de Gauss-Markov[166]. Estas condiciones se refieren, en particular, a los parámetros de la distribución de muestreo de los coeficientes estimados b_j. Este teorema establece, por lo tanto, los fundamentos de los tests de hipótesis aplicables a los coeficientes estimados b_j .

Cuando las hipótesis H1 hasta H4 se respetan, entonces los resultados del método de los menores cuadrados tienen las propiedades siguientes:

1. El método de los menores cuadrados produce estimadores no sesgados de los parámetros β_j.. En otras palabras, cada uno de los coeficientes estimados b_j posee una distribución de muestreo cuyo promedio (la espe-

[166] El matemático Carl Friedrich Gauss (1777-1855) es el inventor de la distribución normal y del método de los menores cuadrados (en 1794 Gauss aplicó el método por primera vez para la estimación en 1801 de la trayectoria del asteroide Ceres); Andreï Andrelevitch Markov (1856-1922) es, en particular, el autor de un teorema límite central.

ranza matemática) es igual al "verdadero" valor del coeficiente β_j.

2. El método de los mínimos cuadrados produce, también, un estimador no sesgado de σ^2 el cual corresponde a la varianza común de los términos aleatorios. La variable aleatoria

$$\frac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \frac{SSR}{n-k}$$

es un estimador no sesgado de σ^2.

3. El método de los menores cuadrados produce, también, un estimador no sesgado de la varianza de muestreo $\sigma_{b_j}{}^2$ de cada uno de los coeficientes estimados b_j [167] y de la covarianza de muestreo de cada par de coeficientes estimados $\sigma_{b_j b_h}$.[168]

4. En el conjunto de todos los estimadores de los β_j que son lineales y que no son sesgados, el estimador de los mínimos cuadrados b_j es el "mejor", o sea posee la más alta eficacia relativa (con relación a las propiedades deseables de los estimadores, vea el capítulo 2-2). En otras palabras, es el estimador cuya distribución de muestreo posee la más pequeña varianza; es cuando decimos que el estimador de los mínimos cuadrados es BLUE ("Best Linear Unblased Estimate").

3-2.2.3 El modelo clásico de la regresión lineal normal

Vimos que el modelo de muestreo que se define combinando el modelo lineal general con las hipótesis H1 hasta H4 es in-

[167] Los logiciales de aplicación estadística procuran automáticamente estos valores estimados.

[168] La matriz estimada de las varianzas-covarianzas se define con $(X'X)^{-1}\ SSR\ /\ (n-k)$

completo. Así, antes de efectuar cualquier test de hipótesis es necesario complementar la especificación del modelo de muestreo; esto es, tenemos que especificar la forma de la distribución de los términos aleatorios u_i.

H5: Cada uno de los términos aleatorios posee una distribución normal.

Puesto que la distribución normal posee únicamente dos parámetros (el promedio y la varianza), al combinar H1, H2 y H5, se alcanza una especificación casi completa de la distribución de los términos aleatorios:

Los términos aleatorios poseen distribuciones normales idénticas con un promedio nulo; además, son independientes entre sí, es decir, el único parámetro desconocido es su varianza común σ^2.

Basándose en las hipótesis H1 hasta H5, deducimos que se obtienen los valores de los términos aleatorios de una misma población.

Claro está, la selección de la distribución normal es muy cómoda. Para empezar, en estas condiciones, los estimadores poseen también una distribución normal,[169] puesto que una combinación lineal de variables normales es también una variable normal y que los estimadores de los menores cuadrados son lineales con relación a los y_i (y, por lo tanto, con relación a los términos aleatorios u_i). Luego, la distribución normal es de relativo fácil manejo por encerrar únicamente dos parámetros.

Pero, ¿en qué nos basamos al momento de pretender que la distribución de términos aleatorios es, efectivamente, normal? Esta pretensión es, a veces, solamente una implicación del modelo teórico o de la naturaleza del fenómeno estudia-

[169] Es esta propiedad, en particular, la que permite aplicar el test *t* de Student.

do. Sin embargo, la justificación de más rigor se refiere al teorema límite central. Basándose en algunas variantes de este teorema, si el término aleatorio de la regresión representa la influencia combinada de un gran número de variables faltantes en el modelo, entonces se puede considerar que la distribución normal es una aproximación razonable de la distribución de la influencia combinada de las variables faltantes.[170]

La suma de la hipótesis H5 completa la especificación del modelo de muestreo, conocido también como "modelo clásico de regresión lineal normal" (es decir, el modelo clásico de regresión lineal, más la hipótesis de normalidad de los términos aleatorios). Cuando se respetan las hipótesis H1 hasta H5, la estadística facilita al investigador toda una gama de variables-tests que permiten efectuar diversos tests de hipótesis de diferentes tipos. En particular, ya vimos cómo se podía aplicar el test t de Student y el test *F* de Fisher.

No obstante, es responsabilidad del investigador decidir si las condiciones H1 hasta H5 que constituyen el modelo de muestreo son aceptables y si, por consiguiente, los tests que dependen del modelo son válidos.[171] En efecto, al tratar con los tests clásicos, como el test de Stndent del cual dimos algunos ejemplos, el mismo modelo de muestreo no se discute. La decisión de aceptar o no las hipótesis H1 hasta H5 puede basarse en consideraciones a priori. Sin embargo, es posible validarlas luego con, para empezar, un examen visual de los residuos de la regresión y luego, con más rigor, con la aplicación de tests de diagnóstico. Estos tests de hipótesis son, por

[170] Gujarati (1992, p. 93); Theil (1971, p 368-370), Freund (1962, pp. 185-188); Malinvaud (1969, pp. 268-271).

[171] Es la razón de ser de la observación que podemos leer al pie de la p. 19 de Lemelin y Polèse (1995): "Stricly speaking, however, the classical hypotheses under which the tests are exact are not fully realized [...]".

así decirlo, tests de "nivel superior" que se aplican, justamente, a ciertos aspectos del modelo de muestreo.[172]

Resumen: especificación de un modelo aleatorio

Condiciones del modelo clásico de regresión lineal:

H1: Para cada observación, el valor del término aleatorio es sorteado de una población teórica de promedio nulo:
$E(u_i) = 0$ para todo i.

H2a: Para todas las observaciones, las poblaciones teóricas de donde se sortean los valores de los términos aleatorios tienen la misma varianza:
$\sigma_i^2 = \sigma^2$ para todo i.

H2b: Para cada observación, el valor del término aleatorio es estadísticamente independiente de los valores de los términos aleatorios de las demás observaciones:
$\sigma_{ij} = 0$ para todas las combinaciones i,j cuando $i \neq j$.

H3: Las variables independientes x_{ij} son no aleatorias (en particular sus mediciones se efectúan sin errores).

H4: Existen menos parámetros para estimar qué observaciones y no hay redundancias entre las variables independientes.

Propiedades del estimador de los menores cuadrados en el modelo clásico de la regresión lineal: el teorema de Gauss-Markov

1. El estimador de los menores cuadrados de β_j es no sesgado:
   el promedio de la distribución de muestreo de b_j es igual a β_j .

[172] Por ejemplo, Heikkila *et al*. (1989) examinan con la ayuda de tests de diagnósticos el problema de la multicolinealidad espacial entre diversas variables de distancia (p. 228).

2. $\frac{1}{n-k}\sum_i (y_i - \hat{y}_i)^2 = \frac{SSR}{n-k}$ es un estimador no sesgado de σ^2.
3. El método de los menores cuadrados produce, también, un estimador no sesgado de la varianza de muestreo $\sigma_{b_j}^2$ de cada uno de los coeficientes estimados b_j, y de la covarianza de muestreo de cada par de coeficientes estimados $\sigma_{b_j b_h}$.

El estimador de los menores cuadrados es el estimador que posee la más grande eficacia relativa (la más pequeña varianza muestral): es cuando decimos que el estimador de los menores cuadrados es BLUE ("Best Linear Unblased Estimate").

Condición suplementaria del modelo clásico de regresión lineal normal

H4: Existen menos parámetros para estimar qué observaciones y no hay redundancias entre las variables independientes.

3-2.3 ¿SE RESPETAN LAS HIPÓTESIS DEL MODELO DE MUESTREO?¿Y EN CASO CONTRARIO, QUÉ SUCEDE?*

El momento de efectuar los tests de hipótesis que requieren de variables-tests, como es el caso del *t* Student o del *f* de fisher, nunca se cuestiona el modelo de muestreo en sí (vea 3-2.1). Sin embargo, es importante que el analista lo ponga en tela de juicio.

Existen procedimientos estadísticos formales para probar algunos aspectos del modelo de muestreo. No obstante, en muchos de los casos, es posible establecer un diagnóstico preliminar con un simple examen visual del gráfico de los residuos. Por lo general, cualquier motivo geométrico, cual-

* Referencias: Wonnacott y Wonnacott (1992, pp. 524-527)

quier apariencia de organización debería llamarnos la atención.

Los principales problemas que puede revelar un examen de los residuos son:

4. Una inadecuada especificación del modelo teórico.
5. La autocorrelación de los términos aleatorios.
6. La heteroescedasticidad.
7. Observaciones excéntricas.

Examinemos concretamente en qué consiste cada uno de estos cuatro problemas. Luego comentaremos algo sobre la multicolinearidad.

3-2.3.1 Error de especificación del modelo teórico

Especificar un modelo consiste en construir una lista de las variables independientes y definir la forma de la relación entre éstas y la variable dependiente. El error de especificación más frecuente se comete al omitir una de las variables independientes que debería aparecer en el modelo. De la misma manera, se comete un error de especificación en cuanto a la forma, al instituir una relación lineal entre las variables cuando una relación lineal entre sus logaritmos se imponía. Es evidente que al cometer estos tipos de errores de especificación, se compromete, en su totalidad, a las hipótesis del modelo clásico de regresión lineal, empezando por la misma relación lineal.

Existe una multitud de posibilidades de cometer errores de especificación de un modelo. Nos contentaremos con dar una ilustración de lo dicho. Supongamos que el "verdadero" modelo se defina con

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + u_i$$

La variable dependiente y_i es una función de x_i y del cuadrado de x_i. Contando la constante, tenemos tres variables independientes. Aunque no sea lineal, es posible "linealizarlo"

fácilmente con solo considerar x_i y x_i^2 como dos variables diferentes.

Supongamos ahora que estimamos el modelo incompleto:

$$y_i = \alpha_0 + \alpha_1 x_i + u_i$$

El término faltante $\alpha_2 x_i^2$ se manifestará entonces en los residuos. Al momento de examinar el gráfico de la relación entre los residuos y la variable independiente x_i, será posible detectar entre ellos una relación sistemática que tomará la forma de una curva. Se ilustra esta situación con las figuras 6 a 8.

Ilustración geométrica de un error de especificación

Figura 6 – Observasiones y regresiones: Una relación cuadrática

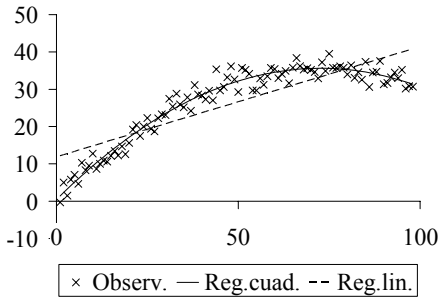


Figura 7 – Residuos de la regresión cuadrática
Residuos sin error de especificación

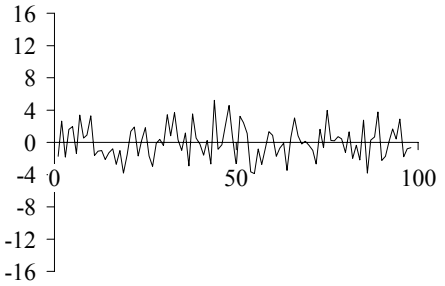
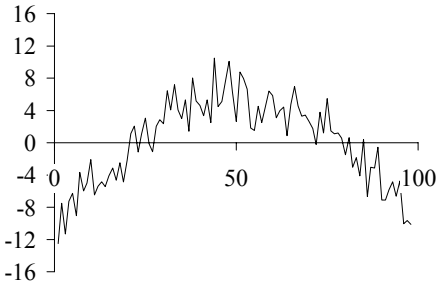


Figura 8 – Residuos de la regresión lineal
Residuos con error de especificación



3-2.3.2 Autocorrelación de los términos aleatorios

Surge una contradicción entre la autocorrelación de los términos aleatorios y la hipótesis H2b del modelo clásico de la regresión lineal, la cual nos indica que los términos aleatorios de las diferentes observaciones son independientes entre sí ($\sigma_{ij} = 0$ para todas las combinaciones i,j donde $i \neq j$).

La auto-correlación es frecuente cuando las observaciones se efectúan en momentos sucesivos en el tiempo. Datos de

esta naturaleza se llaman series temporales o series cronológicas. No obstante, con las series temporales, sucede a menudo que, a causa de cierta inercia, las desviaciones aleatorias necesiten algún tiempo para desaparecer[173]. Así, si u_t, valor del término aleatorio en el periodo t, es positivo entonces el promedio de u_{t+1} *teniendo* $u_t>0$ (la esperanza matemática condicional) no será nula, pero positiva. En estas condiciones, tenemos

$$\sigma_{t,t-1} \neq 0$$

lo que contradice la hipótesis H2b del modelo clásico de la regresión lineal.

A menudo, es posible detectar cierta auto-correlación en las series temporales con examinar la grafica de los residuos en función del tiempo. Podremos, entonces, observar que los errores sucesivos parecen encadenarse los unos con los otros al lugar de efectuar saltos desordenados. Las figuras 9 y 10 ilustran este fenómeno.

Se generaron los datos subyacentes a los residuos de regresiones que se presentan en las figuras 8 y 10 con la ecuación

$$y_t = x_t + 10\ \eta_t$$

donde se generó η_t gracias a un proceso autorregresivo del tipo

$$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$$

[173] Consideremos, por ejemplo, el fenómeno de la evolución de los precios: existe en el funcionamiento de la economía cierta "rigidez institucional" (tal es el caso con los contratos a largo plazo como las convenciones colectivas) la cual provoca que los precios no reaccionen inmediatamente a los cambios en los factores fundamentales. En la relación entre los precios y los factores fundamentales, esto se traduce con la aparición de auto-correlación.

teniendo ε_t una distribución normal. Se repiten las figuras tres veces, cada una con valores diferentes del parámetro α: 0.9, 0.6 y 0.

Es posible encontrar también autocorrelación al manejar datos espaciales. Su detección se complica aún más; en efecto, cuando el tiempo no posee más que una dimensión, el espacio posee dos dimensiones, lo que impide trazar gráficas como se efectuó arriba.

Consecuencias

Los estimadores de los mínimos cuadrados siguen insesgados, sin embargo su varianza es fuerte (los estimadores son menos precisos). Además, las fórmulas que se exhibieron en el caso anterior para estimar la varianza de los estimadores subestiman la verdadera varianza (o sea que nos deja creer que hay más precisión); de aquí que los tests estadísticos ya no sean válidos.

Test de detección

Existe un test de detección de la autocorrelación temporal conocido como el test de Durbin-Watson (vea también Kennedy, 1992, p. 128).

Remedios

En cuanto exista autocorrelación, es necesario aplicar el remedio que consiste en completar el modelo emitiendo hipótesis sobre el mecanismo de autocorrelación, lo que permite usar un método llamado método de los mínimos cuadrados generalizados.

Ilustración geométrica de la autocorrelación
(autocorrelación fuerte)

Figura 9a – Observasiones y regresión:
Autocorelación de los términos aleatorios

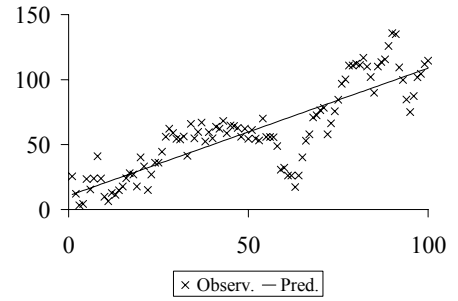
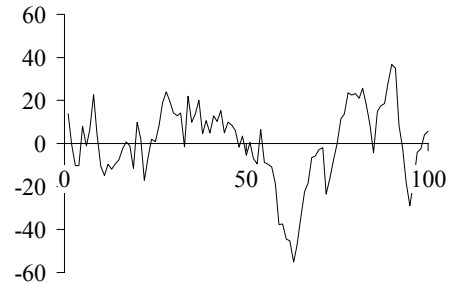


Figura 10a – Residuos de la regresión:
Autocorelación de los términos aleatorios



Términos aleatorios η_t generados con la fórmula
$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$, con $\alpha = 0.9$

Ilustración geométrica de la autocorrelación
(autocorrelación mediana)

Figura 9b – Observasiones y regresión:
Autocorelación de los términos aleatorios

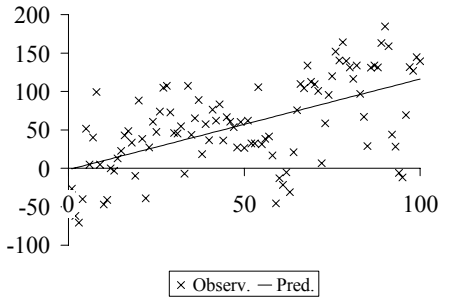
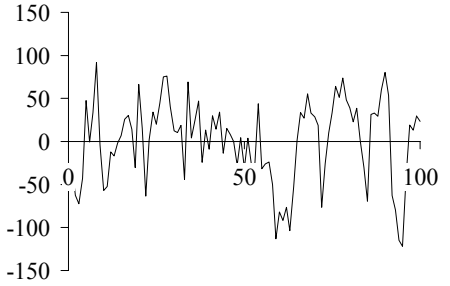


Figura 10b – Residuos de la regresión:
Autocorelación de los términos aleatorios



Términos aleatorios η_t generados con la fórmula
$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$, con $\alpha = 0.6$

Ilustración geométrica de la autocorrelación
(ninguna autocorrelación)

Figura 9c – Observasiones y regresión:
Autocorelación de los términos aleatorios

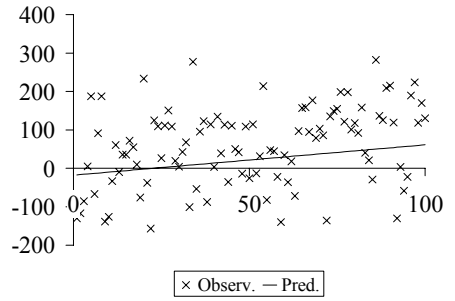
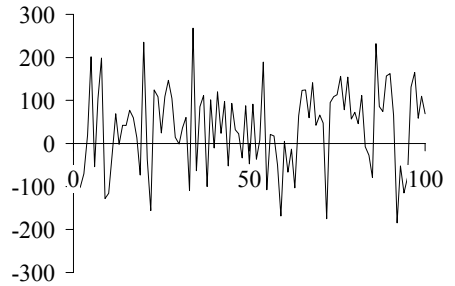


Figura 10c – Residuos de la regresión:
Autocorelación de los términos aleatorios



Términos aleatorios η_t generados con la fórmula
$\eta_t = (1 - \alpha)\ \varepsilon_t + \alpha\ \eta_{t-1}$, con $\alpha = 0$

3-2.3.3 Heteroscedasticidad

La heteroscedasticidad es el contrario de la homoscedasticidad. La homoscedasticidad es la hipótesis H2a del modelo clásico de la regresión lineal, la cual nos indica que la varian-

za del error es la misma para todas las observaciones ($\sigma_i^2 = \sigma^2$ para todos los i)

Se encuentra la heteroscedasticidad a menudo, y particularmente en los estudios en corte transversal donde es posible que los términos aleatorios sean proporcionales al "tamaño" del sujeto observado.

> Por ejemplo, en un estudio sobre los gastos de vivienda de los hogares, es posible que la variabilidad de éstos aumente al mismo tiempo que el ingreso, es decir, los hogares con bajos ingresos se ven obligados a limitar sus gastos de vivienda cuando, por lo contrario, entre los hogares con altos ingresos, algunos optan por usar gran parte de sus ingresos en una vivienda lujosa, y otros prefieren encontrar una vivienda confortable aunque sin grandes lujos, con el fin de gastar su dinero de otra manera. En este particular caso, las variaciones aleatorias debidas a los diferentes gustos[174] son mayores para los hogares más holgados.

En un gráfico de residuos, la heteroscedasticidad podría aparecer como un motivo en forma de trompeta o cono, siempre y cuando ordenemos las observaciones en orden creciente de la variable dependiente o de una de las variables independientes. En lugar de tener en abscisa solamente los números de orden de las observaciones, es posible construir también una gráfica donde se representen los residuos en función de los valores correspondientes de la variable dependiente o de una de las variables independientes en abscisa.

Se generaron los datos subyacentes a los residuos de regresiones que se presentan en las figuras 11 y 12 con la ecuación

$$y_i = x_i + 100\ \eta_i$$

[174] Este es el ejemplo de un modelo al cual le faltan algunas variables inobservables (los gustos) cuyo efecto se representa con el término aleatorio.

donde se generó η_i con la ecuación

$$\eta_i = 0{,}1\left(\varepsilon_i \sqrt{x_i}\right)$$

con ε_i la cual tiene una distribución normal.

Es importante observar que $x_i = i$ de tal manera que las observaciones son automáticamente ordenadas en orden creciente de la variable independiente x_i.

Consecuencias

La precisión del estimador no es tan buena y los tests de hipótesis no son válidos.

Test de detección

Test de Goldfield y Quandt (Theil, 1971, pp. 196-199; vea también Kennedy, 1992, p. 126).

Remedios

Se puede corregir el modelo de muestreo con la transformación de los datos. Por ejemplo, si nos damos cuenta que la varianza depende de una de las variables independientes, digamos x_{ik}, y que es posible representar de manera aproximativa esta relación con

$$\sigma_i^2 = x_{ik}\,\sigma^2$$

entonces es posible recrear la homoscedasticidad y las condiciones de Gauss-Markov con sólo aplicar a las variables la transformación que sigue:

$$y_i' = \frac{y_i}{\sqrt{x_{ik}}} \text{ y } x_{ij}' = \frac{x_{ij}}{\sqrt{x_{ik}}}$$

Ilustración geométrica de la heteroscedasticidad

Figura 11 – Observasiones y regresión: Heteroscedasticidad

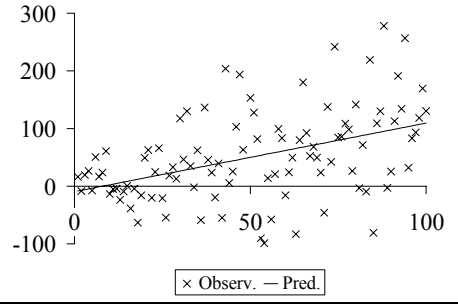
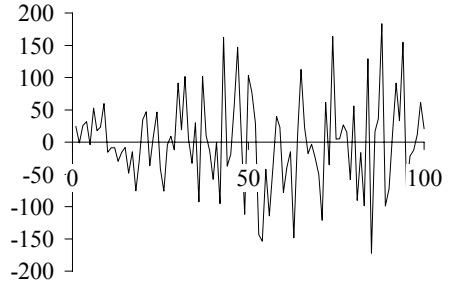


Figura 12 – Residuos de la regresión: Heteroscedasticidad



Términos aleatorios η_i generados con la fórmula

$$\eta_i = 0{,}1\left(\varepsilon_i \sqrt{x_i}\right)$$

3-2.3.4 Observaciones excéntricas

Las observaciones excéntricas (outliers) provienen a veces de situaciones donde factores que no se tomaron en cuenta en el modelo intervinieron. Aunque las observaciones excéntricas no interfieren con las hipótesis del modelo clásico de regresión lineal, es posible que falseen los resultados de la regre-

sión. Un examen de los residuos permite detectar las observaciones excéntricas. Luego, es posible buscar si factores particulares son la causa de tales observaciones y decidir dejarlos a un lado. Sin embargo, es importante evitar eliminar observaciones ad hoc, por simple comodidad... Las figuras que mostramos a continuación nos dan un ejemplo visual de residuos con la presencia de observaciones excéntricas.

Ilustración geométrica de
la presencia de observaciones excéntricas

Figura 13 – Observasiones y regresión:
Observaciones excéntricas (*outliers*)

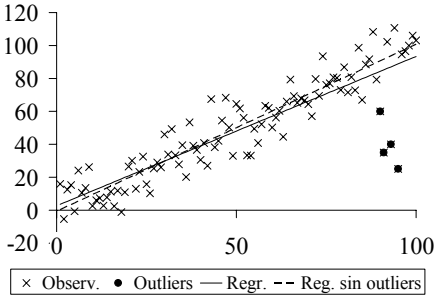


Figura 14 – Residuos de la regresión
con observaciones excéntricas (*outliers*)

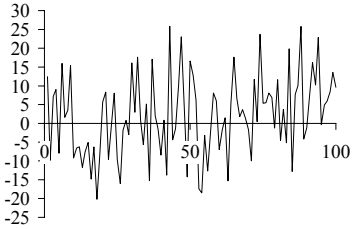
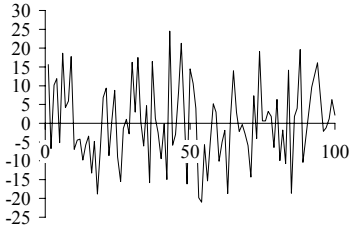


Figura 15 – Residuos de la regresión sin observaciones excéntricas (*outliers*)



3-2.3.5 Multicolinealidad[175]

Definición

Se hace la distinción entre la multicolinealidad estricta y la multicolinealidad aproximativa.

Multicolinealidad estricta. Surge una contradicción entre la multicolinearidad estricta y la hipótesis H4 del modelo clásico de la regresión lineal, la cual nos indica que no debe existir redundancia entre las variables independientes[176]. La multicolinealidad estricta es muy poco común con datos reales cuando puede ser el resultado de un error de especificación si el modelo contiene variables mudas (dummy variables). Trataremos nuevamente este problema cuando estudiaremos el análisis de varianza por medio de la regresión múltiple (capítulo 4-2).

