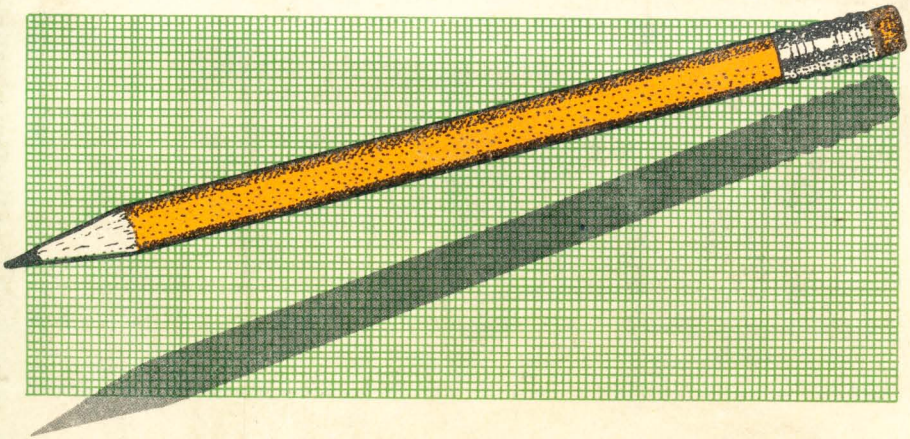


Estadística básica para planificación

con una parte nueva sobre **muestreo**

Arturo Núñez del Prado Benavente

14a. edición corregida y aumentada

siglo veintiuno editores

ESTADÍSTICA BÁSICA PARA PLANIFICACIÓN

por

ARTURO NÚÑEZ DEL PRADO BENAVENTE

12 DIC 1996

siglo veintiuno editores

MÉXICO
ESPAÑA
ARGENTINA
COLOMBIA

900044515 - BIBLIOTECA CEPAL

siglo veintiuno editores, sa de cv
CERRO DEL AGUA 248, DELEGACIÓN COYOACÁN, 04310 MÉXICO, D.F.

siglo veintiuno de españa editores, sa
C/PLAZA 5, MADRID 33, ESPAÑA

siglo veintiuno argentina editores, sa

siglo veintiuno de colombia, ltda
AV. 3a. 17-73 PRIMER PISO, BOGOTÁ, D.E. COLOMBIA

primera edición, 1971
decimocuarta edición, corregida y aumentada, 1987
© siglo xxi editores, s.a. de c.v.
ISBN 968-23-1384-8

derechos reservados conforme a la ley
impreso y hecho en méxico/printed and made in mexico

338.90182/N972/1987
c.2

economía
y
demografía

12 DIC 1996

63576

Textos *del* INSTITUTO LATINOAMERICANO DE PLANIFICACIÓN ECONÓMICA Y SOCIAL

El Instituto Latinoamericano de Planificación Económica y Social (ILPES) es un organismo autónomo creado bajo la égida de la Comisión Económica para América Latina (CEPAL) y establecido el 1 de julio de 1962 en Santiago de Chile como proyecto del Fondo Especial de las Naciones Unidas con amplio apoyo de los países de la región y de diversos organismos internacionales y privados.

Su objeto principal es proporcionar, a solicitud de los gobiernos, servicios de capacitación y asesoramiento en América Latina y realizar investigaciones sobre desarrollo y planificación. Desde su fundación, el Instituto ha venido ampliando y haciendo más profunda la obra de la CEPAL en el campo de la planificación, merced al esfuerzo conjunto de un grupo de economistas y sociólogos distinguidos de América Latina, entregados por completo al estudio y solución de los problemas fundamentales que preocupan en la actualidad a los países de esta parte del mundo.

Desde su creación el Instituto ha realizado una labor de gran significación dentro de las funciones que se le encomendaron. A fin de difundirla debidamente en el ámbito latinoamericano, se ha llegado a un acuerdo con Siglo XXI de México, para que vaya publicando y distribuyendo los trabajos del Instituto.

ÍNDICE

PRÓLOGO A LA QUINTA EDICIÓN

La naturaleza de los trabajos que me correspondió realizar primero en el ILPES, luego como funcionario de gobierno y actualmente en la CEPAL, me ha brindado la oportunidad de verificar en el terreno que los objetivos que se tuvieron en mente al preparar la primera edición de este libro se han visto holgadamente cumplidos. No soy el más indicado para evaluar sus bondades, si es que las tuviera. Pero no me parece falta de modestia destacar que el objetivo que más gravitó en la concepción del texto —cual fue el de inculcar en el lector y, principalmente, en el alumno de los cursos del ILPES un espíritu crítico guiado por la búsqueda de coherencias cuantitativas— es cada vez más legítimo.

En mis conversaciones con profesores universitarios, funcionarios de oficinas de planificación y ex alumnos del ILPES diseminados en América Latina, invariablemente se toca el tema de la fidelidad de la información cuantitativa. La información estadística disponible en América Latina no siempre, por decir lo menos, satisface las pruebas básicas de coherencia. Es muy frecuente encontrar estimaciones que, debiendo guardar ciertos grados de consistencia con otras variables, aparecen publicadas contraviniendo elementales pruebas de compatibilidad.

A medida que se progresa en los trabajos cuantitativos en la región, resulta más necesario realizar diversas pruebas de coherencia de las estimaciones que se han de utilizar en los trabajos sobre planificación. Por mucho que ellas provengan de listados de computadoras o de sofisticados modelos matemáticos, es prudente someterlas a los contrastes indispensables. Sólo una vez que hayan resistido y pasado la prueba de formar parte de esquemas cuantitativos razonablemente admisibles e integrados, puede pensarse en utilizarlas en los trabajos que la planificación implica en sus distintos plazos y niveles de agregación.

Los diversos grados y sentidos de interdependencia entre las variables que conforman los procesos de planificación y política económica facilitan la configuración de esquemas cuantitativos coherentes. Las ventas de un sector son compras de otros, los ingresos de ciertos grupos provienen de otros, los excesos de gasto de los menos merman el consumo de los más, etc. No resulta difícil vincular una estimación sobre cierta variable a otras con las cuales, dado el funcionamiento de la actividad económica, debiera mantener ciertas relaciones razonables.

Estas reflexiones dirigidas principalmente a quienes enseñan la estadística básica tienen su sustento en reiteradas comprobaciones de lo indispensable que resulta dudar de estimaciones aisladas. Cuántas horas de esfuerzo, cuánto dinero, cuántas decisiones erróneas no se hubieran evitado si se tuviera como inclinación permanente someter a pruebas de coherencia las estimaciones que se utilizan. Esta práctica no sólo favorece el trabajo cuantitativo, también tiene enormes ventajas en la interpretación de los fenómenos económicos. Interpretar magnitudes que se relacionan con otras es más fácil —y más útil en la identificación de sus implicaciones— que considerarlas aisladamente. De la misma manera que en medicina el conocer la temperatura o la cantidad de glóbulos blancos no es suficiente para aventurar un diagnóstico, en economía las estimaciones aisladas apenas sugieren la existencia de eventuales fenómenos, y en todo caso no son más que el inicio de una investigación que en su desarrollo debe alcanzar niveles de rigor que avalen sus conclusiones.

A la larga, los economistas que trabajan con instrumentos cuantitativos desarrollan capacidades, muchas veces sobre la base de una aguda intuición, que les permiten calibrar las estimaciones y, utilizando su experiencia, establecer las pruebas pertinentes. Sin embargo, si en la etapa de formación de un planificador puede desarrollarse esa tendencia de manera que su inquietud natural sea la de verificar coherencias, se habrá ganado tiempo y se habrá conseguido aumentar su productividad potencial.

La lectura de los distintos capítulos de este libro y especialmente de los ejercicios, problemas y temas de discusión, permite comprobar la intención permanente de acicatear al lector para que se incline, antes de utilizar una información, por la verificación de consistencias fundamentales.

ARTURO NÚÑEZ DEL PRADO

agosto de 1976

PRÓLOGO

FRANCISCO AZORÍN

Mucho me honra y complace presentar una obra de tan sobresaliente mérito como es la *Estadística básica para planificación,* del profesor Arturo Núñez del Prado. Es evidente la necesidad de información estadística para la planificación del desarrollo económico y social; no lo es menos la conveniencia de planificar a su vez las operaciones estadísticas, para que sirvan a su finalidad con eficacia. Este libro viene a satisfacer con suma oportunidad tan urgentes requerimientos.

Es grande el número de publicaciones disponibles sobre estadística matemática o aplicada e incluso de las dedicadas a las investigaciones económicas y sociales, pero apenas existen obras que tengan en cuenta el aspecto específico de la planificación. El profesor Núñez del Prado ha estado en posición privilegiada para preparar esta obra, por su extensa experiencia en las diversas fases de la planificación y por haber dirigido cursos y seminarios para planificadores, lo que no le ha impedido mantenerse fiel a su vocación original de estadístico. Así ocurre que el planificador incipiente o experto que consulte este libro se encuentra ante un estadístico que le habla en su propio lenguaje. Desde el comienzo se trata del diagnóstico y de la prognosis, de la delimitación de la estrategia, así como de los procedimientos censales, muestrales y de experimentación numérica, y partiendo, como es ineludible, de los elementos del estudio, se procede con ritmo mesurado a recorrer el trecho desde los estadígrafos descriptivos estáticos a los problemas bidimensionales de correlación y regresión, pero siempre atendiendo al objetivo principal de la tarea. Por ejemplo, en la primera parte se destaca la utilización de los indicadores de la programación y de los números índices, y en la segunda el empleo de los coeficientes de elasticidad.

Atención especial merecen los temas de discusión, que con los problemas y ejercicios facilitan la comprensión y el dominio de las técnicas y suscitan el interés por nuevos caminos, distintos

de los que se siguen en las ciencias fisicoquímicas o en la ingeniería clásica.

Este libro tendrá sin duda un efecto positivo y fecundo en la integración de los estudios de planificación y de estadística, y será una buena herramienta para quienes trabajan unidos por el común deseo de acelerar el desarrollo de los pueblos.

Circula, tanto en castellano como en otros idiomas, una profusa bibliografía sobre estadística; la conocen, frecuentan y utilizan, provechosamente por cierto, los interesados. Por ello, quizá cabría preguntarse qué sentido tiene aumentar, con un nuevo trabajo, el número ya abundante de publicaciones sobre la materia. Sin embargo, las siguientes consideraciones parecen justificar el esfuerzo.

En primer término, la bibliografía sobre estadística, en su gran mayoría, presenta métodos estadísticos puros, ya sea por su tratamiento matemático o desde un punto de vista descriptivo, menos riguroso, aunque más accesible a los investigadores en general. Por ello no es fácil encontrar un texto que resuma un conjunto de métodos estadísticos necesarios para el investigador en materia de planificación. Uno de los propósitos de este trabajo consiste, precisamente, en realizar una selección de temas, los más útiles, que permitan al planificador disponer de un instrumental indispensable. Por otra parte, es necesario advertir que el texto fue concebido para estudiantes de cursos intensivos de planificación, donde a la estadística como asignatura se le concede un número reducido de clases y seminarios. Los diferentes temas presentados en el texto, se complementan con una serie de ejercicios de seminarios, a través de los cuales se pretende mostrar problemas y soluciones concretas sobre aspectos que, habitualmente, se presentan durante la realización de planes de desarrollo.

Tampoco puede dejar de considerarse la gran heterogeneidad de alumnos que asisten a estos cursos, pues el tema de la planificación económica y social compete a muchas disciplinas. El tratamiento de los diferentes temas, en consecuencia, debe hacerse accesible a esta especial categoría de alumnado. La anterior observación se refiere principalmente a los métodos matemáticos utilizados. Para la comprensión de los planeamientos metodológicos ofrecidos, es necesario un manejo ágil del álgebra superior y una cabal comprensión de los conceptos básicos de geometría analítica y cálculo diferencial e integral.

Otro aspecto que se tuvo en cuenta al redactar este texto es que la asignatura estadística brinda un instrumental previo que

permite un mejor tratamiento de otras asignaturas de los cursos de planificación. Se hizo, por lo tanto, un esfuerzo de compatibilización con las necesidades de instrumental que tienen asignaturas como análisis económico, contabilidad social, evaluación de proyectos, planificación, etc.

Los puntos anteriores constituyen un conjunto de condiciones, en función de las cuales se ha estructurado un curso especial y se redactó un texto básico; para exponer todos y cada uno de los temas presentados, se tuvieron en cuenta las mencionadas restricciones.

Este libro corresponde a una versión revisada de los apuntes de clase utilizados en los cursos que se dictan en el Instituto Latinoamericano de Planificación Económica y Social, cuya amplia acogida corroboran sus varias ediciones mimeografiadas y el interés que suscitó su publicación en forma de Cuadernos. De todas maneras parece necesario advertir que el trabajo que se presenta en modo alguno pretende ser un manual de estadística económica. Tiene tan sólo un propósito didáctico y constituye un texto de iniciación para el estudio de esta técnica cuantitativa. Está destinado a facilitar el dominio de un instrumental mínimo indispensable en materia de planificación, proporcionando un conjunto de conceptos que aseguren al planificador la posibilidad de percibir tanto las ventajas como las limitaciones del empleo de indicadores estadísticos, y permitan interpretar cabalmente los estadígrafos de uso más frecuente. Interesa al mismo tiempo que el planificador pueda establecer, en estudios más profundos, relaciones de trabajo eficientes con estadísticos y econometristas.

Es preciso destacar que este texto está constituido en buena parte por resúmenes y adaptaciones de temas tratados en otros, como los *Apuntes de estadística* del profesor Pedro Vuskovic, el *Curso general de estadística* del profesor Enrique Cansado, la *Introducción a la estadística* de G. V. Yulé y M. G. Kendall, y la *Estadística general aplicada* de F. Croxton y D. Cowden. El objetivo perseguido no es la originalidad conceptual; antes bien, se pretende brindar un texto didáctico, que incluya un conjunto de temas íntimamente relacionados con la planificación, complementados con ejercicios que ilustren conceptos de manejo cotidiano en la práctica.

Debo hacer constar aquí mi reconocimiento a los señores Jorge Carvajal, Sergio Chaigneau, Leonardo Navarro, Jacinto Vaello y Max Vildósola, quienes como profesores de seminario han colaborado conmigo en las clases de estadística para planificación

dictadas en el curso básico del Programa de Capacitación del Instituto. Este libro recoge sus valiosas sugerencias. Varios capítulos, especialmente aquellos que tienen estrecha vinculación con planificación y contabilidad nacional, han sido discutidos con el señor Pedro Sainz, quien ha hecho aportes de inestimable utilidad. El señor Gregorio Weinberg ha revisado con mucha paciencia el texto de esta primera edición. Su colaboración y la de otras personas cuyo enunciado sería interminable comprometen mi gratitud.

ARTURO NÚÑEZ DEL PRADO

PRIMERA PARTE

ESTADÍSTICA DESCRIPTIVA

I

ESTADÍSTICA Y PLANIFICACIÓN

A. LAS NECESIDADES DE INFORMACIÓN

Tomar decisiones racionales supone disponer de informaciones fieles, en cantidad suficiente, y con la oportunidad debida. Las decisiones erróneas se deben tanto a la falta de información como a deficientes evaluaciones de ésta. Cabe reconocer, desde un comienzo, que una buena proporción de las informaciones que debe manejar un planificador son de tipo cuantitativo. La estadística convencional presenta métodos que facilitan el análisis sobre variables cuantitativas; para el análisis de variables cualitativas, la estadística no paramétrica ya alcanzó un grado de desarrollo que permite tratamientos serios y de verdadera utilidad. Existe, por lo tanto, un cuerpo de conocimientos que posibilita el análisis tanto de variables cuantitativas como de las cualitativas. Sin embargo, las decisiones en planificación implican evaluaciones no estadísticas para ciertos aspectos del complejo problema. Conviene tener presente, por lo que antes se ha dicho, que la estadística es un instrumento útil que permite analizar una parte de los fenómenos que condicionan las decisiones en planificación, donde intervienen aspectos económicos, sociales y políticos con todas sus interacciones. Puede lograrse el conocimiento de la estructura de importaciones, la distribución del ingreso y la composición de fuerzas políticas, empleando instrumentos estadísticos; pero estimar la reacción de las clases populares ante una cierta tasa de crecimiento del consumo, exige una evaluación subjetiva e implica, en parte, juicios no cuantitativos.

Ha surgido una falsa controversia sobre las necesidades de información. Por una parte, reducir los problemas que aparecen en planificación a términos puramente cuantitativos, sería simplificar en exceso el problema; por otra, desconocer la utilidad de los instrumentos, significaría circunscribir la discusión a un marco muy general y peligrosamente confuso. Estas dos posiciones no constituyen alternativas; y tampoco es posible encontrar defensores de una u otra. Identificarse con alguna de estas posiciones extremas, significaría no comprender realmente qué sig-

nifica la planificación. Por lo demás, tampoco parece posible disociar el planteamiento de problemas generales, de una necesaria cuantificación. Así, por ejemplo, proponer una cierta redistribución del ingreso, supone conocer con bastante detalle cuantitativo, la distribución existente y la redistribución deseada; señalar la necesidad de distribuir ingresos, es apenas indicar vagamente un problema y un objetivo, cuya factibilidad y profundidad no pueden juzgarse si se desconocen cuantitativamente los tramos de ingreso y las proporciones de la población que los capta. Este ejemplo en modo alguno pretende postular que las informaciones deben ser necesariamente precisas; en planificación se puede trabajar con estimaciones, que en general suponen aproximaciones. Tampoco es indispensable una precisión rigurosa; saber, por ejemplo, que el coeficiente de inversión es del 10 o del 12 por ciento no establece una significativa diferencia para calificar esta situación. Vale decir, se pueden tolerar desvíos razonables, mas es preciso tener conciencia de qué significa un orden de magnitud, un intervalo razonable o, en general, la utilidad de una aproximación. Es indispensable, cuando se trabaja con estimaciones, plantear un intervalo donde, con una probabilidad cercana al 100 por ciento, se encuentre el valor verdadero; ahora bien, este intervalo tendrá una amplitud razonable, siempre que las conclusiones no sean significativamente distintas en uno y otro extremo de dicho intervalo. Admitido en planificación este criterio puede permitirse, y a veces no hay otra alternativa, la cuantificación aproximada, sobre la base de estimadores. Juicios de valor y opiniones generales no bastan; es preciso que las justificaciones obedezcan a deducciones lógicas, identificando magnitudes, proporciones e indicadores en general, pues todos ellos ayudan, y a veces en forma insustituible, a calificar fenómenos y obtener conclusiones objetivas y consistentes.

Dentro de este contexto general la acumulación de informaciones y su uso racional es inherente a un proceso de planificación. A continuación se detalla el empleo de los métodos estadísticos durante las diversas etapas de un proceso de planificación.

1] *En el diagnóstico*

Identificar los principales problemas de un sistema socioeconómico implica, por una parte, la investigación de una perspectiva histórica, la identificación de la estructura de poder, el comportamiento de estos grupos y los resultados que producen en las principales variables de evaluación: distribución del ingreso, es-

tructura del comercio exterior, estructura y magnitud de la inversión, estructura del consumo, etc. Ahora bien, puede ser útil, aunque no suficiente, conformarse con impresiones cualitativas sobre los aspectos citados; esta clase de información, si bien admite cierto tipo de calificaciones, no permite una comprensión cabal del funcionamiento del sistema considerado. Un diagnóstico completo debería incluir una descripción de su funcionamiento.

Para responder a este punto, resulta casi innecesario insistir sobre la urgencia de contar con información básica y con métodos estadísticos que permitan tratar y evaluar dicha información; se necesitan tanto los análisis de corte transversal, como los de procesos dinámicos. En otras palabras, la estadística descriptiva, vale decir, los indicadores de posición, dispersión y asimetría hacen posible caracterizar situaciones en el tiempo desde diferentes puntos de vista. Además el análisis de series cronológicas, los análisis de regresión y correlación se utilizan para analizar el comportamiento y la interrelación de las variables en el tiempo. Como es evidente se trata de análisis parciales, por etapas, que necesariamente deben compatibilizarse para obtener conclusiones consistentes. Cada estadígrafo o indicador muestra un aspecto, una faceta del problema; es necesario disponer, pues, de un conjunto de indicadores, estáticos y dinámicos, para así poder analizar el problema desde ángulos diferentes que den un marco integral al estudio.

Sin embargo, aun disponiendo de indicadores estáticos y dinámicos, tampoco es posible dar por terminada esta etapa sin antes disponer de una descripción detallada del funcionamiento de la actividad socioeconómica; y esta descripción puede ser literal o matemática. Esta última forma tiene evidentes ventajas, en cuanto a consistencia, claridad, precisión y el planteamiento explícito de supuestos. La estadística juega un papel fundamental en la determinación de funciones y en la especificación de valores de los parámetros que intervienen en las diferentes relaciones.

2] *En la prognosis*

En las proyecciones de las variables socioeconómicas resulta particularmente importante la aplicación de métodos de regresión y correlación y las estimaciones por razón, proporción y elasticidad, cuando se acepta el cumplimiento de los supuestos implícitos en las tendencias históricas, y se admiten reacciones similares a las

del pasado, frente a cambios en determinados factores cuyo comportamiento futuro pueda preverse, o sobre el cual pueda influirse deliberadamente. Esta visión anticipada del futuro permite comprender la magnitud de los problemas en una dimensión potencial. En los diagnósticos suelen percibirse graves problemas, pero una proyección deja ver el empeoramiento de esas situaciones, y los efectos que tendrían, si no se adoptan decisiones que cambien el sentido de esas tendencias. Las extrapolaciones estadísticas son precisamente los instrumentos apropiados para establecer prognosis.

3] *En la configuración de la imagen*

Si por la imagen a largo plazo que se puede hacer de un país se entiende un conjunto de intenciones condicionadas por expectativas de variables no controlables, que refleje los cambios deseables para las situaciones percibidas en el diagnóstico, parece lícito convenir que dicha imagen no sólo debiera incluir características cualitativas, sino que admitiría y sería beneficioso que en la imagen se precisen hasta donde la información permita caracterizaciones cuantitativas de importancia. Si en el diagnóstico, por ejemplo, se advierte que la estructura de exportaciones e importaciones es perniciosa, y por lo tanto debe ser modificada, postular que se debe cambiar esa situación en favor de una estructura de exportaciones que conceda mayor importancia al rubro manufacturas a expensas de una menor participación de materias primas, es sólo formular una intención general, que no puede ser calificada por su factibilidad ni por la importancia del objetivo y su compatibilidad con otros objetivos que plantea la imagen. Para cumplir con los requisitos mencionados, no se puede negar la necesidad de disponer de un conjunto de datos cuantitativos y cualitativos que pueden obtenerse aplicando métodos estadísticos idóneos. La imagen debería ser la culminación de un plan a largo plazo, y como tal, factible y consistente. ¿Cómo calificar factibilidad y consistencia en un plano muy general de ideas? No se pretende justificar la caracterización de una imagen bajo una maraña de datos que harían perder la visión central de las metas propuestas. Es a todas luces evidente que una imagen, por el plazo dentro del cual se la sitúa y por las informaciones que se disponen en el momento de configurarla, no puede abarcar detalles si se desea hacer un trabajo serio y honesto. Pero sí debe contener un conjunto de variables que permitan juzgar la audacia o la cautela en el planteamiento de objetivos y su compatibi-

lidad intrínseca. El estudio de series cronológicas, las proyecciones por elasticidad, los modelos de regresión y correlación y, en general, los modelos de planificación, constituyen herramientas de enorme utilidad.

4] *En la delimitación de la estrategia*

Constituye la estrategia el vínculo entre diagnóstico e imagen; que en la práctica significa decidirse por alternativas referentes a medidas muy generales de política económica que conducen al logro de los objetivos planteados en la imagen.

Naturalmente existe una relación íntima entre diagnóstico, imagen y estrategia; y estos conceptos del proceso de planificación están mutuamente condicionados. A los efectos de su presentación se los considera por separado, pero eso no debe hacer suponer que se trata de elementos disociados; antes bien, supone un encadenamiento de acciones en el tiempo. Se dijo antes que la elección de una estrategia resulta de tomar en cuenta alternativas; supuesto esto, es necesario evaluar dichas alternativas en la medida que las informaciones lo permitan. Necesariamente es indispensable definir grandes proyectos ligados a las alternativas de estrategia e imagen planteadas. Es fundamental un mínimo de cuantificaciones, proyecciones y estimaciones que sustenten las evaluaciones. Cambios sustantivos y concretos, implícitos o explícitos, en la imagen y estrategia deberían evaluarse tanto desde puntos de vista cualitativos, como cuantitativos. Nuevamente parece oportuno insistir sobre el hecho que es indispensable un mínimo conocimiento de métodos estadísticos para enfrentar este complejo problema. Es evidente que la base de sustentación de imágenes y estrategias realistas está constituida por el acopio de informaciones respecto de perspectivas en el avance tecnológico, en las prospecciones de recursos naturales, en las variaciones de las estructuras sociales, económicas, institucionales, políticas, en la composición de la estructura de poder, etc. Supone, por lo tanto, mecanismos ágiles de captación de informaciones que permitan anticipar la toma de decisiones, de acuerdo a las expectativas que se tengan.

La obtención de informaciones durante un proceso de planificación, debería ser una tarea continua y no esporádica; y por otra parte, la información primaria que se obtenga debe ser depurada, clasificada, resumida y analizada, aplicando adecuadas técnicas estadísticas.

5] *En el plan a mediano plazo*

Consiste, por una parte, en la actualización de la imagen a un plazo que media entre los tres y diez años, con mucho mayores detalles y especificaciones. Por otra, implica un conjunto de decisiones de política económica mucho más concretas e individualizadas que en el plan a largo plazo; se definen proyectos sectoriales y regionales que dan contenido físico a la estrategia; se hace indispensable una evaluación más precisa de objetivos y metas. Las técnicas de proyección y los modelos de programación constituyen los instrumentos cuantitativos más utilizados durante esta etapa. Estimaciones por regresión y elasticidad, la utilización de matrices de insumo producto, los balances de materiales, etc., son instrumentos a los cuales nuevamente se apela, por lo general, en forma intensiva. La confección de una matriz de insumo producto exige el manejo de una serie de métodos estadísticos; las actualizaciones de matrices obsoletas implican correcciones de coeficientes que suponen una racional utilización de números índices y de técnicas muestrales. Proyecciones a precios constantes y a precios variables que alcancen una estructura deseada, justifican una cabal comprensión de los temas señalados. Parece pues innecesario destacar las necesidades de estimación estadística de parámetros que suponen descripciones más detalladas del funcionamiento del sistema socioeconómico necesarias a mediano plazo.

6] *En el plan a corto plazo*

Para esta etapa es necesario concretar medidas específicas de política económica; pues parece que debería haber una superposición entre las intenciones y las actuaciones. Los mecanismos de evaluación que permitirán calificar la transformación de las intenciones en acciones, requieren una descripción muy detallada del funcionamiento del sistema económico. Es fundamental disponer, en un plano sectorial, institucional, y sociopolítico, más detallado, de funciones que expresan el comportamiento de variables trascendentes, por ejemplo funciones de producción, funciones de ingreso, funciones de consumo, funciones de precios, etcétera. Todo esto supone que se dispone de una cantidad de informaciones y se utilizan métodos estadísticos que permiten determinar los valores de los parámetros.

7] *En el control de avance y en la reformulación de los planes*

Sabido es que el proceso de planificación constituye una continua revisión de alternativas a la luz de las nuevas informaciones que

se van obteniendo a medida que transcurre el tiempo. Por otra parte, es indispensable estar informado sobre la realidad de la actividad socioeconómica para compararla con las intenciones que reflejan los diferentes tipos de plan. Estos dos hechos, entre otros, obligan a realizar sondeos periódicos para ir percibiendo las posibles distorsiones y para cerciorarse del grado de avance en el cumplimiento de las metas y objetivos del plan. Además es necesario subrayar que el acopio de información debe ser oportuno; verificar hechos históricos siempre será interesante para una serie de propósitos; pero verificar hechos en el momento que se producen es indispensable para las tareas de planificación. El empleo de técnicas muestrales, aplicadas tanto a la estimación de variables cuantitativas como a la de variables cualitativas, por sus indudables ventajas, parece ser una herramienta verdaderamente útil, y esto sobre todo si se piensa en el costo y la oportunidad con que se entregan los resultados. Así, entre los estudios coyunturales que se realizan en Francia, figuran encuestas mensuales sobre el desarrollo de la actividad industrial, sobre las condiciones de vida de las familias, etc., informaciones éstas que se recogen con extraordinaria frecuencia y periodicidad. En los países latinoamericanos, donde las ideas de cambio y reforma, en lo social y en lo económico, implican tareas reflejadas en una cantidad de planes, es fundamental plantear métodos oportunos para captar información y sistemas que permitan evaluarla.

B. MÉTODOS DE OBTENCIÓN DE INFORMACIONES

Para las distintas etapas del proceso de planificación enumeradas en páginas anteriores, es posible mencionar los siguientes métodos que permiten recoger la información necesaria.

1] *Censo*

Como es sabido constituye una indagación completa, sobre las variables que interesa investigar, de los elementos que componen una población claramente definida. El conocimiento censal de una población asegura la posibilidad de obtener datos fehacientes, siempre que no se cometan errores en la recopilación y en el tratamiento de la masa de datos. En general es muy difícil que un censo, sobre todo cuando la población es muy amplia y diversa, esté exento de alguno de los errores señalados. Mientras éstos no distorsionen significativamente las características reales de las poblaciones censadas, pueden pasarse por alto, des-

de el punto de vista de la planificación, desviaciones razonables respecto de los valores verdaderos.

Determinan las desventajas más serias de este método, el elevado costo que significa trabajar con volúmenes de información muy grandes y la demora consiguiente en obtener resultados concretos. Sin embargo, son indispensables investigaciones censales, pese a estas desventajas, por lo menos cada cierto número de años, para posibilitar la utilización de otros métodos durante los períodos intermedios y para disponer periódicamente de informaciones completas.

2] *Técnica muestral*

Debe entenderse como una indagación parcial sobre las variables que interesa investigar, de los elementos que componen una población. Es parcial, puesto que se considera una fracción, una muestra, de la población; sin embargo esta fracción poblacional debe ser calificada por su representatividad, es decir, debe asegurar que refleja, con alguna aproximación, las características poblacionales que interesa investigar. Este método, en cierta medida debe ser considerado como un complemento de las investigaciones censales; y complemento en un doble sentido: para intercalar estimaciones entre los períodos censales y para desglosar y agregar otras variables a las investigadas por métodos censales.

Sus grandes ventajas radican en el costo que, en general, es muy inferior al de un censo; en la oportunidad con que se entregan las estimaciones, y también en la posibilidad de realizar indagaciones exhaustivas sobre fenómenos concretos. Tal vez la mayor desventaja de esta técnica la determine la necesidad de trabajar con márgenes de probabilidad inferiores al 100 por ciento, es decir, sin la certeza absoluta que las estimaciones son válidas. Ahora bien, esta desventaja implica riesgos que, por lo general, tienen una probabilidad de ocurrencia inferior al 10 por ciento y muchas veces menores al 5 por ciento. Esta probabilidad puede ser tan pequeña como se quiera; pero su reducción tiene como contrapartida un crecimiento del tamaño de la muestra. El objetivo es trabajar con probabilidades pequeñas de error y con tamaños de muestra que no conviertan en prohibitiva la investigación por razones de costo y tiempo.

3] *Estudios de casos típicos*

Puede considerarse este método como un límite de muestras pequeñas dirigidas. Consiste en seleccionar algunos elementos representativos de grupos homogéneos de la población estudiada. El análisis de estos casos, que constituyen puntos importantes del abanico completo de la población, puede entregar informaciones que aunque incompletas, representen por lo menos un punto de partida para indagaciones más precisas. Este método debe ser interpretado como una investigación preliminar, como una prueba de factibilidad de posteriores investigaciones muestrales o censales. Sus ventajas en materia de costo y tiempo son evidentes. Y su gran desventaja estriba en el hecho que incorpora cierta dosis de arbitrariedad en la calificación de lo que es un caso típico o representativo; pero como se dijo, su principal aplicación responde a su factibilidad. Piénsese, por ejemplo, en diferentes alternativas de impuestos progresivos y exenciones; si se seleccionan algunos casos representativos: familias de un obrero no calificado, de uno calificado, de un empleado público, de un empleado particular, de un profesional, de un gerente, de un rentista, etc., en cada uno de estos casos podrá probarse cada una de las alternativas planteadas. Algunas de ellas serán fácilmente descalificadas y la discusión puede llegar a circunscribirse a muy pocas alternativas. Una pequeña muestra puede dilucidar la discusión cuando el número de alternativas se ha reducido. Es necesario admitir que este método supone algún conocimiento realista de la población para elegir los casos realmente representativos.

4] *Experimentación numérica*

El surgimiento y la utilización generalizada de computadores electrónicos permite la posibilidad de tanteos, pruebas de ensayo y error, con una velocidad extraordinaria, empleando un conjunto numeroso de alternativas. Para la planificación a corto plazo, disponer de una descripción detallada del funcionamiento de la actividad económica es una herramienta de innegable utilidad; ahora bien, las descripciones detalladas suponen que se dispone de una extraordinaria cantidad de información. Para ello se hace necesario destinar una buena cantidad de tiempo y esfuerzo a la recopilación de datos, aunque sean aproximados, que permitan alimentar el modelo que representa la formalización matemática de las ideas que se tengan respecto del funcionamiento del sistema económico. Las relaciones del modelo, de definición y compor-

tamiento, implican una enorme cantidad de parámetros, muchos de ellos desconocidos; por ello, una alternativa podría ser el intento de reproducir la historia reciente a través del modelo.

El modelo se alimenta con los datos disponibles sobre variables exógenas y parámetros y se dan valores intuitivos respecto de los parámetros desconocidos. Con ese conjunto de datos, unos fieles y otros estimados, se trata de reproducir la historia, por ejemplo de los últimos dos años, cuyas variables exógenas y las principales variables de resultado son conocidas. De la comparación entre las variables de resultado efectivas y las variables de resultado dadas por el modelo, surgen ideas para reformular las ecuaciones de comportamiento y para modificar los valores intuitivos de los parámetros. Se pueden repetir muchas veces los experimentos hasta encontrar satisfactoria la reproducción que entrega el modelo. Una reproducción aproximada de las variables de resultado es el punto de partida para realizar pruebas de sensibilidad, las que consisten en modificar ligeramente los valores de cada uno de los parámetros y analizar su efecto sobre las variables de resultado; la utilización de coeficientes de elasticidad permite calificar los parámetros como críticos y neutros, según el valor de dichos coeficientes. Esta calificación da una pauta para realizar investigaciones estadísticas y econométricas que permitan estimar aquellos parámetros críticos con métodos más rigurosos y confiables.

Dada la interrelación que tienen las variables en un modelo de este tipo, también será posible estimar, por medio de la experimentación numérica, variables y parámetros que puedan despejarse de otras relaciones de la descripción. Evidentemente esto supone una gran confianza en el tipo de funciones elegidas y en los valores de los parámetros conocidos.

C. LA DISPONIBILIDAD DE INFORMACIÓN EN AMÉRICA LATINA

Una apresurada generalización sobre las disponibilidades de información estadística en América Latina corre el riesgo de no tener una validez total. La gran heterogeneidad de países que caben dentro del común denominador de subdesarrollados también se manifiesta en la disponibilidad de datos. Con todo, es posible exponer algunas consideraciones que poseen un cierto carácter general.

En primer lugar, es necesario aceptar como hecho evidente la insuficiencia de información básica sobre varios aspectos de importancia para la planificación. Pocos son los países que poseen

informaciones sobre distribución de ingreso y de riqueza, y aparentemente no hay alguno que efectúe este tipo de investigaciones en forma continuada a nivel nacional.

La carencia de datos en una buena parte de los países latinoamericanos llega a ser realmente crítica; en algunos apenas se dispone de uno o dos índices de precios, y donde como agravante se notan características inflacionarias permanentes. Esta carencia de datos sobre precios ofrece un doble inconveniente: la dificultad de plantear políticas de precios y la imposibilidad de trabajar con valores reales, ya sea poder de compra o expresiones físicas; dos aspectos que interesan fundamentalmente a toda tarea de planificación. No es difícil seguir enumerando una serie de deficiencias en materia de datos: se desconocen estructuras de consumo, de inversión, de capital, de producción, de costos, etcétera.

En segundo lugar, una vez admitido el hecho que hay una aguda escasez de datos, tampoco se puede dejar de reconocer que, aun así, no siempre se hace un aprovechamiento óptimo de la escasa información disponible. Además existen problemas de interpretación de datos, provocados por deficiencias de las publicaciones y por falta de entrenamiento en el análisis estadístico.

Sin desconocer que en general la información es escasa, conviene tener en cuenta que hay una mayor cantidad de datos de la que habitualmente se supone, pues también existen investigaciones cuyos resultados no revisten carácter oficial. Es indispensable, por tanto, realizar inventarios de las estadísticas no publicadas. Además, existe una gran cantidad de información en estado primario, cuyo traspaso a tarjetas perforadas o a cintas o discos magnéticos, bastaría para obtener clasificaciones o resúmenes que, a su vez, pueden transformar una masa de datos casi inútil en informaciones aprovechables para propósitos múltiples.

En tercer lugar, no se puede dejar de subrayar que al problema de la escasez, se agrega el del atraso con que se dan a conocer informaciones que pueden tener enorme importancia en un determinado momento, pero que después sólo son útiles para verificar hechos pretéritos, que ya pertenecen a la historia. La idea de oportunidad en materia de planificación adquiere una significativa importancia.

En cuarto lugar, parece útil un breve comentario sobre la calidad de las informaciones. Con frecuencia muchas estadísticas provienen de muestras insuficientes o fueron obtenidas a través de cuestionarios que, involuntaria o inadvertidamente, inducen a cierto tipo de respuestas. Un conjunto de informaciones socio-

económicas, difícilmente puede superar todas las pruebas de consistencia que pueden plantearse.

Finalmente, dada la exigua disponibilidad de recursos para enfrentar investigaciones estadísticas, puede observarse una curiosa asignación de prioridades de tareas. Los organismos públicos y privados, plantean muchas veces investigaciones en forma independiente, sin preocuparse por armonizar criterios de los usuarios ni por dispendiosas duplicaciones.

D. EL CONTROL DE LA CALIDAD DE LAS INFORMACIONES ESTADÍSTICAS

Del aserto que en planificación no cabe exigir precisión rigurosa, no se sigue necesariamente que se pueda trabajar con cualquier calidad de información recogida; se advirtió ya que las desviaciones debían ser razonables y no exceder ciertos límites de tolerancia fuera de los cuales las conclusiones pueden ser ambiguas. Las siguientes consideraciones quizá puedan facilitar la comprobación de la calidad.

1] *Hipótesis*

Cuando se plantea una investigación, es fundamental establecer hipótesis previas acerca de los posibles valores que tendrían las variables investigadas. Confrontar los resultados efectivos con las hipótesis previas, constituye una primera prueba que ayuda en la evaluación. Si ambos tipos de datos apuntan en el mismo sentido, aumenta la confianza para admitirlos; si difieren en forma significativa será necesario averiguar la causa de las diferencias, por lo tanto, habría que revisar las hipótesis previas y la metodología que se aplicó para obtener informaciones.

2] *Consistencia*

Es conveniente establecer pruebas de consistencia interna, dentro del conjunto de los investigadores que se obtienen como resultado de la investigación, y también consistencia con otros antecedentes disponibles. Todo esto permite calificar mejor los datos con los cuales se trabaja. Recuérdese que, en planificación, uno de los puntos más delicados es lograr compatibilizaciones en distintos sentidos.

3] *Representatividad*

La evaluación de la representatividad de un indicador es parte de la etapa de control; un ejemplo concreto de falta de una evaluación adecuada, es la jerarquización del nivel de desarrollo de un país basándose sobre los ingresos por habitante.

En primer lugar, un promedio no siempre es representativo; para que lo sea, todos los valores tendrían que estar muy próximos a dicho promedio; el caso de los ingresos por habitante está muy lejos de acercarse a esta situación. En segundo lugar, hay un problema de cómputo de los ingresos reales: evaluaciones del autoconsumo y utilización de sistemas de deflactación muy diferentes entre los países, hacen todavía menos comparables los referidos promedios. A pesar de todo esto hay una tendencia a utilizarlos en las comparaciones internacionales sin especificar sus limitaciones. Cuando se manejan estadígrafos o indicadores que, en general, resumen de manera imperfecta una masa de datos, parece recomendable detenerse en el análisis de su representatividad y especificar qué faceta del problema puede abordarse a través de ellos y qué otros aspectos quedan al margen de tales indicaciones.

4] *Sesgos*

Cuando se recolectan informaciones por medio de encuestas, es particularmente importante registrar y eliminar, o por lo menos reducir, la posibilidad de trabajar con informaciones que contengan sesgos. Por sesgo se entiende una desviación sistemática de las características observadas respecto de las características reales, por consiguiente parecería casi innecesario subrayar el peligro que esto significa durante la etapa de obtención de conclusiones. Un mal diseño de cuestionarios puede, en cierto modo, inducir las respuestas pedidas. Los encuestadores, a veces, no mantienen una posición neutra e inclinan, en un cierto sentido, a quienes responden. Una encuesta sobre ingresos y consumos generalmente evidencia subestimaciones en el monto de las rentas y distorsiones en la estructura del gasto, sobre todo en los estratos de ingresos altos, y esto por razones fáciles de imaginar. La realización de confrontaciones utilizando muestras yuxtapuestas, permite comprobar la existencia de estos desvíos y, a veces, en algún sentido cuantificarlos. Con todo, tampoco puede desconocerse que el proceso de reducción de estos errores supone esfuerzos que encarecen sustancialmente la investigación.

E. ESTADÍSTICA PARA PLANIFICACIÓN

Respecto de las tareas inmediatas que parecen ineludibles, admitida la necesidad de establecer procesos continuos de planificación, se estima conveniente discutir las siguientes ideas:

1] *Investigaciones simultáneas*

No puede aguardarse que estén disponibles los datos para comenzar una serie de investigaciones socioeconómicas. Parecería más bien que la decisión de iniciar una investigación, debería adoptarse admitiendo dos criterios: por un lado, estructurar la parte conceptual y sustantiva de la investigación; y, por otra, simultáneamente, la investigación estadística pertinente. Muchas investigaciones se postergan por falta de información, cuando todo retardo causa un perjuicio que se magnifica a medida que transcurre el tiempo. Durante una primera etapa, los organismos de investigación deberían incorporar unidades cuyo objetivo esencial fuese la captación de información, en particular centros de muestreo, para no depender de otros organismos en la disponibilidad oportuna de datos. En ese sentido, las investigaciones interdisciplinarias parecen ofrecer facilidades en el planteamiento de investigaciones simultáneas; desde luego, no es ésta una solución integral del problema de la información; antes bien, se corre el riesgo de duplicar esfuerzos y malgastar recursos por falta de comunicación y coordinación entre los organismos. Sin embargo, como etapa transitoria y mientras se organizan planes nacionales de estadística, parecería que podrían evitarse así mayores postergaciones. Por lo demás no sería demasiado oneroso algún grado de duplicación, dada la exigüidad de los recursos destinados a la investigación estadística. En todo caso un mínimo de coordinación entre los diferentes organismos sería suficiente durante esta etapa transitoria.

2] *Organización institucional*

Un tema sobre el que se insistió mucho y sobre el que poco se hizo, es la adecuación de la estructura institucional que facilitaría el establecimiento de los procesos de planificación. En este sentido, y desde el punto de vista que se está considerando parece realmente conveniente establecer una estrecha relación entre las oficinas de estadística y los organismos de planificación. La cantidad de recursos disponibles para captar información obliga a una detenida jerarquización de las diferentes investigaciones. El

análisis de las distintas publicaciones en materia estadística que dan a conocer los países revela que existe una apreciable cantidad de información, que aunque importante es susceptible de ser sustituida o postergada en beneficio de otra más urgente. Es posible que este tipo de consideración esté determinada en parte por algún interés específico en materia de planificación, pero en todo caso el diseño de un plan nacional de estadística supondría evaluar diferentes asignaciones de recursos, considerar los intereses de los usuarios públicos y privados, evitaría innecesarias duplicaciones, delimitaría las responsabilidades de los diferentes organismos para la entrega de la información y, lo que es más importante, establecería plazos concretos para dar por terminadas investigaciones que muchas veces se arrastran durante lapsos prolongados.

3] *Requisitos concretos de información estadística en materia de planificación*

Parecería realmente fuera de lugar aquí elaborar, por interminable, una lista exhaustiva; sin embargo, enumerar los principales campos donde parece urgente concretar investigaciones mostraría la enorme e impostergable tarea que se tiene por delante. (Se parte del supuesto que existen adecuados sistemas de contabilidad nacional: cuentas nacionales, esquemas de fuentes y usos de fondos, matrices de insumo producto, balanza de pagos, etc.)

a) Tal vez uno de los principales campos corresponda al análisis de la distribución del ingreso y de la riqueza. Distinguir entre asalariados y no asalariados supone dos categorías demasiado elementales y heterogéneas en exceso. Es preciso detallar escalas de ingreso que abarquen grupos homogéneos de personas; pero convéngase también que cualquier indagación sobre esta materia ofrece muchas dificultades. La veracidad de los datos sobre ingresos personales (sobre todo en el grupo no asalariado) debería quedar asegurada por investigaciones complementarias sobre la propiedad, el consumo, la tributación, etc. Estudios con controles cruzados son indispensables desde el punto de vista de la calidad de los datos.

Tampoco parece exagerado admitir que informaciones sobre estos aspectos deberían constituir el punto de partida de la elaboración de cualquier plan que especifique acciones concretas.

b) Otro aspecto, muy ligado al anterior, es el reconocimiento de la estructura de preferencias de los consumidores según sus diferentes niveles de ingreso. La compatibilidad entre una estruc-

tura de producción y la demanda final, implica trabajar con coeficientes de elasticidad ingreso y elasticidad gasto a un nivel de desagregación bastante grande; la obtención de funciones consumo por tipos de productos homogéneos, permite dar coherencia a las variables. Por otra parte, investigaciones de este tipo permitirían verificar la representatividad de las ponderaciones efectuadas en los diferentes índices de precios al consumidor.

c) Dentro de las tareas urgentes en materia de recopilación estadística para la planificación, tiene especial importancia determinar el valor del capital por sectores, subsectores o ramas de actividad, que es, por cierto, un problema en extremo delicado. Requiere estudios muy prolijos confeccionar un inventario con especificación de antigüedad, y que al mismo tiempo evalúe el contenido.

d) Las estructuras de precios vigentes no siempre se ajustan a los objetivos sociales y económicos que plantean los planes. Estudiar la formación de los precios, sus componentes, la especulación y las duplicaciones en las distintas etapas del proceso de distribución y comercialización, son tareas de tipo estadístico que también deben tener prioridad en el futuro inmediato. La disponibilidad de índices de precios, en general, es insuficiente para cumplir con algún éxito tareas de cuantificación económica a través del tiempo; hay países, por ejemplo, donde sólo se dispone de un índice de los llamados de "costo de vida" confeccionado sobre una base obsoleta y para un sector escasamente representativo de la población.

e) Dentro de los países pueden delimitarse zonas geoeconómicas bastante diferenciadas, la necesidad de establecer determinado tipo de desagregación de la contabilidad social que tome en cuenta algún tipo de regionalización, permitiría dar, desde este punto de vista, una mayor coherencia a los objetivos perseguidos, a los proyectos elegidos y a las medidas de política económica.

f) El análisis de factibilidad necesario para imponer ciertas medidas de política económica, exige identificar los centros de decisión y la gravitación e influencia que tienen sobre la actividad económica; tales investigaciones tampoco son ajenas al empleo de métodos estadísticos. Lo que se dio en llamar "sociometría" apunta en este sentido.

g) Numerosas variables macroeconómicas se calculan tomando como referencia un período equivalente al año; sin embargo, una desagregación semestral o trimestral, permitiría adoptar medidas oportunas, si los hechos contingentes así lo requieren. Este tipo de desagregación temporal, tendría que basarse, naturalmente,

sobre estimaciones algo elementales. Estimaciones sobre producto, consumo, inversión, a un nivel de detalle sectorial parecen factibles utilizando muestras e indicadores parciales. Los indicadores a corto plazo cobran actualmente una significativa importancia.

F. PLAN NACIONAL DE ESTADÍSTICA

Las anteriores consideraciones sobre escasez, oportunidad, duplicaciones, necesidad de establecer prioridades en materia de investigaciones estadísticas y la fijación de plazos para realizarlas, constituyen fundamentos sólidos para confeccionar planes de estadística a nivel nacional. La realización y puesta en marcha de un plan de esta naturaleza implica la necesidad de hacer un inventario crítico, en materias de calidad y oportunidad, de las informaciones estadísticas existentes, tanto de las publicadas como de las que se reservan para el uso interno de diferentes organismos. Disponer de un catálogo semejante significaría una considerable ayuda para los investigadores; permitiría realizar un diagnóstico eficiente sobre la disponibilidad de informaciones.

La confección de un plan de estadística supone determinar prioridades respecto de los diferentes tipos de información, de la periodicidad y oportunidad de su publicación. Una encuesta a los principales usuarios, actuales y potenciales, parecería un método adecuado para preparar una lista del conjunto más general de informaciones necesarias. La frecuencia con que se requiere cada tipo de información, permitiría fijar prioridades a las diferentes investigaciones. Es evidente que los países empeñados en un proceso de planificación, tendrían que asignar recursos y especificar investigaciones, con una orientación clara en ese sentido.

A esta altura del trabajo se hace necesario repetir que la incorporación de técnicas muestrales para recoger información y utilizar computadores electrónicos en el procesamiento y cálculo de datos, conducen a un ahorro de tiempo y a veces de costo particularmente significativos.

La asignación de responsabilidad en cada una de las investigaciones programadas es un punto delicado durante la elaboración de un plan; en este sentido deberán precisarse las responsabilidades de las oficinas de estadística, así como también de cada uno de los demás organismos públicos, las universidades y centros de investigación y de las instituciones privadas.

La Oficina de Estadística dependiente del organismo de plani-

ficación, o por lo menos estrechamente vinculada a él, debería asumir la responsabilidad central de la coordinación administrativa y técnica del plan. Es de gran importancia, una vez asignadas las responsabilidades, discutir a profundidad los posibles métodos para captar información, el sistema de control, el tipo de clasificaciones, los indicadores resultantes y sus métodos de cómputo.

La fijación de plazos, entre el comienzo y el término de las investigaciones, es fundamental para garantizar el cumplimiento del programa de trabajo.

Estimaciones de costo, confección de presupuestos por investigación y el logro de financiamientos oportunos son tareas que completan este conjunto de ideas que pretende ser un esquema primario de discusión.

Nada fácil es imponer la idea de trabajar con seriedad y perspectiva, cuando se programan tareas de investigación estadística, perfilando un verdadero plan. Los problemas a corto plazo, en general, postergan asignaciones de recursos para esta clase de trabajos; cuando por ejemplo existen presiones políticas, económicas y sociales, por captar fracciones de presupuestos exiguos, es difícil imponer un plan de estadística que demanda recursos nada desdeñables y que parece no ofrecer utilidad inmediata. Sólo una evaluación a largo plazo, y una comparación con otras asignaciones de recursos que no siempre puede defenderse, permiten favorecer esta propuesta.

Una de las tareas de los planificadores es precisamente, llamar la atención sobre la urgencia de un plan de esta naturaleza, indispensable para cualquier proceso de planificación.

II

ESTADÍGRAFOS DESCRIPTIVOS ESTÁTICOS

A. DISTRIBUCIONES DE FRECUENCIA

1] *Generalidades*

Un conjunto de datos, o masa estadística, puede ser resumido y clasificado de acuerdo a criterios convenientes. Provengan las informaciones de censos o de muestras relativamente grandes, siempre serán útiles para el análisis, ya que difícilmente podrán obtenerse conclusiones válidas de una masa estadística no clasificada.

Los tipos de variables fundamentales, por lo menos para este trabajo, serán los siguientes:

a) variables cardinales: susceptibles de medición cuantitativa; y las que a su vez comprenden:
   i) *continuas:* variables que pueden tomar cualquier valor dentro de un intervalo (ingresos, estaturas, distancias, etc.)
   ii) *discretas:* variables que sólo toman algunos valores dentro de un intervalo (número de hijos por familia, número de accidentes de tránsito por día, etc.)

b) variables ordinales: sólo susceptibles de ordenación pero no de medición cuantitativa (grado de cultura de una persona: muy culta, regularmente culta, poco culta, inculta).

Para cada uno de estos tipos de variables, un conjunto de observaciones puede dar origen a una distribución de frecuencias; y ésta debe entenderse como un cuadro o tabla resumen de los datos originales.

En el caso de variables continuas será necesario fijar intervalos de frecuencia para llegar a un resumen efectivo de la información original. El punto medio de cada intervalo se denominará marca de clase y constituirá el valor representativo de cada intervalo. El número de observaciones que correspondan a cada intervalo se denominará frecuencias absolutas.

Una tabla de distribución de frecuencia para variable continua y sus símbolos correspondientes se presenta de la siguiente forma:

Ingresos de profesionales			*Número de profesionales*
Intervalos		*Marcas de clase*	*Frecuencias absolutas*
Y'_{i-1}	Y'_i	Y_i	n_i
Y'_0	Y'_1	Y_1	n_1
Y'_1	Y'_2	Y_2	n_2
Y'_2	Y'_3	Y_3	n_3
.	.	.	.
.	.	.	.
Y'_{m-1}	Y'_m	Y_m	n_m

donde:

$$Y_i = \frac{Y'_{i-1} + Y'_i}{2} \quad : \quad \text{marca de clase}$$

$$n = \sum_{i=1}^{m} n_i \quad : \quad \text{número de observaciones}$$

$$c_i = Y'_i - Y'_{i-1} \quad : \quad \text{amplitud del intervalo}$$

Estas tablas pueden ser de amplitud constante o de amplitud variable, según los valores que tome c_i.

Cuando se trata de variable discreta o discontinua, la tabla de distribución de frecuencias adquiere la forma siguiente:

Y_i	n_i
Y_1	n_1
Y_2	n_2
Y_3	n_3
.	.
.	.
Y_m	n_m

Cabe destacar que cuando la variable adquiere numerosos valores distintos para abreviar el trabajo, con cierta arbitrariedad y con alguna pérdida de precisión, puede tratarse como una variable continua, formando intervalos de clase.

Por último, en el caso de variables no mensurables, dicha tabla adoptará una forma como la siguiente:

Variable	*Frecuencia*
Característica A	n_A
Característica B	n_B
.	.
.	.
Característica Z	n_Z

Él lector advertirá que las tablas de distribución de frecuencias facilitan enormemente el análisis. Es muy ventajoso disponer de informaciones clasificadas en intervalos o en valores específicos de la variable, ya que, de esta manera, es posible obtener conclusiones primarias acerca de la variable que se investiga.

Respecto de las frecuencias, es posible y generalmente útil presentarlas en términos relativos, calculando la proporción que corresponde a cada intervalo o marca de clase sobre el total de observaciones. Se denominan frecuencias relativas, y se simbolizarán por h_i:

$$h_i = \frac{n_i}{n}$$

Tanto las frecuencias absolutas como las relativas son susceptibles de acumulación respecto de los intervalos o marcas de clase. Las frecuencias absolutas acumuladas se simbolizarán por N_i y se definen:

$$N_i = \sum_{i=1}^{i} n_i$$

Las frecuencias relativas acumuladas se simbolizarán por H_i y se definen:

$$H_i = \sum_{i=1}^{i} h_i$$

En general este tipo de frecuencias se acumulan en sentido creciente de la variable, y una frecuencia acumulada N_i indica el número de casos u observaciones donde la variable toma valores a lo sumo iguales a Y_i, en el caso de variable discreta, y a Y'_i en el caso de variable continua. Sin embargo, para ciertos análisis también es necesario acumular en sentido inverso; de ahí que se hable de frecuencias acumuladas hacia arriba o hacia abajo.

2] *Representación gráfica*

En general, la representación gráfica de una tabla de distribución de frecuencias, permite percibir con mayor claridad algunas características de la masa de datos que se investiga; por ello resulta bastante más fácil trasmitir conclusiones a personas no habituadas a la interpretación de distribuciones de frecuencia, cuando se utilizan gráficos estadísticos.

a) Representación gráfica de variable continua. Si se utiliza un par de ejes coordenados, en el eje de las abscisas se representará la variable estudiada, en tanto que en el eje de las ordenadas, se representará las frecuencias correspondientes. Recuérdese que en este tipo de variables la frecuencia corresponde a un intervalo y por esto se representa mediante una superficie.

Con un ejemplo se ilustrarán estas ideas; admítase, en este sentido, la siguiente tabla correspondiente a las edades de los participantes de un curso de estadística:

Edades		*Alumnos*	*Amplitud de intervalo*
Y'_{i-1}	Y'_i	n_i	C_i
18	22	10	4
22	26	20	4
26	30	16	4
30	38	12	8
38	40	1	2

Puesto que la amplitud más frecuente es 4, puede adoptársela como amplitud unitaria; así el cuarto intervalo tendrá dos veces la amplitud unitaria elegida y el quinto intervalo tendrá la mitad de dicha amplitud. La representación gráfica se hará según la gráfica 1 de la página siguiente.

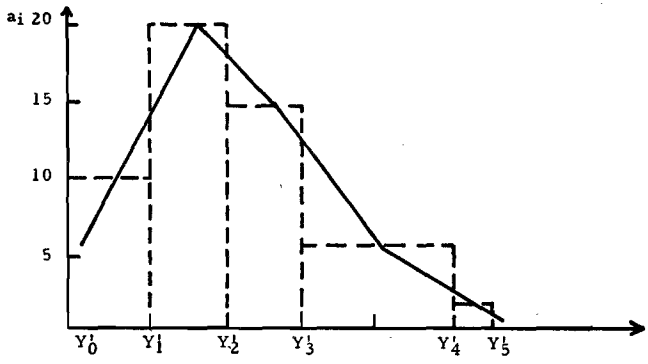
En la gráfica, para calcular la altura de cada rectángulo se plantea la relación de superficie siguiente:

$$\text{superficie} = \text{base} \times \text{altura}$$

$$n_i = c'_i \times a_i$$

donde c'_i es la amplitud unitaria elegida con el objeto de diseñar una gráfica adecuada. Desde luego que también pudo haberse trabajado con las amplitudes originales, aunque habría sido algo más laborioso.

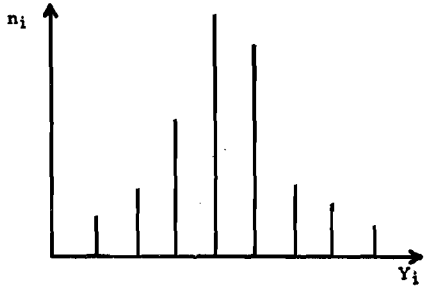
GRÁFICA 1



Este tipo de gráficas recibe el nombre de histogramas y la línea quebrada que une los puntos medios de los lados superiores de los rectángulos se denomina polígono de frecuencias.

b) Representación gráfica de variable discreta. En este caso la frecuencia correspondiente a cada valor de la variable estará representada por una barra vertical.

GRÁFICA 2



Naturalmente, se puede construir, en forma similar, gráficas que relacionen la variable con cualquiera de los tipos de frecuencias que se han visto, relativas, acumuladas, etcétera.

A continuación se presentará un ejemplo donde se seguirán todos los pasos necesarios para llegar a una tabla completa de distribución de frecuencias. Supóngase que se dispone de las siguientes informaciones acerca de los sueldos de los obreros de una fábrica (en dólares por mes):

68	48	53	73	100	80	40	55	65	95	85	35	110	120	60
90	70	40	80	100	70	50	55	70	65	45	80	60	90	50
55	60	30	110	110	90	70	60	45	65	80	85	90	68	72
50	40	45	90	105	108	35	45	50	70	82	84	66	38	48

Una de las primeras decisiones que deben adoptarse es determinar el número de intervalos que tendrá la tabla. Para ello es necesario considerar el objetivo que se persigue con el estudio de la variable, qué tipo de diferenciaciones o agrupamientos interesaría conocer; por otra parte, es indispensable determinar el recorrido de la variable, es decir, el menor y mayor valor entre los datos que se analizan. Por último, el número de observaciones, de manera que los diferentes intervalos tengan frecuencias en alguna medida significativas. Cuando existan valores escasos, muy alejados de lo que podría llamarse una concentración central, puede optarse por dejar los intervalos extremos abiertos. Supóngase que, tomando en cuenta las consideraciones anteriores, se decide clasificar los datos originales en 9 intervalos de amplitud constante. Dado que la diferencia entre los valores extremos (30 y 120) es de 90, la amplitud de los intervalos será igual a 10.

Y'_{i-1}	Y'_i	*Observaciones*	n_i	h_i	N_i	H_i	N_i*	H_i*	Y_i
30.0	40	𝍸 //	7	7/60	7	7/60	60	60/60	35
40.1	50	𝍸 𝍸	10	10/60	17	17/60	53	53/60	45
50.1	60	𝍸 ///	8	8/60	25	25/60	43	43/60	55
60.1	70	𝍸 𝍸 /	11	11/60	36	36/60	35	35/60	65
70.1	80	𝍸 /	6	6/60	42	42/60	24	24/60	75
80.1	90	𝍸 ////	9	9/60	51	51/60	18	18/60	85
90.1	100	///	3	3/60	54	54/60	9	9/60	95
100.1	110	𝍸	5	5/60	59	59/60	6	6/60	105
110.1	120	/	1	1/60	60	60/60	1	1/60	115
			60	1					

* Acumuladas "hacia arriba".

Para evitar situaciones ambiguas, cuando se presentan observaciones con valores de la variable que corresponden a los límites de los intervalos, puede seguirse el criterio planteado en el ejemplo, es decir, agregar un decimal a la columna de límites inferiores, aunque esto tenga utilidad solamente para clasificar las

observaciones, ya que en los cálculos que posteriormente se tratará, tales decimales son despreciados.

Con lo visto hasta el momento, es posible realizar los primeros análisis de un conjunto dado de datos. Tanto la representación gráfica como la tabulación de las distintas clases de frecuencias ayudan a ensayar los primeros juicios. Es necesario insistir sobre la necesidad de tomar en cuenta, en todas las decisiones respecto de la tabla de frecuencias, la naturaleza del fenómeno que se investiga; sobre todo en lo que se refiere al número de intervalos y a sus amplitudes: constantes o variables. Naturalmente que, una vez clasificados los datos originales, será preciso realizar análisis con mayor profundidad, utilizando los instrumentos estadísticos que se detallarán en las próximas páginas.

B. ESTADÍGRAFOS DE TENDENCIA CENTRAL

Una vez conseguida la clasificación de los datos originales, cuyas características más esenciales se destacan, será preciso calcular un conjunto de indicadores que caractericen en forma algo más precisa la distribución que se está estudiando. Interesa, en primer término, disponer de estadígrafos que representen valores centrales en torno de los cuales se agrupan las observaciones, en general se los designa como promedios, y son de extraordinaria utilidad tanto en el análisis de una distribución, como en la comparación entre distribuciones.

1] *Media aritmética*

Es sin duda el estadígrafo más utilizado, sobre todo en la cuantificación de variables económicas. Se simbolizará por $\bar{Y}$ o $M[Y_i]$, y se definirá como:

$$\bar{Y} = M[Y_i] = \frac{\sum_{i=1}^{m} Y_i \, n_i}{n}$$

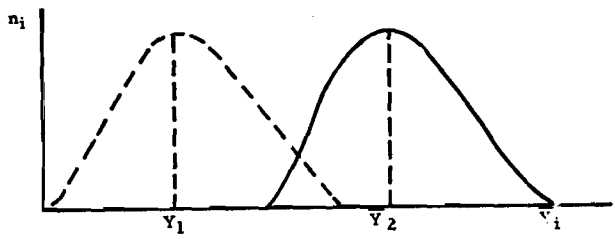
Puede observarse que a cada valor de la variable o marca de clase se atribuye una importancia o peso equivalente a la frecuencia absoluta correspondiente. Esta fórmula de cálculo es para datos agrupados en forma de una distribución de frecuencias. Cuando se desea calcular una media aritmética de datos no agrupados, todas las frecuencias absolutas serán iguales a la unidad, se simbolizará por $\bar{X}$ o $M[X_i]$ y se definirá como

$$\bar{X} = M[X_i] = \frac{\sum_{i=1}^{n} X_i}{n}$$

En general cuando el número de observaciones es relativamente grande conviene la agrupación de frecuencias en intervalos. Es evidente que cuando se resume un conjunto de datos en un número dado de intervalos, se pierde precisión; esta pérdida estará relacionada con la amplitud del intervalo, cuanto mayor sea ésta, menos preciso será el cálculo. Por ello, en general, para un mismo conjunto de datos, la media aritmética obtenida de los datos originales, que es un cálculo exacto, diferirá de la obtenida de una tabla de distribución de frecuencias. La razón estriba en el supuesto de uniformidad de la distribución de frecuencias dentro de cada intervalo, supuesto que generalmente no se cumple; mas, en todo caso, esa pérdida de precisión está más que compensada por las ventajas que significa tener una tabla de frecuencias. Por lo demás en ciencias sociales, la precisión necesaria autoriza, dentro de ciertos límites, a desentenderse de una rigurosidad extrema.

En la gráfica que a continuación se presenta, se tienen dos distribuciones de frecuencias (supuesto un gran número de intervalos pequeños) muy similares, y sin embargo con medias aritméticas muy distintas.

GRÁFICA 3



La media aritmética como estadígrafo de tendencia central indica la posición de la distribución. Sobre este estadígrafo cabe advertir su alta sensibilidad a valores extremos de la variable. Un valor muy alejado de los valores centrales, aunque poco representativo por ser único, puede hacer variar significativamente el promedio. Por ello, cuando se está utilizando este indicador

en un análisis, vale la pena advertir la representatividad de los valores extremos y la influencia que éstos tienen sobre el resultado. Muchas veces se concluye que es preferible estratificar previamente los datos originales en dos o tres categorías, realizando cálculos de medias aritméticas en forma separada para cada grupo.

a) Propiedades. Se presentarán las propiedades más importantes de la media aritmética.

i) Primera propiedad: La suma de las desviaciones ponderadas de los valores de la variable respecto de la media aritmética es cero

$$\sum_{i=1}^{m} (Y_i - \bar{Y})\, n_i = 0$$

$$\sum_{i=1}^{m} Y_i\, n_i - n\bar{Y} = 0$$

$$n\,\bar{Y} - n\bar{Y} = 0$$

ii) Segunda propiedad: La suma de los cuadrados de las desviaciones ponderadas de los valores de la variables es un mínimo, cuando se toman respecto de la media aritmética. Se iniciará la demostración tomando desviaciones respecto a un valor cualquiera P, para luego concluir que P forzosamente tendrá que ser la media aritmética

$$\sum_{i=1}^{m} (Y_i - P)^2\, n_i \quad \text{mínimo}$$

La derivada de esa expresión respecto de P, se iguala a cero.

$$-2 \sum_{i=1}^{m} (Y_i - P)\, n_i = 0$$

$$\sum_{i=1}^{m} Y_i\, n_i - nP = 0$$

$$P = \frac{\sum_{i=1}^{m} Y_i\, n_i}{n}$$

Es seguro que igualando la primera derivada a cero se obtiene un mínimo, porque se llega a un valor concreto. Para que fuera un máximo, P tendría que ser infinito.

iii) Tercera propiedad: La media aritmética de una variable más (menos) una constante es igual a la media de la variable más (menos) la constante.

$$M\,[Y_i \pm K] = M\,[Y_i] \pm K$$

$$\sum_{i=1}^{m} \frac{(Y_i \pm K)\,n_i}{n} = M\,[Y_i] \pm K$$

$$\sum_{i=1}^{m} \frac{Y_i\,n_i}{n} \pm \frac{n\,K}{n} = M\,[Y_i] \pm K$$

$$M\,[Y_i] \pm K = M\,[Y_i] \pm K$$

iv) Cuarta propiedad: La media aritmética de una variable multiplicada (dividida) por una constante, es igual a la constante que multiplica (divide) a la media de la variable.

$$M\,[Y_i\,K] = K\,M\,[Y_i]$$

$$\sum_{i=1}^{m} \frac{Y_i\,K\,n_i}{n} = K\,M\,[Y_i]$$

$$K \sum_{i=1}^{m} \frac{Y_i\,n_i}{n} = K\,M\,[Y_i]$$

$$K\,M\,[Y_i] = K\,M\,[Y_i]$$

b) Métodos abreviados de cálculo. Se presentarán estos métodos no tanto por el ahorro de tiempo que puede significar su aplicación, como porque destacan algunos detalles sobre la media aritmética, y muestran procedimientos de trabajo con cambio de variable.

i) Primer método abreviado: Se trata de reducir la magnitud de la variable, en términos de desviaciones respecto de un origen de trabajo O_t arbitrariamente elegido. En cuanto a la elección de O_t vale la pena, para que el método sea realmente abreviado, tomar como origen de trabajo un

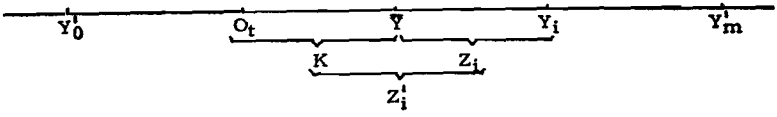
valor o marca de clase central de la distribución de frecuencias.