La multicolinealidad estricta no representa ningún problema de detección, puesto que se diagnostica su presencia

[175] Wonnacott y Wonnacott (1992, pp. 568-572).

[176] Técnicamente, cuando existe multicolinealidad estricta, el rango de la matriz X es inferior a k, lo que implica que la matriz inversa $(X'X)^{-1}$ no exista y el estimador tampoco.

con el software de aplicación estadística (por la imposibilidad de los cálculos de estimación).

Multicolinealidad aproximativa. La multicolinearidad aproximativa es mucho más frecuente. Acontece cuando una de las variables independientes se correlaciona fuertemente con otra o una combinación lineal de las demás. Esta variable, aunque no estrictamente redundante, puede considerarse como "casi" redundante, es decir, su aportación de información no es relevante comparada a la información que aportan las otras variables.

Consecuencias

La precisión de los estimadores es baja, o sea que sus varianzas de muestreo $s_{b_j}{}^2$ son elevadas. No es posible diferenciar correctamente la influencia de las variables que están correlacionadas entre sí.

De manera concreta, en el caso de dos variables, esto puede manifestarse como sigue: ninguna de las dos variables posee un coeficiente significativamente diferente de cero, sin embargo, al quitar las dos variables, el modelo resultante no aprueba el test F.

Test de detección

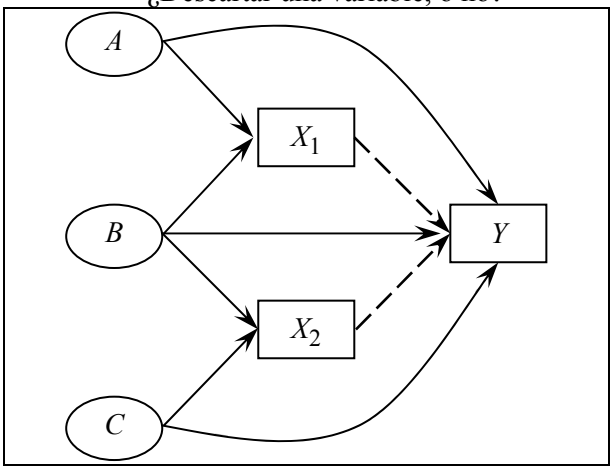
En el caso de la correlación entre dos variables, es posible examinar los coeficientes de correlación simple de las variables independientes entre ellas. Pero la multicolinealidad es, a menudo, mucho más compleja y es necesario recurrir a tipos de análisis más sofisticados (vea Kennedy, 1992, p. 180, con relación a *condición index*).

Remedios

En algunos casos existe quizás una o varias variables independientes que sobran en el modelo. No obstante, casi siempre debemos aceptar "vivir con" y renunciar separar la influencia de las variables correlacionadas. En algunos casos, descartar una variable por motivo de multicolinealidad podría hasta revelarse un grave error.

La figura que presentamos a continuación ilustra tal situación. Los factores inobservables *A*, *B* y *C* influencian la variable dependiente *Y*. Los factores inobservables no pueden figurar en el modelo. Para reemplazarlos, el modelo contiene dos variables independientes observables X_1 y X_2: la primera es influenciada por *A* y *B*, y la segunda, por *B* y *C*. A causa de la influencia compartida de *B*, X_1 y X_2. son correlacionadas. Sin embargo, al descartar una de las dos variables, descartaríamos al mismo tiempo el factor subyacente *A* o *C*.

¿Descartar una variable, o no?



CONCLUSIÓN DE LA TERCERA PARTE

En la tercera parte de la obra abordamos el análisis de regresión lineal. En primer lugar, se hizo hincapié en la regresión cuando sirve para operacionalizar un modelo de una relación entre una variable dependiente y una o varias variables independientes. Pues solamente en el contexto de este trámite de operacionalización se puede dar un sentido a los resultados de la regresión. Desde este punto de vista, todos los modelos de análisis multivariado son como el análisis de regresión: no se pueden interpretar sus resultados, sino a la luz del modelo conceptual subyacente.

Además, la regresión lineal es una herramienta muy polivalente del análisis multivariado. A partir de ella se pueden abordar muchos otros métodos, más avanzados o especializados. Así expusimos las bases del funcionamiento del estimador de los mínimos cuadrados y examinamos el coeficiente de determinación múltiple como medición de ajuste para evaluar las prestaciones de un modelo.

Abordamos desde tres puntos de vista los métodos de inducción aplicados a la regresión múltiple. Primero, se examinó el papel de lo aleatorio en la regresión múltiple, con el fin de demostrar que los tests de hipótesis aplicados a la regresión múltiple se fundamentan en los mismos principios epistemológicos establecidos en la segunda parte de la obra. Sin embargo, en la inducción estadística a partir de datos de

muestra, lo aleatorio es inherente al vínculo entre la muestra y la población. Aquí lo aleatorio traduce más bien la naturaleza aproximada del modelo, lo cual, en su parte determinista, es análogo a las ideas de Platón: la realidad observable –los datos– no son más que un reflejo imperfecto del modelo. Lo aleatorio es esa imperfección.

Se presentaron después los métodos de inducción aplicados a la regresión múltiple desde un punto de vista pragmático. Gracias a ejemplos, vimos en particular de qué puede servir el test t de Student y cómo utilizarlo. Finalmente, se enunciaron las hipótesis del modelo clásico de la regresión lineal y se ilustraron sus consecuencias concretas a través del análisis diagnóstico de los residuos.

ANEXO 3-A
LA LECTURA DE UNA ESPECIE DE COMPUTADORA

La presente sección busca echar un vistazo en el mundo del análisis de regresión. Presentamos, pues, tres extractos de salidas de computadora que provienen de los resultados de Lemelin y Polèse (1995). El programa que usamos es SAS. El primer extracto contiene los resultados relativos a la ecuación (6) de Lemelin y Polèse (1995):

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i + d (\ln GNPC_i) \qquad (6)$$

No se hace mención de los resultados del segundo extracto en Lemelin y Polèse (1995). Se trata del modelo truncado:

$$\ln PURB_i = a + b \ln PTOT_i + c \ln GNPC_i$$

Finalmente, el tercer extracto contiene los resultados relativos a la ecuación (1) de Lemelin y Polèse (1995):

$$\ln PLAR_i = \ln K + h \ln PURB_i \qquad (1)$$

Entre los elementos que podemos encontrar las salidas de computadora, mencionemos:

1. La variable dependiente de la regresión es *LPURB*, o sea ln *PURB*.
2. Este cuadro contiene los resultados de la estimación de los parámetros *a*, *b*, *c* y *d*.

3. Los parámetros se identifican por el nombre de la variable de la cual son coeficiente (en SAS, la variable INTERCEP, del inglés "intercept", refiera a la constante del modelo).
4. Los valores estimados de los parámetros y parámetros estándar *Beta*.

Los parámetros *Beta* son los valores que tendrían los coeficientes si se reemplazara las variables del modelo por variables estándar. Por ejemplo, se reemplazaría:

$\ln PURB_i$ por $ZLPURB_i = \dfrac{\ln PURB_i - \overline{\ln PURB}}{s_{\ln PURB}}$,

con $s_{\ln PURB} = \sqrt{\dfrac{1}{n-1}\sum_i \left(\ln PURB_i - \overline{\ln PURB}\right)^2}$

$\ln PTOT_i$ por $ZLPTOT_i = \dfrac{\ln PTOT_i - \overline{\ln PTOT}}{s_{\ln PTOT}}$,

con $s_{\ln PTOT} = \sqrt{\dfrac{1}{n-1}\sum_i \left(\ln PTOT_i - \overline{\ln PTOT}\right)^2}$

etc.

y el parámetro *Beta* relacionado a ln *PTOT* se da con

$$Beta_{\ln PTOT} = \hat{b}\frac{s_{\ln PTOT}}{s_{\ln PURB}}$$

5. Error tipo (desviación estándar de muestro) de los valores estimados de los parámetros: $s_{\hat{a}}$, $s_{\hat{b}}$, $s_{\hat{c}}$ y $s_{\hat{d}}$.
6. Valor del *t* de Student para la hipótesis que el valor del parámetro sea cero; tenemos $t = \dfrac{\hat{a}}{s_{\hat{a}}}$, etc.
7. Probabilidad crítica correspondiente al valor del *t*: es el umbral de significanza que se tendría que tomar para que el valor crítico correspondiente sea igual al valor

calculado del *t*; es equivalente al resultado de la función `TDIST` en Excel.

8. El cuadro *Analysis of variance* (análisis de varianza) detalle el cálculo del R^2.
9. Valores de las sumas de cuadrados que entran en el cálculo del R^2 y sirven para construir tests *F* de Fisher:
   - Línea "Model": *SSM* = variabilidad de la cual el modelo de cuenta
   - Línea "Error": *SSR* = variabilidad residual
   - Línea "C Total": *SST* = variabilidad total
10. Número de grados de libertad (*DF*, "degrees of freedom") asociado a cada suma de cuadrados:
    - *SSM*: $k-1$
    - *SSR*: $n-k$
    - *SST*: $n-1$
11. Desviaciones cuadráticas promedias ("Mean squares"); se cálculan con división entre el número de grados de libertad correspondiente:
    - Modelo: $\frac{SSM}{k-1}$
    - Residuos: $MSE = \frac{SSR}{n-k}$; es el valor estimado de la varianza del término aleatorio, σ^2.
12. Valor del *F* de Fisher para la hipótesis nula que todos los parámetros sean cero, excepto la constante. Se cálcula con

$$F(k-1; n-k) = \frac{SSM \big/ (k-1)}{SSR \big/ (n-k)}$$

13. Probabilidad crítica correspondiente al valor calculado del *F*.
14. Coeficiente de determinación multiple:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

15. Coeficiente de determinación multiple ajustado:

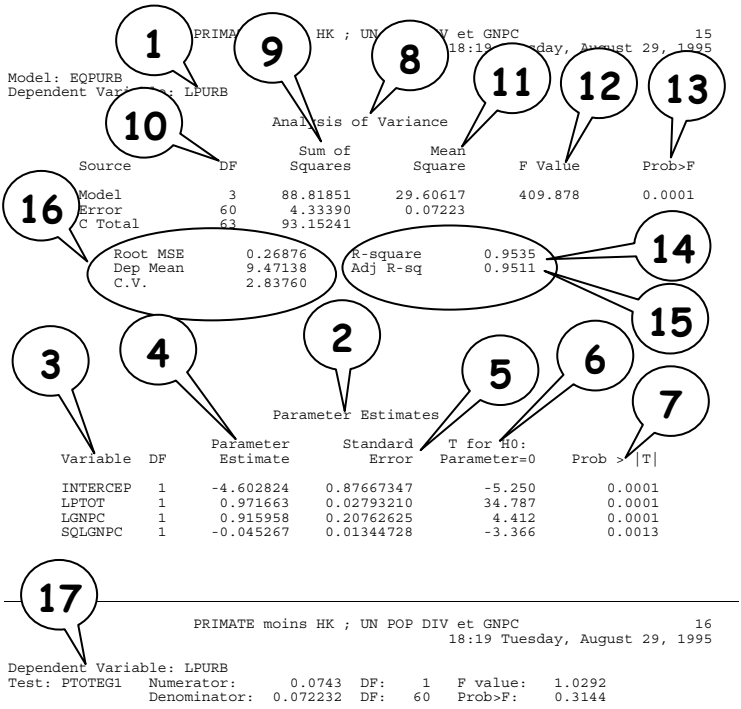
$$\overline{R}^2 = 1 - \frac{SSR \big/ (n-k)}{SST \big/ (n-1)}$$

16. Evaluación de la capacidad de predicción del modelo:
   - Raíz cuadrada del error cuadrático promedio "Root MSE" = $\sqrt{MSE}$
   - Valor promedio de la variable dependiente: m_y.
   - "Coeficiente de variación" (C.V.); en realidad, se trata del coeficiente de variación en porcentage:

$$= 100 \frac{\sqrt{MSE}}{m_y}$$

17. Test de la hipótesis que el coeficiente de ln *PTOT* sea igual a 1. Este test se hace con el *F* de Fisher. Puesto que el test *F* y el test *t* de Student se basan en el mismo modelo muestral (modelo clásico de la regresión lineal normal), el resultado es absolutamente idéntico: la probabilidad crítica es 0.3144.

Especie del logicial SAS



DIGRESIÓN: EL ASPECTO DE LA RELACIÓN ENTRE LA POBLACIÓN URBANA Y EL PIB PER CÁPITA

La ecuación (6) es una relación cuadrática entre los logaritmos de las variables. Sin embargo, ¿cuál es el aspecto real de la relación entre la población urbana y el PIB per cápita?

Antes de la era de las hojas de cálculo electrónicas (como Excel), cuando se deseaba conocer el aspecto de una función sin tener que calcular un gran número de valores, se recurría al cálculo diferencial y álgebra. Era necesario conocer las derivadas primeras y segundas de la función, luego resolver las ecuaciones que permitían determinar los puntos de intersec-

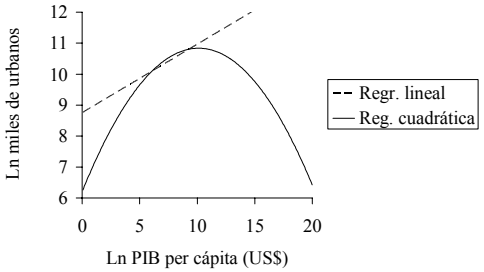
ción con el eje de las abscisas, sus máximas y mínimas, sus puntos de inflexión, para finalmente determinar el signo de la función y de sus derivadas entre estos puntos de referencia. Hoy en día, sólo basta simular en una hoja de cálculo.

Por lo que trata de la ecuación (6), los resultados presentados más arriba nos muestran que el coeficiente *b* es aproximadamente igual a 1. Esto implica que, todo siendo igual de un lado (particularmente el PIB per cápita), la población urbana es más o menos proporcional a la población total. Por lo tanto, nos concentraremos en el aspecto de la relación entre el PIB per cápita y la población urbana, dada una población total. Fijamos la cifra de la población en su valor promedio en la muestra (67.6 millones de habitantes) y hicimos variar el PIB per cápita entre cero y el valor inverosímil de US$ 100.000. Para cada valor, calculamos cuál fue la población urbana predicha por el modelo. Los resultados se reproducen a continuación en graficos.

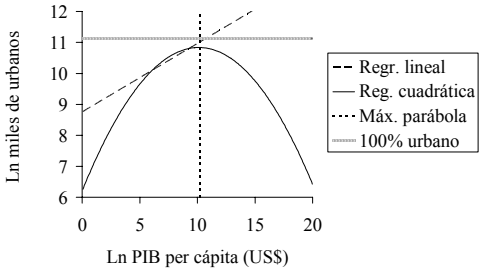
Los puntos que queremos principalmente destacar son:

- El modelo que representa la ecuación (6) es un modelo descriptivo el cual es válido solo al interior de un cierto campo de variación.
- El aspecto visual de la curva depende de la selección de las escalas (natural o logarítmica).

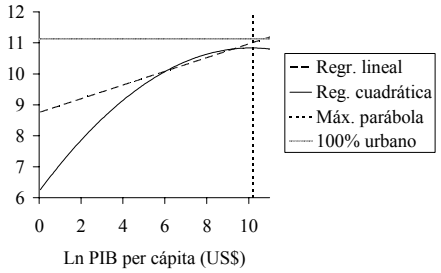
Población urbana calculada[177]



Población urbana calculada

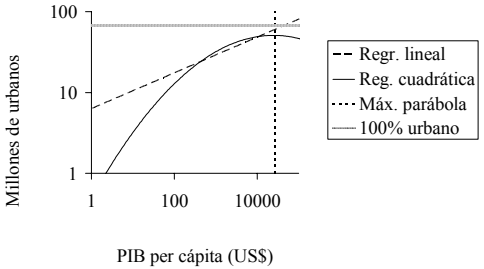


Población urbana calculada

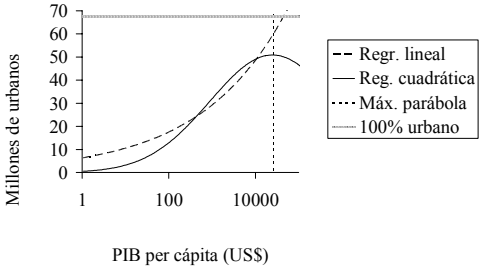


[177] Cálculo hecho con una población total de 67.6 millones.

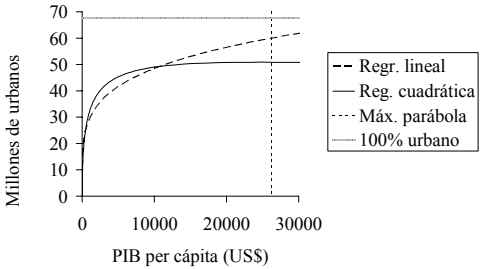
Población urbana calculada (escalas logarítmicas)



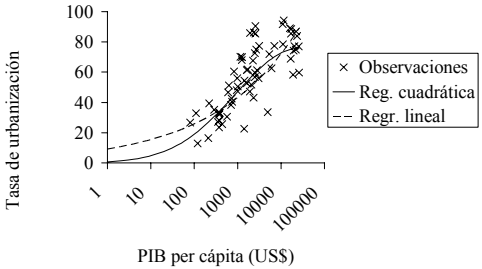
Población urbana calculada (escala horizontal logarítmica)



Población urbana calculada



Tasas de urbanización observadas y calculadas



ANEXO 3-B
LA ALEGORÍA DE LA CAVERNA DE PLATÓN[*]

RESUMEN

Robert Baccou resume de esta manera la alegoría de la Caverna:

> Una alegoría nos enseñará, ahora, la posición de los hombres en relación con la verdadera luz. Imaginemos a unos cautivos encadenados en una caverna subterránea encarando la pared opuesta a la entrada, de tal manera que no pueden ver otra cosa que no sea esa pared. Esta pared es iluminada por la luz de una hoguera que arde frente a la entrada en ba-

[*] Referencias:
Platon, *La République*, Garnier-Flammarion, Paris, 1966; introducción, traducción y notas por Robert Baccou.
Platón, *Diálogos*, *La república o de lo justo*, Libro VII, pp. 551 – 554 y p. 563; Editorial Porrúa, México, D.F. 1975.
NdT: En la versión original, el autor reproduce parte del diálogo del Libro VII refiriéndose a la traducción francesa que menciona y de donde sustrae el mismo resumen de la alegoría de la Caverna escrito por Robert Baccou.

Es posible encontrar el texto de la alegoría de la caverna en los sitios siguientes:
http://www.virtualistes.org/platon.htm.
http://www.hermanubis.com.br/Artigos/FR/ARFRLaCavernedePlaton.htm.
http://plato-dialogues.org/fr/tetra_4/republic/caverne.htm.
http://www.cyberphilo.com/ref/caverne.html.

jada pero del otro lado de un camino bordeado de un pequeño muro. Por detrás del muro caminan unas personas que levantan sobre sus espaldas objetos diversos como estatuillas de hombres, de animales, etcétera. Los cautivos no pueden ver más de esos objetos que sombras proyectadas en el fondo de la caverna, así mismo, lo único que oyen es el eco de las palabras que intercambian los que cargan. Acostumbrados desde su nacimiento a contemplar esas vanas imágenes, a escuchar esos sonidos confusos, cuyo origen ignoran, viven en un mundo de sombras que creen es la realidad.

Ahora bien, imaginemos que uno de ellos sea liberado de sus cadenas y sea arrastrado hacia la luz; al principio la luz lo cegará y no será capaz de distinguir nada a su alrededor, por instinto buscará mirar nuevamente hacia las sombras que no lastiman sus ojos y, durante algún tiempo, las creerá más reales que los objetos del nuevo mundo, pero, en cuanto sus ojos se hayan acostumbrado al ambiente luminoso, podrá percibir el reflejo de esos objetos en las aguas para luego poder mirarlos directamente. Durante la noche, contemplará la luna y las constelaciones y finalmente será capaz de sostener el resplandor del sol. Entonces se dará cuenta que su vida anterior no era más que un sueño sombrío, y se compadecerá de sus antiguos compañeros de cautiverio. No obstante, si baja a verlos nuevamente para instruirlos y para enseñarles la vacuidad de esas sombras de la caverna y les describe el mundo de la luz, ¿quién podrá escucharlo sin reír?, ¿quién, sobre todo, dará por cierto su divina revelación? Hasta los más sabios lo considerarán loco hasta el punto de amenazarlo de muerte si persiste en su generoso intento.

Entendemos sin ninguna pena el significado de esta alegoría. Los hombres son aquí en esta tierra esclavos de sus sentidos: en la oscuridad del mundo de la materia, en perpetuo cambio, sólo captan sombras o vagos reflejos. Pero los modelos de estas sombras y, aún más, la fuente luminosa de estos reflejos, les son desconocidos hasta el punto de, ni siquiera, entrever su existencia. La única ciencia –o lo que denominan con este nombre– consiste en descubrir un cierto orden en las apariencias, una secuencia prevista en el inter-

minable desfile de las sombras, las cuales pasan sin cesar frente a ellos moviéndose en un telón misterioso. Sólo quien haya roto sus cadenas y se haya elevado fuera de las tinieblas de la caverna hasta el reino del sol, sólo éste podrá contemplar y fijar en su alma el puro esplendor de las esencias, pero cuando haya vivido un buen rato en este reino, sus ojos, acostumbrados a estos ideales esplendores, no podrá más distinguir las sombras de abajo.

DIÁLOGO

Sócrates y Glaucón son los dos protagonistas del diálogo del Libro VII

– Represéntate ahora el estado de la naturaleza humana respecto de la ciencia y de la ignorancia, según el cuadro que de él voy a trazarte. Imagina un antro subterráneo que tiene todo a lo largo una abertura que deja libre a la luz el paso, y, en ese antro, unos hombres encadenados desde su infancia, de suerte que no puedan cambiar de lugar ni volver la cabeza, por causa de las cadenas, que les sujetan las piernas y el cuello, pudiendo solamente ver los objetos que tengan delante. A su espalda, a cierta distancia y a cierta altura, hay un fuego cuyo fulgor les alumbra, y entre ese fuego y los cautivos se halla un camino escarpado. A lo largo de ese camino, imagina un muro semejante a esas vallas que los charlatanes ponen entre ellos y los espectadores, para ocultar a éstos el juego y los secretos trucos de las maravillas que les muestran.
– Todo eso me representa.
– Figúrate unos hombres que pasan a lo largo de ese muro, porteando objetos de todas clases, figuras de hombres y de animales de madera o de piedra, de suerte que todo ello se aparezca por encima del muro. Los que los portean, unos hablan entre sí, otros pasan sin decir nada.

– ¡Extraño cuadro y extraños prisioneros!
– Sin embargo, se nos parecen punto por punto. Y, ante todo, ¿crees que verán otra cosa, de sí mismos y de los que se hallan a su lado, más que las sombras que van a producirse frente a ellos al fondo de la caverna?
– ¿Qué más pueden ver, puesto que desde su nacimiento se hallan forzados a tener siempre inmóvil la cabeza?
– ¿Verán asimismo, otra cosa que las sombras de los objetos que pasen por detrás de ellos?
– No.
– Si pudiesen conversar entre sí, ¿no convendrían en dar a las sombras que ven los nombres de esas mismas cosas?
– Indudablemente.
– Y si al fondo de su prisión hubiese un eco que repitiese las palabras de los que pasan, ¿no se figurarían que oían hablar a las sombras mismas que pasan por delante de sus ojos?
– Sí.
– Finalmente, no creerían que existiese nada real fuera de las sombras.
– Sin duda.
– Mira ahora lo que naturalmente habrá de sucederles, si son libertados de sus hierros y se les cura de su error. Desátese a uno de esos cautivos y oblíguesele inmediatamente a levantarse, a volver la cabeza, a caminar y a mirar hacia la luz; nada de eso hará sin infinito trabajo; la luz le abrasará los ojos, y el deslumbramiento que le produzca le impedirá distinguir los objetos cuyas sombras veía antes. ¿Qué crees que respondería si se dijesen que hasta entonces no ha visto más que fantasmas, que ahora tiene ante los ojos objetos más reales y más próximos a la verdad? Si se le muestran luego las cosas a medida que vayan presentándose, y se le obliga, en fuerza de preguntas, a decir qué es cada una de ellas, ¿no se le sumirá en perplejidad,

y no se persuadirá a que lo que antes veía era más real que lo que ahora se le muestra?

– Sin duda.

– Y si se le obligase a mirar al fuego, ¿no enfermaría de los ojos? ¿No desviaría sus miradas para dirigirlas a la sombra, que afronta sin esfuerzo? ¿No estimaría que esa sombra posee algo más claro y distinto que todo lo que se le hace ver?

– Seguramente.

– Si ahora se le arranca de la caverna, y se le arrastra, por el sendero áspero y escarpado, hasta la claridad del sol, ¡qué suplicio no será para él ser así arrastrado! ¡Qué furor el suyo! Y cuando haya llegado a la luz libre, ofuscados con su fulgor los ojos, ¿podría ver nada de la multitud de objetos que llamamos seres reales?

– Le sería imposible, al primer pronto.

– Necesitaría tiempo, sin duda, para acostumbrarse a eIlo. Lo que mejor distinguiría, sería, primero, las sombras; luego, las imágenes de los hombres y de los demás objetos, pintadas en la superficie de las aguas; finalmente, los objetos mismos. De ahí dirigiría sus miradas al cielo, cuya vista sostendría con mayor facilidad durante la noche, al claror de la luna y de las estrellas, que por el día y a la luz del sol.

– Sin duda.