Se definirá esta variable reducida en la forma siguiente:

$$Z'_i = Y_i - O_t$$

Gráficamente se fijan los siguientes puntos dentro del recorrido de una variable:

GRÁFICA 4



Además, las desviaciones respecto de la media aritmética se simbolizarán por Z_i, es decir:

$$Z_i = Y_i - \bar{Y}$$

La diferencia entre la media aritmética y el origen de trabajo arbitrariamente elegido se designará por K, es decir.

$$K = \bar{Y} - O_t$$

Por este hecho, observando la gráfica puede concluirse que

$$\bar{Y} = O_t + K$$

Se trata de encontrar una expresión para K, en función de desviaciones Z'_i, para disponer de una fórmula de cálculo. En efecto,

$$Z'_i = \overline{Z_i} + K \text{ (multiplicando por } n_i)$$

$$Z'_i\ n_i = Z_i\ n_i + K\ n_i \text{ (aplicando sumatoria)}$$

$$\sum_{i=1}^{m} Z'_i\ n_i = \sum_{i=1}^{m} \overline{Z_i\ n_i} + nK$$

Recuérdese que en la primera propiedad de la media aritmética se demostró que:

$$\Sigma\ Z_i\ n_i = \Sigma\ (Y_i - \bar{Y})\ n_i = 0$$

Luego

$$K = \frac{\Sigma\, Z'_i\ n_i}{n}$$

La fórmula de cálculo abreviado será

$$\bar{Y} = O_t + \frac{\Sigma\, Z'_i\ n_i}{n}$$

En la siguiente tabla de distribución de frecuencias se aplica este método:

Intervalo $Y'_{i-1} - Y'_i$	*Marca de clase* Y_i	*Frecuencia* n_i	*Desviaciones* $Z'_i = Y_i - O_t$*	*Desviaciones ponderadas* $Z'_i\ n_i$
0 - 8	4	8	—19	—152
8 - 20	14	10	—9	—90
20 - 26	23	30	0	0
26 - 30	28	9	5	45
30 - 40	35	3	12	36
		60		—161

* Se elige $O_t = 23$.

$$\bar{Y} = 23 - \frac{161}{60} = 20.32$$

ii) Segundo método abreviado. Este método en general sólo se aplica con ventaja, cuando es constante la amplitud de los intervalos. Como en el anterior, se trata de trabajar en términos de desviaciones, pero además en este método dichas desviaciones se expresan en unidades de intervalo (dividiéndolas por C).

Esta nueva variable es en consecuencia,

$$Z''_i = \frac{Y_i - O_t}{C} = \frac{Z'_i}{C}$$

de donde

$$Z'_i = CZ''_i$$

En la fórmula del primer método se remplaza Z'_i y se tiene

$$Y = O_t + C \frac{\sum_{i=1}^{m} Z''_i \, n_i}{n}$$

Cabe destacar que este método permite obtener Z''_i en forma totalmente mecánica; basta fijar el origen de trabajo para completar todos los valores de Z''_i.

En el siguiente ejemplo podrá apreciarse las ventajas de esta forma de cálculo.

Intervalo $Y'_{i-1} - Y'_i$	*Marca de clase* Y_i	*Frecuencia* n_i	*Desviaciones* Z''_i	*Desviaciones ponderadas* $Z''_i \, n_i$
2 - 6	4	20	−3	−60
6 - 10	8	40	−2	−80
10 - 14	12	50	−1	−50
14 - 18	16*	90	0	0
18 - 22	20	60	1	60
22 - 26	24	40	2	80
		300		−50

* Eligiendo $O_t = 16$.

Aplicando la fórmula del segundo método abreviado se tiene

$$\bar{Y} = O_t + C \frac{\sum_{i=1}^{m} Z''_i \, n_i}{n}$$

$$\bar{Y} = 16 - 4\,\frac{50}{300} = 15.33$$

Obsérvese que una vez fijado el origen de trabajo las Z_i'' se colocan en sucesión decreciente para las marcas de clase menores que O_t y en sucesión creciente para las marcas de clase mayores.

2] *Mediana* (M_e)

Se trata de otro estadígrafo de tendencia central de aplicación muy frecuente. Se define como el valor de la variable que supera a no más de la mitad de las observaciones y es superado por no más de la mitad de dichas observaciones. Es un estadígrafo menos sensible que la media aritmética ante valores extremos de la variable, y puede ser calculado aun en variables de tipo ordinal.

Cuando las observaciones no están agrupadas en forma de una tabla de distribución de frecuencias, su cómputo es en extremo sencillo. Basta disponer los valores en orden creciente y ubicar el valor central. Por ejemplo, supóngase que se tienen ordenados los siguientes valores de gastos en consumo de 7 familias (en dólares por mes).

$$40 - 47 - 60 - 70 - 78 - 80 - 90$$

La mediana será 70 dólares ya que este valor supera a 3 observaciones (40, 47 y 60) que no son más que la mitad (la mitad es 3.5) y a su vez es superada por 3 observaciones (78, 80 y 90) que tampoco son más de la mitad.

Cuando el número de observaciones es par, existen dos valores centrales que satisfacen la definición de mediana. Si en el ejemplo anterior se agrega una familia adicional se tiene:

$$40 - 47 - 60 - 70 - 78 - 80 - 90 - 180$$

En este caso tanto 70 como 78 son valores medianos. La mitad de las observaciones es 4 y 70 supera a 3 que no son más de la mitad y es superado por 4 que tampoco es más de la mitad, pues es exactamente la mitad. Igual cosa ocurre con 78; para evitar ambigüedades se toma como mediana en estos casos el punto medio entre los dos valores medianos. En el ejemplo, la mediana definitiva sería 74 dólares. Obsérvese la poca sensibi-

lidad de este estadígrafo a los valores extremos. Al agregar la 8a. observación con un valor bastante más alto que el resto, la mediana ha experimentado apenas un ligero crecimiento; más todavía aunque en vez de 180 aquel valor hubiese sido de 5 000, la mediana siempre habría sido 74. En cambio no ocurre lo mismo para la media aritmética, que es sumamente sensible a ese tipo de valores extremos; intente el lector su cálculo en uno y otro ejemplo, y verificará una fuerte variabilidad.

Si los datos se agrupan en una tabla de frecuencias, el cálculo de la mediana implica computar previamente las frecuencias acumuladas.

a) Variable discreta, en este caso bastará con identificar la frecuencia acumulada que es inmediatamente mayor a la mitad de las observaciones. La mediana será aquel valor de la variable que corresponda a dicha frecuencia acumulada. Ejemplo:

Número de predios por persona Y_i	*Número de propietarios* n_i	*Frecuencia acumulada* N_i
1	200	200
2	160	360
3	150	510
4	100	610
5	80	690
6	40	730
7 y más	20	750
	750	

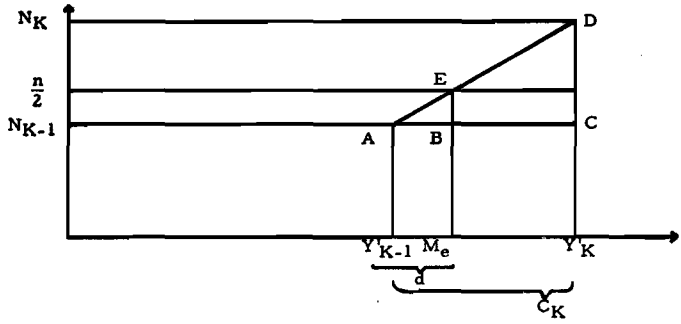
Siendo $\frac{n}{2} = 375$, la menor frecuencia acumulada que supera este valor es 510, que corresponde al valor 3 de la variable, siendo éste el valor mediano. Dicho valor supera a 360 observaciones que no son más de la mitad y es superado por 260 que tampoco son más de la mitad, satisfaciendo la definición de mediana.

b) Variable continua, en este caso el problema consiste en determinar un punto dentro del intervalo en que está comprendida la mediana. La identificación del intervalo donde se halla la mediana, es exactamente igual al caso de variable discreta: el intervalo será aquel que corresponde a la fre-

cuencia acumulada inmediatamente superior a la mitad de las observaciones. Como se dijo, la preocupación consiste en fijar un punto dentro de este intervalo, que corresponde a la mediana. Para ello se adoptará el supuesto que las observaciones se distribuyen linealmente dentro del mencionado intervalo.

Gráficamente se tiene:

GRÁFICA 5



Sea $Y'_{K-1} - Y'_K$ el intervalo donde se halla la mediana. Luego.

$$M_e = Y'_{K-1} + d$$

Será necesario encontrar una expresión para "d" en función de frecuencias que son los valores conocidos que se dispone y por los cuales está determinado este estadígrafo.

Por semejanza de triángulos se tiene que

$$\frac{AB}{BE} = \frac{AC}{CD}$$

Pero

$$AB = d$$

$$BE = \frac{n}{2} - N_{K-1}$$

$$AC = C_K \text{ (amplitud del intervalo K-ésimo)}$$

$$CD = N_K - N_{K-1} = n_K$$

Remplazando, se tiene

$$\frac{d}{\frac{n}{2} - N_{K-1}} = \frac{C_K}{n_K}$$

de donde

$$d = C_K \frac{\frac{n}{2} - N_{K-1}}{n_K}$$

Luego

$$M_e = Y'_{K-1} + C_K \frac{\left(\frac{n}{2} - N_{K-1}\right)}{n_K}$$

Ejemplo: la siguiente tabla muestra la distribución de los coeficientes producto-capital de 310 empresas industriales:

Coeficientes $Y'_{i-1} - Y'_i$	*Empresas* n_i	*Frecuencias acumuladas* N_i
0.15 - 0.20	40	40
0.20 - 0.30	80	120
0.30 - 0.42	100	220
0.42 - 0.50	60	280
0.50 - 0.70	30	310
	310	

$$\frac{n}{2} = \frac{310}{2} = 155$$

La menor frecuencia acumulada que supera a 155 es $N_K = 220$. Luego, $n_K = 100$,

$Y'_{K-1} = 0.30$, $C_K = 0.12$ y $N_{K-1} = 120$ (donde $K = 3$)

Remplazando estos valores en la fórmula

$$M_e = 0.30 + 0.12 \frac{155 - 120}{100} = 0.30 + 0.042 = 0.342$$

Como una extensión de este estadígrafo, será fácil ampliar el concepto a otros indicadores que dividen la masa de informaciones en otras proporciones y no sólo en mitades como lo hace la mediana.

Se tiene el caso de los cuartiles que dividen las observaciones en cuartas partes; así, el primer cuartil Q_1 es un valor de la variable que supera a no más de un cuarto de las observaciones, y es superado por no más de tres cuartos de ellas. Para identificar el intervalo donde se halla Q_1, habrá que determinar la frecuencia acumulada inmediatamente superior a $\frac{n}{4}$. El cálculo es similar al de la mediana:

$$Q_1 = Y'_{K-1} + C_K \frac{\left(\frac{n}{4} - N_{K-1}\right)}{n_K}$$

Para el tercer cuartil, sucede otro tanto

$$Q_3 = Y'_{K-1} + C_K \frac{\left(3\frac{n}{4} - N_{K-1}\right)}{n_K}$$

Para identificar el intervalo que comprende a Q_3, habrá que averiguar cuál es la frecuencia acumulada N_K, inmediatamente superior a $3\frac{n}{4}$; no es necesario detallar el segundo cuartil, porque coincide con la mediana. En forma similar se pueden encontrar estadígrafos que dividan al total de observaciones en décimas partes (deciles), en centésimas partes (percentiles), etc. Las fórmulas correspondientes pueden deducirse por analogía con las de los cuartiles.

Así, el 7º decil estará dado por

$$D_7 = Y'_{K-1} + C_K \frac{\frac{7n}{10} - N_{K-1}}{n_K}$$

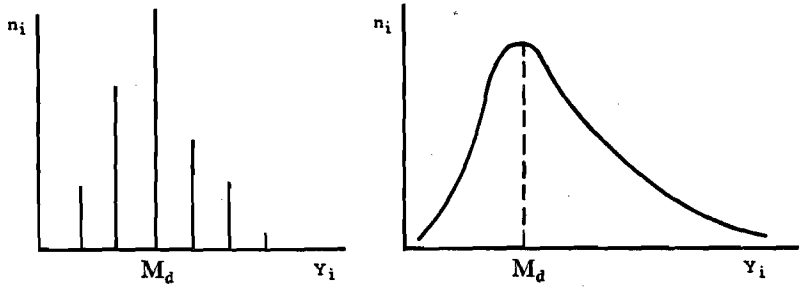
El 35 percentil estará dado por

$$P_{35} = Y'_{K-1} + C_K \frac{\frac{35n}{100} - N_{K-1}}{n_K}$$

3] *Moda o valor modal*

Se trata de otro estadígrafo de tendencia central. Tiene un significado bastante preciso y es de extraordinaria utilidad, aunque inexplicablemente poco utilizado en estudios socioeconómicos. Se simbolizará por M_d y se definirá como aquel valor de la variable al que corresponde la máxima frecuencia. Es muy corriente que se confunda el valor modal con la frecuencia máxima; recuérdese que es un valor de la variable y por lo mismo se le representa en el eje de las abscisas. Está dado por la frecuencia máxima, pero no se trata de una frecuencia.

GRÁFICA 6



a) Variable discreta, una vez agrupados los datos, es posible determinar inmediatamente el valor modal; bastará con fijar el valor de la variable que más se repite.

Ejemplo:

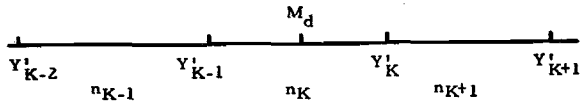
Número de cargas familiares Y_i	*Número de familias* n_i
0	80
1	120
2	210
3	380
4	180
5	60
6 o más	40
	1 070

La frecuencia máxima es 380, que corresponde al cuarto valor de la variable. El valor modal en consecuencia, es 3; este valor modal será tanto más representativo cuanto mayor sea la frecuencia máxima. Se presentarán algunos casos donde el valor modal pierde significación; es el caso donde hay varios valores de las variables que tienen frecuencias similares. Es así como se califica a las distribuciones de unimodales, bimodales, multimodales, etcétera.

b) Variable continua, de la misma manera que en el cálculo de la mediana, primero es necesario determinar el intervalo donde se halla comprendido el valor modal; en este caso bastará ver cuál es el intervalo que tiene la frecuencia máxima. El paso siguiente es determinar un punto dentro de ese intervalo. Existen algunos criterios, un tanto arbitrarios, para deducir fórmulas del valor modal. Uno de esos criterios toma en cuenta la magnitud de las frecuencias de los intervalos contiguos (mayor y menor) al que comprende el valor modal; en otras palabras, dividirá al intervalo en partes inversamente proporcionales a las frecuencias de los intervalos contiguos.

Gráficamente:

GRÁFICA 7



La moda estará más cerca del intervalo contiguo que tenga mayor frecuencia.

$$\frac{M_d - Y'_{K-1}}{Y'_K - M_d} = \frac{n_{K+1}}{n_{K-1}}$$

Despejando M_d de la relación anterior, se tiene:

$$M_d\, n_{K-1} - Y'_{K-1}\, n_{K-1} = Y'^{*}_K\, n_{K+1} - M_d\, n_{K+1}$$

pero,

$$Y'_K = Y'_{K-1} + C_K$$

$$M_d\,[n_{K-1} + n_{K+1}] = Y'_{K-1}\,[n_{K+1} + n_{K-1}] + C_K\, n_{K+1}$$

$$M_d = Y'_{K-1} + \frac{C_K\, n_{K+1}}{n_{K-1} + n_{K+1}}$$

Es necesario destacar que la deducción anterior no toma en cuenta la amplitud de los intervalos contiguos; en caso de amplitudes muy diferentes, puede distorsionarse el valor de este estadígrafo.

Ejemplo:

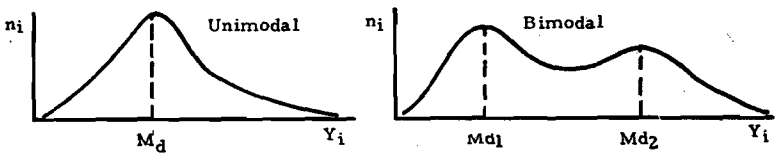
Ingresos de profesionales $Y'_{i-1} - Y'_i$	*Profesionales* n_i
0 - 20	25
20 - 40	45
40 - 60	80
60 - 80	60
80 - 100	40
100 - 120	15
120 y más	5

Inmediatamente se puede adelantar que la moda se encontrará en el tercer intervalo: 40 — 60. Aplicando la fórmula

$$M_d = 40 + \frac{20\,(60)}{45 + 60} = 40 + \frac{1\,200}{105} = 51.43$$

Si se supone una gran cantidad de intervalos pequeños para una cierta distribución, gráficamente el valor modal estaría así representado:

GRÁFICA 8



Este estadígrafo, al igual que la mediana, puede determinarse para variables cualitativas, ya que basta con encontrar la frecuencia máxima. El siguiente ejemplo ilustra esta posibilidad:

Color de automóviles preferido por los clientes	*Clientes encuestados*
Blanco	18
Azul	22
Verde	40
Amarillo	25
Rojo	75

El "valor" modal, en este caso, es el rojo, ya que por tener la máxima frecuencia es el preferido por los clientes.

Al iniciar el estudio de este estadígrafo, se decía que inexplicablemente era poco utilizado en los análisis. Es evidente que requiere, para su cómputo, más información que la media aritmética; ello podría explicar en forma parcial su poco uso, pero aun cuando se dispone de información muchas veces se cree suficiente calcular por ejemplo un ingreso por habitante y quedarse con un análisis parcial. Sin duda, para caracterizar adecuadamente una distribución, se requiere una serie de estadígrafos cuyas indicaciones se complementen. Por ejemplo, saber que dos países tienen ingresos por habitante de 150 y 200 dólares al año puede permitir cierto tipo de conclusiones, pero saber además que los valores modales de los ingresos anuales por habitante son de 140 y 130 dólares respectivamente permite obtener conclusiones bastante más objetivas. Naturalmente, son necesarios muchos otros antecedentes e indicadores, que se presentarán a continuación, para realizar análisis más completos. Interesa destacar la necesidad de buscar un conjunto de indicadores que haga posible el análisis de las distintas facetas de un fenómeno sometido a estudio.

4] *Media geométrica*

Este estadígrafo se define como la raíz de orden n del producto de los n valores de la variable.

Cuando los datos no están agrupados, su fórmula de cálculo es

$$M_g = \sqrt[n]{X_1 \; X_2 \; X_3 \; \ldots \; X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

Para fines prácticos es preferible calcular el logaritmo de la media geométrica y luego el antilogaritmo de ésta:

$$\log M_g = \frac{1}{n} \sum_{i=1}^{n} \log X_i$$

Si los datos aparecen grupados, es decir, si las marcas de clase tienen frecuencias superiores a la unidad, se tendrá la siguiente fórmula.

$$M_g = \sqrt[n]{Y_1 \ldots Y_1 \; Y_2 \ldots Y_2 \; Y_m \ldots Y_m}^{n_1 \qquad n_2 \qquad n_m}$$

$$= \sqrt[n]{Y_1^{n_1} Y_2^{n_2} \cdots Y_m^{n_m}}$$

$$M_g = \sqrt[n]{\prod_{i=1}^{m} Y_1^{n_i}}$$

$$\log M_g = \frac{1}{n} \sum_{i=1}^{m} (\log Y_i) \, n_i$$

El estadígrafo que se estudia, aparte del inconveniente que significa el engorro de su cálculo, está además limitado porque los valores de la variable deben ser positivos para que pueda ser interpretado. Si algún valor de la variable es cero, la media geométrica será cero; igualmente si aparece algún valor negativo el estadígrafo toma un valor imaginario. Pese a estos inconvenientes, para cierto tipo de variables, en especial las cronológicas, que sigan una tendencia exponencial, se hace indispensable su uso si se desea calcular valores intermedios, es decir, si se desea interpolar no linealmente. Por ejemplo, si cierta población en 1940 era de 2.5 millones y en 1960 alcanza a 4 millones, para calcular la población en 1955 sería indispensable el empleo de la media geométrica, si se admite el crecimiento exponencial de la población a una tasa constante.

$$P_{1950} = \sqrt{(2.5) \quad (4.0)} \doteq 3.15 \text{ millones}$$

$$P_{1955} = \sqrt{(3.15) \quad (4.0)} \doteq 3.55 \text{ millones}$$

5] *Media armónica*

El último de los estadígrafos de tendencia central que aquí se abordará se define como el recíproco de la media aritmética de los valores recíprocos de la variable.

Para datos no agrupados:

$$M_h = \frac{1}{\frac{\Sigma \frac{1}{x_i}}{n}} = \frac{n}{\Sigma \frac{1}{x_i}}$$

Para datos agrupados:

$$M_h = \frac{n}{\Sigma \frac{n_i}{Y_i}}$$

Ejemplo: Un grupo de trabajadores construyen los primeros 120 metros de una avenida con una productividad de 12 metros diarios, en cambio los siguientes 120 metros lo hacen a razón de 18 metros por día. Se trata de determinar la productividad diaria durante todo el trabajo.

Si se decidiera calcular la media aritmética se tendría:

$$\bar{Y} = \frac{12 + 18}{2} = 15 \text{ metros diarios}$$

Por otra parte los primeros 120 metros requieren 10 días y los siguientes 120 metros 6.67 días; es decir, todo el trabajo lo harían en 16.67 días. Si la productividad diaria es de 15 metros, en los 16.67 días construirían un total de 250.05 metros, lo que es inconsistente, ya que el trabajo total es de sólo 240 metros. Si en cambio se utiliza la media armónica.

$$M_h = \frac{2}{\frac{1}{12} + \frac{1}{18}} = \frac{72}{3 + 2} = \frac{72}{5} = 14.4 \text{ metros}$$

Trabajando con una productividad media de 14.4 metros por día, en 16.67 días se construirán 240 metros.

Como pudo advertirse, la media armónica se aplica cuando se presenta una relación inversa entre las variables implícitas; en el caso del ejemplo la relación inversa aparece entre la productividad y el tiempo:

e: espacio

p: productividad $\qquad e = p \times t \qquad p = e \cdot \frac{1}{t}$

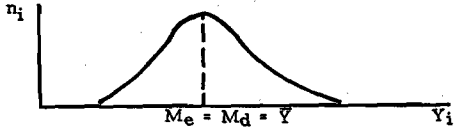
t: tiempo

C. EVALUACIÓN DE LOS ESTADÍGRAFOS DE TENDENCIA CENTRAL

En más de una oportunidad se insistió sobre la necesidad de disponer de un conjunto de indicadores sobre la variable que se está estudiando. Los indicadores presentados, que en general se denominan de tendencia central, tienen definiciones precisas; por ello muestran aspectos particulares del fenómeno que se estudia. Se trata de un conjunto de estadígrafos complementarios; las conclusiones a que en último término den lugar, deberán ser producto de la consideración simultánea de los valores que alcanzan dichos indicadores.

Al analizar la bondad de cada uno de estos indicadores, es preciso tener presente el volumen de observaciones tomadas en cuenta para su cálculo y las limitaciones de cada uno de ellos. Siempre es conveniente complementar el análisis de las cifras, con una representación gráfica de la distribución de frecuencias de la variable; le interesa destacar la posición relativa de la media aritmética, la mediana y el valor modal. La posición relativa de estos estadígrafos depende de la forma de la distribución; de esta suerte si la distribución es simétrica, es decir, si se observa perfecta simetría respecto de un eje central, los tres estadígrafos coinciden.

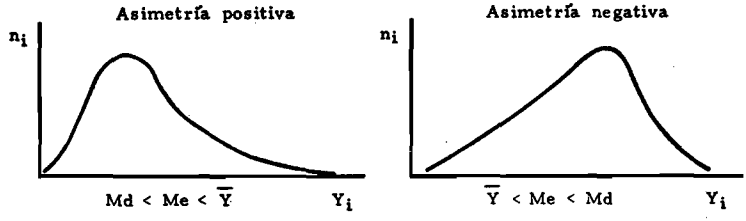
GRÁFICA 9



En el caso de distribuciones no simétricas, la posición relativa de los estadígrafos depende del tipo de asimetría. De esta manera, si la asimetría es positiva, es decir, si la distribución tiene su rama o manto más extendido hacia valores positivos de la variable, la moda será menor que la media aritmética. La mediana, por el hecho de dividir la masa de observaciones en dos partes, que-

dará comprendida entre ambas. Si la asimetría es negativa, es decir, cuando la distribución se extienda suavemente hacia valores negativos de la variable, la moda superará a la media aritmética, permaneciendo la mediana, y por la misma razón dada en el otro caso, comprendida entre ambos indicadores. Gráficamente

GRÁFICA 10

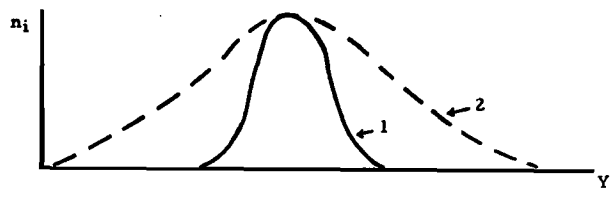


Recuérdese que la media aritmética es un estadígrafo muy sensible a valores extremos de la variable, de allí que en un caso sea el mayor de los tres estadígrafos y en otro caso el menor de ellos. La moda, como el valor de la variable que más se repite, tiene en general una clara ubicación. Habría que agregar que su valor depende sobremanera de la amplitud de intervalo elegida y su representatividad sólo se garantiza cuando existe una clara concentración de frecuencias en un intervalo dado.

D. ESTADÍGRAFOS DE DISPERSIÓN

Una vez caracterizada la distribución a través de estadígrafos de tendencia central y conociendo el tipo de asimetría, interesa tener indicaciones acerca del grado de heterogeneidad con que la variable se distribuye en un conjunto de observaciones. Dos distribuciones pueden tener iguales estadígrafos de tendencia central, sin embargo pueden mostrar grados de dispersión diferentes, como puede observarse en la gráfica que a continuación se muestra.

GRÁFICA 11



Evidentemente en la primera distribución (línea continua) los valores aparecen más concentrados en torno al eje central, en anto que en la otra aparecen mucho más dispersos. Si ambas listribuciones representaran ingresos de dos poblaciones, se conluiría que en la primera distribución los ingresos son más homoéneos, mientras que en la segunda se observaría gran disparidad ntre ingresos altos, medios y bajos.

Parecería innecesario destacar la importancia que tiene contar on indicadores que pudieran mostrar este tipo de características n una distribución; sobre todo en lo que se refiere a distribuión de ingresos, ahora de tanta actualidad, es indispensable conar con indicaciones adecuadas en este sentido.

] *Recorrido de la variable*

Cuando se aborda el problema de la dispersión, lo primero que e piensa es el campo de recorrido de la variable: la diferencia ntre el mayor y menor valor de ella. Si bien brinda una primera dea acerca de la heterogeneidad, tiene el inconveniente que sólo oma en cuenta los dos valores extremos, descuidando el conjuno de valores intermedios. Puede suceder que uno de los valores xtremos esté accidentalmente desplazado y no constituya por anto un valor representativo; en este caso el recorrido sería exaerado y la dispersión aparecería distorsionada. Para iniciar el nálisis es conveniente considerar el recorrido, pero en ningún aso es suficiente.

Para variable discreta $R = Y_m - Y_0$

Para variable continua $R = Y'_m - Y'_0$

] *Recorrido intercuartílico*

Como una manera de subsanar el inconveniente de los valores xtremos que presentaba el estadígrafo anterior, se define un uevo indicador, que toma en cuenta el recorrido entre el prier y tercer cuartil.

$$D_q = Q_3 - Q_1$$

Si bien es cierto que este indicador representa un adelanto especto del anterior, no lo es menos que siempre toma dos vaores de la variable, dejando de lado el resto, y en consecuencia a influencia de valores extremos puede, aunque en menor me-

dida, originar algún tipo de deformación en cuanto al grado de dispersión.

3] *Varianza*

Se define este estadígrafo en virtud de la propiedad de la media aritmética que minimiza la suma de las desviaciones al cuadrado. Se simbolizará por σ^2 ó $V[Y_i]$.

a) Para datos no agrupados

$$V[X_i] = \sigma^2 = \frac{\Sigma (X_i - \bar{X})^2}{n} = \frac{\Sigma X_i^2}{n} - \bar{X}^2$$

b) Para datos agrupados

$$V[Y_i] = \sigma^2 = \frac{\Sigma (Y_i - \bar{Y})^2 n_i}{n} = \frac{\Sigma Y_i^2 n_i}{n} - \bar{Y}^2$$

Si bien la varianza no tiene un fin *per se* sino que se utiliza en materias que se presentarán posteriormente, da origen a un estadígrafo que sí tiene utilidad e interpretación práctica. Se trata de la desviación típica o estándar que se define como la raíz cuadrada positiva de la varianza.

$$\sigma = +\sqrt{\sigma^2}$$

Mientras más dispersa sea la variable, mayor será la magnitud de la desviación típica puesto que mayores serán los desvíos respecto de la media aritmética, sin posibilidad de compensación de desvíos por tratarse de suma de cuadrados. Este estadígrafo se expresa en las mismas unidades de la variable estudiada, en tanto que la varianza se expresa en el cuadrado de la unidad de medida.

Como puede observarse en la fórmula, este indicador de dispersión toma en cuenta todos los valores de la variable con sus correspondientes frecuencias o pesos relativos, sin embargo, siempre es sensible a valores extremos. Por ello es conveniente, antes de calcular los estadígrafos, hacer un análisis previo de la tabla de distribución de frecuencias, para percibir la representatividad de valores extremos y sus posibles efectos sobre los valores de los estadígrafos.

c) Propiedades de la varianza

i) Primera propiedad: La varianza de una variable a la cual se le suma (resta) una constante, es igual a la varianza de la variable original.

$$V[Y_i \pm K] = V[Y_i]$$

Aplicando la definición de varianza a la variable $Y_i + K$, se tiene:

$$\frac{\sum_{i=1}^{m} (Y_i \pm K - M[Y_i \pm K])^2}{n} n_i = V[Y_i]$$

pero se vio que

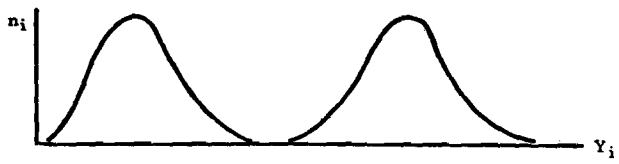
$$M[Y_i + K] = K + M[Y_i]$$

$$\frac{\sum_{i=1}^{m} (Y_i \pm K - M[Y_i] \pm K)^2}{n} n_i = V[Y_i]$$

$$\frac{\sum_{i=1}^{m} (Y_i - \bar{Y})^2 n_i}{n} = V[Y_i]$$

Gráficamente, las dos distribuciones tienen la misma varianza, pese a estar desplazadas en el eje de las abscisas.

GRÁFICA 12



ii) Segunda propiedad: La varianza del producto de una constante por una variable, es igual al cuadrado de la constante por la varianza de la variable.

$$V[K\ Y_i] = K^2\ V[Y_i]$$

$$\frac{\sum_{i=1}^{m} (K\ Y_i - K\ \bar{Y})^2\ n_i}{n} = K^2\ V[Y_i]$$

$$\frac{\sum_{i=1}^{m} [K^2\ (Y_i - \bar{Y})^2]\ n_i}{n} = K^2\ V[Y_i]$$

$$\frac{K^2 \sum_{i=1}^{m} (Y_i - \bar{Y})^2\ n_i}{n} = K^2\ V[Y_i]$$

$$K^2\ V[Y_i] = K^2\ V[Y_i]$$

d) Componentes de la varianza. En el caso que un conjunto de datos haya sido dividido previamente en grandes categorías o estratos, es posible desglosar la varianza en dos componentes muy útiles para el análisis. Admítase que una masa de datos ha sido dividida en L estratos; cada estrato tendrá una media aritmética, una varianza y un número de observaciones que expresa la importancia de cada uno de estos estratos. En este caso la variabilidad total puede deberse tanto a variabilidad dentro de cada estrato como a variabilidad entre los diferentes estratos.

i) Intervarianza: Estadígrafo que representa la variabilidad entre los estratos; se define como la varianza entre las medias de los estratos.

$$\sigma_b^2 = V[\bar{Y}_h] = \frac{\sum_{h=1}^{L} (\bar{Y}_h - \bar{Y})^2\ n_h}{n}$$

donde

$\bar{Y}_h$ es la media aritmética del estrato h

$\bar{Y}$ es la media aritmética general

n_h es el número de observaciones o tamaño de cada estrato.

ii) Intravarianza: Estadígrafo que representa la variabilidad dentro de los estratos; se define como el promedio de las varianzas de los estratos.

$$\sigma_w^2 = M[\sigma_h^2] = \frac{\sum_{h=1}^{L} \sigma_h^2 \, n_h}{n}$$

donde σ_h^2 es la varianza del estrato h, y n_h y n obedecen a las mismas definiciones del caso anterior.

Dado que los dos estadígrafos estudiados son partes componentes de la varianza, a continuación se presenta la correspondiente demostración.

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

$$\frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi} - \bar{Y})^2}{n} = \frac{\sum_{h=1}^{L} (\bar{Y}_h - \bar{Y})^2 \, n_h}{n} + \frac{\sum_{h=1}^{L} \sigma_h^2 \, n_h}{n}$$

pero

$$\sigma_h^2 = \frac{\Sigma (Y_{hi} - \bar{Y}_h)^2}{n_h}$$

remplazando

$$\sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi} - \bar{Y})^2 = \sum_{h=1}^{L} (\bar{Y}_h - \bar{Y})^2 \, n_h + \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{(Y_{hi} - \bar{Y}_h)^2 \, n_h}{n_h}$$

Elevando al cuadrado los correspondientes binomios

$$\sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi}^2 - 2\bar{Y} \, Y_{hi} + \bar{Y}^2) = \sum_{h=1}^{L} (\bar{Y}_h^2 - 2\bar{Y} \, \bar{Y}_h + \bar{Y}^2) \, n_h +$$

$$+ \sum_{h=1}^{L} \sum_{i=1}^{n_h} (Y_{hi}^2 - 2\bar{Y}_h \, Y_{hi} + \bar{Y}_h^2)$$

Aplicando las propiedades de la sumatoria

$$\sum_{h=1}^{L} [\sum_{i=1}^{n_h} (Y_{hi}^2) - 2\bar{Y}\, n_h\, \bar{Y}_h + n_h\, \bar{Y}^2)] = \sum_{h=1}^{L} \bar{Y}_h^2\, n_h -$$

$$- n\, \bar{Y}^2 + \sum_{h=1}^{L} \sum_{i=1}^{n_h} [(Y_{hi}^2) - n_h\, \bar{Y}_n^2]$$

$$\sum_{h=1}^{L} \sum_{i=1}^{n_h} Y_{hi}^2 - n\, \bar{Y}^2 = \sum_{h=1}^{L} \bar{Y}_h^2\, n_h - n\, \bar{Y}^2 + \sum_{h=1}^{L} \sum_{i=1}^{n_h} Y_{hi}^2 - \sum_{h=1}^{L} n_h\, \bar{Y}_h^2$$

Simplificando términos queda

$$\sum_{h=1}^{L} \sum_{i=1}^{n_h} Y_{hi}^2 = \sum_{h=1}^{L} \sum_{t=1}^{n_h} Y_{hi}^2$$

Luego

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

Ejemplo: Piénsese en los sueldos y salarios pagados por una fábrica, que tienen una varianza de 137 600. Si las observaciones se clasifican por estratos: obreros, empleados administrativos y técnicos, será posible analizar más a fondo la distribución de ingresos.

Las informaciones disponibles serían:

Estratos (h)	*Tamaño estrato* n_h	*Media por estrato* $\bar{Y}_h$	*Varianza por estrato* σ_h^2
Obreros	300	400	160 000
Empleados administrativos	100	400	160 000
Técnicos	100	500	40 000

La media aritmética general será:

$$\bar{Y} = \frac{\sum_{h=1}^{L} \bar{Y}_h\, n_h}{n} = \frac{2\,100}{5} = 420$$

$$\sigma_w^2 = \frac{\sum_{h=1}^{L} \sigma_h^2 \, n_h}{n} = \frac{680\,000}{5} = 136\,000$$

$$\sigma_b^2 = \frac{\sum_{h=1}^{L} (\bar{Y}_n - \bar{Y})^2 \, n_h}{n} = \frac{8\,000}{5} = 1\,600$$

El cálculo de los anteriores estadígrafos permite concluir que la variabilidad se debe principalmente a heterogeneidad en las remuneraciones dentro de los estratos y no así a diferencias entre estratos; en otros términos las remuneraciones promedio de cada estrato, son bastante homogéneas ya que la intervarianza es pequeña, mientras que las remuneraciones dentro de cada estrato son muy heterogéneas puesto que la intravarianza es bastante grande. (Ver Anexo.)

e) Métodos abreviados de cálculo.

1) Primer método abreviado. Se trata de encontrar una fórmula que reduzca el volumen de operaciones; como en el caso de la media aritmética, la variable se expresará en términos de desvíos respecto de un origen de trabajo. Recuérdese que:

$$\sigma^2 = \frac{\sum_{i=1}^{m} (Y_i - \bar{Y})^2 \, n_i}{n} = \frac{\sum_{i=1}^{m} Z_i^2 \, n_i}{n}$$

ya que

$$Z_i = Y_i - \bar{Y}$$

Si se desarrolla el cuadrado del binomio dentro de la sumatoria, se tiene:

$$\sigma^2 = \frac{\sum_{i=1}^{m} Y_i^2 \, n_i - 2\bar{Y} \sum_{i=1}^{m} Y_i \, n_i + n\bar{Y}^2}{n}$$

pero

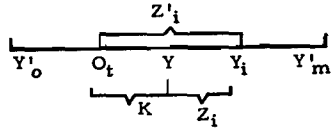
$$\sum_{i=1}^{m} Y_i \, n_i = n\bar{Y}$$

luego,

$$\sigma^2 = \frac{\sum_{i=1}^{m} Y_i^2 \; n_i}{n} - \bar{Y}^2$$

Observando la siguiente gráfica, resulta inmediata la deducción de la fórmula abreviada:

GRÁFICA 13



$$Z'_i = Z_i + K$$

Elevando al cuadrado

$$Z'^2_i = Z_i^2 + 2K\, Z_i + K^2$$

Ponderando por n_i

$$Z'^2_i \; n_i = Z_i^2 \; n_i + 2K\, Z_i \; n_i + K^2 \; n_i$$

Aplicando sumatoria

$$\sum_{i=1}^{m} Z'^2_i \; n_i = \sum_{i=1}^{m} Z_i^2 \; n_i + 2K \sum_{i=1}^{m} Z_i \; n_i + n\, K^2$$

Pero la primera propiedad de la media aritmética decía:

$$\sum_{i=1}^{m} Z_i \; n_i = 0 \quad ,$$

luego

$$\sum_{i=1}^{m} Z'^2_i \; n_i = \sum_{i=1}^{m} Z_i^2 \; n_i + n\, K^2$$

pero

$$\sum_{i=1}^{m} Z_i^2 \; n_i = n\, \sigma^2$$

y

$nK = \sum_{i=1}^{m} Z'_i \; n_i$ (en el cálculo abreviado de la media aritmética).

Luego

$$\sigma^2 = \frac{\sum_{i=1}^{m} Z'^2_i \; n_i}{n} - \left(\frac{\sum_{i=1}^{m} Z'_i \; n_i}{n} \right)^2$$

ii) Segundo método abreviado. Consiste, como se recordará, en expresar las desviaciones en términos de unidades de intervalo (dividiendo por la amplitud del intervalo).

$$\frac{Y_i - O_t}{c} = \frac{Z'_i}{c} = Z''_i$$

es decir que $Z'_i = c \; Z''_i$ y remplazando en la fórmula anterior se obtiene:

$$\sigma^2 = c^2 \left\{ \frac{\sum_{i=1}^{m} Z''^{\,2}_i \; n_i}{n} - \left(\frac{\Sigma \, Z''_i \; n_i}{n} \right)^2 \right\}$$

En el siguiente ejemplo se podrá comprobar el ahorro de tiempo que significa la aplicación de estas fórmulas.

En el caso del primer método abreviado:

Impuestos por contribuyente			*Número de contribuyentes*			
Y'_{i-1}	Y'_i	Y_i	n_i	Z'_i	$Z'_i \, n_i$	$Z'^{\,2}_i n_i$
0	20	10	20	—40	—800	32 000
20	40	30	15	—20	—300	6 000
40	60	50*	10	0	0	0
60	80	70	8	20	160	3 200
80	100	90	5	40	200	8 000
			58		—740	49 200

* $O_t = 50$.

Remplazando estos valores en la fórmula, se obtiene:

$$\sigma^2 = \frac{49\,200}{58} - \left(\frac{-740}{58}\right) = 848.3 - 162.8 = 685.5$$

En el caso del segundo método abreviado se tiene:

Y'_{i-1}	Y'_i	Y_i	n_i	Z''_i	$Z''_i\, n_i$	$Z''^2_i n_i$
0	20	10	20	−2	−40	80
20	40	30	15	−1	−15	15
40	60	50*	10	0	0	0
60	80	70	8	1	8	8
80	100	90	5	2	10	20
			58		−37	123

* $O_t = 50$.

Aplicando la fórmula respectiva:

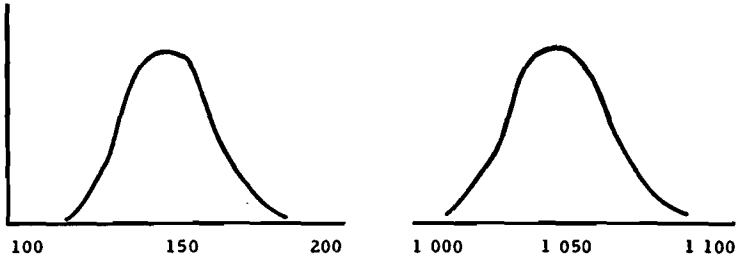
$$\sigma^2 = 20^2 \left\{\frac{123}{58} - \left(\frac{-37}{58}\right)^2\right\} = 400\ \{2.12 - 0.407\} = 685.5$$

4] *Coeficiente de variabilidad*

Tanto la varianza como la desviación típica tienen el inconveniente de los estadígrafos absolutos, ya que en el caso de indicaciones sobre dispersión, tiene mucha importancia no tomar en cuenta la posición de la distribución. Estos estadígrafos, sobre todo al comparar distribuciones, pueden deformar las conclusiones.

Obsérvense las dos distribuciones que aparecen a continuación.

GRÁFICA 14



Ambas distribuciones muestran la misma dispersión en torno a la media, es decir, tienen igual varianza y desviación típica; sin embargo, en términos relativos, una distribución donde el menor ingreso es 1 000 y el mayor es 1 100, es mucho más homogénea que otra distribución donde el menor ingreso es 100 y el mayor 200. En un caso la diferencia entre el mayor y menor ingreso es 10%, mientras que en el otro es de 100%.

Surge, por consiguiente, la necesidad de disponer de un estadígrafo que tome en cuenta la tendencia central de la distribución. Se define así el coeficiente de variabilidad, como la razón entre la desviación típica y la media aritmética.

$$CV = \frac{\sigma}{\bar{Y}}$$

En el ejemplo anterior, si ambas distribuciones tuvieran, por ejemplo, una desviación típica de 60, los coeficientes de variabilidad serían:

$$CV_1 = \frac{\sigma 1}{\bar{Y}_1} = \frac{60}{150} = 0.4 = 40\%$$

$$CV_2 = \frac{\sigma 2}{\bar{Y}_2} = \frac{60}{1\,050} = 0.057 = 5.7\%$$

Estos estadígrafos permiten llegar a conclusiones más realistas y ciertas.

El coeficiente de variabilidad, o desviación típica relativa como también se le llama, puede tomar valores tan grandes como se quiera, ya que no hay una relación de dependencia limitante entre σ y $\bar{Y}$. Por otra parte, en el caso de una distribución donde la media aritmética fuera negativa no tiene sentido considerar el signo para calificar la dispersión. Por ello este estadígrafo podría definirse como el valor absoluto del cociente entre la desviación típica y la media aritmética.

Puesto que las propiedades de la media aritmética y la desviación típica ya fueron analizadas, las propiedades del coeficiente de variabilidad serán el resultado de las propiedades de los indicadores componentes.

Si bien es cierto que para calificar la dispersión de una distri-

bución es más apropiado el coeficiente de variabilidad, de esto no debe deducirse que la varianza y la desviación típica carecen de utilidad; por el contrario, son muy útiles en el tratamiento de materias que se estudiarán posteriormente.

E. UTILIZACIÓN DE INDICADORES DE LA PROGRAMACIÓN

Con mucha frecuencia se escuchan quejas respecto de la escasez de informaciones estadísticas básicas para los países en vías de desarrollo. Admitida la deficiencia, es conveniente también reconocer que no siempre se hace un uso óptimo de la escasa información disponible. Si se reconoce como etapa primaria durante un proceso de planificación la posibilidad de realizar diagnósticos, será necesario destacar la enorme utilidad de la estadística descriptiva en lo que se refiere a caracterización de fenómenos en un momento dado. Por una parte, el diagnóstico requiere disponer de una perspectiva histórica sobre las variables estratégicas; por otra, esa misma perspectiva histórica debe complementarse con análisis cuantitativos en profundidad durante períodos que presenten cambios de orientación y/o ritmo en las tendencias observadas, aparte de una cuantificación detallada para el momento "cero" de un plan. Es para esos puntos para lo que el instrumental de estadística descriptiva debe ser puesto a disposición del analista. Se ha insistido sobre la urgente necesidad de contar con *un juego de indicadores* para las principales variables; cada indicador mostrará una faceta de la variable estudiada, y un conjunto de ellos permitirá una adecuada e integral calificación de las variables que interesan para el diagnóstico.

En general, un conjunto de indicadores permite realizar análisis de consistencia; de ese modo puede llegarse a una primera evaluación acerca de la calidad y fidelidad de la información que se pretende utilizar.

TEMAS DE DISCUSIÓN

Indique si las siguientes afirmaciones son verdaderas o falsas y justifique su opinión.

1] En toda distribución con media nula el percentil 38 es igual, en valor absoluto al percentil 62.

2] Las siguientes fórmulas dan resultados exactamente iguales para la media aritmética.

$$\frac{\sum_{i=1}^{n} X_i}{n}, \quad \frac{\sum_{i=1}^{m} Y_i n_i}{n}, \quad O_t + c \frac{\Sigma Z_i'' n_i}{n}$$

3] La media armónica de una constante es igual al reciproco de la constante.

4] Los siguientes datos son consistentes.

$$h_2 = 0.4 \qquad h_1 = 0.2 \qquad H_3 = 0.8 \qquad n = 50$$

$$n_4 = 5 \qquad \bar{Y} = 25 \qquad M_d = 30$$

5] Para que la intervarianza fuese igual a la varianza, las medias de los estratos deberían ser iguales.

6] En una distribución asimétrica, el valor de la mediana coincide con el del segundo cuartil.

7] En una distribución normal tipificada el sexto percentil es igual a -0.15.

8] Los siguientes datos son consistentes:

$$\sigma_b^2 = 780$$

$$\bar{Y} = 150$$

$$CV = 20\%$$

$$\sigma_w^2 = 110$$

9] En una distribución cualquiera, se debe verificar que:

$$\bar{Y} + \sigma \leqslant Y'_m$$

10] Los siguientes datos son consistentes:

$$m = 6 \qquad h_1 = 0.2 \qquad h_4 = 0.2$$

$$H_2 = 0.6 \qquad H_3 + H_4 = 1.9$$

11] El número de intervalos en que se clasifica una masa de datos, depende de la cantidad de éstos.

12] La moda es un mejor indicador que la media aritmética, porque es poco sensible a valores extremos.

13] En una distribución normal, donde la mediana es 8 y la media cuadrática 10, el coeficiente de variabilidad será inferior al 50%.

14] Los siguientes datos son consistentes:

$$M_g = 0 \qquad M_d = -10$$

15] La mediana de una distribución es 50. Si se multiplican los valores de la variable por 1.6, el valor de la mediana será 80.

16] Los siguientes datos son consistentes para una población dividida en dos estratos de igual tamaño.

$$\bar{Y}_1 = 10 \qquad \bar{Y}_2 = 10$$

$$\sigma_1 = 8 \qquad \sigma_2 = 10$$

$\sigma^2 = 81$ (varianza para toda la población)

17] Si las edades de los alumnos siguen una distribución normal con media 28 y desviación típica 2, la proporción de alumnos menores a 23 años es del orden de 10 por ciento.

18] En el sector servicios el sueldo promedio es de 200 u. m. Si los varones constituyen el 70 por ciento de la población remunerada, es factible que su ingreso promedio mensual sea de 300 u. m.

19] Dada una población con una cierta varianza, la magnitud de la intervarianza y de la intravarianza depende del criterio de estratificación.

20] El valor modal está condicionado por el número de intervalos en que se tabule una masa de datos.

21] Los siguientes datos son consistentes en una distribución simétrica.

Recorrido intercuartílico:	80
Percentil Nº 80:	200
Media aritmética:	160

22] La media geométrica de los valores de una variable multiplicados por una constante, es igual a la media geométrica de la variable original.

23] Los siguientes datos son consistentes:

$$m = 5 \qquad \sum_{i=1}^{5} Y_i^2 \, n_i = 400 \qquad \bar{Y} = 5$$

$$h_5 = 0.1 \qquad H_3 = 0.5 \qquad n_4 = 8$$

24] El valor modal carece de adecuada representatividad, cuando las frecuencias de los intervalos contiguos al que contiene a este estadígrafo, representan fracciones pequeñas del total de observaciones.

25] Si la varianza de los ingresos de los obreros de una industria es 400, y la intravarianza y la intervarianza son iguales, estos componentes no alteran su participación relativa dentro de la varianza total si se reajusta a todos los obreros en un 20%.

26] La desviación típica de las utilidades reinvertidas de las sociedades anónimas de cierto país, es de 25 000 unidades monetarias. Si solamente se consideran las sociedades anónimas industriales, su desviación típica no podrá exceder de 25 000 unidades monetarias.

27] Al comparar las regiones A y B, se tiene que las desviaciones típicas de los ingresos familiares son de 600 y 450 unidades monetarias respectivamente. Se puede concluir en consecuencia, que en la región B el ingreso está más uniformemente distribuido.

PROBLEMAS PROPUESTOS

1] Elabore una distribución de frecuencias de 7 intervalos, de modo que el coeficiente de variabilidad sea de 80%.

2] El sector asalariado de una región ha sido dividido en dos estratos: obreros y empleados. Se trataba de analizar los efectos de una política de redistribución de ingresos. Los datos disponibles, antes y después de aplicar tal política, fueron los siguientes:

	Antes	*Después*
Coeficiente de variabilidad general	60%	50%
$\bar{Y}$ (obreros)	15	?
$\bar{Y}$ (empleados)	?	30
Proporción de obreros	0.6	0.6
σ_b^2	150	24
σ^2	225	144

Se le pide enumerar qué conclusiones le merece la política de redistribución tanto a nivel general como a nivel de estrato. Justifique sus opiniones, deduciendo aquellos estadígrafos que le parezcan pertinentes.

3] Compare el grado de heterogeneidad en los salarios de los obreros de la construcción en dos países.

Venezuela	*Argentina*
σ = 650 bolívares	σ = 50 000 pesos
Salario promedio anual: 1000 dólares	Salario promedio anual: 780 dólares
Tipo de cambio: 5 bolívares por dólar	Tipo de cambio: 350 pesos por dólar

¿Qué podría concluir en cuanto a la distribución de ingresos en ambos países?

4] En cierta comunidad el impuesto total recaudado a 100 000 contribuyentes durante el año 1966 fue de 13 millones de unidades monetarias. El coeficiente de variabilidad de los tributos individuales es de 120%, se sabe que hay 60 000 contribuyentes obreros que cancelaron en conjunto 3 millones de unidades monetarias. Calcule la intravarianza de esa población estratificada en obreros y no obreros, y comente brevemente la homogeneidad de la distribución de los tributos.

5] El sector agrícola de un país se divide en dos estratos: cultivos intensivos y cultivos extensivos; respecto de los ingresos de los trabajadores del sector, se tienen los siguientes indicadores:

$$\bar{Y} = 100 \text{ u. m.} \qquad CV = 50\% \qquad \sigma_b^2 = 400$$

Se decide reajustar los ingresos de los trabajadores en 20% y darles además una bonificación de 30 u. m. a cada uno de ellos. Analice qué cambios en la distribución del ingreso provoca el reajuste.

EJERCICIOS

1] La inversión real anual de un grupo de industrias pesqueras se detalla a continuación:

Miles de dólares

10	12	8	40	6	8	10	30	2	8	6	14
16	20	25	28	30	26	30	4	6	10	18	17
13	17	21	7	6	8	14	7	15	19	27	22
0	14	6	8	9	11	13	15	18	20	30	60
12	6	5	5	6	8	7	12	15	36	39	52

Se pide:

a) Forme una tabla de distribución de frecuencias, con ocho intervalos de amplitud constante;

b) Calcule las frecuencias relativas;

c) Calcule las frecuencias absolutas y relativas acumuladas;

d) Represente gráficamente la distribución de frecuencias (histograma y polígono de frecuencias);

e) Calcule los siguientes estadígrafos:

i) Media aritmética (datos originales y métodos abreviados);

ii) Mediana;

iii) Moda;

iv) Media armónica.

2] Las distribuciones de ingreso de dos países son las siguientes:

País A		*País B*	
Ingresos anuales por hab.: dólares	*Población remunerada*	*Ingresos anuales por hab.: dólares*	*Población remunerada*
80 — 100	30 000	60 — 90	10 000
100 — 120	80 000	90 — 120	20 000
120 — 140	40 000	120 — 150	50 000
140 — 160	10 000	150 — 180	20 000
160 — 200	4 000	180 — 210	15 000
200 — 220	1 000	210 — 240	10 000
		240 — 270	4 000

a) Fundamente, empleando el cálculo de estadígrafos que crea conveniente, el grado de desarrollo comparado de ambos países. ¿Necesita además otros antecedentes para calificar a estos países con mayor rigor?

b) Calcule el ingreso promedio del 30% de la población de mayores ingresos en cada país. Compare los resultados y discuta sus conclusiones. Realice el mismo cálculo para el 40% de la población de menores ingresos.

c) ¿Cuál es el ingreso promedio de los dos países en conjunto?

d) Suponiendo que el país B devalúa su moneda en 30% y el país A en 5%, calcule los nuevos ingresos promedio en ambos países.

3] El ingreso por habitante de un país es de 310 dólares al año; su sector obrero, que constituye el 59 por ciento de la población, percibe 1/5 del ingreso total. Calcule el ingreso por habitante de este sector.

4] Un automovilista se dirige de Santiago a Talca, ciudades que distan 300 km entre sí; los primeros 200 km los recorre a una velocidad de 120 km/hora y el resto a 80 km/hora. Utilizando un estadígrafo de tendencia central, calcule la velocidad media.

5] La media aritmética entre dos números es 8 y su media geométrica 2. Calcule la media armónica.

6] Si la población de un país es al 31 de diciembre de 1950 de 5 800 000 personas, y al 31 de diciembre de 1960 de 7 200 000, calcule la población en 1955.

7] Si se tiene una distribución de frecuencias simétrica, con 6 intervalos de amplitud constante, y los siguientes datos:

$$n = 150 \qquad Y'_5 = 60 \qquad n_2 = n_1 + 5$$
$$n_3 = 30 \qquad Q_1 = 43.5$$

calcule el 6º decil

8] La siguiente tabla de distribución de frecuencias representa los impuestos personales de un conjunto de profesionales.

Impuestos (u. m.)		*Profesionales*
Y'_{i-1}	Y'_i	n_i *(miles)*
0	20	30
20	40	25
40	60	15
60	80	13
80	100	12
100	120	5

Se le pide que:

a) Calcule la varianza por los tres métodos que conoce.

b) Calcule el coeficiente de variabilidad.

c) Si se cobra un impuesto adicional de 5 u. m. por persona, ¿cuál es la desviación típica?

d) Si se reajustan los impuestos por persona en 20%, ¿cuál es la varianza?

9] El C. V. de los ingresos de 200 empleados de una compañía es 57 por ciento. Después de reajustar, según ley, todos los sueldos en E° 11, este C. V. es ahora de 50 por ciento. Sin embargo, la gerencia fija un sueldo mínimo de E° 71. Antes del reajuste, había 35 personas que tenían un sueldo medio de E° 40 y todos ellos ganaban menos de E° 60; con la nueva política de la gerencia, sus sueldos serán elevados a E° 71. Determine la cantidad de dinero que necesitará mensualmente la compañía, para pagar los sueldos después de hacer efectivos los reajustes.

10] En una distribución de frecuencias se multiplican los valores de variable por 3, y se obtiene una media aritmética de 54; si se suma 5 a los valores de la variable, se obtiene una media cuadrática de 24. Calcule el coeficiente de variabilidad de la distribución.

11] Se clasificó a los trabajadores de un mineral en dos categorías: mayores y menores de 25 años, y se extrajo la siguiente información:

	Número de obreros	*Productividad media*	*Desviación típica*
	n_h	$\bar{Y}_h$	σ_h
Mayores de 25 años	200	40	70
Menores de 25 años	300	60	40

Calcule la varianza de todos los obreros del mineral.

12] La media aritmética y la desviación típica de los porcentajes de reinversión de utilidades de las empresas constructoras son de 40% y 20% respectivamente. Si se supone que esta variable sigue aproximadamente una distribución normal,* se le pide que

a) Calcule cuál es el porcentaje de empresas que reinvierten más del 70%;

b) Calcule el porcentaje de empresas que reinvierten entre 20 y 50%;

c) Calcule cuál es la proporción de empresas que reinvierten cuando mucho 35% de sus utilidades.

13] Se sabe que los consumos familiares en bienes esenciales, están relacionados con los consumos familiares de energía eléctrica mediante la función:

$$\text{C. Es.} = 4.5 \text{ C. En.} - 5$$

Si la media aritmética y la desviación típica de los consumos de energía se estiman en E° 40 y E°15 respectivamente, calcule el coeficiente de variabilidad de los consumos esenciales.

SOLUCIÓN DE EJERCICIOS

1] Dado que el menor valor es 0, el máximo 60, y se especifican 8 intervalos de amplitud constante, se clasifican los datos en la siguiente forma; quedan así resueltas las partes a, b y c.

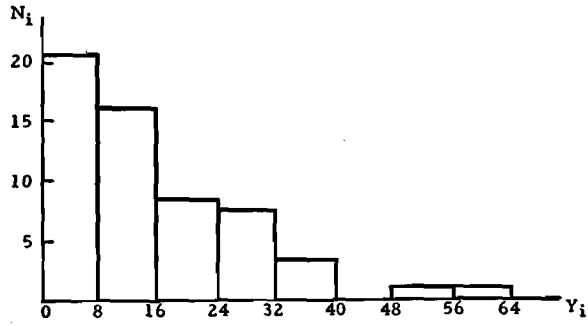
$Y'_{i-1} - Y'_i$	n_i	h_i	$H_i \downarrow$	$N_i \downarrow$	$H_i \uparrow$	$N_i \uparrow$
0 — 8	21	21/60	21/60	21	60/60	60
8.1 — 16	17	17/60	38/60	38	39/60	39
16.1 — 24	9	9/60	47/60	47	22/60	22
24.1 — 32	8	8/60	55/60	55	13/60	13
32.1 — 40	3	3/60	58/60	58	5/60	5
40.1 — 48	0	0	58/60	58	2/60	2
48.1 — 56	1	1/60	59/60	59	2/60	2
56.1 — 64	1	1/60	60/60	60	1/60	1
	60	1				

Obsérvese que la amplitud del primer intervalo, es ligeramente superior a la del resto.

* Véase *infra* el anexo sobre distribución normal.

d) La representación gráfica sería la siguiente:

GRÁFICA 15



e) Los valores de los estadígrafos serían los siguientes:

i) Media aritmética; existen varias alternativas

— datos originales

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{967}{60} = 16.12$$

— Marcas de clase

$$\bar{Y} = \frac{\sum_{i=1}^{m} Y_i\, n_i}{n} = \sum_{i=1}^{m} Y_i\, h_i = \frac{912}{60} = 15.2$$

Y_i	h_i	$Y_i\, h_i$
4	21/60	84/60
12	17/60	204/60
20	9/60	180/60
28	8/60	224/60
36	3/60	108/60
44	0	0
52	1/60	52/60
60	1/60	60/60
		912/60

— Métodos abreviados

Haciendo $O_i = 20$

Y_i	n_i	Z'_i	$Z'_i\ n_i$	Z''_i	$Z''_i\ n_i$
4	21	−16	−336	−2	−42
12	17	−8	−136	−1	−17
20	9	0	0	0	0
28	8	8	64	1	8
36	3	16	48	2	6
44	0	24	0	3	0
52	1	32	32	4	4
60	1	40	40	5	5
	60		−288		−36

Por el primer método abreviado

$$\bar{Y} = O_t + \frac{\Sigma Z'_i\ n_i}{n} = 20 + \left(\frac{-288}{60}\right) = 20 - 4.8 = 15.2$$

Por el segundo método abreviado

$$\bar{Y} = O_t + c\,\frac{\Sigma Z''_i\ n_i}{n} = 20 + 8\left(\frac{-36}{60}\right) = 20 - 4.8 = 15.2$$

ii) Mediana:

$$M_e = Y'_{k-1} + C_k\,\frac{\frac{n}{2} - N_{k-1}}{n_k}$$

$\frac{n}{2} = \frac{60}{2} = 30$. La menor frecuencia acumulada que supera este valor es $N_2 = 38$; luego en el segundo intervalo estará comprendida la mediana. (Véase tabla 1.1.)

$$M_e = 8 + 8\,\frac{30-21}{17} = 8 + \frac{72}{17} = 12.23$$

iii) Moda: Puesto que la frecuencia máxima corresponde al primer intervalo, se presenta un caso particular.

$$M_d = Y'_{k-1} + \frac{C_k\,n_{k+1}}{n_{k-1} + n_{k+1}} = \frac{8 \cdot 17}{17} = 8$$

Nótese que este estadígrafo adquiere el valor del límite superior del primer intervalo, porque no existe intervalo inferior.

iv) Media armónica

Y_i	4.00	12.00	20.00	28.00	36.00	44.00	52.00	60.00
n_i	21.00	17.00	9.00	8.00	3.00	0.00	1.00	1.00
n_i/Y_i	5.25	1.41	0.45	0.28	0.08	0.00	0.02	0.01

$$M_h = \frac{n}{\sum_{i=1}^{m} \frac{n_i}{Y_i}} = \frac{60}{7.50} = 8$$

2] *a)* Un primer indicador es el ingreso promedio (la población se expresa en miles de personas).

País A				*País B*					
$Y'_{i-1} - Y'_i$	Y_i	n_i	$Y_i n_i$	$Y'_{i-1} - Y'_i$	Y_i	n_i	Z''_i	Z''_i	n_i
80 — 100	90	30	2 700	60 — 90	75	10	—2	—20	
100 — 120	110	80	8 800	90 — 120	105	20	—1	—20	
120 — 140	130	40	5 200	120 — 150	135	50	0	0	
140 — 160	150	10	1 500	150 — 180	165	20	1	20	
160 — 200	180	4	720	180 — 210	195	15	2	30	
200 — 220	210	1	210	210 — 240	225	10	3	30	
				240 — 270	255	4	4	16	
		165	19 130			129		56	

$$\bar{Y}_A = \frac{\sum Y_i n_i}{n} = \frac{19\,130}{165} = 115.94$$

$$\bar{Y}_B = O_t + C \frac{\sum Z''_i n_i}{n} = 135 + 30 \frac{56}{129} = 148.02$$

Un primer indicio que el país B tendría mayor grado de desarrollo, está dado por el mayor valor de su ingreso por habitante; sin embargo, es necesario calcular otros estadígrafos y conocer otro tipo de informaciones, para calificar aproximadamente ambas situaciones.

El cálculo de los valores modales ya indica que la diferencia entre ambos países, no parece tan importante como lo indicarían los ingresos promedios.

$$M_{d_A} = Y'_{k-1} + \frac{C_k\, n_{k+1}}{n_{k-1} + n_{k+1}} = 100 + \frac{20 \cdot 40}{70} = 111.4$$

$M_{d_B} = 135$ corresponderá con la marca de clase que tiene la frecuencia máxima, puesto que los intervalos contiguos tienen igual frecuencia. No se puede pretender resumir una situación tan compleja en un par de estadígrafos, pues cada indicador constituye un análisis parcial de aquello que es susceptible de cuantificación.

b) Se trata de acumular las frecuencias hacia arriba, hasta completar el 30% de n. Como las frecuencias no coinciden perfectamente con esta cifra, se hace necesario dividir el intervalo de manera que la frecuencia acumulada $N_i \uparrow$ sea el 30% de n.

País A

$$n = 165$$
$$0.30\ n = 49.5$$

La frecuencia absoluta deberá ser 34.5 para satisfacer las condiciones del problema; 34.5 representa el 86.25% de 40 y en esta proporción se divide el intervalo, quedando

$Y'_{i-1} - Y'_i$	n_i	Y_i	$Y_i\, n_i$
122.75 – 140	34.5	131.38	4 532
140.00 – 160	10.0	150.00	1 500
160.00 – 200	4.0	180.00	720
200.00 – 220	1.0	210.00	210
	49.5		6 962

$$\bar{Y}_A \quad \frac{6\,962}{49.5} = 140.6$$

País B

Siguiendo el mismo proceso se llega a la siguiente tabla.

$Y'_{i-1} - Y'_i$	n_i	Y_i	$Y_i\, n_i$
165.5 – 180	9.7	172.7	1 675.2
180.0 – 210	15.0	195.0	2 925.0
210.0 – 240	10.0	225.0	2 250.0
240.0 – 270	4.0	255.0	1 020.0
	38.7		7 870.2

$$\bar{Y}_B = \frac{7\,870.2}{38.7} = 203.4$$

Para el 40% de la población de ingresos menores, el procedimiento es similar, se llega a las siguientes tablas estimadas.

País A

$Y'_{i-1} - Y'_i$	n_i	Y_i	$Y_i n_i$
80 — 100	30	90.0	2 700
100 — 109	36	104.5	3 762
	66		6 462

$$\bar{Y}_A = \frac{6\,462}{66} = 97.9$$

País B

$Y'_{i-1} - Y'_i$	n_i	Y_i	$Y_i n_i$
60 — 90	10.0	75.0	750.0
90 — 120	20.0	105.0	2 100.0
120 — 133	21.6	126.5	2 732.4
	51.6		5 582.4

$$\bar{Y}_B = \frac{5\,582.4}{51.6} = 108.2$$

c) $\bar{Y}$ (Conjunto) $= \dfrac{n_A \bar{Y}_A + n_B \bar{Y}_B}{n_A + n_B} =$

$$= \frac{(165)\ (115.9) + (129)\ (148.0)}{294} = 130$$

d) **La devaluación implica una tasa de cambio más alta, en moneda del país por dólar. Luego, para el país A:**

$$M\left[\frac{Y_i}{1.30}\right] = \frac{1}{1.30} M\,[Y_i] = \frac{1}{1.30}\ 115.94 = 89.2$$

Para el país B

$$M\left[\frac{Y_i}{1.05}\right] = \frac{1}{1.05} M\,[Y_i] = \frac{1}{1.05} \cdot 148.02 = 141.0$$

3] Se sabe que

$$\frac{\text{Ingreso total}}{\text{Población}} = \text{Ingreso por habitante}$$

$$\frac{Y}{P} = 310$$

Para el sector obrero, el ingreso por habitante será

$$\bar{Y}\ \text{(obrero)} = \frac{1/5\,Y}{0.59\,P}$$

pero $Y = 310 \cdot P$, luego

$$\bar{Y}\ \text{(obrero)} = \frac{1/5\ (310 \cdot P)}{0.59 \cdot P} = \frac{62}{0.59} = 105.08$$

4] El estadígrafo indicado es la media armónica

$$M_h = \frac{n}{\Sigma \frac{n_i}{Y_i}} = \frac{300}{\frac{200}{120} + \frac{100}{80}} = \frac{720}{7} = 103$$

5] Sean x e y los números:

$$\frac{x+y}{2} = 8$$

$$\sqrt{xy} = 2$$

$$M_h = \frac{2}{\frac{1}{x} + \frac{1}{y}} = \frac{2}{\frac{x+y}{xy}} = \frac{2xy}{x+y} = \frac{xy}{\frac{x+y}{2}} = \frac{4}{8} = \frac{1}{2}$$

6] Si se supone un crecimiento exponencial a una tasa constante, es factible aplicar la media geométrica, que dará la población a mitad del período.

$$M_g = \sqrt{5.8 \cdot 7.2} = \sqrt{41.76} = 6.459$$

7] Los datos disponibles, pueden disponerse así en una tabla de frecuencias.