– Finalmente, se hallaría en condiciones, no sólo de ver la imagen del sol en las aguas y en todo aquello en que se refleja, sino de fijar en él la mirada, de contemplar al verdadero sol en verdadero lugar.

– Sí.

– Después de esto, dándose a razonar, llegará a concluir que el sol es quien hace las estaciones y los años, quien lo rige todo en el mundo visible, y que es en cierto modo causa de lo que se veía en la caverna.

– Es evidente que llegaría por grados hasta hacerse esas reflexiones.
– Si llegase entonces a recordar su primera morada, la idea que en ella se tiene la sabiduría, y a sus compañeros de esclavitud, ¿no se alborozaría de su mudanza, y no tendría compasión de la desdicha de aquellos?
– Seguramente.
– ¿Crees que sintiese todavía celos de los honores, de las alabanzas y recompensas allí otorgados al que más rápidamente captase las sombras a su paso, al que recordase con mayor seguridad las que iban delante, detrás o juntas, y que por tal razón sería el más hábil en adivinar su aparición, o que envidiase la condición de los que en la prisión eran más poderosos y más honrados? ¿No preferiría, como Aquiles en Homero, pasarse la vida al servicio de un pobre labrador y sufrirlo todo, antes que volver a su primer estado y a sus ilusiones primeras?
– No dudo que estaría dispuesto a soportar todos los males del mundo, mejor que vivir de tal suerte.
– Pues pon atención a esto otro: si de nuevo tornase a su prisión, para volver a ocupar en ella su antiguo puesto, ¿no se encontraría como enceguecido, en el súbito tránsito de la luz del día a la oscuridad?
– Sí.
– Y si mientras aún no distingue nada, y antes de que sus ojos se hayan repuesto, cosa que no podría suceder sino después de pasado bastante tiempo, tuviese que discutir con los demás prisioneros sobre esas sombras, ¿no daría qué reír a los demás, que dirían de él que, por haber subido a lo alto, ha perdido la vista, añadiendo que sería una locura que ellos quisiesen salir del lugar en que se hallan, y que si a alguien se le ocurriese querer sacarlos de allí y llevarlos a la región superior, habría que apoderarse de él y darle muerte?

– Indiscutiblemente.
– Pues ésa es precisamente, mi querido Glaucón, la imagen de la condición humana. El antro subterráneo es este mundo visible; el fuego que lo ilumina, la luz del sol; el cautivo que sube a la región superior y la contempla, es el alma que se eleva hasta la esfera inteligible. He aquí, a lo menos, mi pensamiento, puesto que quieres saberlo. Dios sabe si es cierto. Por mi parte, la cosa me parece tal como voy a decir. En los últimos límites del mundo inteligible está la idea del bien, que se percibe con trabajo, pero que no puede ser percibida sin concluir que ella es la causa primera de cuanto hay de bueno y de bello en el universo; que ella, en este mundo visible, produce la luz y el astro de quien la luz viene directamente; que, en el mundo invisible, engendra la verdad y la inteligencia; que es preciso, en fin, tener puestos los ojos en esa idea, si queremos conducirnos cuerdamente en la vida pública y privada.
– Soy de tu parecer, en cuanto puedo comprender tu pensamiento.
– Consiente, pues, asimismo, en no extrañarte de que los que han llegado a esa sublime contemplación desdeñen la intervención de los asuntos humanos, y que sus almas aspiren sin tregua establecerse en ese eminente lugar. La cosa debe ser así, si es conforme a la pintura alegórica que de ella he trazado.
– Así debe ser.
– ¿Es de extrañar que un hombre, al pasar de esa divina contemplación a la de los miserables objetos que nos ocupan, se turbe y parezca ridículo cuando, antes de haberse familiarizado con las tinieblas que le rodean, se ve obligado a disputar ante los tribunales, o en algún otro lugar, acerca de sombras y fantasmas de justicia, y a explicar en qué forma los concibe ante personas que jamás vieron a la propia justicia?

– Nada de sorprendente veo en ello.
– Un hombre sensato se hará la reflexión de que la vista puede ser turbada de dos maneras y por dos causas opuestas: por el paso de la luz a la oscuridad, o por el de la oscuridad a la luz; y aplicando a los ojos del alma lo que acontece a los del cuerpo, cuando la vea turbada y embarazada para distinguir ciertos objetos, en lugar de reírse sin razón de semejante perplejidad, examinará si proviene de que descienda de un estado más luminoso, o si es porque, pasando de la ignorancia a la luz, quede ofuscada por su fulgor excesivo. En el segundo caso, la felicitará por su perplejidad; en el primero, compadecerá su suerte o, si quiere reírse a costa suya, sus burlas serán menos ridículas que si se dirigiesen al alma que vuelve a descender de la morada de la luz.
– Sensatísimo es lo que dices.
– Ahora bien, si todo esto es cierto, fuerza es concluir de ello que la ciencia no enseña en la forma en que cierta gente pretende. Se alaban de hacerla penetrar en un alma en que nada hay de ella, aproximadamente como, podría darse vista a unos ojos ciegos.
– A voz en cuello lo dicen.
– Pero el presente discurso nos hace ver que todos poseen en su alma la facultad de aprender, con un órgano a ello destinado; que todo el secreto consiste en apartar a ese órgano, con toda el alma, de la visión de lo que nace, hacia la contemplación de lo que es, hasta que pueda fijar sus miradas en lo que hay de más luminoso en el ser; es decir, según nosotros, en el bien; del mismo modo que, si el ojo no estuviese dotado de movimiento propio, ocurriría por fuerza que todo el cuerpo habría de girar con él, en el tránsito de las tinieblas a la luz, ¿no es así?
– En efecto.

– En esa evolución que se obliga a hacer al alma, todo el arte consiste, pues, en hacerla girar de la manera más fácil y más útil, No se trata de conferirle facultad de ver, que ya tiene; pero su órgano está orientado en mala dirección, no mira adonde es debido, y eso es lo que hay que corregir.
– Me parece que no hay otro secreto.

(Libro VII, pp. 551-554)

[...]

– Recuerda al hombre de la caverna que decíamos: empieza por ser libertado de sus cadenas; después, dejando las sombras, se vuelve hacia las figuras, artificiales y hacia el fuego que las ilumina. Finalmente, sale de ese lugar subterráneo para subir hasta los lugares que el sol alumbra; y como quiera que sus débiles ojos no pueden al principio fijarse en los animales, ni en las plantas, ni en el sol, recurre a sus imágenes pintadas en la superficie de las aguas, y a sus sombras; pero estas sombras pertenecen a seres reales, y no a objetos artificiales como en la caverna, y no se han formado gracias a la luz que el prisionero tomaba por el sol. El estudio de las ciencias de que hemos hablado produce el mismo efecto. Eleva la parte más noble del alma hasta la contemplación del más excelente de todos los seres, como, en el otro caso, el órgano más agudo del cuerpo humano se eleva hasta la contemplación de lo más luminoso que existe en el mundo material y visible.

(Libro VII, p. 563)

INTRODUCCIÓN A LA CUARTA PARTE: EL ANÁLISIS CUANTITATIVO DE DATOS CUALITATIVOS

Vimos ya cómo, de algún modo, era posible medir propiedades cualitativas como la nacionalidad de una persona (capítulo 1-1). Sin embargo, aunque se admita que una propiedad cualitativa se pueda medir, la medición de tal propiedad parece imperfecta al momento de compararla con la medición de propiedades como la superficie o el ingreso. En efecto, al considerar las cuatro relaciones que distinguen la definición axiomática de una medición ($=$, $\neq$, $<$, $>$), no podemos más que constatar que para una propiedad cualitativa como la nacionalidad, no es posible tomar una decisión con dos de estas relaciones. Es la razón por la cual se distinguen varios tipos de variables según la escala de medición que se les asocia: variables categóricas, ordinales, de intervalo y racionales.

Se puede considerar la mayor parte de las variables de intervalo o racionales, así como algunas variables ordinales, como variables continuas. Por lo contrario, las variables categóricas y un gran número de variables ordinales son variables discretas. Ahora bien, vimos con relación al modelo lineal general y al análisis de regresión, que es un método de análisis de los datos, que es posible aplicar solamente si se basa en un modelo teórico formalizado a partir de una relación entre una variable dependiente continua y una o varias

variables independientes continuas o discretas. En esta cuarta parte del curso, nos dedicamos a examinar métodos que se aplican a variables categóricas u ordinales discretas, como las que se acostumbra emplear para medir propiedades cualitativas.

Con el análisis de las tablas de contingencias, es posible examinar las relaciones entre varias variables categóricas. En un análisis de las tablas de contingencias, ninguna variable puede tomar el lugar de una variable dependiente. El análisis de varianza se basa en un modelo teórico formalizado a partir de una relación entre una variable dependiente continua y una o varias variables independientes, todas discretas. Por lo tanto, el análisis de varianza parece ser un caso particular del análisis de regresión (aunque el formato tradicional para presentar los resultados sea diferente). Hacemos, pues, un especial énfasis en la manera de adaptar el modelo de regresión al análisis de varianza. Finalmente, los modelos con variable independiente cualitativa se refieren a métodos como el logit o el probit.

CAPÍTULO 4-1
EL ANÁLISIS DE LAS TABLAS DE CONTINGENCIA

4-1.1. INTRODUCCIÓN

4-1.1.1. ¿Qué es una tabla de contingencia?

Una tabla de contingencia es una manera de presentar datos de enumeración (de conteo) de individuos previamente clasificados en categorías. Es de constatar, por lo tanto, que una tabla de contingencia es ya el resultado de un tratamiento de datos puesto que los individuos (observaciones) tuvieron que ser el objeto de una previa clasificación y de un previo conteo.

Se determina el formato de la tabla dependiendo del modelo de clasificación que se emplea. Es necesario que el modelo de clasificación se constituya de categorías mutuamente exclusivas y que sea exhaustivo, es decir que en el lenguaje de la teoría de conjuntos, las categorías deben constituir una partición del universo de modo que cada individuo pertenezca a una y a una sola categoría.

Se definen las categorías por medio de una o varias variables de clasificación (variables categóricas) que corresponden a cuantos atributos (dimensiones) tienen los individuos. Se describe a los individuos observados con el fin de clasificarlos en función de los valores de sus atributos. Se efectúa un

conteo de todos los individuos que tengan la misma descripción (los mismos valores de atributos) para luego inscribir su número en la celda correspondiente de la tabla de contingencia resultado de esta clasificación. La tabla de contingencia tiene tantas dimensiones como haya variables de clasificación y tantas celdas como haya combinaciones de categorías.

Examinemos un pequeño ejemplo de construcción de una tabla de contingencia a partir de datos brutos. Teniendo la tabla de observaciones siguiente, donde las observaciones ya están ordenadas por sexo, pues por color de ojos:

	Nombre	Sexo	Color de ojos
1	Dolores	M	Azules
6	Juan	H	Azules
4	Marco	H	Negros
2	Maria	M	Azules
3	Pedro	H	Negros
5	Guadalupe	M	Negros

Se ve en la tabla arriba la estructura matricial de los datos brutos. A partir de estos datos, es posible deducir una tabla de contingencia del color de los ojos en función del sexo:

	Sexo		
Color de ojos	M	H	Total
Azules	2	1	3
Negros	1	2	3
Total	3	3	6

Esa tabla de contingencia tiene también una estructura matricial, pero ya no se trata de datos brutos: la tabla de contingencia es el resultado de un tratamiento. Se nota que una tabla de contingencia de una sola variable es nada más una tabla de frecuencias.

Tabla 1: Población activa empleada
en la región metropolitana de Montreal, 1991
Zona de residencia según el sexo y la profesión

Zona de residencia	Profesiones					
	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Mujeres						
Montreal[178]	24,025	58,204	76,450	24,385	28,825	*211,889*
Resto CUM[179]	22,575	42,207	70,003	14,065	17,435	*166,285*
Anillo norte	16,785	31,699	63,491	11,975	18,630	*142,580*
Anillo sur	18,365	35,674	65,290	10,485	19,380	*149,194*
Fuera RMR[180]	3,265	7,535	11,089	3,190	3,565	*28,644*
Total Mujeres	*85,015*	*175,319*	*286,323*	*64,100*	*87,835*	*698,592*
Hombres						
Montreal	32,336	55,045	43,546	65,340	46,850	*243,117*
Resto de CUM	39,146	39,920	37,819	46,173	28,749	*191,807*
Anillo Norte	33,287	27,560	31,170	62,852	29,329	*184,198*
Anillo Sur	36,006	32,464	30,600	58,778	29,721	*187,569*
Fuera de RMR	8,270	8,590	8,270	22,305	9,099	*56,534*
Total Hombres	*149,045*	*163,579*	*151,405*	*255,448*	*143,748*	*863,225*
Total hombres y mujeres						
Montreal	56,361	113,249	119,996	89,725	75,675	*455,006*
Resto de CUM	61,721	82,127	107,822	60,238	46,184	*358,092*
Anillo Norte	50,072	59,259	94,661	74,827	47,959	*326,778*
Anillo Sur	54,371	68,138	95,890	69,263	49,101	*336,763*
Fuera de RMR	11,535	16,125	19,359	25,495	12,664	*85,178*
Total H + M	*234,060*	*338,898*	*437,728*	*319,548*	*231,583*	*1,561,817*

Fuente: Statistique Canada, Censo de 1991.

[178] Se refiere aquí al municipio de Montreal tal como era definido hasta la fusión en 2002 de todos los municipios que antes formaban la CUM (ver la próxima nota).

[179] CUM = Comunidad Urbana de Montreal: todos los municipios de la Isla de Montreal. La CUM volvió con la fusión del 2002 en el Municipio del Gran Montreal.

[180] RMR = Región Metropolitana de Censo.

Esta tabla de contingencia posee tres dimensiones: la zona de residencia, la profesión y el sexo. La zona de residencia tiene 5 categorías, la profesión, 5 y hay 2 sexos. La tabla contiene, por lo tanto, 5 x 5 x 2 = 50 celdas a las cuales se suman las líneas y columnas de los totales y subtotales.

4-1.1.2. El análisis de las tablas de contingencias entre los métodos de análisis multivariado

De manera general, el análisis mutivariado designa el conjunto de los métodos de análisis estadístico que tratan simultáneamente con más de una variable. En particular, se recurre al análisis multivariado para:

- medir el grado de asociación entre dos o más variables;
- estimar los parámetros de una relación entre dos o más variables;
- evaluar hasta qué punto las diferencias entre dos o varios grupos de observaciones son significativas;
- intentar predecir a cuál grupo pertenece un individuo a partir de sus demás características;
- buscar reconocer una estructura en un conjunto de datos.

Varias técnicas de análisis multivariado permiten distinguir entre las variables dependientes y las variables independientes. Las variables dependientes son las variables cuyo valor se quiere predecir; las otras variables son las independientes.[181] Es posible clasificar los métodos de análisis multi-

[181] Se tomaron los términos "variable dependiente" y "variable independiente" del área de las ciencias experimentales, cuando el investigador fija de manera "independiente" el valor de ciertas variables (como, por ejemplo, la dosificación de un tratamiento) para, luego, observar su efecto en la variable "dependiente". Se da a veces a las variables independientes el nombre de variables "explicativas". Sin embargo, hay que tener sumo cuidado con esta expresión por la connotación de causalidad que transmite. En un modelo con una sola ecuación, la variable dependiente se llama también "endógena", es decir que se

variado en función del número de variables dependientes e independientes y según si las unas o las otras son variables discretas o continuas.[182]

La tabla que sigue presenta una clasificación de los métodos que examinamos en el marco de este curso.

<table>
<tr><th colspan="2">Variable dependiente</th><th>Variables independientes</th><th colspan="2">Método</th></tr>
<tr><td colspan="2" rowspan="2">Ninguna</td><td>2 variables categóricas</td><td rowspan="2">Análisis de tabla de contingencia</td><td>... con 2 dimensiones</td></tr>
<tr><td>Más de 2 variables categóricas</td><td>... con más de 2 dimensiones</td></tr>
<tr><td colspan="2" rowspan="2">Continua</td><td>Discretas (categóricas)</td><td colspan="2">Análisis de varianza o Regresión múltiple</td></tr>
<tr><td>Continuas y/o discretas</td><td colspan="2">Regresión múltiple</td></tr>
<tr><td rowspan="2">Categórica</td><td>2 categorías</td><td rowspan="2">Continuas y/o discretas</td><td rowspan="2">Logit o probit</td><td>... binomial</td></tr>
<tr><td>Más de 2 categorías</td><td>... multinomial</td></tr>
</table>

Este capítulo trata del análisis de las tablas de contingencia. Con este método, es posible examinar las relaciones entre

---

determina al interior del modelo, mientras que las variables independientes son "exógenas", es decir que se determinan al exterior del modelo. Las variables independientes se llaman también "estímulos", y entonces las dependientes son "respuestas". En inglés, es posible encontrar las parejas predictor/criterion, stimulus/response, task/perfomance, input/output.

[182] Se infiere esto de la escala de medición asociada a cada variable, es decir, las variables categóricas son discretas mientras que se considera frecuentemente las variables racionales y de intervalo como variables continuas. En cuanto a las variables ordinales, existen pocos métodos que se adaptan específicamente a ellas; en la practica, se consideran continuas. Sin embargo, en tales condiciones, la interpretación de los resultados debe tomar en cuenta la naturaleza ordinal de las variables.

varias variables categóricas. En el análisis de las tablas de contingencia, ninguna variable toma el papel de variable dependiente.

4-1.1.3. Reglas de presentación de una tabla de contingencia

El principio general de presentación de una tabla de contingencia no difiere de cualquier otra tabla: todo debe encaminarse para que el lector sepa perfectamente de lo que trata.

Las principales reglas de presentación que conviene por lo general respetar, son las siguientes:

1. La tabla se encabeza con un titulo que identifica la población o, si fuera el caso, la muestra a la cual se refiere la tabla (en nuestro ejemplo, la población activa empleada en la RMR de Montreal en 1991); note que la identificación de la población contiene, cuando el caso lo requiere una referencia a la zona geográfica y al periodo de tiempo; indica cuales son las unidades de medición empleadas (miles de personas, millones de dólares o...; se puede omitir este elemento en nuestro ejemplo puesto que se trata de número de personas); identifica las dimensiones de la tabla (variables categóricas de clasificación; aquí, la zona de residencia, el sexo y la profesión).
2. Algunos subtítulos indican a cuál variable corresponden las diferentes dimensiones de la tabla (aquí, las líneas corresponden a las zonas de residencia, las columnas a las profesiones y la tabla se divide en partes en función de la tercera dimensión, el sexo).
3. El encabezado de cada columna, línea o parte de la tabla indica a cuál categoría de la variable corresponde esta columna, línea o parte de la tabla.
4. La tabla contiene líneas y columnas de totales así como del gran total (1,561,817); las líneas y las columnas

se identifican con claridad y resaltan (en nuestro ejemplo, con caracteres en negrillas).
5. Finalmente, se indica la fuente de los datos (aquí en términos generales, cuando lo ideal es procurar una referencia bibliográfica completa).

Es posible que una tabla de contingencia contenga también los elementos siguientes:

- proporciones o porcentajes;
- subtotales;
- llamadas y notas correspondientes.

Si la tabla contiene proporciones o porcentajes, es importante poder evidenciar con toda claridad si se trata de proporciones (fracciones contenidas entre cero y uno) o de porcentajes (contenidos entre cero y cien). Además, es necesario indicar claramente la razón de los porcentajes o proporciones (¿porcentaje de qué?). Una manera de llevar a cabo esta tarea es escribiendo "100%" donde conviene hacerlo. Finalmente, es importante no sobrecargar una tabla hasta el punto de dificultar su lectura; puede ser preferible, pues, presentar dos tablas, una para los números y la segunda para los porcentajes.

Lo anterior es también válido para los subtotales. Por ejemplo, en la tabla exhibida más arriba, podríamos pensar que fuese útil presentar el subtotal de la CUM (suma de las dos primeras líneas de cada parte). No obstante, se deben formular los subtotales para que el lector reconozca con claridad lo que se sumó. Además, es importante no sobrecargar la tabla y para este efecto, cabe a menudo presentar los mismos datos en dos tablas: una primera tabla detallada (a veces en anexo) y una segunda, más agregada, que de hecho presenta subtotales.

Para aclarar puntos de los títulos, subtítulos o encabezados sin necesidad de alargarlos indebidamente, se usan notas, las mismas que se pueden emplear para dar definiciones de

algunos términos o para enunciar fórmulas que permitieron calcular las cifras de la tabla.

Finalmente, existen varias maneras de estructurar un tabla con más de dos dimensiones. En el ejemplo anterior, la tercera dimensión, el sexo, corresponde a las diferentes partes de la tabla. Es posible también usar la técnica de la subdivisión de las líneas o de las columnas, técnica que se ilustra con una figura más abajo.

Figura de una tabla
de contingencia con subdivisión de las columnas

	Profesiones, sexo								
Zona de residencia	Directores, gerentes, administradores y similares			Profesionales docentes y cuellos blancos especializados			...	TOTAL todas las profesiones		
	M	H	T	M	H	T		M	H	T
Montreal							...			
					⋮					
Total										

En cuanto se usa, estos métodos de representación, se recomienda acercar hacia la parte interna las variables cuya interacción se desea examinar. La estructura que representamos justo arriba convendría perfectamente para el estudio de la interacción entre sexo y zona de residencia considerando que, en estas condiciones, la profesión toma el papel de una variable de control (se examinan las variables de control más abajo, en el apartado 6); el formato anterior se adapta mejor al examen de la relación entre profesión y zona de residencia mientras que el sexo toma, entonces, el papel de variable de control.

4-1.2 FRECUENCIAS RELATIVAS Y PROBABILIDADES EN UNA TABLA DE CONTINGENCIA

Aunque se puedan generalizar los métodos que a continuación se presentan para las tablas de más de dos dimensiones, porque es más simple, nos limitaremos ahora, a analizar las tablas de dos dimensiones. Con este propósito, retomaremos la tabla anterior pero omitiendo la dimensión del sexo.[183] Para esto, solo basta sumar los hombres y las mujeres, como se efectúa en la tabla siguiente cuyos números son los mismos que aparecen en la tercera parte de la tabla 1.

Tabla 2: Población activa empleada
en la región metropolitana de Montreal, 1991
Zona de residencia según la profesión

	Profesiones						
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trab. especial. en servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones	Repartición $p_{i\bullet}$
Montreal	56,361	113,249	119,996	89,725	75,675	*455,006*	0.29
Resto CUM	61,721	82,127	107,822	60,238	46,184	*358,092*	0.23
Anillo Norte	50,072	59,259	94,661	74,827	47,959	*326,778*	0.21
Anillo Sur	54,371	68,138	95,890	69,263	49,101	*336,763*	0.22
Fuera RMR	11,535	16,125	19,359	25,495	12,664	*85,178*	0.05
Total	*234,060*	*338,898*	*437,728*	*319,548*	*231,583*	*1,561,817*	1.00
Repartic. $p_{\bullet j}$	0.15	0.22	0.28	0.20	0.15	1.00	

[183] Es importante darse cuenta que este procedimiento destruye parte de la información. No se aconseja, por lo tanto, de ninguna manera, esta práctica que se lleva a cabo en este momento sólo por razones pedagógicas.

Usaremos la simbología que sigue:

x_{ij}	número de observaciones de la columna *j* en la línea *i*
$x_{\bullet j} = \sum_i x_{ij}$	número total de observaciones de la columna *j*
$x_{i\bullet} = \sum_j x_{ij}$	número total de observaciones de la línea *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	número total de observaciones en la tabla

Con esta simbología se aplica la convención de que sumar sobre cualquiera de las dos dimensiones puede representarse reemplazando el índice correspondiente con un punto grueso. Por ejemplo, en la tabla de población activa por zona de residencia y por profesión, tenemos:

x_{23} = 107,822, el número de empleados de oficina que viven en la CUM fuera de Montreal;

$x_{\bullet 3}$ = 437,728, el número de empleados de oficina empleados en la RMR;

$x_{2\bullet}$ = 358,092, el número de personas empleadas en la RMR que viven en la CUM fuera de Montreal;

$x_{\bullet\bullet}$ = 1,561,817, el número total de personas empleadas en la RMR.

El análisis de una tabla de contingencia se refiere mucho más a la estructura de los datos que a las magnitudes de los números. Por esta razón se formulan, por lo general, los análisis en términos de las frecuencias relativas que se calculan simplemente con dividir los números por el total pertinente.

Se interpretan las frecuencias relativas como probabilidades. De esta manera, $p_{34} = \frac{74{,}827}{1{,}561{,}817} = 0.048$ corresponde a la probabilidad que un individuo, sorteado entre las 1,561,817 personas empleadas en la RMR, forme parte de la profesión Obrero y viva en el Anillo Norte. Por consiguiente, en el denominador, encontramos el número de individuos donde se efectúa el sorteo (1,561,817) y en el numerador encontramos el número de individuos que reúne la o las características que se pretende examinar (74,827).

Diferentes cálculos de frecuencias relativas corresponden a los diferentes conceptos de probabilidad. Así,

$p_{ij} = {}^{x_{ij}}\!/_{x_{\bullet\bullet}}$ es la probabilidad conjunta de pertenecer al mismo tiempo a *i* y a *j*.

Ejemplo: $p_{34} = \frac{74{,}827}{1{,}561{,}817} = 0.048$ es la probabilidad de formar parte de la profesión Obrero y de vivir en el Anillo Norte.

$p_{i\bullet} = {}^{x_{i\bullet}}\!/_{x_{\bullet\bullet}} = \sum_j {}^{x_{ij}}\!/_{x_{\bullet\bullet}} = \sum_j p_{ij}$ es la probabilidad marginal de pertenecer a *i*.

Ejemplo: $p_{3\bullet} = \frac{326{,}778}{1{,}561{,}817} = 0.209$ es la probabilidad de vivir en el Anillo Norte independientemente de la categoría profesional.

$p_{\bullet j} = {}^{x_{\bullet j}}\!/_{x_{\bullet\bullet}} = \sum_i {}^{x_{ij}}\!/_{x_{\bullet\bullet}} = \sum_i p_{ij}$ es la probabilidad marginal de pertenecer a *j*.

Ejemplo: $p_{\bullet 4} = \frac{319{,}548}{1{,}561{,}817} = 0.205$ es la probabilidad de pertenecer a la profesión Obrero independiente de la zona de residencia.

$p_{j/i\bullet} = {x_{ij}}\big/{x_{i\bullet}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{i\bullet}/x_{\bullet\bullet}} = {p_{ij}}\big/{p_{i\bullet}}$ es la probabilidad condicional de pertenecer a *j* dado que se pertenece a *i*.

Ejemplo: $p_{4/3\bullet} = \frac{74{,}827}{326{,}778} = 0.229$ es la probabilidad de pertenecer a la profesión Obrero dado que la zona de residencia es el Anillo Norte.

$p_{i/\bullet j} = {x_{ij}}\big/{x_{\bullet j}} = \frac{x_{ij}/x_{\bullet\bullet}}{x_{\bullet j}/x_{\bullet\bullet}} = {p_{ij}}\big/{p_{\bullet j}}$ es la probabilidad condicional de pertenecer a *i* dado que pertenece a *j*.

Ejemplo: $p_{3/\bullet 4} = \frac{74{,}827}{319{,}548} = 0.234$ es la probabilidad de vivir en el Anillo Norte dado que se pertenece a la profesión Obrero.

Es obvio que, al sumar todas las probabilidades o frecuencias relativas posibles, el resultado es 1:

$$\sum_i \sum_j p_{ij} = \sum_i p_{i\bullet} = \sum_j p_{\bullet j} = 1$$

$$\sum_j p_{j/i\bullet} = {\sum_j x_{ij}}\big/{x_{i\bullet}} = 1$$

$$\sum_i p_{i/\bullet j} = {\sum_i x_{ij}}\big/{x_{\bullet j}} = 1$$

4-1.3 TEST DE HIPÓTESIS DE INDEPENDENCIA EN UNA TABLA DE CONTINGENCIA

4-1.3.1 Presentación intuitiva

Para cada profesión, la tabla 3 presenta la distribución de los individuos entre las zonas de residencia. Sin que sea una gran sorpresa, constatamos que los individuos de profesiones diferentes se reparten de manera diferente en el espacio, entre las zonas de residencia.

Tabla 3: Población activa empleada
en la región metropolitana de Montreal, 1991
Repartición entre las zonas de residencia según la profesión

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	0.241	0.334	0.274	0.281	0.327	*0.291*
Resto CUM	0.264	0.242	0.246	0.189	0.199	*0.229*
Anillo norte	0.214	0.175	0.216	0.234	0.207	*0.209*
Anillo sur	0.232	0.201	0.219	0.217	0.212	*0.216*
Fuera RMR	0.049	0.048	0.044	0.080	0.055	*0.055*
Total	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*

Sin embargo, ¿son significativas estas diferencias? Para examinar este problema, se comparan las distribuciones observadas con una distribución que sería, de manera hipotética, la misma para todas las profesiones; esta distribución hipotética es simplemente la distribución del total (última columna de la tabla).

Pero, ¿cómo decidir si estas diferencias son o no son "significativas"? Para esto se procede a un test de hipótesis (para más detalles sobre los tests de hipótesis, vea el capítulo 2-3). La hipótesis que pretendemos probar es que las distribuciones son idénticas y las diferencias observadas no son más que accidentes, productos del azar.

Hay tres etapas en este test:

1. Medir la desviación estándar entre lo que se observó y la hipótesis;
2. Determinar cuál es la probabilidad de que una diferencia tan grande sea el producto del azar (cuanto más grande es esta diferencia, menos probable es que sea producto del azar).
3. Tomar una decisión.

Primera etapa: medir la diferencia

Para medir la diferencia entre las observaciones y la hipótesis, es necesario, primero, tener una representación exacta de la hipótesis. Por lo tanto, se calculan las frecuencias que, teóricamente, se obtendrían si las distribuciones fuesen idénticas (tabla 4).

Tabla 4: Población activa empleada
en la Región Metropolitana de Montreal
Frecuencias teóricas
en la hipótesis de distribuciones idénticas

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	68,189.0	98,731.6	127,523.8	93,094.3	67,467.4	*455,006.0*
Resto CUM	53,665.1	77,702.2	100,361.9	73,265.7	53,097.1	*358,092.0*
Anillo norte	48,972.2	70,907.4	91,585.6	66,858.8	48,454.0	*326,778.0*
Anillo sur	50,468.6	73,074.1	94,384.0	68,901.8	49,934.5	*336,763.0*
Fuera RMR	12,765.1	18,482.7	23,872.7	17,427.4	12,630.0	*85,178.0*
Total	*234,060.0*	*338,898.0*	*437,728.0*	*319,548.0*	*231,583.0*	*1,561,817.0*

Se calculan estas frecuencias teóricas con sólo multiplicar el total de cada columna con la distribución de la totalidad (última columna de la tabla de las reparticiones): $x^*_{ij} = x_{\bullet j}\, p_{i\bullet}$ donde el asterisco sirve para distinguir las frecuencias teóricas de las frecuencias observadas. Por ejemplo:[184]

$$x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319548 \times 0.0545378 = 17427.4$$

Podemos observar que los totales de las líneas y de las columnas de la tabla 4 y los mismos totales de la tabla 2 de los valores observados son iguales. Esto no es casual y se deduce de la fórmula de cálculo

[184] En la fórmula que sigue y con el fin de obtener frecuencias teóricas exactas, se debe tomar el valor de la probabilidad con 7 decimales puesto que el multiplicador es del orden de cientos de miles. Se busca tal precisión en este contexto para lograr claridad en el desarrollo que sigue, sin embargo esta precisión no es necesaria en la práctica.

$$\sum_i x_{ij}^* = \sum_i x_{\bullet j}\, p_{i\bullet} = \sum_i x_{\bullet j}\left(\frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = x_{\bullet j}\,\frac{\sum_i x_{i\bullet}}{x_{\bullet\bullet}}$$

$$\sum_i x_{ij}^* = x_{\bullet j}\,\frac{x_{\bullet\bullet}}{x_{\bullet\bullet}} = x_{\bullet j}$$

$$\sum_j x_{ij}^* = \sum_j x_{\bullet j}\, p_{i\bullet} = p_{i\bullet}\sum_j x_{\bullet j} = p_{i\bullet}\, x_{\bullet\bullet}$$

$$\sum_j x_{ij}^* = \frac{x_{i\bullet}}{x_{\bullet\bullet}}\, x_{\bullet\bullet} = x_{i\bullet}$$

Después de calcular las frecuencias teóricas, hay que medir la diferencia entre el conjunto de las frecuencias teóricas y el conjunto de las frecuencias observadas. Para esto, se aplica la fórmula

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x_{ij}^*\right)^2}{x_{ij}^*}$$

Esta estadística es conocida como el Khi-dos (Ji o Chi cuadrado) de Pearson y con el símbolo X^2 tal y como aparece en la fórmula.

Tabla 5: Población activa empleada
en la Región Metropolitana de Montreal
Cálculo del Chi-cuadrado

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	2,051.7	2,134.6	444.4	121.9	998.5	*5,751.1*
Resto CUM	1,209.3	252.0	554.5	2,316.5	900.1	*5,232.4*
Anillo norte	24.7	1,913.6	103.3	949.6	5.1	*2,996.2*
Anillo sur	301.7	333.4	24.0	1.9	13.9	*675.0*
Fuera RMR	118.5	300.8	853.4	3,734.7	0.1	*5,007.5*
Total	*3,706.0*	*4,934.4*	*1,979.6*	*7,124.6*	*1,917.6*	*19,662.2*

Los valores de la tabla 5 son las contribuciones de las celdas individuales al Chi-cuadrado

Así, en el caso de la quinta celda de la cuarta columna

$$\frac{(25{,}495 - 17{,}427.4)^2}{17{,}427.4} = 3{,}734.7$$

El Chi-cuadrado es simplemente la suma de todos los elementos de esta tabla: 19,662.2.

Segunda etapa: determinar la probabilidad

¿Por qué se emplea esta fórmula y no otra? Encontramos la repuesta a esta pregunta en la teoría de la inducción estadística. Se emplea esta fórmula porque, gracias a la estadística matemática, se conoce la distribución de probabilidad del Chi-cuadrado que se calculó de esta manera. En efecto, el Chi-cuadrado posee una distribución asintótica muy conocida: es la ley del χ^2 (decimos “Chi-cuadrado” puesto que el símbolo χ es la letra griega “Chi”). Es posible aplicar este re-

sultado siempre y cuando se use un cierto modelo de muestreo; esto es, un modelo de muestreo es un modelo que describe el proceso aleatorio por el cual, suponemos, se generan las diferencias entre las frecuencias observadas y las frecuencias teóricas (vea capítulo 2-2). No estudiaremos este modelo en este momento pero sí notaremos que este modelo de muestreo es lo suficientemente general como para poder aplicar el test de hipótesis del Chi-cuadrado de Pearson a una gran variedad de situaciones (vea más abajo, 4-1.4).

Es importante, ahora, notar que al momento de usar la ley del χ^2, es necesario tomar en cuenta lo que conocemos como el número de grados de libertad del cual dependen las probabilidades que la ley del χ^2 nos da. Para el test de hipótesis del Chi-cuadrado de Pearson, el número de grados de libertad es igual a

$$(C - 1)\,(L - 1),$$

donde C es el número de columnas y L, el número de líneas en la tabla.

En nuestro ejemplo (tablas 2 a 5), C corresponde al número de profesiones y L, al número de zonas; por consiguiente, en número de grados de libertad es igual a

$$(5 - 1)\,(5 - 1) = 16$$

Hagamos un pequeño paréntesis sobre el número de grados de libertad. Se representa la Ley del Chi-cuadrado con una curva cuya forma varía con el número de valores que el azar es libre de perturbar, por así decirlo. En una tabla de contingencia, los totales de líneas y columnas son fijos, así que en cada una de las C columnas, una vez que este tremendo azar haya "libremente" perturbado $(L - 1)$ valores, el último valor de la columna es determinado por la diferencia entre el total y los otros $(L - 1)$ valores; de igual manera, en cada una de las L líneas, una vez que se haya "libremente" perturbado $(C - 1)$ valores, el último valor de la columna es determinado por la diferencia entre el total y los otros $(C - 1)$ valores. En consecuencia, en la tabla completa, una vez que

introducidas $(C - 1)$ $(L - 1)$ “perturbaciones”, se determinan los demás valores por la necesidad de respetar los totales marginales.

Por medio de una tabla del Chi cuadrado o de la función Ley.Khidos (CHIDIST en inglés o Prueba Chi) del tabulador Excel X^2, es posible ahora determinar la probabilidad de que la diferencia medida entre las frecuencias observadas y las frecuencias teóricas sea tan grande; en particular, el valor de Prueba Chi[185] (19662;16) es inferior a 2.4×10^{-300}.

Tercera etapa: tomar una decisión

Una probabilidad de 2.4×10^{-300} es una probabilidad tan pequeña que es en extremo improbable que las desviaciones de las frecuencias observadas con relación a las frecuencias teóricas se deben únicamente al azar. De hecho, es tan improbable que, al menos que surjan circunstancias excepcionales, la buena decisión que se debe tomar es rechazar esta hipótesis y, por lo contrario, concluir que existe ciertamente una relación entre la profesión y la zona de residencia.

4-1.3.2 ¿¡Datos idénticos, nueva pregunta... respuesta idéntica?!

Acabamos de examinar si era significativo que los individuos con profesiones diferentes se repartieran entre zonas de residencia diferentes. Ahora pretendemos saber si la composición profesional de la población empleada es, de manera significativa, diferente de una zona de residencia a otra. Los datos pertinentes se encuentran en la tabla 6, más abajo, la cual procura, para cada zona de residencia, la distribución de los individuos entre las profesiones.

[185] Prueba Chi es la función correspondiente a Xi cuadrado en la versión en español de Excel.

Para ser más precisos, queremos probar la hipótesis de que no existen diferencias significativas entre las zonas con relación a la composición profesional de las personas empleadas que ahí viven. En la tabla 6 se compara, por lo tanto, las diferentes distribuciones con aquellas que, teóricamente, encontraríamos si las distribuciones fueran idénticas, lo que corresponde a la distribución de la totalidad (último renglón).

Tabla 6: Población activa empleada
en la Región Metropolitana de Montreal
Composición profesional de las zonas de residencia, 1991

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	0.124	0.249	0.264	0.197	0.166	*1.000*
Resto CUM	0.172	0.229	0.301	0.168	0.129	*1.000*
Anillo norte	0.153	0.181	0.290	0.229	0.147	*1.000*
Anillo sur	0.161	0.202	0.285	0.206	0.146	*1.000*
Fuera RMR	0.135	0.189	0.227	0.299	0.149	*1.000*
Total	*0.150*	*0.217*	*0.280*	*0.205*	*0.148*	*1.000*

Es posible constatar que existen, de hecho diferencias entre las zonas de residencia en cuanto a la composición profesional. Para saber si estas diferencias son significativas, procedemos de la misma manera que anteriormente, es decir, empezamos por calcular las frecuencias teóricas. Sin embargo, ¡qué sorpresa!, las frecuencias teóricas que arroja el cálculo son las mismas que las frecuencias teóricas calculadas cuando examinábamos el problema de la distribución entre las zonas de las diferentes profesiones (el lector puede verifi-

carlo por sí mismo). Es inútil seguir, pues llegaremos forzosamente a la misma conclusión.

Claro está que esto no es el resultado del azar. En el primer caso, tenemos

$$x^*_{ij} = x_{\bullet j} \times p_{i\bullet}$$

Por ejemplo:

$$x^*_{54} = x_{\bullet 4} \times p_{5\bullet} = 319{,}548 \times 0.0545378 = 17{,}427.4$$

En el caso presente,

$$x^*_{ij} = x_{i\bullet} \times p_{\bullet j}$$

Por ejemplo:

$$x^*_{54} = x_{5\bullet} \times p_{\bullet 4} = 85{,}178 \times 0.2046002 = 17{,}427.4$$

Las dos fórmulas permiten llegar al mismo resultado numérico por ser completamente equivalentes:

$$x^*_{ij} = x_{\bullet j}\ p_{i\bullet} = x_{\bullet j} \frac{x_{i\bullet}}{x_{\bullet\bullet}} = x_{i\bullet} \frac{x_{\bullet j}}{x_{\bullet\bullet}} = x_{i\bullet}\ p_{\bullet j}$$

De por sí, en la práctica cambiamos por lo general de la tabla de las frecuencias observadas a la tabla de las frecuencias teóricas por medio de la fórmula

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$$

La tabla de las frecuencias teóricas es, por lo tanto, biproporcional, es decir, las columnas son proporcionales entre sí, así como las líneas.

4-1.3.3 Generalización: la independencia estadística en una tabla de contingencia

El análisis de una tabla de contingencia se basa en el postulado de que el número de individuos observados en las celdas de la tabla depende de una estructura subyacente. El análisis tiene como objetivo descubrir esta estructura. Evocaremos

más tarde y de manera breve el modelo log-lineal que sirve para representar esta estructura. En este instante, sólo nos interesa un aspecto particular de esta estructura: la independencia estadística.

¿Qué es la independencia estadística? En la teoría de las probabilidades, un evento aleatorio *A* es independiente de otro evento *B* cuando la probabilidad que el evento *A* suceda siga la misma de que el evento *B* suceda o no. Por ejemplo, en la tabla de la población activa por zona de residencia y por profesión, hay independencia entre las variables zona de residencia y profesión si, por un individuo sorteado, la probabilidad de vivir en una zona dada es la misma no importando la profesión de este individuo. De manera simétrica, hay independencia si la probabilidad de pertenecer a un grupo de profesión dada es la misma no importando la zona de residencia del individuo.

Por ejemplo, digamos que el evento *A* es "el individuo vive en el Anillo Norte" y el evento *B* es "el individuo es empleado de oficina": si hubiera independencia, la probabilidad de que un individuo sorteado viva en el Anillo Norte (probabilidad del evento *A*) sería la misma no importando que este individuo fuera empleado de oficina (evento *B*) o no.

Examinemos de más cerca cómo se manifiesta la independencia en una tabla de contingencia. Con este objetivo, es importante empezar por interpretar las frecuencias relativas de la tabla como si fueran probabilidades observadas, las mismas que se confrontarán en contra de las probabilidades teóricas del modelo o de la hipótesis. Así, para un individuo sorteado entre 1,561,817 trabajadores censados de la RMR, la probabilidad de que sea un obrero y que viva en el Anillo Sur se define con

$$p_{44} = \frac{x_{44}}{x_{\bullet\bullet}} = \frac{69{,}263}{1{,}561{,}817} = 0.044$$

De la misma manera se calculan las probabilidades marginales observadas. Así, para un individuo sorteado entre

1,561,817 trabajadores censados de la RMR, la probabilidad de que sea un obrero se define con

$$p_{\bullet 4} = \sum_i p_{i4} = \sum_i \left(\frac{x_{i4}}{x_{\bullet\bullet}} \right) = \frac{x_{\bullet 4}}{x_{\bullet\bullet}} = \frac{319{,}548}{1{,}561{,}817} = 0.205$$

Además, la probabilidad que un individuo sorteado entre 1,561,817 trabajadores censados de la RMR, viva en el Anillo Sur se define con

$$p_{4\bullet} = \sum_j p_{4j} = \sum_j \left(\frac{x_{4j}}{x_{\bullet\bullet}} \right) = \frac{x_{4\bullet}}{x_{\bullet\bullet}} = \frac{336{,}763}{1{,}561{,}817} = 0.216$$

Y con todo esto, ¿dónde está la independencia? Si la probabilidad de ser un obrero es independiente de la zona de residencia entonces la fracción de obreros en cada zona debería ser igual a $p_{\bullet 4}$, o sea a 20.5%. Como sabemos que la fracción de trabajadores que viven en la Anillo Sur es igual a $p_{4\bullet}$, o sea a 21.6%, entonces, entre los 1,561,817 trabajadores censados de la RMR, los que son obreros y viven en la Anillo Sur deberían representar 20.5% de 21.6% del total, o sea

$$p_{\bullet 4} \times p_{4\bullet} = 0.205 \times 0.216 = 0.044$$

Esto, lo recordamos, en caso que los dos eventos (ser obrero y vivir en la Couronne Sud) sean independientes. En este particular, pasa que el resultado es muy cercano al valor de p_{44}, lo que permite creer que en efecto los dos eventos podrían ser independientes.

Sin embargo, este análisis es incompleto porque cada una de las dos variables, zona de residencia y profesión, contiene más de dos categorías. Queriendo, por lo tanto, generalizar, si dos variables son independientes, es de esperarse que la probabilidad observada de pertenecer al mismo tiempo a la categoría *i* de la primera variable y a la categoría *j* de la segunda sea igual al producto de las probabilidades marginales:

$p_{ij} = p_{i\bullet} \times p_{\bullet j}$ para todos los pares *i*,*j*

Podemos llegar a la misma conclusión tomando otro camino aunque partiendo de la misma definición, o sea: "Un evento *A* es independiente de otro evento *B* cuando la probabilidad de que el evento *A* suceda siga la misma de que el evento *B* suceda o no." En el lenguaje de la teoría de las probabilidades, este enunciado equivale a decir que la probabilidad condicional de *A* es igual a su probabilidad marginal, es decir, en una tabla de contingencia de 2 dimensiones:

$$p_{i/\bullet j} = p_{i\bullet}$$

Puesto que $p_{i/\bullet j} = p_{ij} \Big/ p_{\bullet j}$, esto implica $p_{i\bullet} = p_{ij} \Big/ p_{\bullet j}$, o sea $p_{ij} = p_{i\bullet}\ p_{\bullet j}$

De una manera equivalente, las 2 variables categóricas son independientes si:

$$p_{j/i\bullet} = p_{\bullet j}$$

Puesto que $p_{j/i\bullet} = p_{ij} \Big/ p_{i\bullet}$, esto implica $p_{\bullet j} = p_{ij} \Big/ p_{i\bullet}$, o sea $p_{ij} = p_{i\bullet}\ p_{\bullet j}$.

Es ésta la definición exacta de la independencia estadística entre dos variables categóricas. Resalta a la vista que esta definición es perfectamente simétrica con relación a las dos variables. Además, es posible constatar que las frecuencias teóricas de las cuales se trato más arriba son las probabilidades que se esperaría ver en la hipótesis de la independencia estadística. En efecto:

$$x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}} = \left(\frac{x_{i\bullet}}{x_{\bullet\bullet}}\right)\left(\frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) x_{\bullet\bullet} = p_{i\bullet}\ p_{\bullet j}\ x_{\bullet\bullet}$$

La tabla de frecuencias teóricas (tabla 4) es, por lo tanto, una representación exacta de la hipótesis de independencia.

Pero, ¿por qué nos interesa tanto la hipótesis de la independencia? Porque, si dos variables son independientes, se puede considerar que ninguna de las dos tiene una influencia sobre la otra (observe que, en esta formulación, no se identifica a ninguna de las dos variables para que tomen el papel de variable independiente o "explicativa").

Puede pasar que los datos se conformen perfectamente a la hipótesis de independencia. Sin embargo, es mucho más frecuente que los datos difieran de lo que predice el modelo de independencia. En caso de que no difieran "demasiado", se podrá juzgar el modelo aceptable siempre y cuando admitamos que no es más que una aproximación y que entre la realidad y el modelo, interviene un elemento aleatorio que hemos llamado "perturbación debido al azar". Las hipótesis que emitimos en cuanto a este elemento aleatorio permiten delimitar la incertidumbre en cuanto al "verdadero" modelo.

El test de hipótesis que acabamos de describir es, por lo tanto, un procedimiento de inducción estadística que apoya la decisión de rechazar o no rechazar el modelo de la independencia estadística.

4-1.3.4 Otro test: el test de la relación de verosimilitud

Se usa también la estadística de la relación de verosimilitud (con más exactitud, menos dos veces el logaritmo de la razón de las funciones de verisimilitud). Para una tabla rectangular, esta estadística se define con[186]

[186] La fórmula que enseñamos a continuación es de hecho la correcta. Difiere de la definición informal que da Upton (1981, p. 36, definición del Y^2).

$$G^2 = -2\sum_i \sum_j x_{ij} \ln\left(\frac{x^*_{ij}}{x_{ij}}\right)$$

$$G^2 = 2\sum_i \sum_j x_{ij} \ln\left(\frac{x_{ij}}{x^*_{ij}}\right)$$

Como en el caso del Chi-cuadrado de Pearson, G^2 posee una distribución asintótica χ^2 con, bajo la hipótesis de independencia, $(L - 1)\ (C - 1)$ grados de libertad.[187]

Se da un ejemplo del cálculo de esta variable-test tomando los datos de la tabla de la población activa en función de la profesión y la zona de residencia.

Tabla 7: Población activa empleada
en la Región Metropolitana de Montreal
Cálculo de la estadística de la relación de verosimilitud G^2

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	–10,737.1	15,536.0	–7,301.1	–3,307.6	8,687.8	*2,878.0*
Resto CUM	8,632.4	4,548.4	7,730.8	–11,793.9	–6,442.2	*2,675.5*
Anillo norte	1,112.0	–10,634.5	3,126.5	8,425.2	–492.4	*1,536.8*
Anillo sur	4,049.5	–4,765.5	1,517.9	362.2	–826.5	*337.6*
Fuera RMR	–1,168.8	–2,200.5	–4,057.2	9,699.2	34.0	*2,306.7*
Total	*1,888.0*	*2,484.0*	*1,016.8*	*3,385.1*	*960.7*	*9,734.6*

[187] Si bien X^2 y G^2 tienen la misma distribución asintótica, esto no implica que tengan el mismo valor.

Los valores de la tabla 7 son las contribuciones de las celdas individuales al G^2. Así, para la quinta celda de la cuarta columna,

$$25{,}495 \times \ln\left(\frac{25{,}495}{17{,}427.4}\right) = 9{,}699.2$$

El G^2 es sencillamente igual al doble de la suma de todos los elementos de esta tabla: o sea, 19,469.2. La probabilidad crítica correspondiente es inferior a 2.4×10^{-300}.

4-1.4 UN ESPECIAL VISTAZO SOBRE EL CHI-CUADRADO DE PEARSON

4-1.4.1 Las infinitas aplicaciones del test del Chi-cuadrado de Pearson a las tablas de contingencia

Test sobre una sola celda de la tabla

Es posible interpretar cada uno de los términos de la doble sumatoria que integra el Chi-cuadrado como la "contribución" de la celda correspondiente al Chi-cuadrado. Esto nos permite detectar las celdas más "desviadas" con relación a la hipótesis.

Es también posible probar de manera formal la hipótesis de que una celda especifica de la tabla es significativamente "desviada". Sólo se necesita construir una tabla donde se agregue todas las demás líneas y columnas. Por ejemplo, en caso de querer efectuar el test para la celda [*h*, *k*], se construye una tabla agregada 2×2 como lo muestra el modelo siguiente:

x_{hk}	$\sum_{j} x_{hj}$
$\sum_{i} x_{ik}$	$\sum_{i \neq h} \sum_{j \neq k} x_{ij}$

Luego, aplicamos a esta tabla el test del Chi-cuadrado de Pearson con 1 grado de libertad:

$(L - 1)\ (C - 1) = (2 - 1)\ (2 - 1) = 1$

Consideremos, por ejemplo, la fracción de los empleados que viven al exterior de la RMR y que pertenecen a las profesiones Trabajadores especializados en los servicios, personal de explotación de los transportes, etc. En la tabla 5, se puede observar que esta celda de la tabla no contribuye más que por 0.1 con el valor total del Chi-cuadrado. Podemos probar la hipótesis de que esta desviación no es significativa con relación a la hipótesis de independencia. A partir de la tabla de contingencia, se construye la tabla agregada que sigue:

Tabla 8: Tabla agregada Población activa empleada,
Región Metropolitana de Montreal, 1991
Zona residencial, según el sexo y la profesión

	Todas las professones menos →	Trabajadores especializados en los servicios, personal de explotación de los transportes, etc.	Total
RMR	1,257,720	218,919	*85,178*
Fuera de RMR	72,514	12,664	*1,476,639*
Total	*231,583*	*1,330,234*	*1,561,817*

Fuente: Statistique Canada, Censo de 1991.