Y'_{i-1}	Y'_i	n_i	N_i	*Estadígrafos*
Y'_0	Y'_1	n_1	N_1	
Y'_1	Y'_2	$n_1 + 5$	N_2	$Q_1 = 43.5$
Y'_2	Y'_3	30	75	
Y'_3	Y'_4	30	N_4	
Y'_4	60	$n_6 + 5$	N_5	
60	Y'_6	$\frac{n_6}{150}$	150	

Recuérdese que la distribución es simétrica y, por lo tanto,

$$n_1 = n_6 \; ; \; n_2 = n_5 \; ; \; n_3 = n_4 = 30$$

Se sabe que

$$n_1 + n_2 + n_3 = 75$$

$$n_1 + n_1 + 5 + 30 = 75$$

$$n_1 = 20 \qquad \text{luego}$$

$$n_2 = 25$$

Se trata de encontrar el valor de la amplitud del intervalo C_k. Para ello

$$Q_1 = Y'_1 + C_k \frac{\frac{n}{4} - N_{k-1}}{n_k}$$

Para determinar en qué intervalo se halla Q_1, debe verse cuál es la menor frecuencia acumulada que supera a $\frac{n}{4} = 37.5$. Se comprueba que es $N_2 = 45$; luego Q_1 estará dentro del segundo intervalo:

$$Q_1 = Y'_1 + C_k \frac{37.5 - 20}{25}$$

Pero

$$Y'_5 = Y'_1 + 4\ C_k = 60$$

Luego

$$Y'_1 = 60 - 4\ C_k$$

y remplazando en la fórmula de Q_1, se tiene:

$$43.5 = 60 - 4\ C_k + C_k \frac{17.5}{25} = 60 - 4\ C_k + 0.7\ C_k$$

$$3.3\ C_k = 60 - 43.5 = 16.5$$

$$C_k = \frac{16.5}{3.3} = 5$$

La distribución quedará así:

Y'_{i-1}	Y'_i	n_i	N_i
35	40	20	20
40	45	25	45
45	50	30	75
50	55	30	105
55	60	25	130
60	65	20	150

Una vez completada la tabla, se calcula inmediatamente el 6º decil.

$$D_6 = Y'_{k-1} + C_k \frac{\frac{6n}{10} - N_{k-1}}{n_k}$$

Para determinar k, será necesario calcular la menor frecuencia acumulada que supere a $\frac{6n}{10} = 90$. Se comprueba que es N_4.

$$D_6 = Y'_3 + 5 \frac{90 - 75}{30} = 50 + \frac{75}{30} = 52.5$$

8] *a)* Una fórmula para el cálculo de la varianza es

$$\sigma^2 = \frac{\sum Y_i^2 \; n_i}{n} - \bar{Y}^2$$

Y_i	n_i	$Y_i \, n_i$	$Y_i^2 \, n_i$
10	30	300	3 000
30	25	750	22 500
50	15	750	37 500
70	13	910	63 700
90	12	1 080	97 200
110	5	550	60 500
		4 340	284 400

$$\bar{Y} \; \frac{4\,340}{100} = 43.4$$

$$\sigma^2 = \frac{284\,400}{100} - (43.4)^2 = 2\,844 - 1\,884 \doteq 960$$

Utilizando métodos abreviados con $O_t = 50$ se tiene

Z'_i	$Z'_i \, n_i$	$Z'^2_i \, n_i$	Z''_i	$Z''_i \, n_i$	$Z''^2_i \, n_i$
−40	−1 200	48 000	−2	−60	120
−20	−500	10 000	−1	−25	25
0	0	0	0	0	0
20	260	5 200	1	13	13
40	480	19 200	2	24	48
60	300	18 000	3	15	45
	−660	100 400		−33	251

$$\sigma^2 = \frac{\sum Z''^2_i \; n_i}{n} - \left(\frac{\sum Z'_i \; n_i}{n} \right)^2 = \frac{100\,400}{100} - 43.6 \doteq 960$$

También

$$\sigma^2 = C^2 \left\{ \frac{\sum Z''^2_i \; n_i}{n} - \left(\frac{\sum Z''_i \; n_i}{n} \right)^2 \right\} = 400 \left\{ \frac{251}{100} - 0.11 \right\} \doteq 960$$

b) El C. V. es

$$\text{C. V.} = \frac{\sigma}{\bar{Y}} = \frac{\sqrt{960}}{43.4} = \frac{31}{43.4} = 71.4\%$$

c) Recuérdese que

$$V\ [Y_i + K] = V\ [Y_i]$$

$$V\ [Y_i + 5] = V\ [Y_i] = 960$$

Luego,

$$\sigma = 31$$

d) Recuérdese que

$$V\ [K\ Y_i] = K^2\ V\ [Y_i]$$

$$V\ [1.2\ Y_i] = 1.44\ V\ [Y_i] = 1.44\ (960) = 1\,382.4$$

9] Los datos son

$$\frac{\sigma}{\bar{Y}} = 0.57 \qquad (1)$$

$$\frac{\sigma}{\bar{Y} + 11} = 0.50 \qquad (2)$$

ya que

$$M\ [K + Y_i] = M\ [Y_i] + K$$

$$V\ [K + Y_i] = V\ [Y_i]$$

Luego,

$\sigma = 0.57\ \bar{Y}$ remplazando en (2)

$$0.57\ \bar{Y} = 0.50\ (\bar{Y} + 11)$$

$$0.07\ \bar{Y} = 5.5$$

$\bar{Y} = 78.6$ (antes del reajuste)

Además, esta media estaba compuesta por dos grupos: 35 personas con ingreso medio de 40 y 165 personas con ingreso de $\bar{Y}_1$, que se obtendrá de

$$\bar{Y} = \frac{35\ (40) + 165\ (\bar{Y}_1)}{200}$$

$$78.6\ (200) = 1\,400 + 165\ \bar{Y}_1$$

$$\bar{Y}_1 = \frac{15\,720 - 1\,400}{165} = \frac{14\,360}{165} \doteq 87$$

Las nuevas medias aritméticas después de los reajustes serán:
El primer grupo de 35 personas tendrá un ingreso promedio de E° 71.
El segundo grupo de 165 personas tendrá un ingreso medio de E° 98 (87 + 11).
La cantidad de dinero necesaria será:

$$\text{C. D.} = 35\ (71) + 165\ (98) = 2\,485 + 16\,170 = 18\,655$$

10] Los datos son:

$$M\ [3\ Y_i] = 54$$

$$M_c\ [Y_i + 5] = 24$$

Luego

$$M\ [3\ Y_i] = 3\ M\ [Y_i] = 54$$

$$M\ [Y_i] = 18$$

$$M_c\ [Y_i] = \left[\frac{\Sigma\, Y_i^2\ n_i}{n}\right]^{1/2}$$

$$M_c\ [Y_i + 5] = \left[\frac{\Sigma\,(Y_i + 5)^2\, n_i}{n}\right]^{1/2} = 24$$

$$\frac{\Sigma\,(Y_i^2\ n_i + 10\, Y_i\, n_i + 25\, n_i)}{n} = 576$$

$$\frac{\Sigma\, Y_i^2\ n_i}{n} + 10\,\frac{\Sigma\, Y_i\, n_i}{n} + 25 = 576$$

$$\frac{\Sigma Y_i^2 \; n_i}{n} + 10 \; (18) = 551$$

$$\frac{\Sigma Y_i^2 \; n_i}{n} = 371$$

Siendo

$$\sigma^2 = \frac{\Sigma Y_i^2 \; n_i}{n} - \bar{Y}^2 = 371 - 324 = 47$$

$$\sigma = 6.86$$

$$\text{C. V.} = \frac{6.86}{18} = 38.11\%$$

Aplicando las propiedades de la varianza, es posible llegar al mismo resultado:

$$V[Y_i] = V[Y_i + 5] = M_c^2 - (\bar{Y} + 5)^2 = 24^2 - 23^2 = 47$$

11] La descomposición de σ^2 en σ_b^2 y σ_w^2 es útil para llegar al valor de la varianza del conjunto

$$\sigma^2 = \frac{\Sigma (\bar{Y}_h - \bar{Y})^2 \, n_h}{n} + \frac{\Sigma \sigma_h^2 \; n_h}{n}$$

Será necesario calcular Y:

$$\bar{Y} = \frac{\Sigma Y_h \, n_h}{n} = \frac{40 \; (200) + 60 \; (300)}{500} = 52$$

$$\sigma^2 = \frac{(40 - 52)^2 \, 200 + (60 - 52)^2 \, 300}{500} + \frac{4\,900 \; (200) + 1\,600 \; (300)}{500}$$

$$\sigma^2 = \frac{288 + 192}{5} + \frac{9\,800 + 4\,800}{5} = \frac{15\,080}{5} = 3\,016$$

12] Los datos son:

$$\bar{Y} = 0.4$$

$$\sigma = 0.2$$

a) Se tendrá que *estandarizar* el valor 0.7:

$$t_0 = \frac{0.7 - 0.4}{0.2} = 1.5$$

Luego en la tabla se ve que

$$P\ (t_i \geq t_0) = P\ (t_i \geq 1.5) = 0.5 - 0.433 = 6.7\%$$

b) Estandarizando ambos valores se tiene

$$t_0 = \frac{0.2 - 0.4}{0.2} = -1 \qquad t_1 = \frac{0.5 - 0.4}{0.2} = 0.5$$

$$P\ (t_0 \leqslant t_i \leqslant t_1) = P\ (-1 \leqslant t_i \leqslant 0.5) = 0.341 + 0.195 = 53.6\%$$

c) Estandarizando se tiene

$$t_0 = \frac{0.35 - 0.4}{0.2} = -0.25$$

$$P\ (t_i \leqslant t_0) = P\ (t_i \leqslant -0.25) = 40.1\%$$

13] Si la función es

$$C.\ Es. = 4.5\ C.\ En. - 5$$

aplicando el operador media se tiene:

$$M[C.\ Es.] = 4.5\ M[C.\ En.] - 5$$

remplazando valores

$$M[C.\ Es.] = 4.5\ [40] - 5 = 175$$

Aplicando el operador varianza:

$$V[C.\ Es.] = (4.5)^2\ V[C.\ En]$$

$$V\ C.\ Es. = 20.25\ [225] = 4\,556$$

El coeficiente de variabilidad será:

$$C.\ V. = \frac{+\sqrt{4\,556}}{175} = \frac{67.5}{175} = 38.6\%$$

ANEXO: DISTRIBUCIÓN NORMAL

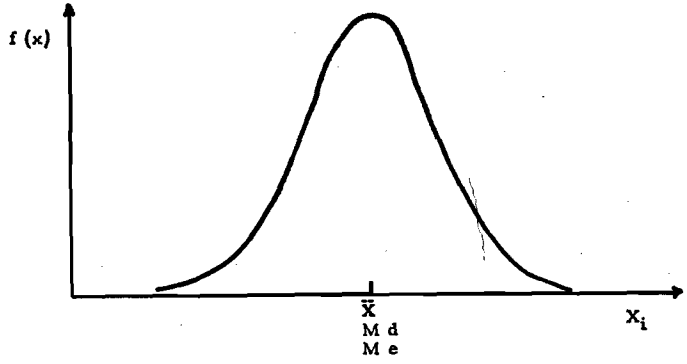
Una de las más utilizadas, dentro de las distribuciones teóricas de variable continua, es la llamada distribución normal o de Gauss. Se trata de una distribución simétrica, unimodal y asintótica al eje de las abscisas. Numerosas variables que se analizan en la investigación socioeconómica corresponden o se aproximan a la distribución normal.

La expresión matemática de esta distribución es la siguiente:

$$f(x) = ce^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}$$

y la representación gráfica de una distribución normal particular es la siguiente:

GRÁFICA 16



Si se desea que el área encerrada por la curva sea igual a 1, bastará con integrar la función entre $-\infty$ y $+\infty$, igualar el resultado a 1 y despejar el valor de c.

$$\int_{-\infty}^{\infty} f(x)\, dx = \sigma\, c\, \sqrt{2\pi} = 1$$

$$c = \frac{1}{\sigma\sqrt{2\pi}}$$

Luego, la expresión matemática de la distribución normal o función de densidad cuando el área que ella encierra es igual a la unidad, es:

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2}$$

La anterior expresión muestra que una distribución normal está completamente determinada cuando se especifica el valor de su media aritmética y de su desviación estándar.

Si se obtiene la primera derivada:

$$f' = -\frac{1}{\sigma^2}(x - \bar{x}) \cdot f(x)$$

igualando a cero, resulta evidente que el único punto máximo se verifica para $x = \bar{x}$. Se trata entonces de una distribución unimodal donde coinciden la mediana, la moda, y la media aritmética.

Si se obtiene la segunda derivada:

$$f'' = -\frac{1}{\sigma^2}\left[1 - \left(\frac{x-\bar{x}}{\sigma}\right)^2\right] f(x)$$

se encuentra dos puntos de inflexión igualando la anterior expresión a 0,

$$x = \bar{x} \pm \sigma$$

Geométricamente la desviación estándar es la distancia que hay entre el eje de simetría y un punto de inflexión.

Toda vez que se desee encontrar el área encerrada por la curva normal entre dos puntos cualesquiera, sería necesario integrar la respectiva función entre los límites deseados; existen tablas estándar donde ya se han realizado los cálculos de las integrales. Es necesario, para hacer uso de dichas tablas, expresar previamente la distribución particular que se tenga, en términos de una variable tipificada, definida de la siguiente manera:

$$t = \frac{x - \bar{x}}{\sigma}$$

donde esta nueva variable tiene media 0 y desviación típica 1. En efecto,

$$M\,[t] = M\left[\frac{x - \bar{x}}{\sigma}\right]$$

aplicando las propiedades de la media aritmética

$$M\,[t] = \frac{1}{\sigma}\,[M\ (x) - \bar{x}] = 0$$

Por otra parte,

$$V\,[t] = V\left[\frac{x - \bar{x}}{\sigma}\right]$$

aplicando propiedades de la varianza

$$V\,[t] = \frac{1}{\sigma^2}\,V\,[x] = \frac{\sigma^2}{\sigma^2} = 1$$

En consecuencia, el área encerrada en una distribución normal tipificada N[0, 1] será

$$\frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{t^2}{2}}\,dt$$

ya que

$$t = \frac{x - \bar{x}}{\sigma}$$

$$dx = \sigma\,dt$$

En cualquier texto de estadística aparecen los valores tabulados de la distribución normal tipificada; una manera muy corriente de tabularla es la siguiente:

Variable tipificada t	*Área comprendida entre 0 y t* $\int_0^t f(t)\,dt$
0.00	0.000
0.05	0.020
0.10	0.040
0.15	0.060
0.20	0.079
0.25	0.099
.	.
.	.
.	.
.	.
.	.
.	.
.	.
0.50	0.195
0.60	0.226
0.80	0.288
0.90	0.316
1.00	0.341
1.20	0.385
1.50	0.433
1.60	0.445
2.00	0.477
2.50	0.494
3.00	0.499
∞	0.500

Es necesario tener cuidado en el uso de estas tablas y cerciorarse de que los límites de la integral correspondan con lo que se desea. Así, hay tablas en las cuales se tabula el valor del área entre $-t$ y $+t$, que corresponderán al doble de área de la tabla mostrada como ejemplo. Se decía que una vez especificados los valores de la media aritmética y la varianza, queda totalmente determinada la función.

Ejemplo: Admítase que las edades de los participantes de un curso siguen aproximadamente una distribución normal, con media 25 años y desviación típica 5. Con estos dos datos será posible calcular la proporción de alumnos en cualquier tramo del eje de las abscisas.

Para encontrar la proporción de alumnos con más de 29

años de edad, previamente se tendría que tipificar este valor. Sea $x_0 = 29$

$$t_0 = \frac{x_0 - x}{\sigma} = \frac{29 - 25}{5} = 0.8$$

Si se observa la tabla de páginas anteriores se tiene que el área entre 0 y 0.8 es de 0.288; luego la proporción de alumnos con más de 29 años será:

$$P\ (x_i > 29) = P\ (t_i > 0.8) = 1 - (0.50 + 0.288) = 0.212$$

Donde P indica proporción o probabilidad. Si se quisiera hallar la proporción de alumnos que tienen edades entre 22 y 26 años, sería necesario tipificar previamente cada uno de estos valores:

Sean $x_0 = 22$ y $x_1 = 26$

$$t_0 = \frac{22 - 25}{5} = -0.6 \qquad t_1 = \frac{26 - 25}{5} = 0.25$$

Recuérdese que la distribución es simétrica y que

$$P\ (t > t_1) = P\ (t < -t_1)$$

O sea que

$$P\ (0 > t > -0.6) = P\ (0 < t < 0.6) = 0.226$$

$$P\ (0 < t < 0.25) = 0.099$$

Sumando las dos áreas se concluye que el 32.5% de los alumnos tienen edades comprendidas entre 22 y 26 años.

III

NÚMEROS ÍNDICE

A. EL PROBLEMA GENERAL

Cuando se define un número índice como un indicador de la tendencia central de un conjunto de elementos que generalmente se expresa como porcentaje, se advierten las limitaciones de todo estadígrafo.

El uso cotidiano que se hace de este instrumento obliga a plantear previamente algunas de sus limitaciones. Es muy útil no olvidar que se trata de un indicador que pretende reflejar el comportamiento de ciertas variables en forma aproximada; en consecuencia no se trata de una medición exacta. Por otra parte es necesario establecer que un número índice plantea una comparación, ya sea en el tiempo o en el espacio, respecto de un punto de referencia denominado base del índice.

A medida que se vayan introduciendo los distintos conceptos que se refieren al conjunto de los números índice, se profundizarán estos planteamientos primarios, ya que la experiencia aconseja que el estudiante tenga ciertas reservas a medida que avanza en este terreno, para evitar posteriormente una utilización indiscriminada y sin las aludidas reservas.

B. CLASES DE NÚMEROS ÍNDICE

Fundamentalmente, dentro de la estadística económica, interesa disponer de indicadores sobre precios, cantidades y valores.

Un índice de precios será un indicador que refleje la variación de los precios de un conjunto de artículos entre dos momentos en el tiempo o dos puntos en el espacio; es el caso de un índice de costo de vida.

Un índice de cantidades será un indicador que refleje la variación en las cantidades de un conjunto de productos entre dos momentos en el tiempo o dos puntos en el espacio; por ejemplo, un índice de producción industrial.

Por último, un índice de valor indica la variación en el valor total de un conjunto de productos entre dos momentos en el

tiempo o dos puntos en el espacio; ejemplo, índice de ventas comerciales.

c. FÓRMULAS DE CÁLCULO

Ocurre que, cuando se trata de analizar la variación, por ejemplo, en el precio de un solo artículo, no es necesario un indicador especial, basta con expresar la variación en términos porcentuales.

Ejemplo:

Período	*Precio del bien*	*Relación porcentual*
1960	20	100
1961	25	125
1962	26	130
1963	30	150

Tomando como punto de referencia el precio del año 1960, y asignándole el valor 100, se calculan, por una regla de tres simple, los índices correspondientes a los otros períodos; así, puede decirse que el precio del bien entre 1960 y 1965 ha experimentado un alza de 50%. El cálculo de un índice, cualquiera que sea éste, referido a un solo bien, no necesita, pues, de un estadígrafo especial.

El problema de los números índice surge cuando se desea averiguar las variaciones de precios o cantidades de un conjunto de artículos; considérese el siguiente ejemplo:

Bienes	*Precios* 1960	1964
A (metro)	10	15
B (quintal)	100	140
C (litro)	50	60
D (tonelada)	8	6
	168	221

Para resumir el incremento en los precios de este conjunto de artículos, una primera solución consistiría en sumar los precios

en ambos períodos y establecer la variación porcentual entre ambos agregados; de esa manera se llegaría a calcular un índice agregativo simple. Si se considera el año 1962 como base, se tiene:

$$\underset{1960}{100} = \frac{168}{168} \cdot 100 \qquad \underset{1964}{131.5} = \frac{221}{168} \cdot 100$$

Podría concluirse que el conjunto de estos precios ha variado en 31.5% durante el período. Sin embargo, el método tiene dos serias limitaciones; por una parte, estará afectado por las unidades a que estén referidos los precios, y por otra, considera igualmente importantes los cuatro bienes, cuando el B puede ser trigo y el bien D comino; no se discrimina, si se emplea este método, la importancia relativa de cada artículo. En cuanto a las unidades de medida, considérese el ejemplo anterior, y supóngase que el precio del bien B se refiere al quintal de trigo; si se tuviera el precio del kilo el resultado del índice sería distinto:

Bienes	*Precios* 1960	1964
A (metro)	10	15
B (kilo)	1	1.4
C (litro)	50	60
D (tonelada)	8	6
	69	82.4

Asignándole siempre a 1960 el valor 100 como base del índice, se tiene que, para 1964, el índice agregativo simple es ahora de 119.42 (100 · 82.4/69). Se llega a un resultado distinto sin que haya habido variación de precios alguna respecto del caso anterior, excepto que en ambos períodos se tomó un precio referido a una unidad distinta.

Por lo que toca a este problema, es posible evitarlo calculando precios relativos. Se asigna a los precios de cada uno de los artículos en el período base, el valor 100 y por regla de tres simple, se calculan los correspondientes al período para el cual interesa conocer el índice.

Si se observan los ejemplos anteriores, se tendrá:

Bienes	*Precios* 1960	1964
A	100	150
B	100	140
C	100	120
D	100	75

Obsérvese que en el bien B, sea cual fuere la unidad de medida, el incremento de su precio es de 40%. Pero aunque en este método denominado de cifras relativas se subsana el problema de las unidades, persisten otros problemas que exigen solución: en primer lugar, la elección de un indicador de tendencia central. Se tienen los precios relativos para 1964, pero es necesario resumirlos por medio de un estadígrafo de posición: media aritmética, mediana, etc. Todos ellos conducirán, en general, a resultados distintos. ¿Cuál de ellos admitir? Si bien en cada caso particular habrá un estadígrafo adecuado que satisfaga las necesidades de representatividad que se tenga, en general se utiliza la media aritmética, principalmente por la facilidad que implica su manejo algebraico. Es necesario cuidar que no hayan valores extremos que distorsionen el estadígrafo. En el último ejemplo, si se toma la media aritmética, el índice para 1960 será de 100 y el de 1964 de 121.25, es decir, el conjunto de artículos habrá experimentado un alza de 21.25% en el período. Obsérvese que la conclusión está referida al conjunto, pues se trata de un promedio. Cada bien en particular acusa variaciones dispares en sus precios, e incluso el bien D muestra decrecimiento.

En consecuencia, tomar cifras relativas salva el inconveniente de las unidades de medida, pero aún subsiste el problema de la ponderación, ya que a cada artículo debe asignársele la importancia debida.

Tomando como punto de partida los precios relativos se ensayarán algunos criterios de ponderación, para obtener los índices más usuales en la investigación económica.

Si los precios relativos

$$\frac{P_n}{P_0}$$

donde p_n es el precio en el período dado y p_0 el precio en el período base, se ponderan por los valores del año base: $p_0\ q_0$,

se obtiene la conocida fórmula de Laspeyres para precios (IPL), es decir:

$$IPL = \frac{\Sigma \frac{p_n}{p_0} \cdot p_0 q_0}{\Sigma p_0 q_0} = \frac{\Sigma p_n q_0}{\Sigma p_0 q_0}$$

La sumatoria se extiende a todos los artículos considerados en el índice. Como en todo promedio aritmético, se divide por la suma de las ponderaciones.

Un índice de precios de Laspeyres debe interpretarse como el nivel que alcanzan los precios en un año dado, respecto de un año base al que se asigna el valor 100, considerando las mismas cantidades del año base en ambos períodos; en otras palabras, se trata de percibir la variación en los precios de una canasta de productos elegidos en el año base y que permanece inalterada durante los períodos sucesivos.

Este índice, por lo tanto, tiene un significado bien concreto. El supuesto que la canasta de productos realmente no registre variaciones significativas, es otro problema; es el analista quien deberá determinar si el supuesto se cumple o no, y por lo tanto, juzgar la conveniencia de utilizar un índice de Laspeyres.

Por otra parte, si los precios relativos $\frac{p_n}{p_0}$, se ponderan por valores híbridos: $p_0\ q_n$, se tiene el índice de precios de Paasche (IPP), que también se utiliza con frecuencia:

$$IPP = \frac{\Sigma \frac{p_n}{p_0} \cdot p_0 q_n}{\Sigma p_0 q_n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}$$

Obsérvese que ahora los precios están multiplicados por las cantidades del año que se calcula (q_n). Por este hecho un índice de precios de Paasche debe interpretarse como la variación de los precios de un conjunto de productos, suponiendo constantes las cantidades del año dado; en otros términos, la canasta de productos que se considera, es la del período que se calcula y se toma esta misma canasta para el año base.

Respecto de los índices de valor, por el significado simple que tienen, no requieren deducciones especiales, ya que son sencillamente el resultado de la división entre los valores del año que se calcula y el año base.

$$IV = \frac{\Sigma\, p_n\, q_n}{\Sigma\, p_0\, q_0}$$

A continuación se considerará un ejemplo donde se calcularán los índices presentados:

	Año 0		*Año 1*		*Año 2*	
Artículos	*p*	*q*	*p*	*q*	*p*	*q*
A	10	4	12	5	20	3
B	4	3	4	3	5	3
C	8	10	8	12	7	15
D	20	2	30	2	40	3

p: precio; q: cantidad.

No hace falta el cálculo en el año 0, base del índice, ya que coinciden precios y cantidades:

$$p_n = p_0 \quad y \quad q_n = q_0$$

Para los índices de Laspeyres se tiene:

$$IPL\ (año\ 1) = \frac{\Sigma\, p_1\, q_0}{\Sigma\, p_0\, q_0} = \frac{48 + 12 + 80 + 60}{40 + 12 + 80 + 40} = \frac{200}{172} = 116.3$$

$$IPL\ (año\ 2) = \frac{\Sigma\, p_2\, q_0}{\Sigma\, p_0\, q_0} = \frac{80 + 15 + 70 + 80}{40 + 12 + 80 + 40} = \frac{245}{172} = 142.4$$

Para los índices de Paasche, suponiendo siempre el año 0 como base del índice:

$$IPP\ (año\ 1) = \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_1} = \frac{60 + 12 + 96 + 60}{50 + 12 + 96 + 40} = \frac{228}{198} = 115.2$$

$$IPP\ (año\ 2) = \frac{\Sigma\, p_2\, q_2}{\Sigma\, p_0\, q_2} = \frac{60 + 15 + 105 + 120}{30 + 12 + 120 + 60} = \frac{300}{222} = 135.1$$

El índice de valor:

$$IV\ (año\ 1) = \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_0} = \frac{60 + 12 + 96 + 60}{40 + 12 + 80 + 40} = \frac{228}{172} = 115.1$$

$$IV \text{ (año 2)} = \frac{\Sigma p_2 q_2}{\Sigma p_0 q_0} = \frac{60 + 15 + 105 + 120}{40 + 12 + 80 + 40} = \frac{300}{172} = 174.4$$

En el siguiente cuadro puede apreciarse la diferencia entre las indicaciones de uno y otro índice

	Índices		
Años	*IPL*	*IPP*	*IV*
0	100.0	100.0	100.0
1	116.3	115.2	115.1
2	142.4	135.1	174 4

Puede apreciarse que el IPL crece más que el IPP. En el primero se considera constante la canasta de productos del período base; en cambio, en el segundo, se considera constante la canasta de productos del período en que se calcula el índice. Por lo tanto ambos índices indican la variación promedio de los precios bajo supuestos diferentes; sin embargo, es muy frecuente confundir el significado de estos indicadores, porque no se consideran los supuestos de uno y otro.

Con referencia a los índices de cantidad, es necesario hacer el mismo tipo de consideraciones, ya que se presentan problemas similares; unidades de medida, ponderaciones, etcétera.

Siguiendo el mismo criterio de los índices de precios, puede obtenerse el índice de cantidades de Laspeyres (IQL).

$$IQL = \frac{\Sigma \frac{q_n}{q_0} \cdot p_0 q_0}{\Sigma q_0 p_0} = \frac{\Sigma q_n p_0}{\Sigma q_0 p_0}$$

El índice de cantidades de Paasche será (IQP).

$$IQP = \frac{\Sigma \frac{q_n}{q_0} \cdot q_0 p_n}{\Sigma q_0 p_n} = \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}$$

Mientras el IQL representa la variación en las cantidades suponiendo constantes los precios del período base, el IQP repre-

senta la variación de las cantidades suponiendo constantes los precios del período calculado. Nuevamente se insiste sobre la necesidad de no descuidar estos supuestos, cuando se interpreta un índice.

Las fórmulas presentadas son, como se dijo, las de uso más frecuente. Hay una cantidad extraordinaria de fórmulas de índices que se diferencian unas de otras según los factores de ponderación utilizados. Por razones de brevedad sólo se mostrarán las más conocidas.

Marshall - Edgeworth para precios

$$IPM = \frac{\Sigma p_n (q_0 + q_n)}{\Sigma p_0 (q_0 + q_n)}$$

Índice de precios de Keynes

$$IPK = \frac{\Sigma p_n (q_0 \wedge q_n)}{\Sigma p_0 (q_0 \wedge q_n)}$$

donde el signo $\wedge$ es ínfimo y quiere indicar que se tome la menor de las cantidades que están a sus costados.

La llamada fórmula "ideal" de Fischer para precios, que es la media geométrica de los índices de Laspeyres y de Paasche.

$$IPF = \sqrt{IPL \cdot IPP} = \sqrt{\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \cdot \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}}$$

En estas últimas fórmulas, bastará remplazar p por q, para obtener las fórmulas correspondientes a índices de cantidad.

D. PRUEBAS SOBRE LOS NÚMEROS ÍNDICE

Irving Fischer, quien plantea la fórmula por él llamada ideal, propuso pruebas para calificar los números índice.

1] *Prueba de reversión de factores*

La prueba se basa sobre un criterio de analogía: lo que es cierto para un producto debiera ser cierto para un conjunto de ellos. Así, para cualquier artículo se tiene que:

$$\text{precio} \times \text{cantidad} = \text{valor}$$

Las fórmulas de Laspeyres y Paasche no satisfacen esta prueba, como puede verse a continuación.

$$\text{IPL} \times \text{IQL} \neq \text{IV}, \text{ en efecto}$$

$$\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \cdot \frac{\Sigma q_n p_0}{\Sigma q_0 p_0} \neq \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

$$\text{IPP} \times \text{IQP} \neq \text{IV}$$

$$\frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \cdot \frac{\Sigma q_n p_n}{\Sigma q_0 p_n} \neq \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

La fórmula de Fischer sí satisface esta prueba

$$\text{IPF} \times \text{IQF} = \text{IV}$$

$$\left(\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \cdot \frac{\Sigma p_n q_n}{\Sigma p_0 q_n}\right)^{1/2} \left(\frac{\Sigma q_n p_0}{\Sigma q_0 p_0} \cdot \frac{\Sigma q_n p_n}{\Sigma q_0 p_n}\right)^{1/2} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

Simplificando términos semejantes se tiene:

$$\frac{\Sigma p_n q_n}{\Sigma p_0 q_0} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

Sin embargo, esta prueba también se satisface con una combinación de índices de precios de Laspeyres y cantidades de Paasche o viceversa, como se comprueba a continuación:

$$\text{IPL} \times \text{IQP} = \text{IV}$$

$$\frac{\Sigma p_n q_0}{\Sigma p_0 q_0} \cdot \frac{\Sigma q_n p_n}{\Sigma q_0 p_n} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

$$\text{IPP} \times \text{IQL} = \text{IV}$$

$$\frac{\Sigma p_n q_n}{\Sigma p_0 q_n} \cdot \frac{\Sigma q_n p_0}{\Sigma q_0 p_0} = \frac{\Sigma p_n q_n}{\Sigma p_0 q_0}$$

Estas dos últimas relaciones merecen retenerse, porque se utilizan con frecuencia.

2] *Pruebas de reversión temporal*

Nuevamente el criterio de analogía; si el precio de un producto es en el período "a" de 40 y en el período "b" de 50, en el primer período se observa que el precio es el 80% del que se da en el período "b", y en éste el 125% del precio del período "a". Lógicamente el producto de estos porcentajes debe dar 1, es decir:

$$\frac{P_n}{p_0} \times \frac{P_0}{p_n} = 1$$

Esta prueba no la cumplen las fórmulas de Laspeyres y Paasche.