El valor del chi cuadrado que se calculó a partir de esta tabla es de 0.11, lo que queda bastante alejado del 19,662.2 obtenido con la tabla detallada. La probabilidad crítica aso-

ciada a 0.11 es de 74%, lo que, claramente, no nos permite rechazar la hipótesis de independencia en la tabla agregada.

Test de homogeneidad entre dos o más grupos o muestras

A menudo sucede que tengamos que analizar tablas que comparan dos grupos de individuos distribuidos en varias categorías. En el caso de dos grupos, la tabla de contingencia tiene el aspecto siguiente:

	Grupo *A*	Grupo *B*	Total *A*+*B*
Categorías			
Total			

Un test de homogeneidad entre dos o más grupos busca determinar si, desde el punto de vista de su repartición entre las categorías de una variable de clasificación dada, ambos son o no significativamente diferentes. Podríamos pretender, por ejemplo, comparar la repartición de los hombres y de las mujeres entre las profesiones. No será muy difícil reconocer que el problema por saber si la repartición de las mujeres entre las profesiones es significativamente diferente de aquella de los hombres no es más que el problema de independencia entre la variable Profesión y la variable Sexo.

Test de homogeneidad entre una subpoblación y el resto de la población

El test de homogeneidad sirve entre otras cosas para comparar un grupo particular con el resto de la población. En particular, se emplea para comparar una muestra con la población

de donde se obtiene para saber si es representativa de algunas características conocidas de la población.

Supongamos, por ejemplo, que efectuemos un sondeo por medio de entrevistas a los residentes de Montreal, los cuales se escogen al azar en el cruce de las calles Sainte-Catherine y Jeanne-Mance. Si consideramos que la variable lingüística es importante para alcanzar el objetivo de tal estudio, tendremos que verificar, al final, si la proporción de francófonos y de anglófonos entrevistados es representativa de la proporción lingüística de Montreal. Con este fin, se construye una tabla basándose en el modelo siguiente:

	Grupo *A*	Resto de la población	Total
Categorías		Calcular por sustracción	
Total			

En nuestro ejemplo, las categorías pertinentes son obviamente Francófono, Anglófono y Otros.[188] Los datos del grupo *A* son aquellos de la muestra o de otro grupo específico del estudio; es posible obtener los datos de la columna Total en

[188] No queremos mencionar ahora la dificultad que existe para definir la pertenencia lingüística de modo operacional y aún más grande dificultad para encontrar en los datos del Censo de *Statistique Canada* la información pertinente (la formulación de las preguntas del censo con relación a la pertenencia lingüística es el objeto de abiertas y fuertes críticas).

cualquier fuente oficial como un censo. Se efectúa el cálculo de las cifras del Resto de la población por sustracción.[189]

Test de la hipótesis de una distribución particular

Generalizando, el test del Chi-cuadrado puede servir para evaluar cualquier hipótesis sobre una distribución de un conjunto de individuos entre categorías.[190] Para esto el conjunto de individuos estudiado es considerado como si fuera una muestra obtenida de una población infinita, la cual debe estar distribuida según la hipótesis que se pretende evaluar.

Por ejemplo, según el censo de la población de 1984 en Costa Rica, este país contaba entonces con 630,995 hombres y 649,619 mujeres (tabla 9). Confrontemos estas cifras con la hipótesis de una distribución 50-50 entre los sexos.

Tabla 9: Población masculina y femenina, Costa Rica, 1984

	Datos del censo	Frecuencias teóricas
Hombres	630 995	640 307
Mujeres	649 619	640 307
Total	*1 280 614*	*1 280 614*

Fuente: http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

Calculemos el valor del Chi-cuadrado como

$$X^2 = \frac{(630{,}995 - 640{,}307)^2}{640{,}307} + \frac{(649{,}619 - 640{,}307)^2}{640{,}307}$$

$$X^2 = 270.85$$

[189] En caso de que el grupo *A* no represente más que una mínima fracción de toda la población, es posible, en la practica, calcular el Khi-dos entre el grupo *A* y la totalidad aunque no sea teóricamente exacto.
[190] Blalock (1972, p. 312), ejercicio núm. 3.

La probabilidad crítica con 1 grado de libertad es igual a 7.4×10^{-61}, lo que lleva a rechazar la hipótesis de que la distribución de la población entre hombres y mujeres no es significativamente diferente de la distribución 50-50.

En apariencia, este procedimiento difiere de aquel empleado hasta el momento. Pero tal no es el caso, puesto que este test se fundamenta en la comparación implícita que se efectúa entre la población estudiada y una población hipotética de tamaño infinito, la cual respeta la distribución hipotética que se pretende probar. De manera explicita, presentamos a continuación la tabla de contingencia subyacente a este test.

Tabla 10: Población masculina y femenina, Costa Rica, 1984

	Población Costa Rica, 1984	Resto	Población hipotética infinita
Hombres	630 995	$0.5 \times Y - 630\,995$	$0.5 \times Y$
Mujeres	649 619	$0.5 \times Y - 649\,619$	$0.5 \times Y$
Total	1 280 614	$Y - 1\,280\,614$	Y

Fuente: http://populi.eest.ucr.ac.cr/observa/estima/cuadro1.htm

Cálculo de las frecuencias teóricas

	Población Costa Rica 1984	Resto
Hombres	$\frac{(630995 - 640307)^2}{640307}$	$\frac{\{(0{,}5\,Y - 630995) - [0{,}5\,(Y - 1280614)]\}^2}{0{,}5\,(Y - 1280614)} = \frac{(640307 - 630995)^2}{0{,}5\,(Y - 1280614)}$
Mujeres	$\frac{(649619 - 640307)^2}{640307}$	$\frac{\{(0{,}5\,Y - 649307) - [0{,}5\,(Y - 1280614)]\}^2}{0{,}5\,(Y - 1280614)} = \frac{(640307 - 649307)^2}{0{,}5\,(Y - 1280614)}$

Si Y es infinitamente grande, la contribución de la columna Resto en el valor del Chi-cuadrado es despreciable (infinitamente pequeño) puesto que el divisor 0.5 (Y – 1,280,614) es infinitamente grande de tal modo que el cálculo equivale a lo hecho anteriormente. Además, el número de grados de libertad es efectivamente igual a $(C - 1)(L - 1)$.[191]

4-1.4.2 Condiciones de validez del test del Chi-cuadrado de Pearson

El test del Chi-cuadrado de Pearson se basa en una aproximación: la distribución del $\chi 2$ es la distribución asintótica de la estadística del Chi-cuadrado de Pearson. Para que sea válido el test, tiene que ser bastante buena la aproximación. Por lo general, se considera que la aproximación es bastante buena y que el test es válido, cuando el número total de observaciones respeta la condición

$$x_{\bullet\bullet} > 10 \times L \times C$$

donde C es el número de columnas y L, el número de líneas de la tabla (Legendre y Legendre, 1998, p. 218).

En la práctica la mayoría de los autores afirman que el test del Chi-cuadrado de Pearson podría no ser válido si existen una o más celdas que contienen menos de 5 observaciones (Freund y Williams, 1973, p. 379).

Según Legendre y Legendre (1998, p. 218), el test podría no ser válido si $x_{\bullet\bullet} < 5 \times L \times C$. Esa condición es muy cercana a la anterior: cuando cumple esa condición, hay necesariamente por lo menos una celda con frecuencia teórica abajo de 5, tal como se demuestra a continuación.

[191] Puede haber situaciones en que el cálculo de las frecuencias teóricas se someta a más de una restricción; en estas condiciones, el número de grados de libertad se calcula de otra manera. Vea Blalock (1972, ejercicio 3, p. 312).

Tenemos :

$\underset{i}{MIN}\left[x_{i\bullet}\right] \leq \frac{x_{\bullet\bullet}}{L}$ et $\underset{j}{MIN}\left[x_{\bullet j}\right] \leq \frac{x_{\bullet\bullet}}{C}$, de manera que

$\underset{i,j}{MIN}\left[x_{i\bullet} x_{\bullet j}\right] \leq \frac{(x_{\bullet\bullet})^2}{L \times C}$, o sea $\underset{i,j}{MIN}\left[\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}\right] \leq \frac{x_{\bullet\bullet}}{L \times C}$

Sigue que, si $x_{\bullet\bullet} < 5 \times L \times C$, o sea si $\frac{x_{\bullet\bullet}}{L \times C} < 5$, entonces la más pequeña frecuencia teórica $x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$ estará abajo de 5.

En cambio, Cochran (1954) y Siegel (1956), a los cuales se refieren Legendre y Legendre (1998, p. 218) emiten las condiciones siguiente, menos restrictivas, que invalidarían el test del Chi-cuadrado:

- Existe una o más celdas *ij* cuya frecuencia teórica x^*_{ij} es inferior a 1.

NOTA: Puesto que $x^*_{ij} = \frac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}}$, esta condición equivale a decir que existe al menos una línea *i* y una columna *j* tales que $x_{i\bullet}\ x_{\bullet j} < x_{\bullet\bullet}$

O bien:

- Existe 20% de las celdas *ij* cuya frecuencia teórica x^*_{ij} es inferior a 5.

Ahora bien, según otros autores, hasta esta última condición parece ser un tanto severa. Legendre y Legendre (1998, p. 218) citan a Fienberg (1950), para quien el test es válido con

un umbral de significación de 5% siempre y cuando todas las frecuencias teóricas sean superiores a 1.

Concretamente, recordaremos que al momento de aplicar el test del Chi-cuadrado de Pearson a una tabla de contingencia, es importante desconfiar de los resultados cuando algunas de las frecuencias teóricas son demasiadas pequeñas.

En caso de tener buenas razones para desconfiar de la validez del test del Chi-cuadrado, ¿qué podemos hacer? Una primera posibilidad es agrupar unas categorías para así fusionar las filas o columnas que sólo contienen un pequeño número de observaciones. Se obtendrán, de esta manera, frecuencias teóricas más altas en las celdas fusionadas. Pero, ¡no basta agrupar categorías de cualquier manera! Agrupar categorías equivale a cambiar el modo de operacionalización de la hipótesis (vea el inicio del capítulo 2-2). Hay que justificarlo con relación al modelo conceptual subyacente a la investigación.

Por otra parte, a menudo es preferible descartar del análisis las categorías que dificultan la interpretación (por ejemplo, las respuestas “No sé” en los datos de encuestas). Se pensará descartar hasta categorías que sí tienen un contenido analítico, pero tienen un pequeño número de observaciones, mientras que no se pueden agrupar con otras para constituir nuevas categorías que sean pertinentes en relación con el modelo conceptual.

4-1.4.3 Algunas propiedades numéricas del test del Chi-cuadrado de Pearson

El Chi-cuadrado de Pearson posee las propiedades que, a continuación enumeramos:

1. Chi-cuadrado es no negativo.
2. Chi-cuadrado es nulo cuando $x_{ij} = x_{ij}^{*}$ para todas las celdas i,j de la tabla.

3. Chi-cuadrado aumenta con el número de observaciones $x_{\bullet\bullet}$.

4. $X^2 \leq x_{\bullet\bullet} Min(L-1, C-1)$

donde C es el número de columnas y L, el número de líneas de la tabla y donde la expresión $Min(L-1, C-1)$ representa el más pequeño valor entre $(L-1)$ y $(C-1)$.

Las dos primeras propiedades son relativamente evidentes si observamos la fórmula de cálculo

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*}$$

Se ilustra la tercera propiedad con el ejemplo que sigue:

Tabla 11: Sensibilidad del Chi-cuadrado al número de observaciones. Ilustración numérica

Tablas de contingencia

			$p_{i\bullet}$
15	10	*25*	0.5
10	15	*25*	0.5
25	*25*	*50*	
0.5	0.5	← $p_{\bullet j}$	

			$p_{i\bullet}$
30	20	*50*	0.5
20	30	*50*	0.5
50	*50*	*100*	
0.5	0.5	← $p_{\bullet j}$	

			$p_{i\bullet}$
60	40	*100*	0,5
40	60	*100*	0,5
100	*100*	*200*	
0.5	0.5	← $p_{\bullet j}$	

Frecuencias teóricas

12.5	12.5	*25*
12.5	12.5	*25*
25	*25*	*50*

25	25	*50*
25	25	*50*
50	*50*	*100*

50	50	*100*
50	50	*100*
100	*100*	*200*

Cálculo del Chi-cuadrado

0.5	0.5
0.5	0.5

Chi-cuadrado = 2
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.157

1	1
1	1

Chi-cuadrado = 4
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.046

2	2
2	2

Chi-cuadrado = 8
núm. de lineas = 2
núm. de col. = 2
grad. de libertad = 1
Prob. crítica = 0.005

Las tres tablas de contingencia de arriba poseen estructuras idénticas. La única cosa que las distingue es el número de observaciones, que son 25, 50 y 100 respectivamente. El test del Chi-cuadrado nos conduce a rechazar la hipótesis de independencia en el tercer caso y, aunque de manera no tan categórica, también en el segundo; por lo contrario, en el primer caso, se tomaría, por lo general, la decisión de no rechazar la hipótesis, al menos dentro de los criterios que se acostumbran usar en ciencias sociales.

Generalizando, cuando para una estructura dada el número de observaciones aumenta en proporción en toda las celdas, el valor del Chi-cuadrado aumenta en la misma proporción. De manera formal, cuando el número de observaciones es multiplicado por α, tenemos:

$$\sum_i \sum_j \frac{\left(\alpha x_{ij}^* - \alpha x_{ij}\right)^2}{\alpha x_{ij}^*} = \sum_i \sum_j \frac{\alpha^2 \left(x_{ij}^* - x_{ij}\right)^2}{\alpha x_{ij}^*}$$

$$\sum_i \sum_j \frac{\left(\alpha x_{ij}^* - \alpha x_{ij}\right)^2}{\alpha x_{ij}^*} = \alpha \left[\sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*} \right]$$

¿Por qué nos interesa esta propiedad? Porque si el número total de observaciones $x_{\bullet\bullet}$ es grande, esto nos puede incitar a rechazar la hipótesis de independencia (y por lo tanto, a considerar que las diferencias entre las distribuciones son estadísticamente significativas) cuando, por lo contrario, no son, de manera segura, científicamente significativas. Al revés, puede suceder que las diferencias reales parezcan estadísticamente no significativas si el número de observaciones es pequeño.

Supongamos, por ejemplo, que en lugar de usar los datos del Censo sobre la profesión y la zona de residencia, tomáramos una muestra de 1 sobre 1000. Supongamos, también, que por una suerte increíble, la muestra sea un buen reflejo de la población de tal manera que las frecuencias observadas fueran iguales a una milésima de las frecuencias observadas del Censo tomando en cuenta el error de redondeo que se podría cometer puesto que es imposible tener fracciones de personas en la muestra. En estas condiciones, obtendríamos la tabla siguiente:

Tabla 12: Muestra ficticia de la población activa empleada en la Región Metropolitana de Montreal
Zona de residencia según la profesión, 1991

	Profesiones					
Zona de residencia	Directores, gerentes, administradores y similares	Profesionales, docentes y cuellos blancos especializados	Empleados de oficina y trabajadores en la venta	Obreros	Trabajadores especializados en los servicios, personal de explotación de transportes, etc.	TOTAL todas las profesiones
Montreal	56	113	120	90	76	455
Resto CUM	62	82	108	60	46	358
Anillo norte	50	59	95	75	48	327
Anillo sur	54	68	96	69	49	337
Fuera RMR	12	16	19	25	13	85
Total	234	339	438	320	232	1,562

Con los datos de esta muestra ficticia pero eminentemente representativa, el valor del Chi-cuadrado de Pearson no es más que 19.79 y la probabilidad correspondiente es de 0.23, por lo tanto ¡no es posible rechazar la hipótesis de independencia!

Regresaremos a este punto cuando tratemos las mediciones de la intensidad de la relación entre dos variables categóricas. De por sí, la cuarta propiedad del Chi-cuadrado de Pearson,

$$X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$$

interviene en la definición de algunas de estas mediciones.

Demostración de que $X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$

La demostración de esta última propiedad requiere una fórmula de cálculo del Chi-cuadrado que difiere de la fórmula que dimos anteriormente. Esta fórmula se deriva de la primera:

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* - x_{ij}\right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left[\left(x_{ij}^*\right)^2 - 2\, x_{ij}^*\, x_{ij} + x_{ij}^2\right]}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j x_{ij}^* - 2\sum_i \sum_j x_{ij} + \sum_i \sum_j \frac{x_{ij}^2}{x_{ij}^*}$$

$$X^2 = x_{\bullet\bullet} - 2x_{\bullet\bullet} + \sum_i \sum_j \frac{x_{ij}^2}{\left(\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}\right)}$$

$$X^2 = x_{\bullet\bullet}\left[\sum_i \sum_j \frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} - 1\right]$$

Para demostrar la cuarta propiedad, sólo basta constatar que,

por una parte $\frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \frac{x_{ij}}{x_{i\bullet}}$

de tal manera que

$$\sum_{i=1}^{L} \sum_{j=1}^{C} \frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \sum_{i=1}^{L} \sum_{j=1}^{C} \frac{x_{ij}}{x_{i\bullet}} = \sum_{i=1}^{L} \frac{x_{i\bullet}}{x_{i\bullet}} = L\,,$$

y por otra, $\frac{x_{ij}^2}{x_{i\bullet} x_{\bullet j}} \leq \frac{x_{ij}}{x_{\bullet j}}$

de tal manera que

$$\sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} \le \sum_{i=1}^{L}\sum_{j=1}^{C}\frac{x_{ij}}{x_{\bullet j}} = \sum_{i=1}^{C}\frac{x_{\bullet j}}{x_{\bullet j}} = C$$

Se deduce que

$$X^2 = x_{\bullet\bullet}\left[\sum_i\sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \le x_{\bullet\bullet}\left[L-1\right]$$

y

$$X^2 = x_{\bullet\bullet}\left[\sum_i\sum_j \frac{x_{ij}^2}{x_{i\bullet}x_{\bullet j}} - 1\right] \le x_{\bullet\bullet}\left[C-1\right]$$

Es lo que queríamos demostrar.

4-1.4.4 Post scriptum: una nueva mirada sobre el cociente de localización

En el capítulo 1-2 presentamos el cociente de localización como un instrumento que sirve para analizar una tabla del empleo por rama y por ciudad o región. Además, mencionamos que este tipo de tabla es una tabla de contingencia (de dos dimensiones). Es posible, por lo tanto, usar el mismo cálculo para cualquier tabla de contingencia que tenga dos dimensiones (aunque el término "localización" sea un tanto incongruente en algunas ocasiones).

Aun más interesante es poder reexaminar el cociente de localización bajo la luz de los tests de hipótesis aplicados a las tablas de contingencias. Concretamente, existe una relación muy simple entre los cocientes de localización y los números que se esperan de la hipótesis de independencia. Se definen estos últimos con

$$x_{ij}^* = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}$$

En cuanto a los cocientes de localización, se calculan con la fórmula

$$QL_{ij} = \frac{\left(x_{ij} \Big/ x_{\bullet j} \right)}{\left(x_{i\bullet} \Big/ x_{\bullet\bullet} \right)},$$

lo que es equivalente, con $QL_{ij} = \dfrac{\left(x_{ij} \Big/ x_{i\bullet} \right)}{\left(x_{\bullet j} \Big/ x_{\bullet\bullet} \right)}$

sea, $QL_{ij} = \dfrac{x_{ij}}{\left(\dfrac{x_{i\bullet}\ x_{\bullet j}}{x_{\bullet\bullet}} \right)} = \dfrac{x_{ij}}{x^*_{ij}}$

El cociente de localización es, por consiguiente, la razón de la frecuencia observada entre la frecuencia teórica bajo la hipótesis de independencia; como lo vimos, esta hipótesis se traduce con una tabla biproporcional. Esta relación permite también expresar el Chi-cuadrado de Pearson en función de los cocientes de localización. En efecto, puesto que tenemos

$$x^*_{ij}\ QL_{ij} = x_{ij}$$

tenemos también

$$X^2 = \sum_i \sum_j x^*_{ij} \left(QL_{ij} - 1 \right)^2$$

Demostración de $X^2 = \sum_i \sum_j x^*_{ij} \left(QL_{ij} - 1 \right)^2$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij} - x^*_{ij} \right)^2}{x^*_{ij}}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^* \; QL_{ij} - x_{ij}^*\right)^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left[x_{ij}^* \left(QL_{ij} - 1\right)\right]^2}{x_{ij}^*}$$

$$X^2 = \sum_i \sum_j \frac{\left(x_{ij}^*\right)^2 \left(QL_{ij} - 1\right)^2}{x_{ij}^*}$$

El Chi-cuadrado de Pearson es una suma ponderada de los cuadrados de las desviaciones de los cocientes de localización con relación al valor de referencia 1; en particular, el peso de cada celda es la frecuencia teórica bajo la hipótesis de independencia.

Tratándose del estudio de una tabla de empleo por ramo y por ciudad o región, parece claro que casi siempre el test del Chi-cuadrado desembocará en rechazar de manera categórica la hipótesis de independencia. Así, el verdadero interés de examinar esta relación es poder interpretar cada uno de los términos de la doble sumatoria como la contribución de la celda correspondiente al Chi-cuadrado. En términos relativos, la razón

$$\frac{x_{ij}^* \left(QL_{ij} - 1\right)^2}{X^2}$$

es la parte de la desviación total (con relación a la biproporcionalidad) que se puede atribuir a la celda *i*,*j*.

4-1.5 MEDICIONES DE LA INTENSIDAD DE LA RELACIÓN ENTRE DOS VARIABLES CATEGÓRICAS

Cuando rechazamos la hipótesis de independencia, esto significa que decidimos que existe una relación estadísticamente significativa entre las dos variables. No obstante, esta relación estadísticamente significativa no es necesariamente pertinente o importante desde un punto de vista científico o práctico. Se destaca sobremanera la necesidad de esta distinción, en particular, cuando examinemos la tercera propiedad que analizamos en 4-3 (el Chi-cuadrado aumenta con el número de observaciones). De aquí, lo útil que representa medir la intensidad de la relación entre las dos variables categóricas.

4-1.5.1 Mediciones derivadas del Chi-cuadrado de Pearson

Como pudimos ver, el Chi-cuadrado de Pearson posee algunas propiedades numéricas no deseables como medición de la intensidad de la relación entre dos variables categóricas:

X^2 aumenta con el número de observaciones $x_{\bullet\bullet}$

$$X^2 \leq x_{\bullet\bullet}\, Min(L-1, C-1)$$

donde *C* es el número de columnas y *L*, el número de líneas de la tabla.

Nos interesa una medición que refleje la estructura más que el número de observaciones y que, en condiciones ideales, variaría entre 0 y 1 en lugar de entre 0 y $x_{\bullet\bullet}\, Min(L-1, C-1)$.

A continuación, enlistamos algunas mediciones que se derivan del Chi-cuadrado de Pearson.

1. $\varphi^2 = \frac{X^2}{x_{\bullet\bullet}}$

Este "Fi-cuadrado" (el símbolo φ es la letra griega "Fi") varía entre 0 y 1 para las tablas 2 x 2, pero, en general, su valor máximo es, de manera considerable, mucho más elevado.

2. El T^2 de Tschuprow:

$$T^2 = \frac{\varphi^2}{\sqrt{(L-1)(C-1)}} = \frac{X^2}{x_{\bullet\bullet}\sqrt{(L-1)(C-1)}}$$

Su valor máximo es igual a 1 si $L = C$; de otra manera, es estrictamente inferior a 1.

3. El V^2 de Cramer

$$V^2 = \frac{\varphi^2}{Min(L-1, C-1)} = \frac{X^2}{x_{\bullet\bullet} Min(L-1, C-1)}$$

El V^2 de Cramer es equivalente al T^2 de Tschuprow cuando $L = C$, pero al opuesto de este último, puede tomar el valor máximo de 1 cuando $L \neq C$.

4-1.5.2 Otras mediciones (tau y lambda)

Principio general

El tau de Goodman y Kruskal (del nombre de la letra griega τ, que tiene el valor fonético de "T") y el lambda (del nombre de la letra griega λ que tiene el valor fonético de "L") no son simétricos puesto que sus mediciones se basan en una distinción entre la variable dependiente y la variable independiente. En caso que la variable dependiente corresponde a las

categorías *j* (columnas), el tau y el lambda miden la intensidad de la relación por medio de la reducción relativa promedio de los errores de asignación que se efectúan al momento de predecir a cuál categoría *j* pertenece un individuo cuando se sabe a cuál categoría *i* pertenece. Su expresión general es por lo tanto:

$$1 - \frac{\begin{array}{l}\text{Número promedio de errores de}\\ \text{asignación cuando se conoce } i\end{array}}{\begin{array}{l}\text{Número promedio de errores de}\\ \text{asignación cuando no se conoce } i\end{array}}$$

Las dos mediciones difieren en cuanto a la regla que se respeta para predecir a cual categoría *j* pertenece el individuo.

El tau de Goodman y Kruskal

Regla de asignación. Se distribuyen los individuos entre las categorías *j* de manera proporcional a los $p_{\bullet j}$ cuando no se conoce *i*, y de manera proporcional a los p_{ij} cuando se conoce *i*.

Fórmula.

$$\tau_J = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - \frac{p_{ij}}{p_{i\bullet}}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})} = 1 - \frac{\sum_i \sum_j p_{ij}\left(1 - p_{j/i\bullet}\right)}{\sum_j p_{\bullet j}(1 - p_{\bullet j})}$$

Valores límites. El tau es igual a cero cuando las dos variables son perfectamente independientes, o sea cuando $p_{ij} = p_{i\bullet}\ p_{\bullet j}$. Es igual a 1 cuando, en cada línea *i* de la tabla, existe solamente una sola celda no nula, lo que permite predecir *j* con certeza cuando se conoce *i*.

La lambda

Regla de asignación. En caso de no saber a cuál categoría *i* pertenecen los individuos, se asignan todos en la categoría que contiene el más grande número de observaciones, o sea la categoría con la probabilidad marginal $p_{\bullet j}$ más grande; en caso de conocer *i*, se asignan los individuos en la categoría *j* que contiene el más grande número de observaciones en el interior de la categoría *i*, o sea la categoría con la probabilidad condicional $p_{j/i\bullet}$ más grande.

Fórmula.

$$\lambda_J = 1 - \frac{\sum_i \left(1 - \frac{p_{i,Max}}{p_{i\bullet}}\right) p_{i\bullet}}{(1 - p_{\bullet Max})} = 1 - \frac{\sum_i (1 - p_{Max/i\bullet})\, p_{i\bullet}}{(1 - p_{\bullet k})}$$

,

donde

$$p_{\bullet Max} = \underset{j}{Max}\, p_{\bullet j}\,,\ p_{i,Max} = \underset{j}{Max}\, p_{ij}$$

y

$$p_{Max/i\bullet} = \underset{j}{Max}\, p_{j/i\bullet}$$

Valores límites. El lambda es igual a 1 cuando, en cada línea *i* de la tabla, existe una única celda no nula y que $p_{i,Max} = p_{i\bullet}$, lo que permite predecir *j* con certeza cuando se conoce *i*. Es igual a cero cuando $p_{i,Max} = p_{i\bullet}\ p_{\bullet Max}$ para todo *i*, aunque las variables no sean independientes, es decir, aunque, para las columnas menos aquella donde se encuentra

$p_{i,Max} = \underset{j}{Max}\, p_{ij}$, tengamos $p_{ij} \neq p_{i\bullet}\ p_{\bullet j}$. Es por esta última propiedad que se prefiere tau.

4-1.6 LAS VARIABLES DE CONTROL EN LAS TABLAS CON MÁS DE DOS DIMENSIONES

La tabla 1 tenía tres dimensiones: el sexo, la zona de residencia y la profesión. Hasta el momento, nuestro análisis consideró las dos últimas dimensiones ignorando la posibilidad de que existan diferencias entre los hombres y las mujeres. Sin embargo, es muy probable que la tabla de contingencia zona de residencia-profesión sea muy diferente para las mujeres y los hombres. El rigor científico nos forza a tomar en cuenta esta posibilidad.

Generalizando, cuando se examina la relación entre dos variables categóricas, es importante preguntarse si otras variables no pudiesen influenciar la intensidad o la forma de esta relación. Y si tal fuera el caso, es necesario tomar en cuenta estas nuevas variables, conocidas como variables de control. Esta expresión proviene del lenguaje de las ciencias experimentales, cuando las condiciones del laboratorio permiten "controlar" el nivel de las variables que podrían influenciar la relación en el estudio. Por ejemplo, en caso de probar un medicamento en ratas y de pensar que la alimentación podría influenciar el rendimiento del medicamento, se efectuarán pruebas sobre diferentes grupos con diferentes regímenes alimenticios "controlados".

Un modo simple de tomar en cuenta estas variables de control es examinar la relación en el estudio (con tests de hipótesis y medición de la intensidad de la relación) de manera separada para cada grupo homogéneo de individuos. En nuestro ejemplo, esto significaría examinar dos tablas de contingencias, una para las mujeres y otra para los hombres. Sin embargo, este procedimiento tiene límites. En particular,

cuando existen variables de control con varias categorías cada una, el número de tablas de contingencia para analizar aumenta rápidamente. Por ejemplo, si pretendemos considerar el sexo y la edad, con cinco grupos de edad, es necesario analizar diez tablas de contingencia. Además, cuando el número de observaciones es limitado, es posible que el número de observaciones sea demasiado pequeño como para confiar en la validez de los tests (por ejemplo, con 1000 observaciones, con 5 profesiones, 5 zonas de residencia, 5 grupos de edad y 2 sexos, tendremos frecuencias teóricas de solamente 4 en promedio, y es muy probable que algunas celdas tuvieran frecuencias teóricas abajo de 1).

Visto desde otra perspectiva, el problema es el de la multiplicidad de las interacciones posibles, la cual aumenta rápidamente con el número de variables (dimensiones de la tabla). Así, en una tabla con dos dimensiones, no existe más que una interacción posible y, por consiguiente, existe solamente una hipótesis de independencia para probar. En una tabla con tres dimensiones, existen cuatro interacciones posibles, es decir tres entre pares de variables y una entre las tres variables al mismo tiempo. En el caso de una tabla con cuatro dimensiones, hay 17 interacciones posibles (cuatro por cada una de las tercias posibles entre las cuatro variables, más una cuádruple interacción que implica todas las variables)...

El modelo log-lineal constituye un marco que permite examinar las diferentes interacciones posibles. El modelo "saturado", el cual incluye todas las interacciones posibles, reproduce perfectamente los datos observados. Una versión generalizada del test de hipótesis de independencia permite seleccionar, entre las numerosas interacciones posibles, aquellas que debemos guardar para representar la estructura subyacente.

Para profundizar en este tema...

Es posible consultar Upton (1981) y encontrar una presentación informal y pragmática del modelo log-lineal, así como un ejemplo de su uso en el contexto de las ciencias regionales. Button *et al*. (1995) ofrecen un ejemplo más reciente de uso del modelo log-lineal.

CAPÍTULO 4-2
EL MODELO LINEAL GENERAL Y LA REGRESIÓN MÚLTIPLE APLICADOS AL ANÁLISIS DE VARIANZA[192]

Vimos que el análisis de regresión es un método estadístico, el cual se aplica cuando un modelo teórico propone una relación entre una variable dependiente continua y una o más variables independientes continuas o discretas. En cuanto al análisis de varianza (analisis of variance, ANOVA), se aplica cuando las variables independientes son todas discretas. Sus características sobresalientes son por lo tanto:

- Un modelo del tipo estímulo-reacción.
- Una reacción que se mide por medio de variables continuas.
- Unos estímulos que se miden por medio de variables discretas.[193]

(En caso de medir los estímulos por medio de variables, siendo algunas discretas y otras continuas, se trata de aná-

[192] Wonnacott y Wonnacott (1992, pp. 503-507); Iman y Conover (1989, caps. 16-17).

[193] Dependiendo del contexto y de la disciplina, se usa diferentes términos para designar las variables independientes, como factores, efectos, categorías, variables cualitativas, variable de clasificación (clasification variables), etc.

lisis de covarianza y es necesario, entonces, usar el modelo lineal general.)

La pregunta de investigación que se propone usualmente es: ¿es la reacción al estímulo significativamente diferente entre categorías?

En el caso del análisis de varianza, existen procedimientos específicos y formatos estándares de presentación de los resultados. Sin embargo, es posible efectuar de manera equivalente un análisis de varianza por medio de la regresión lineal. Esto trae consigo algunas ventajas. Primero, y esto no es el caso del análisis de varianza, el análisis de regresión no impone restricciones en cuanto al plan de muestreo (número de observaciones por categorías), Segundo, es posible combinar el análisis de la varianza con variables independientes continuas (como ya mencionamos, este tipo de modelo es, a veces, conocido como un modelo de análisis de covarianza). Finalmente, el análisis de regresión ofrece una más amplia flexibilidad en cuanto a las hipótesis que se pueden someter a tests estadísticos.

4-2.1 UN EJEMPLO

En el marco de la construcción de una matriz de contabilidad social para Quebec, Robichaud *et al*. (1998) estudiaron el ahorro de los hogares de Quebec a partir de los datos contenidos en el archivo de micro-datos de gran difusión de la encuesta de *Statistique Canada* sobre los gastos de las familias en 1992. El archivo contiene 1900 observaciones para Quebec.

Se obtuvieron las seis variables, que presentamos a continuación, del archivo de microdatos de gran difusión de la encuesta de *Statistique Canada* sobre los gastos de las familias en 1992:

1. Composición del hogar
   - Personas solas.

- Parejas[194] sin hijos.
- Parejas con hijos.[195]
- Familias monoparentales.
- Otros hogares.[196]

2. Número de hijos menores de 16 años.
3. Edad de la persona de referencia.[197]
4. Ingreso del hogar con impuestos deducidos.
5. Variación neta del activo y del pasivo.
6. Seguro.[198]

La formulación del modelo lineal cuyos parámetros se estimarán, se basa en el modelo conceptual siguiente. El monto del ahorro de un hogar (variable dependiente) aumenta con el ingreso con impuestos deducidos, pero depende de la edad del hogar (hipótesis del ciclo de vida) y de la presencia de hijos (gastos más elevados); es posible también que el ahorro se vea influenciado por el hecho de que la responsabilidad del hogar esté en manos de una sola persona (lo que implica, por lo general, que hay un solo ingreso y que, en la mayoría de los casos, no existen más adultos que mantener); además, queremos verificar que la categoría heteróclita de los "Otros hogares" es diferente.

---

[194] Casados o juntados.

[195] Con relación a la composición del hogar, nos referimos a hijos de cualquier edad, nunca casados y que viven bajo el mismo techo que sus padres.

[196] Esta categoría contiene las parejas sin hijos que viven con un familiar que no es su hijo, así como los hogares donde vive por lo menos una persona que no es familiar de la "persona de referencia" (vea la nota siguiente). En particular, encontramos en esta categoría heteróclita, los hogares sin hijos con inquilinos y los grupos de estudiantes que comparten un departamento.

[197] En la encuesta sobre los gastos de las familias, la "persona de referencia" es el miembro del hogar que el contestador designa como el principal sostén financiero, lo que corresponde, normalmente, a la persona con el ingreso más elevado.

[198] Primas de seguros de vida, etcétera.

Es importante aclarar que la selección y la definición de las variables independientes fueron dictadas con el objetivo de construir una matriz de contabilidad social, y que el modelo que presentamos aquí, no debe considerarse como un ejemplo de un modelo de comportamiento de ahorro de los hogares. Lo único importante que se debe inferir de este ejemplo, es la manera de tratar las variables independientes categóricas en la regresión lineal. Añadamos que la presencia del ingreso entre las variables independientes vuelve imposible aplicar a este modelo un análisis de la varianza clásico (al menos que cambiemos el ingreso por una variable categórica), lo que, nuevamente, demuestra la más grande polivalencia del análisis de regresión.

Definimos entonces la variable dependiente:

AHORRO = Variación neta del activo y del pasivo más seguros.

Las variables independientes son:

REVAPIMP = Ingreso del hogar con impuestos deducidos.

Edad.

Composición del hogar.

Veremos luego cómo fueron especificadas las variables de edad y de composición del hogar. Combinando las variables categóricas, obtenemos una repartición de las 1900 observaciones según la composición del hogar y la edad de la persona de referencia. Se presenta esta repartición en la tabla que sigue, en la cual podemos observar que la repartición no se conforma a un plan de muestreo "equilibrado" (con el mismo número de observaciones en cada celda) ni tampoco a un plan de muestreo biproporcional. En estas condiciones, efectuar un análisis de varianza clásico se revelaría muy difícil; sin embargo, el uso de la regresión múltiple no encierra restricciones semejantes en cuanto a la estructura de los datos. No obstante, es importante notar que algunas celdas no tienen más que un número pequeño de observaciones, lo que nos pi-

de actuar con mucha prudencia al momento de interpretar los resultados.

Composición del hogar	Menos de 35	35-45	45-65	65 y más	Total
Personas solas	101	77	132	142	452
Parejas sin hijos	95	49	177	133	454
Parejas con hijos	163	267	239	26	695
Familias mono.	30	71	48	10	159
Otros sin hijos	24	17	41	21	103
Otros con hijos	8	16	12	1	37
Total	421	497	649	333	1900

4-2.1.1 Variables independientes de edad

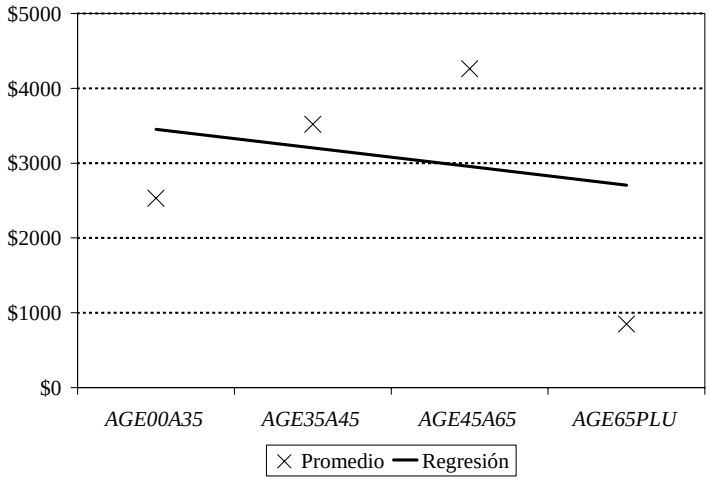
En primer lugar, definimos la variable independiente *GROUPAGE*:

Edad de la persona de referencia	Valor de la variable *GROUPAGE*
Menos de 35 años	1
35 años o más y menos de 45 años	2
45 años o más y menos de 65 años	3
65 años y más	4

La variable *GROUPAGE* es una variable ordinal de orden incompleto (ver cap. 1-1).

Pero esta variable no puede entrar tal cual en la regresión. ¿Por qué? Porque impondría artificialmente una relación lineal entre el grupo de edad y el ahorro, lo cual es contrario a los hechos, como se puede ver en el siguiente gráfico.

Valor promedio del ahorro por grupo de edad
y regresión lineal sobre *GROUPAGE*



Por tanto, no podemos especificar el modelo antes de haber reemplazado la variable GROUPAGE por una serie de variables dicotómicas, puesto que al emplearla tal cual como variable independiente, estaríamos diciendo que el ahorro aumenta (o disminuye) de manera lineal con la categoría edad. Por esta razón, se crean las cuatro variables dicotómicas siguientes:

7. *AGE0A35* = 1 si *GROUPAGE* = 1 (edad < 35); = 0 de otra manera.
8. *AGE35A45* = 1 si *GROUPAGE* = 2 (edad ≥ 35 y < 45); = 0 de otra manera.
9. *AGE045A65* = 1 si *GROUPAGE* = 3 (edad ≥ 45 y < 65); = 0 de otra manera.
10. *AGE65PLU* = 1 si *GROUPAGE* = 4 (edad ≥ 65); = 0 de otra manera.

4-2.1.2 Variables independientes de composición del hogar

La composición del hogar es una variable categórica politómica. Para poder distingüir entre hogares "otros" con y sin hijos, usaremos la información complementaria dada por el número de hijos menores de 16 años. Obtendremos así 6 tipos de hogares:

- Personas solas.
- Parejas sin hijos.
- Parejas con hijos.
- Familias monoparentales.
- Otros hogares sin hijos.
- Otros hogares con hijos.

Sin embargo, hay que reemplazar la variable de composición del hogar, tal como la de edad, por una serie de variables dicotómicas, y por las mismas razones. Es aún más necesario, dado que la composición del hogar no es una variable ordinal (cuando sí lo es *GROUPAGE*).

Presentamos a continuación dos formas de modelización, una que usa 5 variables y otra que usa 3. Ambas formas son representadas por los árboles de clasificación correspondientes.

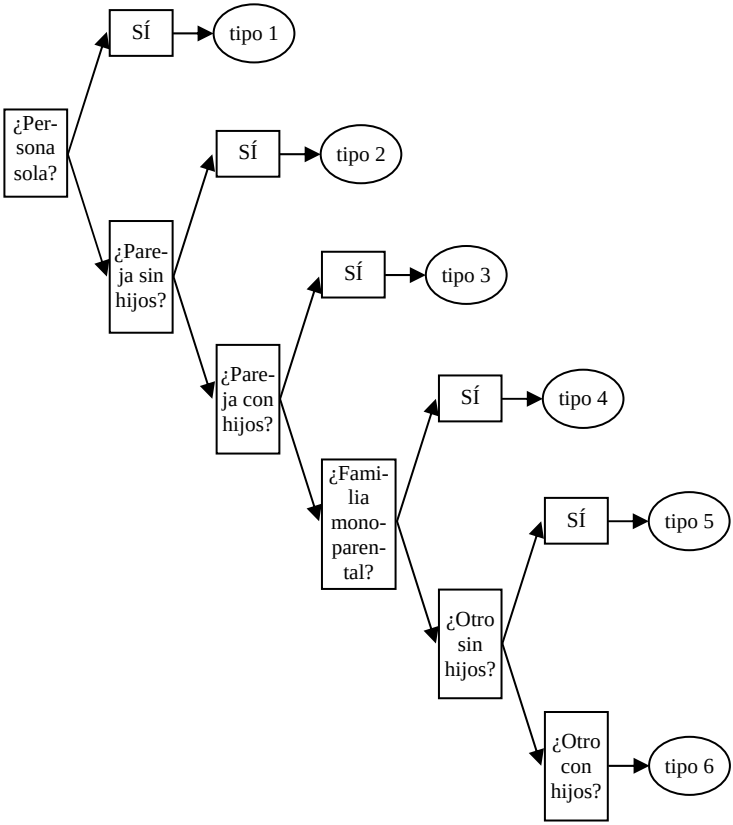
La clasificación con 5 variables consiste sencillamente en definir tantas variables dicotómicas como hay tipos de hogares, sin quitar una porque es redundante (vamos a profundizar eso luego). En cuando al método con 3 variables, se constituye por las siguientes variables dicotómicas:

11. *SEULMONO* = 1 para una persona sola o una familia monoparental; *SEULMONO* = 0 de otra manera
12. *AUTRE* = 1 si el hogar pertenece a la categoría "Otro"; *AUTRE* = 0 de otra manera
13. *ENFANTS* = 1 si el hogar cuenta, al menos, con un hijo menor de 16 años; *ENFANTS* = 0 de otra manera

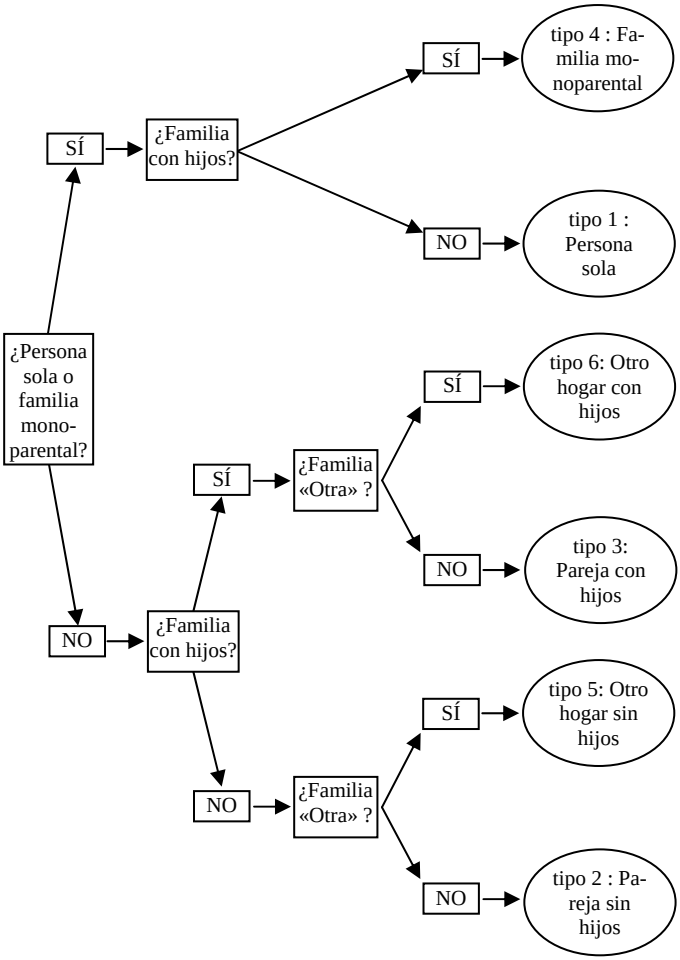
Por hijos entendemos hijos de cualquier edad, nunca casados y que viven con sus padres, salvo para los hogares

"*AUTRE*", para los cuales *ENFANTS* significa solamente hijos menores de 16 años.

Dos árboles de clasificación con variables dicotómicas
I – Clasificación con 5 variables



Dos árboles de clasificación con variables dicotómicas
II – Clasificación con 3 variables



La que usamos fue la de 3 variables. Combinando esas tres variables se obtiene la clasificación que sigue:

Composición de los hogares	Número de hijos	Valor de la variable SEUL-MONO	Valor de la variable AUTRE	Valor de la variable ENFANTS
Personas solas	0	1	0	0
Parejas sin hijos	0	0	0	0
Parejas con hijos	> 0	0	0	1
Familias mono.	> 0	1	0	1
Otros hogares	0	0	1	0
	> 0	0	1	1

Se ve en este cuadro que cada tipo de hogar corresponde a una combinación única de las variables dicotómicas.

¿Cuál es la diferencia entre los dos esquemas de clasificación? De alguna manera, el esquema que escogimos impone una cierta coherencia en el modelo. Por ejemplo, con la tripleta *SEULMONO*, *ENFANTS* y *AUTRE*, el efecto de tener hijos debe ser el mismo independientemente de las demás características del hogar. Esto implica, por lo tanto, restricciones para el modelo. Sin embargo, veremos cómo se pueden evitar estas restricciones con la introducción de variables de interacción (vea 4-2.4).

4-2.2 Eliminación de la redundancia entre las variables independientes

No se deben incluir las cuatro variables dicotómicas de edad juntas entre las variables independientes, porque una de estas variables es redundante. En efecto, si *AGE0A35* = 0 y *AGE45A65* = 0 y *AGE65PLU* = 0, entonces forzosamente *AGE35A45* = 1, es decir, generalizando, que si para una obser-

vación dada, tres de las cuatro variables toman el valor cero, la cuarta toma necesariamente el valor 1. Por consiguiente, es necesario descartar una de las variables del modelo; en estas condiciones, el caso que corresponde a la variable descartada llega a ser el caso de referencia. En nuestro ejemplo, escogemos el grupo de edad de 35 a 45 como caso de referencia.

> De manera formal, al incluir las cuatro variables, estaríamos violando la condición H4 del modelo clásico de la regresión lineal puesto que su suma es siempre igual a 1, es decir igual a la constante del modelo:
>
> *AGE0A35* + *AGE35A45*
> + *AGE45A65* + *AGE65PLU* = 1 = *CONSTANTE*

Es importante observar que, al momento de definir las variables dicotómicas *ENFANTS*, *SEULMONO* y *AUTRE*, eliminamos, de manera implícita, las variables redundantes. En efecto, evitamos definir dos variables correspondientes a una por categoría. Por ejemplo, hubiéramos podido definir

- *AVECENFANTS* = 1 si el hogar cuenta con, por lo menos, un hijo;

  AVECENFANTS = 0 de otra manera;
- *SANSENFANTS* = 0 si el hogar cuenta con, por lo menos, un hijo;

  SANSENFANTS = 1 de otra manera.

No hicimos tal cosa puesto que una de estas dos variables hubiera sido redundante.

Vimos en el apartado 4-2.1 cómo se reemplazaba la variable de composición por la tripleta *SEULMONO*, *ENFANTS* y *AUTRE*. En caso de que hubiéramos querido utilizar el otro esquema de clasificación, tendríamos, por las mismas razones por las cuales se efectuó en el caso de la variable *GROUPAGE,* que haber eliminado la redundancia. Es por eso que el

primer esquema no cuenta con 6 variables dicotómicas, sino con 5.

4-2.3 Especificación de un modelo sin interacción

Estamos listos ahora para enunciar una primera especificación del modelo:

$$EPARGNE = \beta_1 + \beta_2\, REVAPIMP + \beta_3\, SEULMONO + \beta_4\, AUTRE + \beta_5\, ENFANTS + \beta_6\, AGE00A35 + \beta_7\, AGE45A65 + \beta_8\, AGE65PLU$$

Donde es posible notar la ausencia de la variable *AGE35A45* que sería redundante.

Veamos, ahora, lo que este modelo significa para cada una de las 24 posibilidades que alojaron nuestros datos. Los 24 casos se presentan en la siguiente tabla.

Podemos notar en esta tabla que a cada uno de los 24 casos posibles corresponde una combinación única de valores de las variables dicotómicas; esto muestra que no faltan variables puesto que cada caso tiene una representación distinta. Observamos también que el caso de referencia cuando todas las variables dicotómicas son nulas, corresponde a una pareja sin hijos cuya persona de referencia tiene entre 35 y 45 años. Se deduce que los coeficientes de las variables dicotómicas representan las diferencias con relación a este caso de referencia; por ejemplo, el modelo predice que entre el hogar de referencia y una familia monoparental cuya persona de referencia tiene menos de 35 años, teniendo los dos hogares el mismo ingreso, la diferencia será igual a $\beta_3 + \beta_5 + \beta_6$. Podríamos pensar que cada uno de estos tres coeficientes sea negativo; sin embargo, sólo la estimación del modelo podrá aclarar este hecho.