En efecto:

$$IPL_{b,a} \times IPL_{a,b} \neq 1$$

donde el primer subíndice indica el período que se calcula y el segundo el período base.

$$\frac{\Sigma p_b q_a}{\Sigma p_a q_a} \cdot \frac{\Sigma p_a q_b}{\Sigma p_b q_b} \neq 1$$

$$IPP_{b,a} \cdot IPP_{a,b} \neq 1$$

$$\frac{\Sigma p_b q_b}{\Sigma p_a q_b} \cdot \frac{\Sigma p_a q_a}{\Sigma p_b q_a} \neq 1$$

La fórmula de Fischer sí cumple la prueba

$$IPF_{b,a} \times IPF_{a,b} = 1$$

$$\left(\frac{\Sigma p_b q_a}{\Sigma p_a q_a} \cdot \frac{\Sigma p_b q_b}{\Sigma p_a q_b}\right)^{1/2} \left(\frac{\Sigma p_a q_b}{\Sigma p_b q_b} \cdot \frac{\Sigma p_a q_a}{\Sigma p_b q_a}\right)^{1/2} = 1$$

3] *Prueba circular*

Si el precio de un producto es de 10 en el primer período, de 12 en el segundo y de 18 en el tercero, se comprueba que en el segundo período el precio es 120% del precio en el primero

y en el tercer período 150% del precio en el segundo. En consecuencia el precio del tercer período es 180% del registrado en el primero:

$$(120\%)\ (150\%) = 180\%$$

La prueba circular no la cumple ninguna de las fórmulas analizadas. Suponiendo tres períodos para IPL se tiene:

$$IPL_{3,2}\ IPL_{2,1} \neq IPL_{3,1}$$

donde el primer subíndice indica el período que se calcula y el segundo el período base.

$$\frac{\Sigma p_3 q_2}{\Sigma p_2 q_2} \cdot \frac{\Sigma p_2 q_1}{\Sigma p_1 q_1} \neq \frac{\Sigma p_3 q_1}{\Sigma p_1 q_1}$$

Esta relación sólo se cumpliría en el caso que

$$q_1 = q_2 = q_3 \text{ (para todos los artículos).}$$

Sin embargo, la relación planteada se cumple en forma aproximada cuando no existen diferencias significativas en las cantidades durante los distintos períodos.

El lector podrá comprobar que ni la fórmula de Paasche ni la de Fischer cumplen la prueba circular, pudiendo seguir el mismo proceso de comprobación realizado para la fórmula de Laspeyres.

Respecto de estas pruebas, es conveniente aclarar que la existencia de índices que no las cumplan no justifica dejarlos de lado; interesa más el significado concreto del índice, teniendo en cuenta sus alcances y limitaciones. Es así como la fórmula ideal de Fischer, que satisface las pruebas por él planteadas, no es susceptible de ser claramente interpretada, ya que se trata de la combinación de dos índices que, si por separado adquieren cabal significado, al combinarlos ofrecen dificultades para su interpretación.

E. BASE DE UN NÚMERO ÍNDICE

Al definir un número índice se ha destacado que se trata de una comparación de dos momentos en el tiempo o dos puntos en el

espacio. El momento o punto con respecto al cual se establece la comparación recibe el nombre de base de un índice y se le asigna el valor 100, para analizar las variaciones porcentuales. Respecto de la elección del período base hay que tener siempre presente el objetivo que se persigue con el índice; en general se estima que el período base debe ser un período normal. Cabe preguntarse qué se entiende por normalidad en estos casos, cuando en los países en desarrollo los cambios son muy frecuentes y la anormalidad es un denominador común. Tal vez al definir el período base fuese más sensato pensar en un período durante el cual no existan accidentes o cambios violentos. Por lo demás, será necesario cambiar la base del índice cuando los supuestos planteados pierdan validez a medida que pasa el tiempo; es el caso de los índices de costo de vida, cuya base debe modificarse toda vez que la estructura de consumo presente cambios significativos con respecto de la admitida en el período base.

Sobre este mismo asunto, será necesario distinguir dos tipos de base: base fija y base variable. Los índices de base fija son aquellos que mantienen como base un período fijo de referencia, en tanto que los índices de base variable son aquellos que tienen como base el período inmediatamente anterior. Con un índice de base fija puede calcularse el correspondiente de base variable y viceversa; los resultados, en general, diferirán de los que se obtendrían a partir de los datos originales, ya que las fórmulas usuales no cumplen con la prueba circular.

Ejemplo: Supóngase que el índice de Laspeyres para los precios de los materiales de construcción sea el siguiente:

	Índice base 1960 = 100
1960	100
1961	110
1962	120
1963	144
1964	180
1965	190

El correspondiente índice de base variable sería:

Índice de base variable	
1960	—
1961	110.0
1962	109.1
1963	120.0
1964	125.0
1965	105.5

Otra operación que es muy usual respecto de los índices de base fija es la del empalme; se trata, como indica su nombre, de empalmar índices con base distinta. Obsérvese el siguiente ejemplo, donde se tiene un índice para el período 1956-1959 con base 1956, y otro índice de 1959 a 1962 con base 1959, y se pretende tener una serie para todo el período 1956-1962.

Años	*Índice base 1956 = 100*	*Índice base 1959 = 100*
1956	100	
1957	120	
1958	150	
1959	180	100
1960		110
1961		132
1962		150

Mediante una sencilla regla de tres, puede completarse cualesquiera de las dos series, para tener el movimiento del índice durante todo el período.

Años	*Índice base 1956 = 100*	*Índice base 1959 = 100*
1956	100.0	55.6
1957	120.0	66.7
1958	150.0	83.3
1959	180.0	100.0
1960	198.0	110.0
1961	237.6	132.0
1962	270.0	150.0

Debe advertirse que este tipo de empalmes significa sólo una aproximación que puede ser muy defectuosa, dependiendo de la similitud de las bases y de sus supuestos. En todo caso, es posible recurrir a estos empalmes siempre que se tenga conciencia de sus limitaciones.

F. UTILIZACIÓN DE LOS NÚMEROS ÍNDICE

Un número índice indica la evolución de precios, cantidades y valores, para un conjunto de productos. Prestan en consecuencia la utilidad inmediata de reflejar la tendencia de los cambios y ritmos de los conceptos señalados. Ese solo hecho ya justifica su cómputo y su periódica utilización en la investigación socioeconómica. Sin embargo prestan además otros servicios sobre los que es conveniente hacer algunos comentarios.

1] *La deflactación**

Las alteraciones en los sistemas y niveles de precios que se presentan dentro de la actividad económica originan dificultades en la comparación de valores monetarios que corresponden a períodos distanciados. No es mucho lo que se puede deducir de la comparación de valores nominales, es decir, de valores expresados en unidades monetarias de distinto poder adquisitivo. Para poder llegar a conclusiones válidas acerca del comportamiento de una variable que represente "valor", será necesario expresar los montos monetarios nominales en unidades homogéneas; esta transformación recibe el nombre de deflactación, y con ella se pretende eliminar, exclusivamente, el efecto de alteraciones en los precios.

El proceso de deflactación exige disponer de un índice deflactor, es decir, de un indicador que proporcione una pauta de las alteraciones en los precios que tengan relación con la variable que se pretende deflactar. Es de fundamental importancia recordar que no existe un índice deflactor único; cada variable, en rigor, debería tener un deflactor adecuado. La disponibilidad de sólo un reducido número de índices de precios, hace que éstos sean utilizados indiscriminadamente para una serie de pro-

* Se empleará el neologismo "deflactación" para indicar la metodología de transformación de valores expresados en precios corrientes a valores en precios constantes; el propósito es limitar el empleo de "deflación" para caracterizar con esta última expresión el fenómeno económico opuesto a la inflación.

pósitos; aunque este hecho puede adolecer de errores conceptuales, muchas veces se justifica el procedimiento, porque interesa conocer un orden de magnitud antes que un valor exacto, siempre que se tenga conciencia de las limitaciones del método. En todo caso, aun dentro de las escasas disponibilidades de índices de precios, es posible llegar a soluciones aceptables, ya sea eligiendo racionalmente un índice de precios como deflactor o combinando y ponderando en forma adecuada dos o más índices; este último procedimiento, si bien puede no conducir a soluciones ideales, por lo menos puede representar una disminución de las posibles distorsiones.

Antes de profundizar este problema de la deflactación, cuando se desea transformar unidades monetarias heterogéneas (unidades de cada período) en unidades monetarias homogéneas (unidades del período base), y permitir de este modo la comparación en el tiempo, el primer recurso al cual se apela, es expresar los montos monetarios nominales en unidades de moneda extranjera de valor más o menos estable: dólares, libras, etc. Mas sobre este procedimiento caben algunas objeciones. Los gobiernos tienen instrumentos que les permiten fijar los tipos de cambio con las monedas extranjeras en forma arbitraria (arbitraria para los fines que aquí se comentan) y que en general no representan las alteraciones en los niveles de precios. En Chile hay una experiencia reciente; en el transcurso de menos de dos años, la unidad monetaria chilena llegó a sufrir una fuerte devaluación (de E° 1 053 en noviembre de 1962 a E° 3 400 en abril de 1963, por dólar norteamericano tipo de cambio corredor). Si un valor dado en escudos se expresase en dólares en ambas fechas, podría concluirse que entre los dos períodos mencionados dicho valor se redujo a la tercera parte; si bien esto puede ser cierto para una persona que va a gastar sus ingresos en Estados Unidos, no lo es para quien efectúa sus desembolsos en Chile, porque durante ese período ningún índice de precios propiamente tal ha sufrido un aumento de 200 por ciento.

Por otra parte, una segunda objeción a este procedimiento consiste en el hecho de que aun las economías más estables pueden sufrir cierto grado de inflación. Los dos argumentos expuestos descalifican, en general, la conversión a unidades monetarias extranjeras como una alternativa de la deflactación. La mecánica de la deflactación implica dividir los montos monetarios nominales por el índice de precios elegido como deflactor adecuado; y su explicación podrá encontrarse en la siguiente regla de tres. Si en el año n se tiene un valor nominal VN_n y un índice de

precios IP_n, ¿cuál sería este valor expresado en unidades monetarias de igual poder adquisitivo que las del año base? En otros términos, ¿cuál sería este valor si el índice de precios no hubiera variado?

El planteamiento queda así reducido:

$$IP_n \ldots VN_n$$

$$100 \ldots x$$

$$x = \frac{VN_n}{IP_n} \cdot 100 = \text{Valor real}$$

Desde otro punto de vista, se justifica la deflactación pensando en los componentes de un valor: precio por cantidad

$$\text{Índice de precios} \times \text{cantidad} = \text{Valor (100)}$$

$$\text{Cantidad} = \frac{\text{Valor (100)}}{\text{Índice de precios}} = \text{Valor real}$$

Ahora bien, la evolución de las cantidades en el tiempo está libre, por así decirlo, de influencias monetarias directas, y muestra la evolución física, real, de una serie, que es precisamente lo que de pretende con la deflactación.

Los valores reales así obtenidos, están expresados en unidades monetarias que tienen un poder adquisitivo correspondiente al año base del índice deflactor.

Con referencia a este último planteamiento, es necesario distinguir dos tipos de base. Se llamará base propiamente tal al período que corresponde al diseño del índice, donde se perfila la muestra, se establecen las ponderaciones, etc. Por otra parte, se utilizará la expresión "base aritmética" para referirse a cualquier período, que por transformación lineal se le haya asignado el valor 100. Los cambios aritméticos de base implican hasta cierto punto una arbitrariedad que puede originar algún tipo de perturbaciones al establecer las comparaciones. Sus efectos serán tanto mayores, cuanto más distante sea la base propiamente tal del índice deflactor. Cuando hay consenso de que los supuestos de la construcción y diseño del índice siguen siendo válidos, los cambios aritméticos de base no introducirán deformaciones significativas; por ello, antes de deflactar será necesario decidir

de qué año serán las unidades monetarias en que interesa expresar los valores reales. Es recomendable, si se cumple la condición de constancia de los supuestos de la construcción del índice, expresar los valores nominales en unidades monetarias del año más reciente, lo que se consigue mediante un índice deflactor que tenga como base el último año en el sentido cronológico. La utilidad de la sugerencia anotada, radica en el hecho que el investigador tiene una visión reciente y tal vez objetiva del sistema de precios imperante, por lo que es probable que la obtención de conclusiones quede facilitada. Es necesario destacar que un proceso de deflactación conduce a valores reales que pueden tener dos interpretaciones: una expresión física o un poder de compra. Una variable monetaria está compuesta por una suma de valores del tipo $\Sigma p_n\ q_n$; si esta serie se deflacta por un índice de precios de los productos considerados en la serie nominal, el resultado será una expresión física de la serie. En efecto, utilizando un índice deflactor de Paasche, se tiene:

$$\frac{\text{Valor nominal}}{\text{IPP}} = \frac{\Sigma\, p_n\, n_n}{\dfrac{\Sigma\, p_n\, q_n}{\Sigma\, p_0\, q_n}} = \Sigma\, p_0\, q_n$$

El resultado es evidentemente un quantum, es decir, cantidades del período n, valorizadas a precios del período base. Si se hubiera deflactado por un índice de precios de Laspeyres, se tendría:

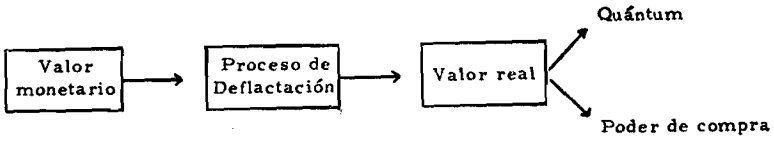
$$\frac{\text{Valor nominal}}{\text{IPL}} = \frac{\Sigma\, p_n\, q_n}{\dfrac{\Sigma\, p_n\, q_0}{\Sigma\, p_0\, q_0}} = \text{IQP} \cdot \Sigma\, p_0\, q_0$$

Resultado que equivale a proyectar un valor en el período base, a través de un índice de cantidades de Paasche. También representa una evolución física de la serie, aunque con connotaciones diferentes al caso anterior. Cabe hacer notar que cuando se desea llegar a una expresión física rigurosa, sólo existe un tipo de deflactor adecuado: el que contienen los productos incluidos en la serie nominal.

Por otra parte, cuando se quiere obtener un poder de compra, es necesario especificar qué uso se dará a un monto monetario; de esta manera, si el sueldo de un empleado en diferentes

períodos se deflacta por un índice de precios al consumidor, el resultado sería un poder de compra en términos de la canasta de productos elegida en el índice deflactor. Si el sueldo de dicho empleado se deflacta por índice de valores bursátiles, el resultado será un poder de compra en términos de acciones y bonos. El uso que se dará a un monto monetario es el que determina el tipo de poder de compra resultante.

Resumiendo esquemáticamente se tiene:



A continuación y a título de ejemplo se presentan los resultados de un proceso de deflactación:

Años	*Sueldo de un empleado (u. m. de c/año)* A	*Índice de precios al consumidor (base 1963 = 100)* B	*Sueldo real (u. m. de 1963)* $\frac{A}{B} \cdot 100$
1958	400	36.5	1 096
1959	480	50.6	946
1960	600	56.5	1 062
1961	680	60.9	1 117
1962	720	69.3	1 039
1963	900	100.0	900

La deflactación presentada implica haber elegido como deflactor el índice de precios al consumidor. Tal vez de los valores reales no pueda deducirse, en forma categórica, que el empleado haya sufrido una merma de esa magnitud en su poder de compra; lo que hubiera ocurrido si dicho empleado gastara todo su ingreso en la forma que lo hace el empleado típico elegido como padrón en el índice de precios al consumidor. Si dicho empleado sólo gasta en manutención el 60 por ciento de sus ingresos y el 40 por ciento restante lo dedica, por ejemplo, a la construcción de una vivienda, la deflactación se realizaría

en dos partes: la primera parte (60%) se deflactaría por el índice de precios al consumidor y el saldo debería deflactarse por un índice de precios de insumos de la construcción; de esta manera se obtendría la expresión real de su poder de compra, en términos del uso que dará a sus ingresos.

Cuando se ha realizado una deflactación, es necesario tener presente que los valores reales así obtenidos son simplemente aproximaciones, que serán tanto mejores cuanto más representativo sea el índice de los precios que se cancelarán con el monto monetario. Esto no quiere decir que sea indispensable "construir" índices si no se dispone de los más adecuados; sólo se pretende destacar la necesidad de seleccionar los índices disponibles, y en todo caso, tener presente las limitaciones que acuse una deflactación obligada por un índice que no sea el mejor. Es responsabilidad de los organismos autorizados y de los principales usuarios la elaboración de índices de precios de la actividad económica.

La deflactación, mecánicamente es un asunto trivial, pero la elección de deflactores adecuados requiere mucha atención. Véase, por ejemplo, qué sucede con los índices de producción y ventas industriales de Chile.* Observando las cifras siguientes debe admitirse que no existe la concordancia que debería existir, tampoco se encuentran razones que expliquen la diferencia; aunque es muy probable que el desajuste radique en la calidad del deflactor utilizado, antes que en las variaciones de existencias.

Período	*Índice de producción industrial Base 1959 = 100*	*Índice de ventas industriales reales Base 1959 = 100*
1959	100.0	100.0
1960	103.5	98.3
1961	106.1	110.6
1962	113.8	127.7
1963		
Enero	108.4	120.9
Abril	122.2	129.8
Agosto	112.8	132.1

* César Molestina, "Los índices de producción industrial manufacturera y ventas reales industriales. Un comentario acerca de su comportamiento", mayo de 1963.

De la simple observación de estas series surgen las contradicciones anotadas. Si se admite que el índice de producción industrial es un indicador confiable, el índice de ventas sería el que adolece de defectos. Estos defectos pueden deberse a dos causas no excluyentes: que el índice de ventas nominales no sea lo suficientemente acertado, por errores inherentes o ajenos al muestreo, o que el deflactor utilizado no tiene la relación apropiada con los precios de los bienes industriales, o una combinación de ambas causas.

2] *El deflactor implícito del producto bruto*

No es necesario citar la imperiosa necesidad que hay de disponer de un sistema de contabilidad social expresada en precios constantes; en este sentido se tratará principalmente la deflactación del producto bruto.

Es importante aclarar la idea de deflactor implícito. Éste aparece como resultado de un proceso de deflactación cuando se trata de cumplir con alguna restricción; se presenta principalmente cuando una variable global ha sido desglosada en componentes, con el objeto de identificar un índice deflactor adecuado con cada componente. En estos casos la restricción aparece como suma de componentes que reproducen la variable global; es necesario agregar que el deflactor implícito se origina en el hecho que la restricción (muchas veces definición) debe ser satisfecha tanto en valores nominales o corrientes como en valores reales o constantes; entonces se dice que el proceso de deflactación es coherente. El deflactor resultante de una deflactación coherente se denomina implícito, porque justamente está implícito en el cumplimiento de la restricción.

Supóngase que X, Y y Z representan valores sujetos a la restricción:

$$X + Y + Z = W$$

restricción que se satisface en valores corrientes. Si además se desea que esta restricción sea satisfecha en valores constantes, se tiene:

$$\bar{X} + \bar{Y} + \bar{Z} = \bar{W}$$

donde la barra sobre el símbolo se utiliza para indicar que se trata de valores reales, es decir, que

$$\bar{X} = \frac{X}{IPX}$$

$$\bar{Y} = \frac{Y}{IPY}$$

$$\bar{Z} = \frac{Z}{IPZ}$$

La suma de $\bar{X} + \bar{Y} + \bar{Z}$ reproduce el valor real de W, es decir, $\bar{W}$. ¿Cuál deberá ser el deflactor de W para que el valor real resultante $\bar{W}$ coincida con la suma de $\bar{X} + \bar{Y} + \bar{Z}$? En efecto, se tiene que

$$\frac{X}{IPX} + \frac{Y}{IPY} + \frac{Z}{IPZ} = \frac{W}{IPW}$$

Basta con despejar IPW de la anterior relación:

$$IPW = \frac{W}{\frac{1}{IPX} \cdot X + \frac{1}{IPY} \cdot Y + \frac{1}{IPZ} \cdot Z}$$

Se comprueba que el deflactor implícito IPW es el promedio armónico de los deflactores componentes, donde los factores de ponderación son justamente los valores nominales; otra forma de calcular el deflactor implícito es comparar $\bar{W}$ que se obtiene como suma de $\bar{X} + \bar{Y} + \bar{Z}$ con W.

$$IPW = \frac{W}{\bar{W}}$$

Para determinar el deflactor implícito del producto bruto, habrá que pensar previamente en alguna definición. Si la restricción es del siguiente tipo:

$$P = C + I + E - M$$

se determinarán los deflactores para consumo, inversión, expor-

taciones e importaciones y se obtendrá como suma el producto real (P).

$$\bar{P} = \frac{C}{IPC} + \frac{I}{IPI} + \frac{E}{IPE} - \frac{M}{IPM}$$

El deflactor implícito del producto (DI), según esta restricción, estará dado por la relación

$$DI = \frac{P}{\bar{P}}$$

Puede verificarse al mismo tiempo que el deflactor implícito es equivalente a la media armónica de los deflactores componentes.

$$DI = \frac{P}{\frac{1}{IPC} \cdot C + \frac{1}{IPI} \cdot I + \frac{1}{IPE} \cdot E - \frac{1}{IPM} \cdot M}$$

Si la restricción fuera de otro tipo, por ejemplo que la suma de los valores agregados sectoriales reproducen el producto bruto, se tendrá otro deflactor implícito sujeto a la restricción propuesta. Habría que discutir previamente la posibilidad de encontrar deflactores adecuados para el valor agregado; una alternativa, por cierto muy discutible, es deflactar el valor agregado de cada sector, por los precios de los bienes que el sector produce. El resultado sería un poder de compra de los valores agregados sectoriales, en términos de los bienes que produce cada sector. Puede argumentarse en favor de este método, sosteniendo que la contrapartida física del valor agregado es justamente la cantidad de bienes producidos, sin embargo, la justificación es en extremo débil. En todo caso, en los cursos de Contabilidad Social se muestran las ventajas y desventajas de este y otros métodos de deflactación. Para ilustrar, en seguida se detalla este procedimiento. Si se desea llegar a expresiones reales, cada rama de actividad debería deflactarse por un índice de precios relacionado directamente con dichas ramas de actividad. Así el valor agregado por el sector industrial, debería deflactarse por un índice de precios de bienes industriales; el valor agregado por el sector agropecuario, se deflactaría por un índice de precios de

bienes agropecuarios. De esta manera se obtendrían valores reales en cada rama que representaría una aproximación a la evolución física de lo producido por cada sector. Sumando los valores reales de todas las ramas de actividad, se tendría el producto bruto en términos reales; y comparando los productos brutos en valores nominales y reales se obtiene el llamado deflactor implícito del producto, que no es otra cosa que un índice general promedio de los precios que rigen en la actividad económica.

Sean:

PN_{ij} el producto nominal del sector "i" en el año "j".

IP_{ij} el índice de precios del sector "i" en el año "j".

PR_{ij} el producto real del sector "i" en el año "j".

De la deflactación resulta

$$PR_{ij} = \frac{PN_{ij}}{IP_{ij}} \quad 100$$

Si se suman todos los productos reales por sector en un año cualquiera "j", se tiene el producto bruto real en el año "j".

$$PR_j = \sum_{i=1}^{L} PR_{ij} = \sum_{i=1}^{L} \frac{PN_{ij}}{IP_{ij}} \cdot 100$$

donde i = 1, 2, . . . L representa el número de sectores considerados. Por otra parte, el producto bruto real PR_j se obtiene deflactando el producto bruto nominal PN_j, por el deflactor implícito DI_j, conceptos todos referidos a un período j, es decir:

$$PR_j = \frac{PN_j}{DI_j} \cdot 100$$

Remplazando en la igualdad anterior, se tiene

$$\frac{PN_j}{DI_j} \cdot 100 = \sum_{i=1}^{L} \frac{PN_{ij}}{IP_{ij}} \cdot 100$$

Despejando DI_j

$$DI_j = \frac{PN_j}{\sum_{i=1}^{L} \frac{PN_{ij}}{IP_{ij}}}$$

que justamente corresponde con la definición de media armónica de los deflactores sectoriales, considerando como ponderaciones los productos nominales de cada sector, ya que

$$PN_j = \sum_{i=1}^{L} PN_{ij}$$

A continuación se presentará un ejemplo para ilustrar el proceso de deflactación de los productos sectoriales.

Supóngase que la actividad económica ha sido dividida en cuatro sectores para los cuales se dispone de las siguientes informaciones:

Producto nominal
(unidades monetarias corrientes)

Sectores	*1960*	*1961*	*1962*
Minería	200	300	400
Agricultura	300	350	400
Industria	200	250	300
Servicios	500	600	700
Producto bruto nominal	1 200	1 500	1 800
Índices de precios (base 1960 = 100)			
Productos mineros	100	130	150
Productos agrícolas	100	110	120
Productos industriales	100	120	130
Servicios	100	110	120

Deflactando el producto de cada sector, por el índice de precios correspondiente, se tendrá los productos reales de cada sector:

Producto real
(unidades monetarias constantes de 1960)

Sectores	*1960*	*1961*	*1962*
Agricultura	200	230.8	266.7
Minería	300	318.2	333.3
Industria	200	208.3	230.8
Servicios	500	545.5	583.3
Producto bruto real	1 200	1 302.8	1 414.1

Para calcular el deflactor implícito, recuérdese que

$$DI_j = \frac{PN_j}{PR_j} \quad 100$$

	1960	*1961*	*1962*
Deflactor implícito	100	115.1	127.3

Como se demostró, estos valores equivalen a los promedios armónicos de los índices deflactores.

Con respecto a la deflactación del producto bruto es conveniente advertir que puede prestarse a presentaciones caprichosas, según sea la base de los índices deflactores. En otras palabras, el producto bruto real mostrará distintas tasas de crecimiento según sea la base de los deflactores. Aparentemente esto no tendría por qué ocurrir, ya que un cambio aritmético de la base del índice deflactor no debería afectar las variaciones reales en el producto bruto. Mediante un ejemplo que pretende exagerar la situación antes que representar un hecho real, se ilustra este planteamiento.

Para abreviar, supóngase que la actividad económica ha sido dividida en dos sectores: primario y secundario. Los datos son los siguientes:

Producto bruto nominal
(u. m. corrientes)

	1956	*1962*
Sector primario	1 000	1 600
Sector secundario	1 000	1 800
Índices deflactores (base 1956 = 100)		
Sector primario	100	110
Sector secundario	100	300

Si se calcula el producto bruto real, en precios constantes de 1956, se tiene:

Producto bruto nominal
(u. m. constantes)

	1956	*1962*
Sector primario	1 000	1 450
Sector secundario	1 000	600
	2 000	2 050

Entre ambos años el crecimiento es 2.5 por ciento, si se computa el producto en precios de 1956. Si aritméticamente se cambia la base de los deflactores y en consecuencia el producto se computa a precios de 1962, se llega a una variación porcentual diametralmente opuesta.

Los nuevos índices deflactores (con base en 1962) serán los siguientes:

Índices deflactores (base 1962 = 100)

Sector primario	91.0	100
Sector secundario	33.3	100

Deflactando los productos nominales, por los índices anteriores se tiene:

Producto bruto real (u. m. constantes)		
Sector primario	1 100	1 600
Sector secundario	3 000	1 800
	4 100	3 400

Puede observarse un decrecimiento de 17 por ciento en el mismo período. Nótese que para los sectores, considerados en forma aislada, no ocurre esta incompatibilidad, la que sólo se presenta cuando se comparan sumas de sectores con crecimientos nominales distintos y cuyos respectivos deflactores también muestran variaciones desiguales. Los resultados serán tanto más contradictorios, en uno y otro caso, cuanto más distintos sean los índices deflactores y mayores sus variaciones. Lo que ocurre es que se está sumando valores que no son rigurosamente homogéneos, dados los cambios aritméticos de la base de los índices. Los resultados a que se ha llegado no deben sorprender si se piensa en la limitación de los mencionados cambios de base. Por otra parte, la estructura del producto bruto nominal es distinta en ambos períodos. En el primer caso los dos índices deflactores son iguales a 100 en 1956 cuando tanto el sector primario como el secundario aportan un 50 por ciento del producto bruto cada uno. En el segundo caso los dos índices deflactores son iguales a 100 en 1962 y el sector primario aporta con un 47 por ciento al producto y con 53 por ciento al sector secundario. Por esta razón no se presentan las inconsistencias anotadas para los sectores considerados en forma aislada, y sí, en cambio, se presentan al comparar sumas deflactadas por índices distintos. La inconsistencia será tanto mayor cuanto más distintas sean las variaciones de los deflactores.

3] *La proyección sobre la base de índices de cantidad*

Un procedimiento corrientemente utilizado para proyectar el producto bruto real por sectores, es el basado sobre índices de cantidad. Consiste en hacer variar el producto de un período dado conforme al índice de producción correspondiente. De esta manera, si se sabe que para 1962 el producto generado por la industria fue de 5 000 unidades monetarias de ese año y se dispone del índice de producción industrial para los años 1962 a

1965, se puede suponer que habrá correspondencia entre las variaciones de dicho producto sectorial y del índice de cantidad.

Años	*Producto industrial (precios de 1962)*	*Índice de producción industrial (base 1962 = 100)*	*Estimación del producto industrial (precios de 1962)*
1962	5 000	100	5 000
1963	—	198	5 400
1964	—	120	6 000
1965	—	135	6 750

Este tipo de estimaciones, para breves períodos parece justificado, pero no hay que perder de vista que el índice de producción muestra más bien las variaciones de la producción global antes que las del producto (valor agregado); la metodología de estimación presentada tiene como supuesto la constancia en los coeficientes de insumo producto, constancia que puede no ser real en una época caracterizada por tanto cambio tecnológico.

G. ÍNDICES DE COMERCIO EXTERIOR

Para abordar el intercambio de bienes y servicios con el resto del mundo, es indispensable cuantificar indicadores acerca de precios, cantidades, valores, etc., relacionados con exportaciones e importaciones; el tipo de problemas que en estos casos debe enfrentarse, son los inherentes a los números índice, más algunas precauciones que deben tomarse por circunstancias que atañen al comercio exterior.

1] *Índices de precios*

En general es necesario homogeneizar los datos básicos, ya que no siempre están sujetos a criterios uniformes de valuación: precios CIF, FOB, precios expresados en monedas diferentes, valores nominales de retorno u otro tipo de valorización, en qué momento se considera que la importación o exportación está consumada: en tránsito, en aduana, etc.

Para el cálculo de índices de precios, puede utilizarse fórmulas de Laspeyres y Paasche. Sin embargo, en caso de utilizar la fórmula de Laspeyres, si algún producto deja de negociarse, im-

plica suponer que su precio bajó a cero, lo que constituye un evidente falseamiento. Habrá que tener la precaución de tomar en cuenta para el cálculo de los productos sólo los negociados tanto durante el período base como durante el período dado. Por otra parte, no es necesario adoptar esta precaución, si se aplica la fórmula de Paasche, ya que en ésta se elimina automáticamente el producto que deja de negociarse y no quedará comprendido en ninguno de los factores $p_n\ q_n$ ni $p_n\ q_0$. Por eso, para determinar índices de precios se suele utilizar fórmulas de Paasche. En estadísticas de comercio exterior, se acostumbra designar a estos índices como índices de valor unitario, por el tipo de informaciones que se tiene. No es el precio y la cantidad de un artículo, sino el precio promedio o valor unitario que correspondería a un artículo proveniente de una partida de artículos no totalmente homogéneos. De este modo, si la importación de 100 automóviles de distintas marcas representa un valor CIF de 400 000 unidades monetarias, el precio promedio por automóvil es de 4 000 u. m. En este sentido es que se prefiere designar a estos índices como índices de valor unitario en vez de precios, aunque metodológicamente no hay diferencias.

2] *Índices de cantidad*

Nuevamente cabe hacer aquí referencia a la denominación especial de índices de quantum que se les da en comercio exterior por razones similares a las anotadas en el punto anterior.

En este caso, si un producto deja de importarse o exportarse quiere decir que la cantidad negociada bajó a cero, lo que queda ahora bien reflejado en la fórmula de Laspeyres. Por esa razón se acostumbra utilizarla en el cálculo de índices de quantum de comercio exterior. Además, utilizar fórmulas de Laspeyres o Paasche para quántum y valor unitario, respectivamente, garantiza el cumplimiento de la prueba de reversión de factores planteada por Fischer. Por ello, basta conocer el valor total de exportaciones o importaciones y uno cualquiera de los dos índices mencionados, para deducir inmediatamente el otro. Recuérdese que

$$\text{IPP} \times \text{IQL} = \text{IV} \qquad \text{IPL} \times \text{IQP} = \text{IV}$$

El trabajo que requiere el cálculo de estos índices induce a la utilización de muestras, en vez de indagaciones totales.

En el caso de índices de precios se supone que los precios de los artículos incluidos en la muestra representan los precios de los

no incluidos. Naturalmente, será necesario calcular los errores de muestreo correspondientes para hacer un uso racional del indicador. Por lo que a los índices de quántum se refiere, el hecho de que en todos los períodos se disponga del valor total de las importaciones, permite cuantificar la representatividad que tenga la muestra utilizada en cada período mediante el cociente.

$$\frac{\Sigma\, p_n\, q_n \text{ (para la muestra)}}{\Sigma\, p_n\, q_n \text{ (para el total)}} = \text{Representatividad}$$

Posteriormente, si se desea calcular un índice de quántum de Laspeyres, se cuantifica $\Sigma\, q_n\, p_0$ en la muestra, y para proyectar al total, se divide por la representatividad (que implica la aplicación de una regla de tres simple). De esa manera se tiene la estimación del numerador de la fórmula para el total poblacional; en cuanto al denominador no hay problema alguno, ya que para cada período se dispone del total de importaciones y exportaciones. Si se desea calcular un índice de quántum de Paasche, será necesario ajustar al 100 por ciento la expresión del denominador de la fórmula, calculada con datos muestrales.

H. ALGUNOS INDICADORES ECONÓMICOS

Se presentarán algunos indicadores de uso muy frecuente en la literatura económica. Si bien es cierto que la mayoría de ellos deberían ser analizados con mucho más profundidad desde el ángulo de la contabilidad social, es conveniente examinarlos desde el punto de vista estadístico.

1] *Indice de la relación de términos del intercambio*

Se define como el cociente entre el índice de precios de exportaciones y el índice de precios de importaciones, referidos ambos a la misma base. En consecuencia, este estadígrafo indica la evolución en el tiempo de la relación de precios registrada durante el período base; en otros términos, representa variaciones en la capacidad de compra de un volumen de exportaciones. Responde a la siguiente interrogante: ¿en qué proporción ha aumentado o disminuido el número de unidades que deben exportarse para financiar la importación del mismo volumen de productos que se introducía al país durante el año base? Se trata de un concepto estrictamente relativo; no se puede concluir que la relación

de intercambio sea buena o mala, sino mejor o peor que la del año base. El índice de la relación de precios del intercambio queda entonces definido como:

$$\text{IRI} = \frac{\text{Índice de precios de exportaciones}}{\text{Índice de precios de importaciones}} = \frac{\text{IPX}}{\text{IPM}}$$

2] *Producto e ingreso bruto*

Cuando estos conceptos se consideran en valores constantes, ambas magnitudes no coinciden. El producto real es una medida del valor de los bienes y servicios debidos al esfuerzo productivo interno. El ingreso real está relacionado, además, con el intercambio con el exterior, puesto que parte de la producción de un país se exporta y parte de los insumos y bienes finales deben adquirirse en el exterior; por consiguiente estas transacciones con el exterior afectan, por la relación de precios del intercambio, la disponibilidad de bienes y servicios que satisfacen la demanda de la población. A medida que se deteriora la relación de intercambio, hay una transferencia al exterior de una parte del esfuerzo productivo interno. En valores constantes, la diferencia entre ambos conceptos, recibe el nombre de efecto de la relación de precios del intercambio, luego:

$$\text{Ingreso bruto} = \text{Producto bruto} \pm \text{Efecto de la relación de precios del intercambio}$$

En símbolos:

$$\text{IB} = \text{PB} \pm \text{EFI}$$

donde

Efecto de la relación de precios del intercambio = Poder de compra de exportaciones − Quántum de exportaciones

Es decir

$$\text{EFI} = \text{PEX} - \text{QX}$$

Además

Poder de compra de exportaciones = Quántum de exportaciones × Índice de la relación de intercambio

$$\text{PEX} = \text{QX} \ (\text{IRI})$$

Dado que el producto del quántum de exportaciones y el índice de precios de exportaciones es equivalente al valor corriente de las exportaciones, el poder de compra también puede definirse como

$$\text{PEX} = \frac{\text{Valor corriente de las exportaciones}}{\text{índice de precios de las importaciones}}$$

en otros términos, el valor nominal o corriente de las exportaciones queda deflactado por el índice de precios de las importaciones. Por otra parte el quántum de las exportaciones es el valor de las exportaciones a precios constantes.

$$\text{QX} = \Sigma\ q_n\ p_0$$

donde la sumatoria se extiende a todos los rubros de exportación, es decir

$$\text{QX} = \frac{\text{Valor corriente de las exportaciones}}{\text{índice de precios de las exportaciones}}$$

remplazando las expresiones correspondientes en la fórmula de EFI se tiene

$$\text{EFI} = \text{QX}\ (\text{IRI}) - \text{QX} = \text{QX}\ [\text{IRI} - 1.00]$$

A continuación se cuantificarán estos conceptos mediante un ejemplo, para destacar su importancia. Supóngase que se tienen las siguientes informaciones:

	1960	*1962*	*1964*
a) Producto bruto (u. m. corrientes)	1 000	1 200	1 500
b) Deflactor implícito (base 1960 = 100)	100	115	140
c) índice de precios de exportaciones	100	110	110
d) índice de precios de importaciones	100	120	130
e) Valor corriente de las exportaciones	200	240	280
Para calcular el EFI, se tiene			
f) PB real (u. m. de 1960) (a/b)	1 000	1 043	1 071
g) IRI (base 1960 = 100) (c/d)	100	92	85
h) QX (u. m. de 1960) (e/c)	200	218	255
i) QX IRI = PEX (h.g)	200	201	217
j) EFI (i-h)	—	—17	—38
k) YB real (f-j)	1 000	1 026	1 033

3] *Capacidad para importar*

Cuando se piensa en términos de valores corrientes, la capacidad para importar no es otra cosa que el valor total de las exportaciones, más el ingreso neto de capitales extranjeros.

La capacidad para importar a precios constantes estará dada por

	Existencia de oro y divisas	EOD
+	Quántum de exportaciones	QX
+	Radicación de capitales extranjeros	RC
+	Efecto de la relación de precios del intercambio	EFI
	Capacidad total de pagos sobre el exterior	
−	Remesas de utilidades e intereses	
−	Salida de capitales extranjeros	
	Capacidad para importar	CPI

La capacidad para importar a precios constantes, también puede determinarse deflactando la capacidad para importar a precios corrientes por el índice de precios de importaciones. La capacidad para importar a precios constantes es:

$$CPI = QX + QX\,[IRI - 1.00] + RC \text{ (neta a precios constantes)}$$

$$CPI = QX\,[IRI] + \frac{RC \text{ (neta a precios corrientes)}}{IPM}$$

$$CPI = QX \cdot \frac{IPX}{IPM} + \frac{RC}{IPM}$$

$$CPI = \frac{\text{Valor corriente de las exportaciones} + EOD + RC \text{ (neta a precios corrientes)}}{\text{Índice de precios de importaciones}}$$

$$CPI = \frac{\text{Capacidad para importar a precios corrientes}}{\text{Índice de precios de importaciones}}$$

4] *Tipo de cambio de paridad*

Como una aplicación de números índice, es conveniente tratar la forma de convertir monedas de distintos países a unidades homogéneas con propósitos de comparación. Utilizar para ello

los tipos de cambio oficiales puede originar distorsiones serias, en la medida que exista más de un tipo de cambio (discriminación de áreas cambiarias), o que el tipo de cambio único esté sobre o subvaluado.

Una alternativa teórica consistiría en la determinación de una canasta de productos, resultando el tipo de cambio entre dos monedas la relación de valores que sería necesario gastar para adquirir dicha canasta. Este método tiene inconveniente de tipo práctico; la determinación de los artículos que conformarían la canasta significa un serio problema, sobre todo si se piensa en la enorme variación de las preferencias de los consumidores en los distintos países. Sin embargo, con algunas restricciones, este método es utilizado en la práctica.

Un método que puede ser utilizado con alguna ventaja es la proyección de un tipo de cambio base, en un período en el que no se haya advertido sobre o subvaluaciones monetarias, sin cambios múltiples y, en general, sin accidentes que descalifiquen a ese período base.

La aludida proyección se efectúa sobre la base de las modificaciones en la relación de precios de los dos países cuyo tipo de cambio se pretende determinar.

Sean los países A y B; se desea estimar el número de unidades monetarias de A, por unidad monetaria de B. El tipo de cambio de paridad en un año cualquiera n, estará dado por

$$\text{Tipo de cambio de paridad año n} = \text{Tipo de cambio año base} \; \frac{\text{Deflactor implícito de A}}{\text{Deflactor implícito de B}}$$

Ejemplo: Supóngase que el tipo de cambio base en el año 0, fue de 5 u. m. de A por 1 u. m. de B, siendo los deflactores implícitos los siguientes:

Año	*D. I. país A*	*D. I. país B*
0	100	100
1	120	110
2	150	120
3	200	140
4	300	150

El tipo de cambio de paridad para los años siguientes será:

Año	*Relación de deflactores* DI_A/DI_B	*Tipo de cambio de paridad*
0	100.0	5.00
1	109.1	5.45
2	125.0	6.25
3	142.9	7.14
4	200.0	10.00

Evidentemente el método da estimaciones que serán tanto más confiables en la medida que el tipo de cambio base haya sido elegido adecuadamente y los deflactores implícitos sean representativos de las variaciones de precios en ambos países.

5] *Transferencias implícitas*

El hecho que en la actividad económica se presenten transacciones intersectoriales donde cada sector tiene precios distintos, origina cierto tipo de transferencias de producto de un sector a otro, implícitas en las transacciones que efectúan cuando se contabilizan a precios corrientes.

Los sectores cuyos precios crecen menos que el promedio general de precios representado por el deflactor implícito, están transfiriendo parte de su producto hacia aquellos sectores que tienen precios que crecen más que el promedio de precios.

Evidentemente que la suma algebraica de las transferencias será nula, por cuanto la ganancia de unos sectores tiene como contrapartida la pérdida de otros.

Se definen las transferencias implícitas para cada sector, como la diferencia entre la producción valorizada a precios del sector, y esa misma producción valorizada a precios promedios representados, como se dijo, por el deflactor implícito:

Transferencias implícitas = Producción a precios corrientes − Producción a precios constantes × Deflactor implícito

en símbolos

$$T\ Imp. = p_n\ q_n - p_0\ q_n\ (DI)$$

donde n y 0 indican el período que se calcula y el período base respectivamente.

$$\sum_{h=1}^{L} T\ Imp\ (h) = \Sigma\ p_n\ q_n - \Sigma\ p_0\ q_n\ (DI) = 0$$

Donde L es el número de sectores considerados,
$h = 1, 2, 3, \ldots L$

$$\sum_{h=1}^{L} T\ Imp\ (h) = \text{Producción nominal} - \text{Producción real} \times \times \frac{\text{Producción nominal}}{\text{Producción real}} = 0$$

Ejemplo:

Sectores	*Producción nominal año 0*	*año 1*	*Índice de cantidad base 0 = 100*	*Producción real*	*Índice de precios base 0 = 100*
1	500	1 080	120	600	180
2	600	800	110	660	121
3	300	350	100	300	117
4	400	420	90	360	117
		2 650		1 920	

El deflactor implícito para el año 1 será:

$$DI = \frac{\text{Producción nominal año 1}}{\text{Producción real año 1}} = \frac{2\,650}{1\,920} = 138$$

Las transferencias implícitas por sector:

Sectores	*Producción nominal* $p_1\ q_1$	*Producción real (DI)* $p_0\ q_1$ *(DI)*	*Transferencias implícitas*
1	1 080	828	+252
2	800	911	−111
3	350	414	−64
4	420	497	−77
	2 650	2 650	0

I. ETAPAS DE LA CONSTRUCCIÓN DE NÚMEROS ÍNDICE

Es interesante e ilustrativo seguir cada uno de los pasos para la confección de indicadores acerca de precios o cantidades. Los pasos que se enumerarán estarán referidos al diseño de un índice de costo de vida por constituir éste un caso bastante general y complejo.

1] *Objetivo del índice*

Es fundamental calificar con toda precisión el objetivo del indicador. En el caso de un índice de costo de vida, es necesario determinar a quiénes estará referido: si a la población en su conjunto, a una ciudad, a profesionales, a obreros, a campesinos, etc. De esto dependerá qué tipo de artículos conformarán la muestra de productos que componen el índice. Es imprescindible no descuidar el objetivo básico: los usos principales que tendrá el indicador.

2] *Determinación de la estructura de consumo*

Es conveniente, cuando se calcula un índice de costo de vida, clasificar los gastos: alimentación, vestuario, vivienda, varios, etc. De esta manera es posible calcular índices, por separado, para cada componente. Este desglose aparte que permite realizar análisis de las variaciones de precios en forma más detallada, es muy útil en la selección de deflactores adecuados. Estos índices desglosados pueden combinarse para obtener el índice general. Las ponderaciones de cada componente se establecen mediante una muestra. Se selecciona una muestra de unidades familiares para el caso del costo de vida, obtenida de la población para la cual se confecciona el índice, por ejemplo la población de obreros y empleados de una ciudad. Esta muestra, en lo posible, debe ser aleatoria y de un tamaño que asegure fidelidad y garantice confianza. En cada una de las unidades elegidas para la muestra, se lleva un registro de los gastos en bienes y servicios por tipo de bien o servicio, donde se recopilan los precios pagados y las cantidades consumidas. Vale la pena llevar este registro durante un tiempo prolongado, por lo general un año, para que permita captar las variaciones en los consumos durante las diferentes estaciones. De esta manera es posible llegar a obtener ponderaciones por componentes y por tipos de bienes y servicios.

3] *Selección de artículos*

En las anotaciones que se hacen en los registros o "libretas de consumo" acerca de los distintos bienes que se consumen, en general aparece una gran cantidad de productos, aunque muchos de ellos responden a consumos esporádicos o accidentales. Es preciso seleccionar los productos más importantes, de consumo habitual y representativos de las preferencias de los consumidores. Para seleccionar estos productos y servicios no es conveniente un método aleatorio, pues es preferible decidir qué productos serán considerados en el índice, tomando en cuenta el volumen del gasto efectuado en cada bien o servicio. En esta forma se llegará a establecer una lista de 100 a 300 productos cuya importancia relativa o ponderación se tiene tabulada.

4] *Formas de valuación*

Otra etapa importante durante el proceso que permite confeccionar índices de precios, es la decisión acerca de los tipos de precios que se considerará. Sabido es que los precios varían enormemente según el lugar donde se adquieren los productos; por otra parte es corriente que para artículos considerados de primera necesidad, hayan precios oficiales y precios reales con grandes diferencias entre sí. Para el caso de índices de costo de vida los precios debieran tomarse en los lugares donde el consumidor adquiere los productos: almacenes, ferias, etc. Sobre el particular es necesario tomar decisiones previas en forma categórica.

5] *Variaciones de calidad de los bienes y servicios*

Otro aspecto fundamental es la especificación precisa, hasta donde sea posible, de la calidad de los productos, de manera que se pueda controlar a través del tiempo la invariabilidad de sus características. Es muy frecuente, sobre todo en los artículos controlados, que se "incrementen" los precios reales, permaneciendo fijos los precios nominales, por una disminución de la calidad de los productos.

6] *Base del índice*

En general cuando se planteó el tema se advirtió sobre la necesidad de elegir como base un período donde estuvieran ausentes las modificaciones y circunstancias violentas. Ahora debe agregarse que, en el caso de índices de costo de vida, el período base debe ser modificado, toda vez que la estructura de consumo haya

cambiado significativamente. En los tiempos actuales, cuando las innovaciones tecnológicas cambian con extraordinaria rapidez, determinando modificaciones en las preferencias de los consumidores, la base de un índice de costo de vida no debería tener una antigüedad superior a los 6 a 8 años.

Sin embargo, en las fórmulas de cálculo, para evitar cambios de base muy seguidos, modificaciones que significan altos costos, se contempla la posibilidad de introducir nuevos artículos, toda vez que se registren cambios en las preferencias de los consumidores.

7] Elección de los métodos de cálculo

En general es necesario tomar decisiones sobre la fórmula de cálculo; corrientemente, cuando se trata de índices de costo de vida se acostumbra utilizar fórmulas de Laspeyres, porque implica considerar constantes las ponderaciones del período base. La utilización de una fórmula de Paasche significaría un esfuerzo muy grande, ya que en cada período (generalmente cada mes) habría que volver a ponderar "hacia atrás", aparte que sería necesario calcular periódicamente estas ponderaciones a medida que pasa el tiempo.

La decisión acerca de qué fórmula utilizar estará condicionada, en primer lugar, al objetivo que se persigue con el índice y a la factibilidad de su diseño desde el punto de vista del costo y la oportunidad con que se entreguen los resultados.

La Dirección de Estadística y Censos de algunos países calcula el índice de costo de vida sobre la base de una modificación a la fórmula de Laspeyres.

$$I_i = I_{i-1}\left(\frac{\Sigma\, q_0\, p_{i-1}\left(\frac{p_i}{p_{i-1}}\right)}{\Sigma\, q_0\, p_{i-1}}\right)$$

Esta fórmula implica el cálculo del índice en un período i, sobre la base del índice en el período anterior, afectándolo por la variación que acusen los precios. La fórmula tiene la ventaja que se pueden introducir nuevos artículos según sus especificaciones, o cambiar la fuente de información.

TEMAS DE DISCUSIÓN

Indique si las siguientes afirmaciones son ciertas o falsas, y justifique su opinión.

1] Poder de compra, valor real y quántum físico, son conceptos equivalentes.

2] La relación de intercambio es desfavorable para los países subdesarrollados, porque los precios de sus importaciones son mayores que los precios de sus exportaciones.

3] La fórmula de Laspeyres no proporciona indicaciones correctas, cuando se trata de averiguar la variación de los valores unitarios de las exportaciones.

4] Deflactar una serie de valores nominales por un índice de precios de Laspeyres, equivale a proyectar el valor del año base ($\Sigma\ p_0\ q_0$) por un índice de cantidades de Paasche.

5] Los siguientes datos son factibles. (Los índices tienen base 1960.)

$$DI_{1967} = 150$$

Transferencias implícitas del sector primario al resto de la economía 200. Índice de precios en 1967 del sector primario 140.

6] El deflactor implícito es único, independiente de la restricción planteada.

7] Las pruebas planteadas por Fischer significan una seria limitación a la utilización de índices de Laspeyres y Paasche.

8] Los siguientes datos son consistentes

$$EFI = -200 \qquad QX = 4\,000 \qquad IPX = 150 \qquad IRI = 80$$

9] Si los índices de precios de Laspeyres y Paasche coinciden en valor numérico, para cierto período, quiere decir que la estructura de ponderaciones es exactamente igual a la del período base.

10] Los índices de precios de Laspeyres no pueden tomar valores negativos.

11] El índice de valor debe ser siempre mayor que el índice de precios.

12] Los siguientes datos son consistentes.

$$PEX = 3\,000 \qquad QX = 2\,400 \qquad IRI = 80$$

13] Si el producto nominal coincide en tres períodos con el producto real, quiere decir que los precios de todos y cada uno de los bienes y servicios han permanecido invariables en los tres períodos.

14] En un período caracterizado por una completa estabilidad de precios, los índices de Laspeyres y Paasche para precios coincidirán necesariamente en valor numérico.

15] Si una economía se divide en dos sectores: primario y secundario, y el índice de precios de productos primarios coincide con el deflactor implícito, quiere decir que las transferencias implícitas de ingresos, son nulas para ambos sectores.

16] La magnitud del efecto de la relación de precios del intercambio, es independiente del período elegido como base de los índices deflactores.

17] Para que el índice de producción industrial refleje cambios reales, debiera ser deflactado por un índice de precios de bienes industriales.

18] El índice de precios de Laspeyres es el que más se presta para el cálculo de variaciones en el costo de vida, principalmente porque toma como ponderaciones las cantidades de un período base considerado como normal.

19] Los índices de base variable, tienen la desventaja de que no puede establecerse comparaciones entre ellos.

20] Si el índice de la relación de precios del intercambio es inferior a 100, está reflejando una situación desfavorable para el país.

21] Si un obrero en 1964 tiene un sueldo superior en 30% al de 1963 y por otra parte, el índice de precios al consumidor para 1963 con base en 1964 es de 70, quiere decir que la situación del obrero en cuanto a poder de compra no ha variado.

PROBLEMAS PROPUESTOS

1] Una fábrica produce 2 bienes (A y B). Si se tienen los siguientes datos incompletos.

Año	*Ventas totales en u. m. corrientes*	*Bien "A"*		*Bien "B"*	
		Precio	*Cantidad*	*Precio*	*Cantidad*
1961	5 000	10	400	100	?
1966	9 000	15	?	400	?

Calcule un índice de cantidad de las ventas de esta fábrica para 1966, con base 1961.

2] Dadas las siguientes informaciones:

Producto nominal del país A en 1968	6 000
Producto nominal del país B en 1968	180 000

Unidades monetarias de A por una unidad monetaria de B en 1962	10
Producto real de A en 1968 en u. m. de 1964	4 000
Producto real de B en 1968 en u. m. de 1962	120 000

El índice general de precios con base variable para el país A, se presenta a continuación:

País A	
1961	—
1962	110
1963	120
1964	110

Se le pide estimar el tipo de cambio de paridad en 1968. Señale brevemente las limitaciones del método.

3] Una economía se divide en tres sectores: primario, secundario y terciario, y se dispone de los siguientes datos:

Producto bruto
(u. m. corrientes)

	1964	*1967*
Sector primario	2 000	4 000
Sector secundario	5 000	6 000
Sector terciario	3 000	5 000

Producto bruto
(u. m. constantes de 1964)

	1964	*1967*
Sector primario	2 000	2 500
Sector secundario	5 000	5 000
Sector terciario	3 000	4 000

Se pide determinar las transferencias implícitas de ingreso entre los sectores.

4] Los índices de precios y producción industrial han sido los siguientes:

Índices	*1958*	*1959*	*1960*	*1965*
Precios	80	100	150	300
Producción	100	110	120	125

Si se sabe que el producto de la industria fue en 1960 de 800 millones de u. m. se pide:

a) Estimar el producto nominal para 1965.
b) Estimar el producto real para 1959, en u. m. de 1965
c) Señalar las limitaciones de los métodos utilizados.

EJERCICIOS

1] Para los artículos A, B, C y D, se tiene los siguientes precios y cantidades en los años que se indican:

	Artículos							
	A		*B*		*C*		*D*	
Años	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*
1959	10	12	4	15	1	10	30	10
1960	12	12	4	15	1	15	30	15
1961	15	20	5	10	2	20	50	20
1962	20	20	5	10	2	30	50	20
1963	30	30	6	15	2	50	60	20

Calcule para todo el período:

a) Un índice de precios de Laspeyres con base 1959;
b) Un índice de cantidad de Paasche con base 1959; y
c) Un índice de valor.

2] El índice de base variable de los precios de los productos agropecuarios muestra el siguiente comportamiento:

Años	*Índice*
1957	–
1958	104
1959	102
1960	108
1961	120
1962	150
1963	130

Calcule el índice tomando 1960 como base fija.

3] Para el índice de producción industrial se tienen los siguientes datos:

Años	*Índice de prod. indust. base 1955*	*Índice de prod. indust. base 1960*
1955	100	–
1956	106	–
1957	114	–
1958	120	–
1960	130	100
1961	–	112
1962	–	120
1963	–	130
1964	–	145

Se le pide que empalme ambos índices y señale claramente las limitaciones del método.

4] Una economía ha sido dividida en tres sectores y sus valores agregados se presentan a continuación (en millones de unidades monetarias corrientes).

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Agricultura	2 000	2 600	3 100	4 000	6 000	8 000
Servicios y otros	4 000	5 600	5 800	7 900	9 000	10 000
Industria	1 600	2 000	2 800	3 800	5 500	6 000
Producto	7 600	9 600	11 700	15 700	20 500	24 000

Por otra parte, se dispone de los siguientes índices para los mismos años.

Índice costo vida	80	100	130	150	190	210
Índice de precios al por mayor	100	120	140	170	210	220
Índice de producción industrial	100	110	125	140	170	200
Índice de precios de la construcción	90	100	120	160	200	210
Índice de sueldos y salarios	100	125	140	160	180	200

Se le pide que:

a) Calcule el producto bruto real para los seis años en unidades monetarias de 1956 y luego de 1962. Compare los resultados.

b) Calcule el índice deflactor implícito del producto bruto y explique su significado.

5] Una familia incrementa sus ahorros nominales en 360 por ciento entre 1955 y 1962. Si en el año 1962 dicha familia desea invertir sus ahorros en materiales de construcción, sólo podrá adquirir una cantidad 20 por ciento mayor que en 1955; en cambio, si destina sus ahorros a la compra de productos agrícolas, no alcanzará a obtener la cantidad que habría comprado en 1955. Pero al mismo tiempo, si invirtiera sus ahorros para comprar ambos tipos de bienos, de tal modo que gastase igual suma en cada uno, podrá adquirir una cantidad 5 por ciento superior al año 1955. Se le pide calcule el índice de precios de productos agrícolas para 1955 con base 1962. ¿Qué limitaciones tendría el método empleado?

6] Se disponen de antecedentes, sobre la base de una muestra de cuatro productos, de las exportaciones de cierto país.

	Muestra de productos de exportación								*Valor total de las exportaciones*
	A		*B*		*C*		*D*		
Años	*p*	*q*	*p*	*q*	*p*	*q*	*p*	*q*	
1958	2	10	20	6	4	12	4	2	250
1959	3	10	20	8	4	14	4	2	320
1960	3	10	25	6	4	16	6	2	420
1961	4	10	30	8	5	20	6	3	450
1962	4	12	30	8	5	20	6	3	480
1963	4	15	25	8	6	25	8	2	500

Se le pide que estime

a) Un índice de cantidad de exportaciones para el total de éstas con base 1958.

b) Un índice de Paasche para los precios de exportación con base 1959.

c) Si el índice de precios de importaciones tiene el siguiente comportamiento:

	1958	*1959*	*1960*	*1961*	*1962*	*1963*
IPM	100	120	160	240	310	430

Se le pide calcule el índice de relación de términos de intercambio con base 1959.

7] Si los precios de los cigarrillos se incrementan en 70 por ciento y, como consecuencia, el índice de costo de vida sube en 1.8 por ciento, ¿qué ponderación dentro del costo de vida tiene este bien?

SOLUCIÓN DE EJERCICIOS

1] *a)* $\text{IPL} = \dfrac{\Sigma\, p_n q_o}{\Sigma\, p_o q_o}$

Años	*A* $p_n q_o$	*B* $p_n q_o$	*C* $p_n q_o$	*D* $p_n q_o$	*Total* $\Sigma\, p_n q_o$	*Índice base 1959=100*
1959	120	60	10	300	490	100.0
1960	144	60	10	300	514	104.9
1961	180	75	20	500	775	158.2
1962	240	75	20	500	835	170.4
1963	360	90	20	600	1 070	218.4

b) $\text{IQP} = \dfrac{\Sigma\, q_n p_n}{\Sigma\, q_o p_n}$

Año	*A* $p_n q_n$	*B* $p_n q_n$	*C* $p_n q_n$	*D* $p_n q_n$	*Total* $\Sigma p_n q_n$	*Índice*
1959	120	60	10	300	490	100.0
1960	144	60	15	450	669	130.2
1961	300	50	40	1 000	1 390	179.4
1962	400	50	60	1 000	1 510	180.8
1963	900	90	100	1 200	2 290	214.0

c) El índice de valor puede obtenerse multiplicando los dos índices antes encontrados; y en virtud de la prueba de reversión de factores:

Años	*IPL*	*IQP*	*IV*
1959	100.0	100.0	100.0
1960	104.9	130.2	136.6
1961	158.2	179.4	283.8
1962	170.4	180.8	308.1
1963	218.4	214.0	467.4

2] Es necesario "desencadenar" el índice previamente:

		Índice base fija 1957=100
1957	100	100.0
1958	100 · 104	104.6
1959	100 · 104 · 102	106.1
1960	100 · 104 · 102 · 108	114.6
1961	100 · 104 · 102 · 108 · 120	137.6
1962	100 · 104 · 102 · 108 · 120 · 150	206.2
1963	100 · 104 · 102 · 108 · 120 · 150 · 130	268.1

Luego será necesario cambiar de base, dividiendo por el valor de 1960 (114.6).

Años	*Índice base fija 1960=100*
1957	87.3
1958	90.8
1959	92.6
1960	100.0
1961	120.0
1962	179.9
1963	233.9

3] Para empalmar el índice con base 1955, bastará aplicar la siguiente regla de tres:

$$\frac{130}{I_{61.55}} \qquad \frac{100}{I_{61.60}}$$

y de igual forma para todos los años. Para empalmar hacia atrás el otro índice, el procedimiento es similar. Los índices empalmados se muestran a continuación:

	Índice base 1955	*Índice base 1960*
1955	100.0	$76.9\left(100 \cdot \frac{1.00}{1.30}\right)$
1956	106.0	$81.5\left(100 \cdot \frac{1.06}{1.30}\right)$
1957	114.0	$87.7\left(100 \cdot \frac{1.14}{1.30}\right)$
1958	120.0	$92.3\left(100 \cdot \frac{1.20}{1.30}\right)$
1960	130.0	100.0
1961	145.6 (130 · 1.12)	112.0
1962	156.0 (130 · 1.20)	120.0
1963	169.0 (130 · 1.30)	130.0
1964	188.5 (130 · 1.45)	145.0

La limitación estriba en el hecho que a lo largo de cualquiera de los dos índices empalmados aparecen implícitamente dos bases: cuanto más diferentes sean los supuestos de una y otra base, tanto más defectuoso será el empalme.

4] *a)* El primer problema por resolver es la elección de los deflactores adecuados, es decir, que exista relación entre el concepto que se desea deflactar y el índice elegido como deflactor. El segundo problema consiste en expresar los deflactores elegidos, con la misma base para disponer de unidades homogéneas. Admítase que después de un estudio se ha decidido deflactar el valor agregado de la agricultura por el índice de costo de vida, el de servicios por un índice de sueldos y salarios y en industria se haría la proyección a través del índice de cantidad respectivo.
Los índices con base 1956 serán los siguientes:

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Deflactor agricultura	100	125	163	188	238	263
Deflactor servicios	100	125	140	160	180	200
Índice de produc. ind.	100	110	125	140	170	200

Los valores agregados reales, en u.m. de 1956, serán:

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Agricultura	2 000	2 080	1 902	2 128	2 521	3 042
Servicios	4 000	4 000	4 143	4 938	5 000	5 000
Industria	1 600	1 760	2 000	2 240	2 720	3 200
Producto real	7 600	7 840	8 045	9 306	10 241	11 242

Para obtener los valores agregados reales, en u.m. de 1962 será necesario expresar los deflactores con base 1962.

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Deflactor agricultura	42	53	68	79	100	111
Deflactor servicios	56	69	78	89	100	111
Índice de prod. ind.	59	65	74	82	100	118

Los valores agregados reales en u.m. de 1962 serán:

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Agricultura	4 762	4 906	4 559	5 063	6 000	7 207
Servicios	7 143	7 246	7 436	8 876	9 000	9 009
Industria	3 245	3 575	4 070	4 510	5 500	6 490
Producto real	15 150	15 727	16 065	18 449	20 500	22 706

Los resultados evidentemente serán distintos, ya que se expresan en unidades distintas; incluso, la tasa de crecimiento del producto real, por las razones citadas, será diferente.

b) **Para calcular el deflactor implícito del producto, se aplicará la**

relación $DI = \dfrac{\text{Producto nominal}}{\text{Producto real}}$

En cuanto al producto real, puede tomarse cualquiera de los ya calculados, según la base del deflactor implícito que se desee. De esta manera, si se pretende que la base sea el año 1956, se tendrá:

	1956	*1957*	*1958*	*1959*	*1962*	*1965*
Producto nominal	7 600	9 600	11 700	15 700	20 500	24 000
Producto real (u.m. de 1956)	7 600	7 840	8 045	9 306	10 241	11 242
Deflactor implícito	100.0	122.4	145.4	168.7	200.2	213.5

5] Con los datos presentados pueden construirse los siguientes índices:

	Índice de valor ahorro nominal	*Índice de cantidad construcción*	*Índice de cantidad mixto*
1955	100	100	100
1962	460	120	105

En virtud de la prueba de reversión de factores pueden obtenerse los siguientes índices de precios:

	Índices de precios construcción	*Índices de precios mixto*
1955	100	100
1962	383	438

El índice de precios mixto, será una combinación lineal de los índices de precios parciales, ponderados por el gasto:

$$\text{IP (mixto)} = \text{IP (const)}\ W_c + \text{IP (agric)}\ W_a$$

$$\text{pero } W_c = W_a = \frac{1}{2}$$

$$438 = \frac{383 + \text{IP (agric)}}{2}$$

$$\text{IP (agric)} = 876 - 383 = 493$$

IP (agric) base 1955 = 100	*IP (agric) base 1962 = 100*
100.0	20.3
493.0	100.0

Los supuestos admitidos fueron los siguientes:

i] Que los índices cumplen con la prueba de reversión de factores;
ii] Que el índice de precios mixto se obtiene a partir de un promedio ponderado de los índices parciales;
iii] Que la estructura del índice de precios agrícolas es la misma entre los años 1955 y 1962.

6] Se calculará un índice de cantidad de Laspeyres

$$IQL = \frac{\Sigma\, q_n\, p_o}{\Sigma\, q_o\, p_o}$$

Años	*A* $q_n p_o$	*B* $q_n p_o$	*C* $q_n p_o$	*D* $q_n p_o$	*Total* $\Sigma q_n p_o$	*Repre-sent. %*	$q_n p_o$ *Ajust. 100%*	*Índice base 1958=100*
1958	20	120	48	8	196	78.4	250	100
1959	20	160	56	8	244	79.8	306	122
1960	20	120	64	8	212	60.9	348	139
1961	20	160	80	12	272	88.4	308	123
1962	24	160	80	12	276	84.6	326	131
1963	30	160	100	8	298	85.2	350	140

b) $IPP = \dfrac{\Sigma\, p_n\, q_n}{\Sigma\, p_o\, q_n}$ donde 0: 1959

Años	*A* $p_o q_n$	*B* $p_o q_n$	*C* $p_o q_n$	*D* $p_o q_n$	$\Sigma p_o q_n$	$\Sigma p_n q_n$	*Índice*
1958	30	120	48	8	206	196	95.1
1959	30	160	56	8	254	254	100.0
1960	30	120	64	8	222	256	115.3
1961	30	160	80	12	282	398	141.1
1962	36	160	80	12	288	406	141.0
1963	45	160	100	8	313	426	136.1

c) El índice de la relación de intercambio está dado por

$$IRI = \frac{IPX}{IPM}$$

Para disponer de los dos índices de igual base, se efectuará un cambio de base aritmético en el IPM.

	1958	*1959*	*1960*	*1961*	*1962*	*1963*
IPX	95.1	100.0	115.4	141.1	141.0	136.1
IPM	85.0	100.0	133.3	200.0	258.3	358.3
IRI	111.9	100.0	86.6	70.6	54.6	38.0

7] Se considerarán dos grupos: uno compuesto por los cigarrillos (C) y otro compuesto por todo el resto de los bienes (R)

$$IC\ (W_C) + IR\ (W_R) = 1.018$$

$$1.7\ W_C + 1.0\ W_R = 1.018$$

$$\text{pero } W_C + W_R = 1$$

$$1.7\ W_C + 1.0\ (1 - W_C) = 1.018$$

$$0.7\ W_C = 0.018$$

$$W_C = \frac{0.018}{0.7} = 0.026$$

Los cigarrillos tienen un 2.6% de ponderación dentro del índice de costo de vida.

SEGUNDA PARTE

ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

ANÁLISIS DE REGRESIÓN

A. MÉTODO DE LOS MÍNIMOS CUADRADOS

Probablemente uno de los temas estadísticos más utilizados en la planificación es el que se refiere al análisis de regresión y correlación.

Es de extraordinaria utilidad conocer en qué forma están relacionadas las variables objeto de análisis, es decir, la función matemática capaz de representar tal relación.

Conociendo tal función, es posible estimar el comportamiento de la variable objeto de estudio, denominada variable dependiente o predictando, de acuerdo a las variaciones de otra u otras variables denominadas independientes o predictoras. De lo anterior se deduce que la regresión debe aplicarse a variables que tengan una relación lógica, es decir, que exista razonablemente dependencia entre las variables. Desde el punto de vista teórico, a cualquier par de variables puede encontrárseles una función matemática o ecuación de regresión que las relacione, pero sólo será de utilidad cuando haya una relación de causalidad entre dichas variables.

Es necesario distinguir dos etapas en el proceso de ajuste por mínimos cuadrados; por una parte, está el problema de elegir la función que relaciona en forma adecuada a las variables; por otra, la necesidad de disponer de un método que permita determinar los valores que asumen los parámetros de la ecuación de regresión. Para solucionar el problema señalado en primer lugar, pueden ser de mucha utilidad las representaciones gráficas y los análisis numéricos de las series de datos. A veces, el propósito es verificar el cumplimiento de ciertas teorías, se trate de una adaptación de teorías ya existentes, o del planteamiento de otras nuevas. En ambos casos la función sujeta a verificación ya está elegida.