Interpretación del modelo sin variables de interacción

Grupo de edad de la persona de referencia	SEULMONO	AUTRE	ENFANTS	AGE00A35	AGE45A65	AGE65PLU	AHORRO predicho por el modelo
Personas solas (Número de hijos = 0)							
<35	1	0	0	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + 0 + \beta_6 + 0 + 0$
≥35 y <45	1	0	0	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + 0 + 0 + 0 + 0$
≥45 y <65	1	0	0	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + 0 + 0 + \beta_7 + 0$
≥65	1	0	0	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + 0 + 0 + 0 + \beta_8$
Parejas sin hijos (Número de hijos = 0)							
<35	0	0	0	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + 0 + \beta_6 + 0 + 0$
≥35 y <45	0	0	0	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + 0 + 0 + 0 + 0$
≥45 y <65	0	0	0	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + 0 + 0 + \beta_7 + 0$
≥65	0	0	0	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + 0 + 0 + 0 + \beta_8$
Parejas con hijos (Número de hijos > 0)							
<35	0	0	1	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + \beta_5 + \beta_6 + 0 + 0$
≥35 y <45	0	0	1	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + \beta_5 + 0 + 0 + 0$
≥45 y <65	0	0	1	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + \beta_5 + 0 + \beta_7 + 0$
≥65	0	0	1	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + 0 + 0 + \beta_5 + 0 + 0 + \beta_8$

Continua...

Interpretación del modelo sin variables de interacción (continuación)

Grupo de edad de la persona de referencia	SEULMONO	AUTRE	ENFANTS	AGE00A35	AGE45A65	AGE65PLU	AHORRO predicho por el modelo
Familias monoparentales (Número de hijos > 0)							
<35	1	0	1	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + \beta_5 + \beta_6 + 0 + 0$
≥35 y <45	1	0	1	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + \beta_5 + 0 + 0 + 0$
≥45 y <65	1	0	1	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + \beta_5 + 0 + \beta_7 + 0$
≥65	1	0	1	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + \beta_3 + 0 + \beta_5 + 0 + 0 + \beta_8$
Otros hogares sin hijos (Número de hijos = 0)							
<35	0	1	0	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + 0 + \beta_6 + 0 + 0$
≥35 y <45	0	1	0	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + 0 + 0 + 0 + 0$
≥45 y <65	0	1	0	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + 0 + 0 + \beta_7 + 0$
≥65	0	1	0	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + 0 + 0 + 0 + \beta_8$
Otros hogares con hijos (Número de hijos > 0)							
<35	0	1	1	1	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + \beta_5 + \beta_6 + 0 + 0$
≥35 y <45	0	1	1	0	0	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + \beta_5 + 0 + 0 + 0$
≥45 y <65	0	1	1	0	1	0	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + \beta_5 + 0 + \beta_7 + 0$
≥65	0	1	1	0	0	1	$\beta_1 + \beta_2\ REVAPIMP + 0 + \beta_4 + \beta_5 + 0 + 0 + \beta_8$

Por otro lado, al observar la tabla, entendemos mejor por qué hubiera sido absurdo incluir en el modelo la variable politómica *GROUPAGE*.

Se presentan los resultados de la estimación en la tabla siguiente.

Variable	Descripción	Símbolo	Coeficiente estimado	*t* de Student	Probabilidad crítica
CONSTANTE		β_1	–7727	–11.062	0.0001
REVAPIMP	Ingreso después de impuestos	β_2	0.340	28.468	0.0001
ENFANTS	Presencia de hjos	β_5	–2260	–4.937	0.0001
SEULMONO	Persona sola o monoparentale	β_3	1903	3.834	0.0001
AUTRE	Hogar « Otro »	β_4	–2578	–3.309	0.0010
AGE00A35	Edad 00-35	β_6	258	0.444	0.6574
AGE45A65	Edad 45-65	β_7	419	0.796	0.4263
AGE65PLU	Edad 65+	β_8	875	1.322	0.1862

$n = 1900$
$R^2 = 0.33$

En particular, constatamos que los coeficientes de las variables correspondientes a la edad no son, de manera significativa, diferentes de cero. ¿Debemos entender, por lo tanto, que la edad no tiene efecto sobre el comportamiento de ahorro?

4-2.4 Introducción de los efectos de interacción

El modelo que se presentó en la tabla anterior no toma en cuenta la posibilidad de efectos de interacción. Existe un gran

número de interacciones posibles. Por tanto, no se encuentra a menudo que un modelo las contenga todas.

Por ejemplo, el modelo predice que el efecto sobre el ahorro de la presencia de hijos es igual a β_5, independientemente de la edad de la persona de referencia y de la composición del hogar. ¿Así sucede en la realidad? En otras palabras, ¿no habrá alguna interación entre la variable *ENFANTS* y las variables *SEULMONO*, *AUTRE*, *AGE0A35*, *AGE45A65* y *AGE65PLU*? Es importante entender que cada uno de los efectos de interacción que evocamos en la frase anterior es simétrico; por ejemplo, en lugar de preguntarse si el efecto de la presencia de hijos (*ENFANTS*) cambia con pertenecer al grupo de los menores de 35 años (*AGE0A35*), es posible preguntarse de manera equivalente si el efecto de pertenecer al grupo de los menores de 35 años cambia con la presencia de hijos.

Para poder incluir la posibilidad de interacción en el modelo, es necesario agregar variables a las ocho que ya tiene el modelo, Así, se define:

9. *MONOMONO* = 1
   si *ENFANTS* = 1 y *SEULMONO* = 1
   MONOMONO = 0 de otra manera

Para ser más conciso, se define matemáticamente[199]

MONOMONO = *ENFANTS* × *SEULMONO*

De la misma manera, tenemos

10. AUTRENFA = ENFANTS × AUTRE
11. *ENFA0035* = *ENFANTS* × *AGEOA35*
12. *ENFA4565* = *ENFANTS* × *AGE45A65*
13. *ENFA65PL* = *ENFANTS* × *AGE65PLU.*
14. *AUTAOO35* = *AUTRE* × *AGEOOA35.*
15. *AUTA4565* = *AUTRE* × *AGE45A65.*

[199] Las variables dicotómicas son variables lógicas o variables de Boole. En álgebra booleania, la conjunción “y” se representa con la multiplicación.

16.*AUTA65PL = AUTRE* × *AGE65PLU.*
17.*SOLA0035 = SEULMONO* × *AGE00A35.*
18.*SOLA4565 = SEULMONO* × *AGE45A65.*
19.*SOLA65PL = SEULMONO* × *AGE65PLU.*

En esta lista podemos notar que no se incluyeron todas las variables posibles (por ejemplo, no hay ninguna variable de interacción con *AGE35A45*) porque, al igual que para otros grupos de variables categóricas, en el caso de las variables de interacción el hecho de incluir en el modelo todas las variables posibles implica redundancia.

Se interpretan los coeficientes de las variables de interacción como unas diferencias. Por ejemplo, vimos en la tabla del apartado 4-2.3 como β_5, el coeficiente de la variable *ENFANTS*, representaba la diferencia, en cuanto al monto del ahorro, entre dos hogares idénticos en todo menos en la presencia de hijos; igualmente, β_7, el coeficiente de la variable *AGE45A65* representa la diferencia, en cuanto al ahorro, entre dos hogares idénticos en todo menos en la edad, puesto que uno pertenece al grupo de edad de referencia (35-45 años) y el otro al grupo de los 45-65 años. En ausencia de variables de interacción, estas diferencias se suman; por ejemplo, el modelo descrito en el apartado 4-2.3 predice que entre un hogar sin hijos del grupo 35-45 años y un hogar con hijos del grupo 45-65 años, la diferencia será igual a $\beta_5 + \beta_7$. Si agregamos a este modelo la variable de interacción *ENFA4565*, esta diferencia será entonces igual a $\beta_5 + \beta_7$, más el coeficiente de la variable de interacción *ENFA4565*.[200]

[200] Es posible efectuar una analogía con la farmacología: el efecto de una combinación de dos medicamentos puede implicar efectos muy diferentes que los efectos de cada uno de estos medicamentos empleados solos. Los medicamentos juntados pueden reforzarse mutuamente o, por lo contrario, anularse el uno al otro.

Además, puede suceder que haya interacción entre una variable categórica y una variable continua. Así, el modelo predice que, independientemente de las características del hogar, un alza de un dólar del ingreso con impuestos retenidos repercutirá en un alza del ahorro de β_2 dólares. ¿Podría este efecto ser diferente para los hogares con hijos? Con el fin de examinar este problema, es necesario incluir, en el modelo, unas variables de interacción. Consideremos, por lo tanto, las tres variables suplementarias siguientes:

20. *REVENFAN* = *REVAPIMP* × *ENFANTS*

21. *REVSELMO* = *REVAPIMP* × *SEULMONO*

22. *REVAUTRE* = *REVAPIMP* × *AUTRE*

Los coeficientes de estas variables se pueden interpretar también como diferencias. Por ejemplo, si comparamos dos hogares idénticos menos en la presencia de hijos, el coeficiente *REVENFAN* representa la diferencia entre los dos hogares en cuanto a su propensión marginal para ahorrar.

Después de incluir unas variables de interacción, el modelo completo se enuncia de la manera siguiente:

$$
\begin{aligned}
EPARGNE = {} & \beta_1 + \beta_2\, REVAPIMP \\
& + \beta_3\, SEULMONO + \beta_4\, AUTRE + \beta_5\, ENFANTS \\
& + \beta_6\, AGE00A35 + \beta_7\, AGE45A65 + \beta_8\, AGE65PLU \\
& + \gamma_1\, MONOMONO + \gamma_2\, AUTRENFA \\
& + \gamma_3\, ENFA0035 + \gamma_4\, ENFA4565 + \gamma_5\, ENFA65PL \\
& + \gamma_6\, AUTA0035 + \gamma_7\, AUTA4565 + \gamma_8\, AUTA65PL \\
& + \gamma_9\, SOLA0035 + \gamma_{10}\, SOLA4565 + \gamma_{11}\, SOLA65PL \\
& + \alpha_1\, REVENFAN + \alpha_2\, REVSELMO + \alpha_3\, REVAUTRE
\end{aligned}
$$

4-2.5 ESTIMACIÓN E INTERPRETACIÓN DEL MODELO

Después de ejecutar el procedimiento backward para eliminar las variables cuyos coeficientes no son significativos, obtenemos los resultados que se presentan en la tabla que sigue.

Variable	Descripción	Símbolo	Coeficiente estimado	Error Estandar	Probabilidad crítica
CONSTANTE		β_1	–10487	729	0.0001
REVAPIMP	Ingresos después de impuestos	β_2	0.400	0.013	0.0001
ENFANTS	Presencia de hijos	β_5	–1927	561	0.0006
SEULMONO	Pers. sola o monoparental	β_3	8233	1000	0.0001
AUTRE	Hogar "Otro"	β_4	7969	1771	0.0001
AGE45A65	Edad 45-65	β_7	1767	779	0.0234
AGE65PLU	Edad 65+	β_8	1513	770	0.0497
Variables de interacción					
ENFA4565	*ENFANTS* × *AGE45A65*	γ_4	–1506	897	0.0932
SOLA4565	*SEULMONO* × *AGE45A65*	γ_{10}	–1983	1008	0.0494
SOLA65PL	*SEULMONO* × *AGE65PLU*	γ_{11}	–1996	1138	0.0796
REVSELMO	*REVAPIMP* × *SEULMONO*	α_2	–0.211	0.030	0.0001
REVAUTRE	*REVAPIMP* × *AUTRE*	α_3	–0.297	0.046	0.0001

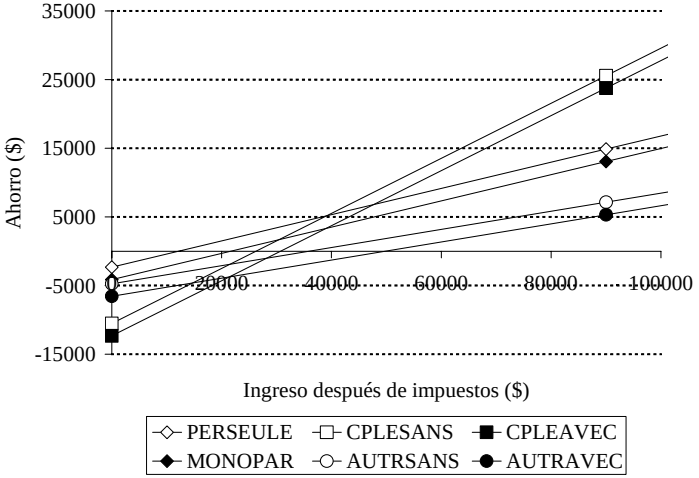
El número de observaciones es de 1900 y el coeficiente de determinación múltiple R^2 es de 0.36.

No es fácil concluir algo claro con todos estos coeficiente y hemos de preguntarnos qué significan realmente. Las figu-

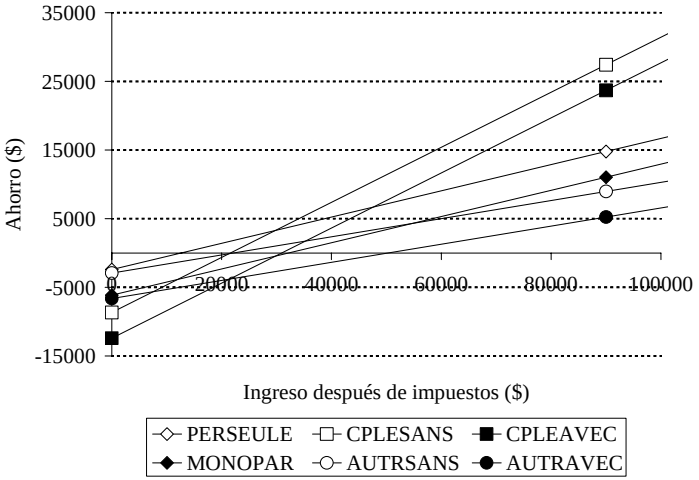
ras que presentamos a continuación ilustran las predicciones del modelo.

Leyenda	
Personas solas	PERSEULE
Parejas sin hijos	CPLESANS
Parejas con hijos	CPLEAVEC
Familias monoparentales	MONOPAR
Otros hogares sin hijos	AUTRSANS
Otros hogares con hijos	AUTRAVEC

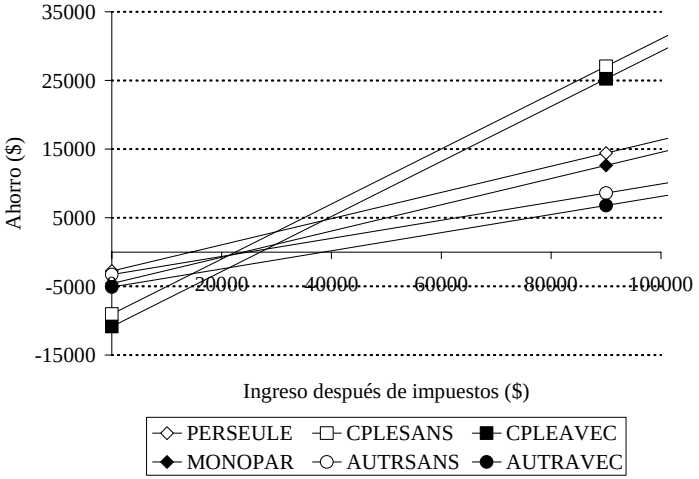
Ahorro de los menos de 45 años



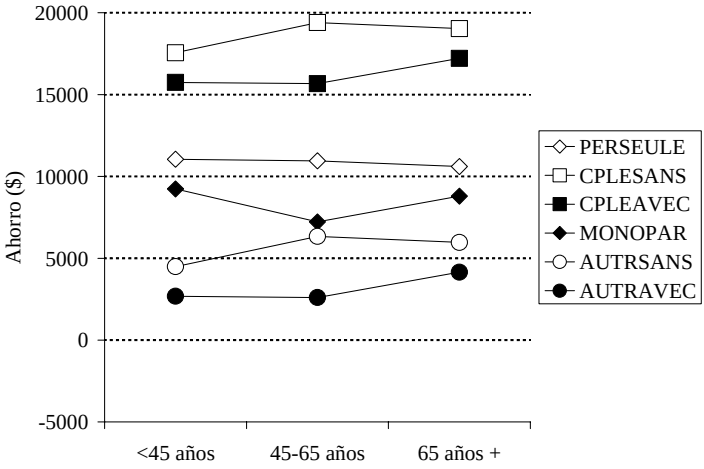
Ahorro de los 45-65 años



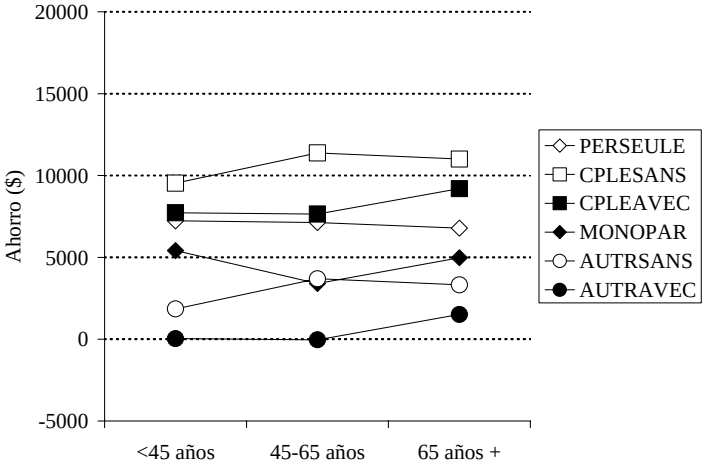
Ahorro de los 65 años y más



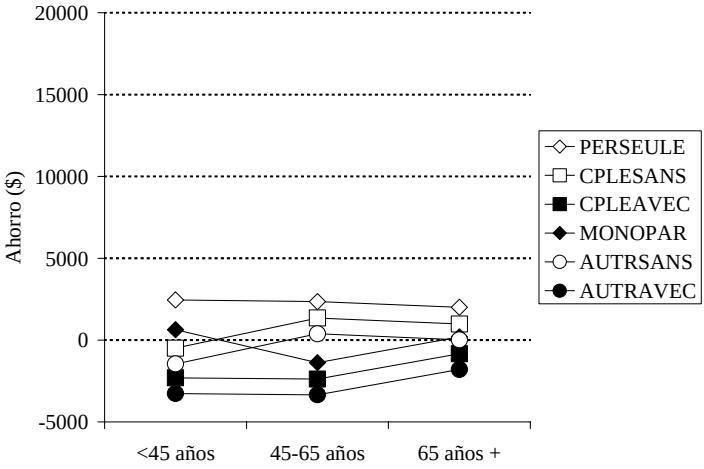
Ahorro según el grupo de edad, con un ingreso de $70,000



Ahorro según el grupo de edad, con un ingreso de $50,000



Ahorro según el grupo de edad, con un ingreso de $25,000



CAPÍTULO 4-3
MODELOS CON VARIABLE DEPENDIENTE CUALITATIVA

4-3.1 MODELOS DE ELECCIÓN BINARIA: LOGIT BINOMIAL Y PROBIT BINOMIAL

4-3.1.1 El problema

Nos proponemos, en este momento, examinar la otra cara, por así decirlo, del análisis de varianza. En el análisis de varianza la variable de reacción es continua, mientras que las variables estímulos son discretas; por lo contrario, ahora, la variable de reacción es discreta, mientras que las variables de estímulo pueden ser continuas.

Es posible deducir el modelo logit a partir del análisis de las tablas de contingencia. Sin embargo, en el análisis de las tablas de contingencia todas las variables de clasificación tienen papeles simétricos cuando, por lo contrario, en el modelo logit una de las variables toma el papel de variable dependiente y las probabilidades de pertenencia a sus diferentes categorías son condicionadas por las demás variables, las cuales, por consiguiente, toman el papel de variables independientes. Además, la posibilidad de encontrar variables continuas entre las variables independientes constituye una

generalización con respecto al logit deducido del análisis de las tablas de contingencia.

Para empezar, consideremos el caso cuando la variable dependiente es dicotómica en lugar de ser politómica; en estas condiciones, sólo existen dos posibilidades:

$y_i = 0$ o 1

Ejemplo: en un estudio sobre la movilidad residencial de los hogares,[201]

- $y_i = 0$ si el hogar *i* no se muda
- $y_i = 1$ si el hogar *i* se muda

Ahora bien, en el modelo estándar

$$y_i = \sum_j x_{ij}\beta_j + u_i$$

el valor de la variable dependiente y_i no puede limitarse a tomar los valores 0 y 1 cuando los términos aleatorios u_i poseen una distribución normal (esto porque una variable normal es continua y su dominio de variación se extiende de $-\infty$ a $+\infty$).

El primer paso que debemos emprender para solucionar esta dificultad es considerar que la verdadera variable dependiente no es la variable dicotómica *y*, más bien la probabilidad que *y* =1:

$$\Pr[y_i = 1] = \sum_j x_{ij}\beta_j + u_i$$

Esto conlleva, sin embargo, dos dificultades:

1. Aunque la nueva variable dependiente $\Pr[y_i = 1]$ sea continua, su campo de variación se limita al intervalo [0, 1], cuando el campo de variación del término aleatorio, en caso de poseer una distribución normal, se extiende de $-\infty$ a $+\infty$.

[201] Para un ejemplo, vea Mongeau (sin fecha).

2. La nueva variable dependiente Pr[$y_i = 1$], no se puede observar, pues lo que observamos no son más que las *realizaciones* ($y_i = 1$ o $y_i = 0$); por lo tanto será necesario recurrir a un método de estimación que sea adecuado a la situación.

4-3.1.2 Modelo de comportamiento

Uno de los fundamentos teóricos posibles del modelo logit binomial es un modelo de comportamiento del tipo estímulo-reacción (stimulus-response), el cual se usa con frecuencia particularmente en biología cuando se somete a los sujetos de un experimento a condiciones (estímulos) que varían de un sujeto a otro. Entonces se observan las reacciones y se busca estimar la relación entre estímulos y reacción.

Es posible presentar este modelo en dos partes. La primera parte constituye el modelo de reacción del sujeto (modelo de comportamiento) y la segunda, el modelo de estímulo total, el cual es el resultado de las diferentes condiciones combinadas a las cuales se somete a un sujeto.

Primera parte: modelo de reacción

No es posible observar directamente el estímulo total al cual se somete al sujeto. Supongamos, no obstante, que exista una variable no observable ("latente") *w* que mide este estímulo, es decir el atractivo o "lo deseable" de una elección (un economista diría la "utilidad"). En el ejemplo del estudio sobre la movilidad residencial, el hogar *i* muda si esta variable latente *w* rebasa un cierto valor crítico, un "umbral de reacción" (*S*):

$y_i = 0$ si $w_i < S$

$y_i = 1$ si $w_i \geq S$

donde *w* es el valor de la variable latente para el hogar *i* y *S* es el umbral crítico más allá del cual el hogar se muda. Es cómodo suponer que se define la variable latente *w* para que *S* sea igual a cero, lo que nos da:

$y_i = 0$ si $w_i < 0$

$y_i = 1$ si $w_i \geq 0$

Esta última hipótesis no impone ninguna restricción al modelo, puesto que *w* es una variable ordinal cuyo cero es arbitrario.

Segunda parte: modelo de estímulo total

Se determina el valor de la variable latente (no observada) con un cierto número de factores medidos con variables independientes apropiadas (los *x*). Por ejemplo, en el estudio sobre la movilidad residencial, las variables independientes podrían tomar en cuenta las características del hogar, las características de la vivienda actual, etcétera.

Si, además, suponemos que la relación entre las variables independientes y la variable latente es lineal, obtenemos el modelo siguiente:

$$w_i = \sum_j x_{ij} \beta_j + u_i$$

donde el término aleatorio u_i representa las variaciones aleatorias entre los sujetos (los hogares). El modelo no emite la hipótesis que todos los sujetos son idénticos sino, más bien que, solamente, el modelo es lo suficiente completo para que podamos considerar las variaciones de comportamientos no explicadas por el modelo como el resultado del azar.

Integración de las dos partes del modelo

Se combina las dos partes del modelo.

$w_i < 0$ equivale a $\sum_j x_{ij}\beta_j + u_i < 0,$

o sea a $u_i < -\sum_j x_{ij}\beta_j$

$w_i \geq 0$ equivale a $\sum_j x_{ij}\beta_j + u_i \geq 0,$

o sea a $u_i \geq -\sum_j x_{ij}\beta_j$

Por lo tanto,

$$y_i = 0 \text{ si } u_i < -\sum_j x_{ij}\beta_j \text{ y } y_i = 1 \text{ si } u_i \geq -\sum_j x_{ij}\beta_j$$

De esta manera, la probabilidad que $y_i = 0$ es igual a la probabilidad de que $u_i < -\sum_j x_{ij}\beta_j$ y la probabilidad que $y_i = 1$ es igual a la probabilidad que $u_i \geq -\sum_j x_{ij}\beta_j$. Estas probabilidades dependen evidentemente de las hipótesis que hacemos en cuanto a la distribución de los u_i. Dependiendo de la hipótesis que hagamos sobre la función de densidad de probabilidad de los u_i, obtendremos un modelo logit o probit:

Hipótesis A: los u_i. poseen distribuciones normales, independientes entre sí, con un promedio nulo y una varianza común σ^2 (hipótesis de Gauss-Markov); éste es el modelo probit.

Hipótesis B: los u_i se distribuyen independientemente uno del otro, y sus funciones de distribución acumulativas se definen con la función logística; éste es el modelo logit.

4-3.1.3 El modelo logit y la inducción estadística

Con un modelo que contiene una variable latente (inobservable), como el anterior descrito, es evidentemente imposible estimar los parámetros β_j por medio del método de los menores cuadrados. Es necesario recurrir al método del máximo de verosimilitud. La primera etapa en la aplicación de este método es construir la función de verosimilitud que, si recordamos bien, es la función de probabilidad de la muestra que expresamos en función de los parámetros.[202] La función de verosimilitud se infiere, por consiguiente, de la hipótesis que se hizo en cuanto a la distribución de los u_i , lo que constituye una especificación completa del modelo aleatorio.

Nos interesa, en este momento y de manera específica, el modelo logit, el cual se infiere de la hipótesis de una distribución logística. La función logística es la función

$$L(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t+1}$$

Si la variable aleatoria *u* posee una función de distribución acumulativa logística, entonces tenemos:

$$\Pr[u \le t] = L(t) = \frac{1}{1+e^{-t}} = \frac{e^t}{e^t+1}$$

Se puede justificar la elección de la función logística con argumentos pragmáticos del siguiente tipo: es relativamente fácil de manejo y es muy semeja a la normal para valores alrededor del promedio (cuando se desemeja para los valores

[202] Al momento de presentar el principio del máximo de verosimilitud (cap. 2-2), definimos la función de verosimilitud como la función de densidad de probabilidad de la muestra. Sin embargo, en nuestro caso presente, la variable observada es discreta (dicotómica con más precisión), por lo tanto, la función de verosimilitud es, sin más, la función de probabilidad.

extremos). No obstante, la elección de la función logística puede basarse en fundamentos teóricos mucho más sólidos.[203]

Los tests de hipótesis que son posibles efectuar se fundamentan en unas distribuciones asintóticas; son, por lo tanto, válidos solamente cuando la muestra es lo suficientemente grande. Dicho de otra manera, más grande serán la muestra, más confiables será estos tests.

Los estimadores b_j de los parámetros β_j poseen distribuciones asintóticas normales y los estimadores de sus diferencias tipo poseen una distribución asintótica χ^2. Habiendo cumplido con estas condiciones, es posible, ahora, aplicar los tests *t* de Student a los parámetros.

4-3.2 HACIA EL LOGIT MULTINOMIAL: UNA GENERALIZACIÓN HEURÍSTICA DEL BINOMIAL

Parece claro que el modelo estímulo-reacción, con umbral de reacción sólo se puede aplicar a una situación de elección binaria. En caso que haya más de dos posibilidades, es necesario un modelo más general donde el sujeto escoge la posibilidad más atrayente para él.

Ejemplo: elección del idioma de uso:

- $y_i = 0$ si el idioma de uso es el francés.
- $y_i = 1$ si el idioma de uso es el inglés.
- $y_i = 2$ si el idioma de uso es el italiano.
- $y_i = 3$ si el idioma de uso es otro idioma.

El modelo teórico de comportamiento subyacente al modelo logit multinomial es el modelo de utilidad aleatoria (random utility). Según este modelo, existe una variable latente que mide la "utilidad" o el atractivo de cada opción en función de los atributos del sujeto y de las características de la opción. El sujeto escoge racionalmente la posibilidad que

[203] Vea, en particular, Ben Akiva y Lerman (1985).

posee el más grande atractivo.[204] No obstante, la utilidad de cada opción no es una función determinista de las variables independientes, más bien, la suma de dos términos la constituye, a saber la utilidad sistemática y un término aleatorio.

En una situación de elección politómica, al lugar de tener una sola variable non observada ("latente") *w*, tantas variables latentes son necesarias como hay posibilidades para que cada una de estas variables mida el atractivo de la posibilidad correspondiente. Escribamos:

w_{ij} : medición del atractivo de la posibilidad *j* para el individuo *i*.

Como primer paso, reformulemos el modelo binario según este esquema. Tenemos como resultado:

$y_i = 0$ si $w_{i0} > w_{i1}$

$y_i = 1$ si $w_{i1} > w_{i0}$

lo que equivale a

$y_i = 0$ si $w_{i0} - w_{i1} > 0$, o sea $w_{i1} - w_{i0} < 0$

$y_i = 1$ si $w_{i1} - w_{i0} > 0$

Por lo tanto, es posible interpretar la medición del estímulo *w* del modelo binario como una medición de la diferencia entre el atractivo de la primera y de la segunda posibilidad. En estas condiciones, el modelo del estímulo se convierte en

$$w_i = w_{i1} - w_{i0} = z_{i1} - z_{i0} + u_i$$

donde $z_{ij} = \sum_h x_{ih}\beta_{hj}$ es la parte determinista del modelo del atractivo de la posibilidad *k*.

[204] Es justamente en este punto que el modelo de utilidad aleatoria se distingue de los modelos con utilidad "constante" donde la probabilidad que una opción sea seleccionada aumenta con su utilidad, pero donde no es seguro que un sujeto escoja la posibilidad más ventajosa para él. En los modelos con utilidad "constante", la elección de la posibilidad más ventajosa es solamente más probable que las otras. Es en este sentido que nos alejamos de un comportamiento racional.

Teniendo:

$p_{ij} = \Pr[y_i = j]$: la probabilidad que el individuo *i* escoja la posibilidad *j*.

Tenemos:

$$p_{i1} = \frac{e^{z_{i1} - z_{i0}}}{e^{z_{i1} - z_{i0}} + 1}$$

$$p_{i0} = 1 - p_{i1} = \frac{1}{e^{z_{i1} - z_{i0}} + 1}$$

lo que implica para los puntos de ventaja (como en apuestas; *odds* en inglés) de $[y_i = 1]$ contra $[y_i = 0]$

$$\frac{p_{i1}}{1 - p_{i1}} = \frac{p_{i1}}{p_{i0}} = e^{z_{i1} - z_{i0}} = \frac{e^{z_{i1}}}{e^{z_{i0}}}$$

Supongamos, sin demostrarlo, que se pueda generalizar este resultado con un número cualquiera de posibilidades y que, para todo par de posibilidad *j*,*k*, tengamos

$$\frac{p_{ij}}{p_{ik}} = e^{z_{ij} - z_{ik}} = \frac{e^{z_{ij}}}{e^{z_{ik}}}$$

Al efectuar la suma de todas las posibilidades *j*, obtenemos

$$\sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

donde $\sum_j p_{ij} = 1$, de tal manera que

$$\sum_j \frac{p_{ij}}{p_{ik}} = \frac{\sum_j p_{ij}}{p_{ik}} = \frac{1}{p_{ik}} = \frac{\sum_j e^{z_{ij}}}{e^{z_{ik}}}$$

de lo que se deduce

$$p_{ik} = \frac{e^{z_{ik}}}{\sum_j e^{z_{ij}}}$$

¡Y aquí tiene usted el modelo logit multinomial!

Los parámetros de este modelo se estiman por medio del método del máximo de verosimilitud, como en el caso del logit binomial. Los procedimientos de inducción estadística son análogos.

CONCLUSIÓN DE LA CUARTA PARTE

Esta cuarta parte es de alcance más práctico para investigadores en ciencias sociales, cuya materia son a menudo datos cualitativos. El objetivo latente aquí es demostrar que es a la vez posible y útil el análisis cuantitativo de datos cualitativos.

El análisis de tablas de contingencia es una técnica de análisis multivariado que se usa con frecuencia. Se hace hincapié en la significación de la hipótesis de independencia, así como en el test correspondiente del Chi-cuadrado de Pearson. Se subraya también la distinción entre una relación *estadísticamente significativa* y una relación *científicamente pertinente*.

El análisis de varianza es también una técnica muy utilizada. Pero no se presenta aquí en su forma clásica. Se eligió más bien mostrar cómo el análisis de regresión permite hacer un análisis de varianza, pero liberándose de las exigencias restrictivas de este último en cuanto al plan muestral, y teniendo la posibilidad de incluir variables independientes continuas (análisis de covarianza). Se hace hincapié en la construcción de las variables booleanas que representan las variables politómicas del análisis de varianza clásico, así como en la interpretación correcta de los resultados.

Finalmente, el capítulo 4-3 presenta brevemente el principio de modelización que hace posible tratar variables dependientes cualitativas (logit, probit).

EPÍLOGO

Al momento de escribir estas palabras, un trimestre llega a su término y siento una mezcla de satisfacción y frustración. Satisfacción, por supuesto, de un profesor cuyos estudiantes, en su mayoría, tuvieron éxito en lograr los objetivos de aprendizaje que les había planteado. Sin embargo, frustración, por los que no tuvieron éxito: ¿se hubiera podido presentar la materia de otra manera, más accesible? Frustración, además, porque se dejó de lado, por falta de tiempo, mucho material, porque hubo que seleccionar entre el contenido de este libro, cuando fue concebido como un todo. Tal o tal punto, que no tuvimos tiempo de ver en clase, ¿faltará en la capacitación de los futuros investigadores?

Con una mezcla igual de satisfacción y frustración pienso en esta obra. Estoy contento de haber podido consignar aquí la cosecha de más de un decenio de esfuerzos pedagógicos. Sin embargo, con cada nueva lectura, encuentro que aquí se tendría que clarificar la exposición, que allá habría que añadir algún complemento, que, en otra parte, los ejemplos podrían ser más convincentes... Incluso a veces, surgen ideas pedagógicas que motivarían a refundir un capítulo entero. En suma, no puedo dejar de pensar que esta obra queda inacabada.

Sin embargo, esperar una versión definitiva no es más que ilusión, ¿verdad? Hay que resignarse a la naturaleza finita del ser humano y compartir ahora lo que siempre quedará imper-

fecto. Eso es lo que decido hacer. Entonces, con toda modestia, presento esta obra a la comunidad científica y universitaria.

André Lemelin
Montreal, abril de 2004

REFERENCIAS

Arriaga, Eduardo (1975) "Selected measures of urbanization", Cap. 2, p. 19-87, en Goldstein, Sidney y Sly, David F., ed. (1975) *The measurement of urbanization and projection of urban population*, International Union for the Scientific Study of Population, Liège.

Ben-Akiva, Moshe E. y Lerman, Steven R. (1985) *Discrete choice analysis: theory and application to travel demand*, MIT Press, Cambridge.

Bishop, Yvonne M. M., Fienberg, Stephen E. y Holland, Paul W. (1975) *Discrete multivariate analysis: theory and practice*, MIT Press, Cambridge, MA.

Blalock, Hubert M. Jr. (1979) *Social statistics*, 2a edición revisada, McGraw-Hill, New York.

Bonnet, Jean (1995) "Les dynamiques régionales et leurs facteurs", *Revue d'Économie Régionale et Urbaine* (1), pp. 3-34.

Braudel, Fernand (1979) *Civilisation matérielle, économie et captalisme, XVe-XVIIIe siècle. 1. Les structures du quotidien*, Armand Collin, Le Livre de Poche Références.

Bryman, Alan y Cramer, Duncan (1990) *Quantitative data analysis for social scientists*, Routledge.

Button, Kenneth J., Scott Leitham, Ronald, W. McQuaid, and John D. Nelson (1995) “Transport and industrial and commercial location”, *Annals of Regional Science*, 29 (2), pp. 189-206.

Coffey, William J. y Polese, Mario (1988) “Locational shifts in Canadian employment, 1971-1981”: Decentralization v. decongestion”, *Geographica*, 32(3), pp. 248-256.

Demaris, Alfred (1992) *Logit modeling: practical applications*, Sage University Papers Series: Quantitative Applications in the Social Sciences.

Duncan, Otis Dudley, y Duncan, Beverly (1955) “A methodological analysis of segregation indexes”, *American Sociological Review*, 20 (2), pp. 210-217.

Freund, John E. (1962) *Mathematical statistics*, Prentice-Hall, Inc., Englewood Cliffs, N.J.

Freund, John E., y Williams, Frank J. (1982) *Elementary business statistics: the modern approach*, 4th ed., Prentice-Hall, Inc., Englewood Cliffs, N.J., pp. 370-388.

Gilles, Alain (1994) *Éléments de méthodologie et d'analyse statistique pour les sciences sociales*, McGraw Hill, Montreal, pp. 227-243

Gujarati, Damodar N. (1992) *Econometria*, 2e ed., traducido de la segunda edición de *Basic Econometrics* (1988), por Mayorga Torrado, Víctor Manuel, McGraw-Hill.

Heikkila, E., Dale-Johnson, D., Gordon, P., Kim, J.I., Peiser, R. y Richardson, H.W. (1989) “What happened to the CBD-distance gradient?”: land values in a polycentric city”, *Environment and Planning* A, 21(2), pp. 221-232.

Huriot, Jean-Marie, y Perreur, Jacky (1990) “Distances, espaces et représentations: une revue”, *Revue d'Économie Régionale et Urbaine* (2), pp. 197-237.

Huriot, Jean-Marie, y Perreur, Jacky (1994) "Espace et distance", Cap. 5 *Encyclopédie d'économie spatiale. Concepts, comportements, organisations, Economica*, Paris.

Ifrah, Georges (1994) *Histoire universelle des chiffres*, tomos 1 y 2, Robert Laffont.

Iman, Ronald L., y Conover, W. J. (1989) *Modern business statistics*, 2nd ed., John Wiley & Sons.

Isard, Walter y Bramhall, David F. (1960) *Methods of regional analysis*: *an introduction to regional science*, MIT Press, Cambridge, MA.

Isserman, Andrew M. (1980) "Estimating export activity in a regional economy: A theoretical and empirical analysis of alternative methods", *International Regional Science Review*, 5(2), pp. 155-184.

Jayet, Hubert (1993) *Analyse spatiale quantitative. Une introduction*, Bibliothèque de Science régionale, Economica, Paris.

Kendall, Maurice G., Stuart, Alan y Ord, J. Keith (1991) *Kendall's advanced theory of statistics*, Oxford University Press.

Kennedy, Peter (1992) *A guide to econometrics*, 3rd ed., MIT Press.

Knapp, Thomas R. (1996) *Learning statistics through playing cards*, Sage Publications.

Kunzmann, Peter, Burkard, Franz-Peter y Wiedmann, Franz (1993), *Atlas de la philosophie*, Librairie Générale Française, La Pochothèque, Le Livre de Poche, Coll. Encyclopedies d'aujourd'hui.

Lafrance, Robert y Schembri, Lawrence (2002) "Parité des pouvoirs d'achat: définition, mesure et interprétation", *Revue De La Banque Du Canada/Bank of Canada Review*, Automne 2002, pp. 29-36.

Lazarsfeld, Paul (1971) "Des concepts aux indices empiriques", texte 1, pp. 27-36, en Boudon, Raymond y

Lazarsfeld, Paul (1971) *Le vocabulaire des sciences sociales: concepts et indices*, Mouton & co., Paris.

Legendre, Louis y Legendre, Pierre (1984) Écologie numérique - Tome 1: Le traitement multiple des données écologiques, segunda edición, Masson et Presses de l'Université du Québec, Paris.

Legendre, Louis y Legendre, Pierre (1984) *Écologie numérique - Tome 2: La structure des données écologiques*, segunda edición, Masson et Presses de l'Université du Québec, Paris.

Legendre, Louis y Legendre, Pierre (1998) *Numerical Ecology*, Coll. Developments in environmental ecology, 20, Deuxième edición en langue anglaise, Elsevier.

Lemelin, André (2000) *Méthodes quantitatives des sciences sociales appliquées aux études urbaines et régionales*, Les Presses de l'Université Laval.

Lemelin, André y Polèse, Mario (1995), "What about the bell-shaped relationship between primacy and development?", *International Regional Science Review*, 18(3), pp. 313-330.

Lemelin, André y Polèse, Mario (avec la collaboration de Pérez Mendoza, Salvador; Rojas Bonilla, Luis, y de Vasquez Lopez, Jaime) (1993), "La localisation de l'emploi est-elle si différente en les pays en développement: Modèles d'urbanisation et analyses comparatives des systèmes urbains canadien et mexicain", *Revue canadienne d'études du développement/Canadian Journal of Development Studies*, vol. XIV No 1, mai 1993, pp. 73-102.

MacLachlan, Ian, y Sawada, Ryo (1997) "Measures of income inequality and social polarization in Canadian metropolitan areas", *The Canadian Geographer/Le Géographe Canadien*, 41(4), pp. 377-97.

Malinvaud, Edmond (1978) *Méthodes statistiques de l'économétrie*, 3a ed., Dunod, Paris (2e ed., 1969).

McKenna, Christopher K. (1980) *Quantitative methods for public decision making*, McGraw-Hill.

Mills, Edwin S., y Hamilton, Bruce W. (1989) *Urban Economics*, Fourth Edición, Scott, Foresman & Co., Glenview, Ill.

Mongeau, Jaël (s. d.) "Discrepancies between housing plans, choices and behaviors in Montréal, 1972-1979".

Norcliffe, G.B. (1983) "Using location quotients to estimate the economic base and trade flows", *Regional Studies*, 17(3), pp. 161-168.

Polèse, Mario (1994) Économie urbaine et régionale: la logique spatiale des mutations économiques, *Economica*, Paris.

Polèse, Mario y Stafford, Robert (1982) "Une estimation des exportations de services des régions urbaines: l'application d'un modèle simple au Canada", *Revue Canadienne des Sciences Régionales*, 5(2), pp. 313-331.

Richardson, H.W., Gordon, P., Jun, M.J., Heikkila, E., Peiser, R. y Dale-Johnson, D. (1990) "Residential property values, the CBD, and multiple nodes: further analyses", *Environment and Planning* A, 22(6), pp. 829-833.

Robert, Serge (1993) *Les mécanismes de la découverte scientifique : une épistémologie interactionniste*, Presses de l'Université d'Ottawa, Ottawa.

Robichaud, Véronique, Lemelin, André, y Fréchette, Pierre (1998) *Construction de la matrice de comptabilité sociale du Québec pour 1992: aspects techniques*, Montréal et Sainte-Foy, INRS-Urbanisation y CRAD, Université Laval, mai, 34 pp. y 131 tableaux.

Rosen, Kenneth T., y Resnick, Mitchel (1980) "The size distribution of cities: an examination of the Pareto law and primacy", *Journal of Urban Economics* 8, pp. 165-186.

Statistique Canada (1996) “Votre guide d’utilisation de l’indice des prix à la consommation”, 62-557-XPB, dic. 1996.

Statistique Canada (1997) “L'indice des prix à la consommation ou comment mesurer l'inflation”, Tendances sociales canadiennes, 1-008-XPF, été 1997.

Taylor, Peter J. (1977) *Quantitative methods in geography*, Houghton Mifflin.

Theil, Henri (1971) *Principles of econometrics*, John Wiley & Sons.

Upton, Graham J.G. (1981) “Log-linear models, screening and regional industrial surveys”, Regional Studies, 15, pp. 33-45.

Upton, Graham J.G. y Fingleton, Bernard (1985) *Spatial Data Analysis by Example*, *Volume I Point pattern and quantitative data*, John Wiley & Sons.

Valeyre, A. (1993) “Mesures de dissemblance et d’inégalité interrégionales: principes, formes et propriétés”, *Revue d’Économie Régionale et Urbaine*, 1, pp. 17-54.

Waldorf, Brigitte S. (1993) “Segregation in urban space: A measurement approach”, *Urban Studies*, 30(7), pp. 1151-1164.

Webber, Michael (1984) *Explanation, prediction and planning: The Lowry model*, Pion, London.

Williamson, J.G. (1965) Regional inequality and the process of national development: a description of the patterns”, *Economic Development and Cultural Change*, Vol. 13, pp. 3-45.

Wonnacott, Thomas H. y Wonnacott, Ronald J. (1991) *Statistique: économie, gestion, sciences, medecine*, 4e ed., Economica.

Índices de precios

Aizcorbe, Ana M. y Jackman, Patrick C. (1993) "The commodity substitution effect in CPI data, 1982-91", US Department of Labor, Bureau of Labor Statistics, *Monthly Labor Review*, déc., pp. 25-33.

Abrahams, Katharine G., Greenlees, John S., y Moulton, Brent R. (1998) "Working to improve the consumer price index", Symposium: Measuring the CPI, *Journal of Economic Perspectives*, Winter, 12(1), pp. 27-36.

Crawford, Allan (1993) "Note technique: les biais de la mesure de l'IPC canadien", *Revue de la Banque du Canada/Bank of Canada Review*, été, pp. 21-36.

Deaton, Angus (1998) "Getting prices right: what should be done?", Symposium: Measuring the CPI, *Journal of Economic Perspectives*, Winter, 12(1), pp 37-46.

Deaton, Angus (1998) "Index number issues in the consumer price index", Symposium: Measuring the CPI, *Journal of Economic Perspectives*, Winter, 12(1), pp. 47-58.

Moulton, Brent R. (1996) "Bias in the consumer price index: What is the evidence?", *Journal of Economic Perspectives*, Fall; 10(4), pp. 159-177.

Nordhaus, William D. (1998) "Quality change in the price indexes", Symposium: Measuring the CPI, *Journal of Economic Perspectives*, Winter, 12(1), pp. 59-68.

Pollak, Robert A. (1998) "The consumer price index: a research agenda and three proposals", Symposium: Measuring the CPI, *Journal of Economic Perspectives*, Winter, 12(1), pp. 69-78.

Statistique Canada (1982) "Document de référence de l'indice des prix à la consommation - Concepts et

procedés (mise à jour fondée sur les dépenses de 1978)", mai, Cat.62-553 hors série.
Statistique Canada (1978) "L'indice de prix à la consommation, nov., Cat. 62-546 hors série.

El IDH del PNUD y los indicadores urbanos

Agostini, S.J. y Richardson S.J. (1997) "A human development index for US cities: Methodological issues and preliminary findings", *Real Estate Economics*, spring, 25(1), pp. 13-41 (Numéro spécial à l'occasion d'Habitat II à Istambul).
Aturupane, Harcha, Glewne, Paul y Isenman, Paul (1994) "Poverty, human development and growth: an emerging consensus", *American Economic Review*, 84(2), pp. 244-249.
Camp, Sharon L., Barberis, Mary, y Hinds, Judith (1990) *Cities. Life in the World's 100 largest metropolitan areas. Statistical appendix*, Population Crisis Committee, Washington, DC.
Collin, Jean-Pierre, Séguin, Anne-Marie y Pelletier, Hermance (1999) "Les indicateurs de positionnement (benchmarking) des métropoles. Besoins et potentialités en contexte montréalais", Actas del debate en Montreal el 29 de octobre de 1998, INRS-Urbanisation, Montréal.
Coombes, M., y Wong C. (1994) Methodological steps in the development of multivariate indexes for urban and regional policy", *Environment and Planning* A, v. 26, 1297-1316.
Flood, Joe (1997) "Urban and housing indicators", *Urban Studies*, 34(10), pp. 1635-1665.
Lemelin, André (1999) "Une contradiction en l'Indicateur de Développement Humain du PNUD", INRS-Urbanisation, Coll. Inedits 99-06, avril, 10 p., texto

revisado y traducido en inglés con el título “Note: An inconsistency in the UNDP’s Human Development Indicator “, INRS-Urbanisation, septiembre, 12 p.

Malpezzi, S., y Mayo, S.K. (1997) “Housing and urban development indicators: A good idea whose time has returned”, *Real Estate Economics*, spring, 25(1), pp. 1-11 (Número especial con motivo del Habitat II en Istambul).

Mayer-Renaud, Micheline (1986) *La distribution de la pauvreté et de la richesse en les régions urbaines du Québec: Portrait de la région de Montréal*, Direction des services professionnels, Centre des services sociaux du Montréal métropolitain.

Mayer-Renaud, Micheline y Renaud, Jean (1989) *La distribution de la pauvreté et de la richesse en la région de Montréal en 1989: Une mise à jour*, Direction des services professionnels, Centre des services sociaux du Montréal métropolitain.

Organisation pour la Coopération et le Développement Economique (OCDE) (1997) *Mieux comprendre nos villes. Le rôle des indicateurs urbains*, OCDE, Paris, 1997 (Titre anglais: *Better understanding our cities. The role of urban indicators*. Síntesis de las comunicaciones presentadas durante una conferencia organizada en Rennes por el OCDE, el OMS, la Comisión Europea y la ciudad de Rennes, 3-4 de abril de1995).

Programme des Nations-Unies pour le Développement (PNUD) (1990) “Définir et mesurer le développement humain”, Cap. 1, pp. 9-18 en Programme des Nations-Unies pour le Développement (PNUD) (1990) *Rapport mondial sur le développement humain*, 1990, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1994) “Un nouveau regard sur l’indicateur de développement humain”, Cap. 5, pp. 96-117, en

Programme des Nations-Unies pour le Développement (PNUD) (1994) *Rapport mondial sur le développement humain*, 1994, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1995) *Rapport mondial sur le développement humain*, 1995, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1996) *Rapport mondial sur le développement humain*, 1996, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1997) *Rapport mondial sur le développement humain*, 1997, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1998) *Rapport mondial sur le développement humain*, 1998, Economica, Paris.

Programme des Nations-Unies pour le Développement (PNUD) (1999) *Rapport mondial sur le développement humain*, 1999, De Boeck & Larcier, Département De Boeck Université, Paris, Bruxelles.

Programme des Nations-Unies pour le Développement (PNUD) (2001) *Rapport mondial sur le développement humain*, 2001, De Boeck & Larcier, Département De Boeck Université, Paris, Bruxelles.

Ravallion, Martin (1997) "Good and bad growth: The Human Development Reports", World Development, 1997, 25(5), pp. 631-638.

Renaud, Jean, Mayer, Francine y Lebeau, Ronald (1996) *Espace urbain, espace social. Portrait de la population de villes du Québec, publié en collaboration avec Les Centres Jeunesse de Montréal*, Institut de Recherche pour le Développement Social des Jeunes, Edición Saint-Martin, Montréal.

Srinivasan, T.N. (1994) "Human development: a new paradigm or reinvention of the wheel?", *American Economic Review*, 84(2), pp. 238-243.

Streeten, Paul (1994) “Human development: means and ends”, *American Economic Review*, 84(2), pp. 232-237.

Sugden, Robert (1993) “Welfare, resources, and capabilities: A review of Inequality Reexamined by Amartya Sen”, *Journal of Economic Literature*, 31(4), pp. 1947-1962.