Una forma de determinar los valores de los parámetros está dada por el método de los mínimos cuadrados, cuyo tratamiento se detalla para cada uno de los casos que se presentan a continuación:

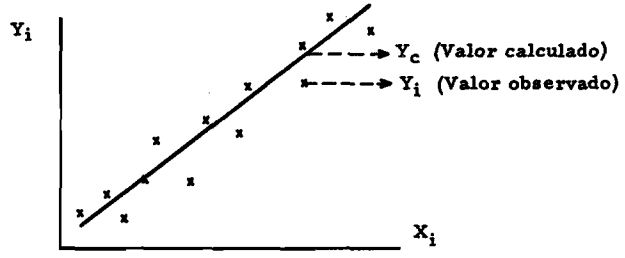
1] *Regresión simple.* Se denomina de esta manera a la metodología que permite obtener ecuaciones, donde sólo intervienen dos variables: una dependiente o predictando y otra independiente o

predictor. Cuando por medio del análisis lógico se ha comprobado la existencia de una relación de causalidad directa o indirecta entre las variables, es necesario determinar cuál es la función matemática que representa adecuadamente la relación. Para ello es indispensable disponer de informaciones acerca de los valores que ha alcanzado cada una de las variables en distintos períodos, si se trata de un análisis histórico cronológico, o en distintos lugares si se trata de un corte transversal en el tiempo. Con las informaciones obtenidas, que deben ser suficientes en número para garantizar un buen ajuste, se construirá una gráfica y se podrá decidir si la función adecuada es una recta, una hipérbola, una parábola, una potencial, una exponencial, etcétera.

Una vez que se ha decidido cuál es la función adecuada para el ajuste de regresión, es posible determinar los parámetros de la función elegida.

a) Línea recta: Si al representar los puntos en una gráfica, éstos muestran un comportamiento rectilíneo como en el ejemplo siguiente:

GRÁEICA 17



es necesario calcular los parámetros o coeficientes de regresión de dicha recta.

$$Y_c = a\,X_i + b,$$

para poder determinar los valores de a y b, se recurre al método de los mínimos cuadrados, que cumple la condición de minimizar la siguiente expresión:

$$\sum_{i=1}^{n} (Y_i - Y_c)^2$$

donde Y_i: es un valor observado;

Y_c: es un valor calculado por la ecuación de regresión;

n: es el número de observaciones.

Si se remplaza Y_c por $a X_i + b$ dentro de la sumatoria, es posible, derivando, encontrar los valores de los coeficientes de regresión a y b que satisfacen la condición. En efecto, llamemos Z a la expresión:

$$Z = \Sigma (Y_i - a X_i - b)^2$$

Se trata de derivar parcialmente respecto de cada uno de los parámetros

$$\frac{\delta Z}{\delta b} = 2 \Sigma (Y_i - a X_i - b) (-1) = 0$$

Aplicando las propiedades de la sumatoria se tiene:

$$\Sigma Y_i = a \Sigma X_i + n b$$

que es la primera ecuación normal.

$$\frac{\delta Z}{\delta a} = 2 \Sigma (Y_i - a X_i - b) (-X_i) = 0$$

Aplicando propiedades de la sumatoria

$$\Sigma Y_i X_i = a \Sigma X_i^2 + z \Sigma X_i$$

que es la segunda ecuación normal.

Obsérvese que se tienen dos ecuaciones normales y dos incógnitas. Se trata de un sistema de ecuaciones que permiten calcular los parámetros o coeficientes de regresión.

$$\left.\begin{array}{l} 1ª \text{ Ecuación normal } \Sigma Y_i = a \Sigma X_i + n b \\ 2ª \text{ Ecuación normal: } \Sigma Y_i X_i = a \Sigma X_i^2 + b \Sigma X_i \end{array}\right\} \text{Sistema}$$

Donde ΣY_i es la suma de los valores observados de la variable dependiente; ΣX_i es la suma de los valores observados de la va-

riable independiente y n es el número de observaciones. En este caso el sistema está formado por dos ecuaciones, porque sólo hay dos parámetros por determinar. El signo del coeficiente de regresión que corresponde con la pendiente de la recta (a), determina si la regresión es directa o inversa. Si "a" es positivo, quiere decir que ante incrementos de la variable predictor, corresponde incrementos de la variable predictando. Si el signo de "a" es negativo, ante incrementos de la variable predictor habrá decrementos de la variable predictando y se dice que la regresión es inversa.

Hasta el momento se estuvo planteando una regresión de "Y en X", es decir, considerando a Y como variable dependiente y a X como variable independiente, cuando se trataba de minimizar:

$$\sum_{i=1}^{n} (Y_i - Y_c)^2$$

Puede perfectamente plantearse una regresión de "X en Y", donde lo que interese minimizar sea:

$$\sum_{i=1}^{n} (X_i - X_c)^2$$

siendo $X_c = a\,Y_i + b$.

Las ecuaciones normales, en este caso, por analogía, serán:

$$\Sigma\, X_i = a\,\Sigma\, Y_i + nb$$

$$\Sigma\, X_i Y_i = a\,\Sigma\, Y_i^2 + b\,\Sigma\, Y_i$$

Téngase presente que los parámetros de la regresión de "Y en X", serán distintos de los parámetros de la regresión de "X en Y". Por ello suele distinguirse a estos parámetros de la siguiente manera:

a_{YX}: coeficiente de regresión de Y en X;

a_{XY}: coeficiente de regresión de X en Y.

En general, cuando se analiza la relación de las variables cuya regresión se pretende determinar, se puede especificar cuál es la variable dependiente y cuál la independiente. Una vez tomada la decisión, se denominará con Y_i a la variable dependiente o pre-

dictando, y con X_i a la variable independiente o predictor, para evitar confusiones. A continuación se plantea un ejemplo que permitirá aclarar algunos aspectos que son difíciles de explicar de otra manera. Como el lenguaje de los símbolos es claro, no permite malos entendidos ni interpretaciones equivocadas.

Ejemplo: Durante los últimos años las ventas de una empresa han crecido por razones de una intensa campaña de promoción de ventas; dichas variables han tenido el siguiente comportamiento en el tiempo.

Año	*Ventas* Y_i	*Gasto en propaganda* X_i
1958	100	10
1959	150	14
1960	200	21
1961	210	22
1962	300	28
1963	500	45
1964	600	55

Interesa determinar la función matemática o ecuación de regresión que relaciona estas variables. Representando estos valores en un gráfico, se concluirá que la recta representa adecuadamente la relación de las variables. Para determinar los parámetros de la recta, se plantean las ecuaciones normales

$$\Sigma Y_i = a \Sigma X_i + nb$$

$$\Sigma Y_i X_i = a \Sigma X_i^2 + b \Sigma X_i$$

Luego es necesario tabular los valores que interesa remplazar en estas ecuaciones normales; a continuación se procede a hacerlo así:

Y_i	X_i	$Y_i X_i$	X_i^2
100	10	1 000	100
150	14	2 100	196
200	21	4 200	441
210	22	4 620	484
300	28	8 400	784
500	45	22 500	2 025
600	55	33 000	3 025
2 060	195	75 820	7 055

Las ecuaciones normales en valores serán:

$$2\,060 = 195\ a + 7\ b$$
$$75\,820 = 7\,055\ a + 195\ b$$

Resolviendo el sistema

$$a \doteq 11.4$$
$$b \doteq -26.1$$

La ecuación de ajuste queda en consecuencia expresada así:

$$Y_c = 11.4\ X_i - 26.1$$

Por medio de esta ecuación se puede determinar valores calculados de la variable dependiente para cualquier valor de la variable independiente. Naturalmente que al realizar estimaciones, por ejemplo para calcular el probable volumen de ventas ante un desembolso en propaganda de 100 ($X_t = 100$), debe tenerse en cuenta el campo de validez de la regresión. No escapará a la atención del lector el hecho que aumentos sucesivos de propaganda no siempre implicarán mayores volúmenes de venta, porque puede darse en un momento determinado la saturación del mercado u otro obstáculo semejante. En consecuencia es necesario que, cuando se realicen estimaciones, se verifique el cumplimiento de los supuestos implícitos en los datos disponibles. Por ello sobre el resultado de una proyección, es indispensable advertir que sólo tendrá validez si se sigue manteniendo la tendencia de los puntos observados durante el período histórico.

b) Potencial: Una función muy utilizada en proyecciones, por su flexibilidad, es la denominada función potencial o de elasticidad. Su expresión matemática es la siguiente:

$$Y_c = b\ X_i^a$$

Para determinar las ecuaciones normales se procede en forma similar al caso de la recta, realizando previamente, mediante la aplicación de logaritmos, una transformación lineal:

$$\log Y_c = \log b + a \log X_i$$

$$\log Y_c = b' + a \log X_i \qquad \text{donde} \qquad b' = \log b$$

En este caso se trata de minimizar la expresión:

$$Z = \sum_{i=1}^{n} (\log Y_i - \log Y_c)^2$$

es decir:

$$Z = \Sigma\ (\log Y_i - a \log X_i - b')^2$$

Derivando respecto de cada uno de los parámetros e igualando los resultados a cero, se obtendrán las dos ecuaciones normales.

$$\frac{\delta Z}{\delta b'} = 2\ \Sigma\ (\log Y_i - a \log X_i - b')\ (-1) = 0$$

$$\frac{\delta Z}{\delta a} = 2\ \Sigma\ (\log Y_i - a \log X_i - b')\ (-\log X_i) = 0$$

Aplicando a ambas derivadas las propiedades de la sumatoria, se tiene:

$$\Sigma \log Y_i = a\ \Sigma \log X_i + n\ b'$$

$$\Sigma \log Y_i \log X_i = a\ \Sigma\ (\log X_i)^2 + b'\ \Sigma \log X_i$$

que forman el sistema de dos ecuaciones normales que permitirán el cálculo de los dos parámetros. Evidentemente el método es un tanto laborioso cuando se tienen muchas observaciones, ya que es necesario trabajar en los logaritmos con por lo menos 5 decimales para evitar aproximaciones que pueden implicar serios desajustes.

c) Exponencial: Cuando se desea calcular tasas de crecimiento tomando en cuenta todos los puntos observados en el período histórico se recurre principalmente a la función:

$$Y = a\ b^t \qquad \text{donde } b = 1 + i$$

t: tiempo en períodos

Aplicando logaritmos a la anterior expresión:

$$\log Y_c = \log a + t_i \log b$$

Como en los casos anteriores interesa minimizar la expresión

$$Z = \sum_{i=1}^{n} (\log Y_i - \log Y_c)^2$$

$$Z = \Sigma\ (\log Y_i - \log a - t_i \log b)^2$$

$$\frac{\delta Z}{\delta \log a} = 2 \Sigma\ (\log Y_i - \log a - t_i \log b)\ (-1) = 0$$

$$\frac{\delta Z}{\delta \log b} = 2 \Sigma\ (\log Y_i - \log a - t_i \log b)\ (-t_i) = 0$$

Aplicando las propiedades de la sumatoria, se obtienen las dos ecuaciones normales

$$\Sigma \log Y_i = n \log a + \log b \Sigma\ t_i$$

$$\Sigma\ t_i \log Y_i = \log a \Sigma\ t_i + \log b \Sigma\ t_i^2$$

El caso general de la función exponencial es el cálculo de tasas de crecimiento cuando se considera el tiempo como variable independiente. Sin embargo, puede considerarse cualquier otra variable independiente y ajustar la función sin hacer referencia a tasas de crecimiento.

En general se obtienen significativas ventajas, cuando se cambia la escala de unidades para la variable t. De este modo, si se tiene una serie con un número impar de datos, se le asigna el valor cero al período central y sucesivamente los primeros dígitos con signo positivo para períodos posteriores y con signo negativo para períodos anteriores. Con ello se consigue que la sumatoria de t sea nula, con lo cual se facilita la resolución del sistema. Cuando el número de datos es par, a los dos períodos centrales se le asignan los valores -1 y $+1$ y para períodos posteriores los primeros números impares con un signo positivo y con uno negativo para los períodos anteriores.

d) Parábola: Esta conocida función se ajusta en forma similar a los casos anteriores.

$$Y_c = a x_i^2 + b X_i + c$$

Dado que la forma general contiene tres parámetros, será necesario determinar tres ecuaciones normales para determinar los va-

lores de a, b, c. Estas tres ecuaciones normales provienen de la derivación parcial respecto de cada uno de dichos parámetros, interesa minimizar la expresión:

$$Z = \sum_{i=1}^{n} (Y_i - Y_c)^2$$

$$Z = \Sigma\ (Y_i - a\ X_i^2 - b\ X_i - c)^2$$

Derivando respecto de a, b y c, se tiene:

$$\frac{\delta Z}{\delta c} = 2\ \Sigma\ (Y_i - a\ X_i^2 - b\ X_i - c)\ (-1) = 0$$

$$\frac{\delta Z}{\delta b} = 2\ \Sigma\ (Y_i - a\ X_i^2 - b\ X_i - c)\ (-X_i) = 0$$

$$\frac{\delta Z}{\delta a} = 2\ \Sigma\ (Y_i - a\ X_i^2 - b\ X_i - c)\ (-X_i^2) = 0$$

Aplicando las propiedades de la sumatoria, se tienen las siguientes ecuaciones normales:

1ª Ecuación normal: $\Sigma\ Y_i = a\ \Sigma\ X_i^2 + b\ \Sigma\ X_i + n\ c$

2ª Ecuación normal: $\Sigma\ Y_i X_i = a\ \Sigma\ X_i^3 + b\ \Sigma\ X_i^2 + c\ \Sigma\ X_i$

3ª Ecuación normal: $\Sigma\ Y_i X_i^2 = a\ \Sigma\ X_i^4 + b\ \Sigma\ X_i^3 + c\ \Sigma\ X_i^2$

Puesto que durante el período histórico se tienen los valores de Y_i y X_i, es necesario tabular todas las sumatorias que aparecen en las ecuaciones normales. Resolviendo el sistema, se tiene determinado el valor de cada uno de los tres parámetros.

e) Hipérbola equilátera: Para el ajuste de algunas funciones de demanda y por la propiedad que tiene que cualquier punto de la función subtiende superficies iguales con los ejes de coordenadas, su aplicación es bastante frecuente. Se trata de un caso particular de la función potencial. Su expresión matemática es:

$$Y_c = \frac{a}{X_i}$$

En vista de que sólo tiene un parámetro, será necesario calcular una ecuación normal, minimizando la expresión:

$$Z = \sum_{i=1}^{n} (Y_i - Y_c)^2$$

$$Z = \Sigma \left(Y_i - \frac{a}{X_i}\right)^2$$

Derivando respecto de a

$$\frac{\delta Z}{\delta a} = 2 \Sigma \left(Y_i - \frac{a}{X_i}\right) \left(-\frac{1}{X_i}\right) = 0$$

Aplicando las propiedades de la sumatoria se tiene:

$$\text{Ecuación normal } \Sigma \frac{Y_i}{X_i} = a \Sigma \frac{1}{X_i^2}$$

f) Otras funciones: Dentro del campo de la investigación económica a veces es preciso ajustar funciones particulares. La metodología de la obtención de ecuaciones normales es similar a los casos considerados. Por ejemplo la función:

$$Y_c = a \log X_i + b$$

Siempre se tratará de minimizar la expresión

$$Z = \sum_{i=1}^{n} (Y_i - Y_c)^2$$

$$Z = \Sigma (Y_i - a \log X_i - b)^2$$

Donde:

$$\frac{\delta Z}{\delta b} = 2 \Sigma (Y_i - a \log X_i - b) (-1) = 0$$

$$\frac{\delta Z}{\delta a} = 2 \Sigma (Y_i - a \log X_i - b) (-\log X_i) = 0$$

Las ecuaciones normales serán:

$$\Sigma Y_i = a \Sigma \log X_i + nb$$

$$\Sigma Y_i \log X_i = a \Sigma (\log X_i)^2 + b \Sigma \log X_i$$

Resolviendo el sistema, es posible determinar el valor de los parámetros.

Siempre es conveniente seguir esta metodología, para funciones cuyas derivadas no compliquen demasiado las expresiones que aparecen en las ecuaciones normales.

2] *Regresión múltiple.* Ocurre que a veces es necesario encontrar funciones donde se relacionen una variable dependiente y dos o más variables independientes, de allí el calificativo de múltiple. En este caso se adoptará una simbología especial. para designar cada una de las variables y parámetros:

X_1: variable dependiente;

X_2, X_3 ... X_p: variables independientes.

Por lo tanto, si se trata de un caso de regresión múltiple donde se consideren dos variables independientes, la función se expresará de la siguiente manera:

$$X_{c\ 1.23} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

donde:

$X_{1.23}$: indica la variable dependiente X_1 que se relaciona con las variables X_2 y X_3; esa es la razón de los subíndices.

$a_{1.23}$: coeficiente de posición (término libre) del plano de regresión donde se consideran la variable dependiente X_1 y las variables independientes X_2 y X_3.

$b_{12.3}$: coeficiente de regresión que multiplica a la variable X_2, cuando además se considera la variable X_3.

$b_{13.2}$: coeficiente de regresión que multiplica a la variable X_3, cuando además se considera la variable X_2.

Es fácil extender esta notación para los casos en que se consideren 3 o más variables independientes. En el caso de tres variables independientes (X_2, X_3, X_4) la función quedará así simbolizada:

$$X_{c\,1.234} = a_{1.234} + b_{12.34}\,X_2 + b_{13:24}\,X_3 + b_{14.23}\,X_4$$

El lector, por analogía con el caso anterior, puede interpretar cada uno de estos símbolos.

Cuando se desea ajustar una función de este tipo a una serie de datos, el método de los mínimos cuadrados implica hacer mínima la expresión.

$$\sum_{i=1}^{n} (X_1 - X_{c\,1.23})^2$$

donde X_1 son los valores observados y $X_{c\,1.23}$ son los valores calculados de la variable dependiente. Por simplificaciones, se suprime el subíndice i. Las ecuaciones normales, en el caso de dos variables independientes, se obtienen minimizando la siguiente expresión:

$$Z = \Sigma\ (X_1 - a_{1.23} - b_{12.3}\,X_2 - b_{13.2}\,X_3)^2$$

Para ello, se deriva parcialmente respecto de cada uno de los parámetros, igualando los resultados a cero.

$$\frac{\delta Z}{\delta a_{1.23}} = 2\,\Sigma\ (X_1 - a_{1.23} - b_{12.3}\,X_2 - b_{13.2}\,X_3)\ (-1) = 0$$

$$\frac{\delta Z}{\delta b_{12.3}} = 2\,\Sigma\ (X_1 - a_{1.23} - b_{12.3}\,X_2 - b_{13.2}\,X_3)\ (-X_2) = 0$$

$$\frac{\delta Z}{\delta b_{13.2}} = 2\,\Sigma\ (X_1 - a_{1.23} - b_{12.3}\,X_2 - b_{13.2}\,X_3)\ (-X_3) = 0$$

Aplicando las propiedades de la sumatoria se tienen las siguientes tres ecuaciones normales que formarán el sistema para calcular el valor de cada uno de los tres parámetros.

$$\Sigma\,X_1 = b_{12.3}\,\Sigma\,X_2 + b_{13.2}\,\Sigma\,X_3 + n\,a_{1.23}$$

$$\Sigma\,X_1\,X_2 = b_{12.3}\,\Sigma\,X_2^2 + b_{13.2}\,\Sigma\,X_2\,X_3 + a_{1.23}\,\Sigma\,X_2$$

$$\Sigma\,X_1\,X_3 = b_{12.3}\,\Sigma\,X_2\,X_3 + b_{13.2}\,\Sigma\,X_3^2 + a_{1.23}\,\Sigma\,X_3$$

Tabulando los valores de las sumatorias que aparecen en el sistema, se podrá resolver para cada parámetro.

B. CONSIDERACIONES PRÁCTICAS

En páginas anteriores se ha detallado la metodología que permite obtener ecuaciones normales por el método de los mínimos cuadrados, para el tipo de funciones usado con más frecuencia en la planificación. Ahora se pretende enunciar algunas de las consideraciones que es necesario hacer, desde el punto de vista práctico, cuando se realizan los mencionados ajustes.

1] *Respecto del tipo de función.* Si se piensa que una de las principales aplicaciones de la regresión es la proyección, en el tiempo o en el espacio, donde no se tienen valores de la variable estudiada, y donde no queda otra alternativa que conformarse con estimaciones provenientes de extrapolación de funciones ajustadas por regresión, deberá admitirse la necesidad de disponer de funciones sencillas que contengan un reducido número de variables y parámetros. Recuérdese que una función complicada, de muchas variables y parámetros, se parecerá más bien a una interpolación, a una función que se aproximará al mayor número de puntos observados. Para determinar tendencia no tiene sentido la interpolación. Recuérdese que para proyectar una variable dependiente, es necesario disponer de estimaciones para todas las variables independientes; pero disponer de estimaciones para muchas variables independientes suele ser en extremo difícil y en todo caso existe alta probabilidad de cometer errores. En cambio una función sencilla, como las analizadas en páginas anteriores, puede representar cabalmente una tendencia de la relación de la variable dependiente con la o las variables independientes.

2] *Respecto del número de observaciones.* Un buen ajuste implica disponer de una cantidad significativa de puntos observados; el conjunto de puntos observados representa una muestra de la relación de las variables en el tiempo o en el espacio. Mientras más grande esta muestra, es decir, mientras mayor número de puntos se posea, tendrá más representatividad y menor será la probabilidad de cometer errores. Cuando se está analizando una ecuación de regresión, una de las primeras cuestiones que se debe aclarar será el número de observaciones, para que con este antecedente se califique en parte la significación de la regresión.

3] *Respecto de la dificultad del cálculo.* El lector comprobará, a través de la realización de ejercicios, el trabajo que exigen los

cálculos de regresión. En la práctica, cuando ya se tiene aclarada la parte conceptual, para lo cual constituye una importante ayuda la realización de ejercicios con calculadoras convencionales, será útil recurrir a los computadores electrónicos, ya que una vez entregadas las informaciones originales, en brevísimo tiempo podrá disponerse de cálculos exactos, ya que los programas de regresión están previamente diseñados. Por otra parte, respecto de la deducción de las ecuaciones normales, puede significar cierta demora obtenerla basándose sobre las derivadas parciales. Existe una regla nemotécnica para hallar ecuaciones normales en funciones lineales respecto de los parámetros. La regla es la siguiente: Para la primera ecuación normal, multiplíquese la función a ajustar por el coeficiente del primer parámetro, y luego aplíquese el operador sumatoria a la función. Para la segunda ecuación, multiplíquese toda la función por el coeficiente del segundo parámetro y luego aplíquese el operador sumatoria. Y así sucesivamente para todas las ecuaciones normales que se deba obtener.

Ejemplos: Se obtendrán las ecuaciones normales de la función:

$$\log Y_c = a \log X_i + \log b$$

Multiplicando ambos miembros de la ecuación por el coeficiente de log b que es 1, y aplicando sumatoria, se tiene la primera ecuación normal:

$$\Sigma \log Y_i = a \Sigma \log X_i + n \log b$$

Multiplicando ambos miembros de la ecuación por log X que es el coeficiente del otro parámetro y aplicando sumatoria, se tiene:

$$\Sigma \log Y_i \log X_i = a \Sigma (\log X_i)^2 + \log b \Sigma \log X_i$$

que es la segunda ecuación normal. Comparando estas dos ecuaciones normales con las obtenidas por derivación parcial en la parte 1, b se concluye que son idénticas.

Si se quisiera obtener la ecuación normal de una recta que pasa por el origen, se tiene:

$$Y_c = a X_i$$

(recta que pasa por el origen ya que no tiene término libre; coeficiente de posición 0).

Multiplicado por X_i, que es el coeficiente del único parámetro, y aplicando sumatoria, se tiene:

$$\Sigma Y_i X_i = a \Sigma X_i^2$$

Por el proceso de derivación, se llega al mismo resultado. En efecto:

$$Z = \sum_{i=1}^{n} (Y_i - Y_c)^2 = \Sigma (Y_i - aX_i)^2$$

$$\frac{\delta Z}{\delta a} = 2 \Sigma (Y_i - aX_i) (-X_i) = 0$$

$$\Sigma Y_i X_i = a \Sigma X_i^2$$

En las páginas siguientes se tratarán conceptos referentes al análisis de correlación, conceptos que permitirán cuantificar el grado de asociación entre las variables estudiadas y la validez de las proyecciones a través de las ecuaciones de regresión.

II

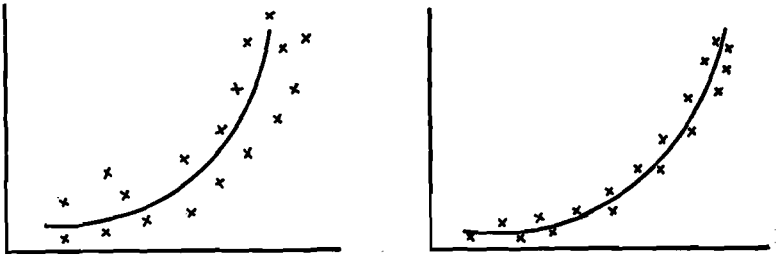
CORRELACIÓN

A. OBJETIVOS DEL ANÁLISIS DE CORRELACIÓN

En el capitulo anterior se presentaron las técnicas del ajuste de funciones por el método de mínimos cuadrados. Una vez determinada la función, es necesario especificar *si hay asociación entre las variables consideradas* y *en qué medido lo están.* En caso de que las variables estén íntimamente asociadas, la ecuación de regresión puede utilizarse para explicar el comportamiento de la variable dependiente (explicada) en términos de las variaciones que experimente la variable independiente (explicativa). Por ejemplo, el incremento del volumen de venta de artefactos eléctricos puede ser explicado por aumentos en los niveles de ingreso, por variaciones en los precios, por modificaciones en los tipos de cambio, etc. Por otra parte, el instrumento de la regresión y correlación puede ser empleado en la estimación de valores de la variable dependiente (predictando), en el entendido que se conocen las variaciones de la variable independiente (predictor). En general, los planes de desarrollo especifican los niveles de ingreso por habitante que se pretende alcanzar en los próximos períodos; con tales datos y la ecuación de regresión del caso, pueden estimarse magnitudes de las variables que muestren un alto grado de asociación con el ingreso, tales como el consumo, la importación de alimentos, la reinversión de utilidades, etc. En todo caso, la validez de una proyección por regresión depende del grado en que están asociadas entre sí las variables; si es alto el grado de la asociación, la estimación tiene base de fundamento, y si la asociación es débil, la proyección no se justifica.

Recuérdese que para determinar la ecuación de regresión es necesario contar con antecedentes sobre los valores que han tomado las variables; la representación gráfica de estos valores ayuda a especificar el tipo de función. En esta etapa ya puede adelantarse algo acerca del grado de asociación. Obsérvese los dos diagramas siguientes:

GRÁFICA 18



En el primer diagrama de dispersión los puntos están más alejados de la función que en el segundo; la proximidad de los puntos observados a la función determina el grado de asociación.

El objetivo básico del análisis de correlación es pues evidente: se trata de disponer de un indicador cuantitativo del grado de asociación que respalde la ecuación de regresión que se pretende utilizar. De hecho, un conjunto de puntos que muestren la relación de un par de variables puede ser representada por cualquier función, pero una representación adecuada sólo se consigue cuando la garantiza una asociación estrecha entre las variables.

B. TIPOS DE CORRELACIÓN

En forma similar a la clasificación de los tipos de regresión presentada en el capítulo anterior, se puede distinguir los siguientes tipos de correlación:

1] Atendiendo al número de variables:
   a) Correlación simple. Cuando se estudia el grado de asociación entre un par de variables: dependiente e independiente.
   b) Correlación múltiple. Cuando se estudia el grado de asociación que simultáneamente existe entre la variable dependiente y dos o más variables independientes.
   c) Correlación parcial. En el caso de correlación múltiple, la cuantificación de la asociación neta entre dos variables, una vez que se elimina estadísticamente la influencia de otras variables independientes.

2] Atendiendo a la forma de la función: Según el tipo de ecua-

ción de regresión se tiene correlación rectilínea, parabólica, potencial, exponencial, logarítmica, etcétera.

3] Atendiendo a la relación de variables:

a) Correlación directa o positiva. Cuando por aumentos en la variable independiente corresponden aumentos de la variable dependiente.

b) Correlación inversa o negativa. Cuando por aumentos en la variable independiente corresponden disminuciones de la variable dependiente.

C. EL COEFICIENTE DE CORRELACIÓN

Definición. Un coeficiente de correlación *indica el grado de asociación* entre las variables; se simbolizará por r y se definirá de la siguiente manera:

$$r = \left(\frac{S^2_{Yc}}{S^2_Y}\right)^{1/2} = \frac{S_{Yc}}{S_Y}$$

Donde S^2_{Yc} representa la varianza explicada, es decir, aquella parte de la varianza total explicada por la ecuación de regresión y S^2_Y representa la varianza total como se la definió en la primera parte del trabajo, es decir:

$$S^2_{Yc} = \frac{\Sigma\,(Y_c - \bar{Y})^2}{n} \qquad (Y_c\text{: valor calculado})$$

$$S^2_Y = \frac{\Sigma\,(Y_i - \bar{Y})^2}{n} \qquad (Y_i\text{: valor observado})$$

Como puede observarse, ambas varianzas expresan un promedio de cuadrados de desviaciones respecto de la media aritmética y su cómputo no difiere del que se realiza para una varianza cualquiera. Lo que ocurre es que la variabilidad total se descompone en *dos fuentes:* la varianza explicada y la varianza no explicada que se define:

$$S^2_{Ys} = \frac{\Sigma\,(Y_i - Y_c)^2}{n}$$

Lógicamente la suma de la varianza explicada y la varianza no explicada reproduce la varianza total, como se demuestra más adelante. La raíz de la varianza no explicada, por el hecho de ser un indicador del grado de dispersión de los puntos observados respecto de los puntos calculados por la ecuación de regresión, recibe el nombre de *error de proyección* y se utiliza para fijar intervalos de confianza.

Observando la fórmula del coeficiente de correlación, éste puede interpretarse como la proporción que representa la desviación típica explicada dentro de la desviación típica total.

D. LIMITACIONES DE LA CORRELACIÓN

La rapidez y sencillez con que ha sido presentado el tema puede hacer que la correlación se interprete sin salvedades y con ilimitados alcances; por ello parece conveniente plantear los siguientes puntos.

1] Un alto coeficiente de correlación no necesariamente determina *causalidad* entre las variables; dos variables pueden aparecer correlacionadas por casualidad y no porque exista una relación de dependencia entre ellas.

2] En cuanto a las variables, es necesario que aparezcan *depuradas* de las influencias de otras variables. Dos series nominales pueden mostrar estrecha asociación porque hay una tercera variable: alzas de precios, que exagera el grado de asociación. Por ello es conveniente trabajar con series reales, por habitante, de manera que haga más significativa la correlación.

3] Dos series pueden también arrojar coeficientes de correlación cercanos a uno, porque el *tamaño de muestra es insuficiente.* En un caso extremo, cuando sólo se tomen dos puntos, el coeficiente de correlación rectilíneo mostrará en general un valor igual a la unidad, pero esto no garantiza la adecuada significación. La calificación del grado de asociación no puede dejar de considerar el *número de puntos* utilizados en el estudio.

4] Desde el punto de vista del tipo de función, sobre todo cuando se tiene por objetivo la proyección de una variable, es conveniente trabajar con *funciones sencillas* capaces de representar la tendencia de la nube de puntos. Si se posee una función complicada con muchos parámetros y muchas variables independientes, posiblemente se obtenga un alto coeficiente de correlación, porque la función, dada su complejidad, pasará muy cerca de los puntos observados. Sin embargo, la correlación pierde validez como ga-

rantía de una adecuada proyección; estimar los valores de las variables independientes, es decir, fijar las variables exógenas se hace más difícil cuando éstas son numerosas.

5] No debe olvidarse que la proyección por regresión y correlación es válida en tanto *sigan en vigencia* los supuestos y circunstancias implícitos en los datos y antecedentes disponibles. Proyectar por regresión la producción agrícola de los próximos períodos, por ejemplo, haciendo caso omiso de una eventual reforma agraria, probablemente conducirá a estimaciones alejadas de la realidad. Es importante, cuando se realizan estimaciones, dejar en claro los supuestos básicos y admitir que cualquier desviación de estos supuestos *exige una revisión* del modelo de proyección o del modelo de análisis según el caso.

6] Por último, merece destacarse que los modelos de regresión y correlación significan una *permanente revisión de supuestos* y acumulación de nuevos antecedentes que permitan ajustar el modelo a las nuevas circunstancias.

E. CORRELACIÓN RECTILÍNEA

Es conveniente presentar en detalle los conceptos expuestos en forma general aplicados al caso específico de la correlación rectilínea.

1] Representación de las magnitudes que determinan las varianzas

Se enunció que la varianza denominada total correspondía exactamente con el concepto utilizado en la primera parte del trabajo. En efecto:

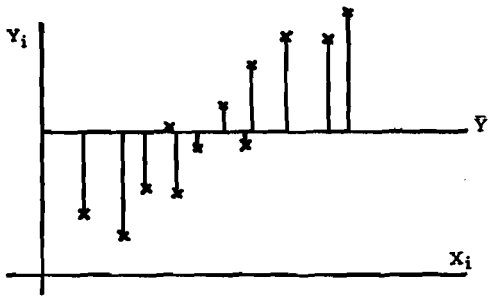
$$S_Y^2 = \frac{\Sigma\ (Y_i - \overline{Y})^2}{n}$$

En la gráfica 19 pueden observarse las desviaciones que toma en cuenta este estadígrafo.

La varianza explicada la determinan las desviaciones de los valores calculados respecto de la media aritmética.

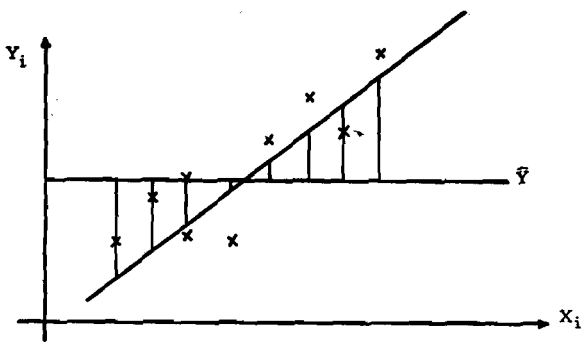
$$S_{Yc}^2 = \frac{\Sigma\ (Y_c - \overline{Y})^2}{n}$$

GRÁFICA 19



Gráficamente:

GRÁFICA 20

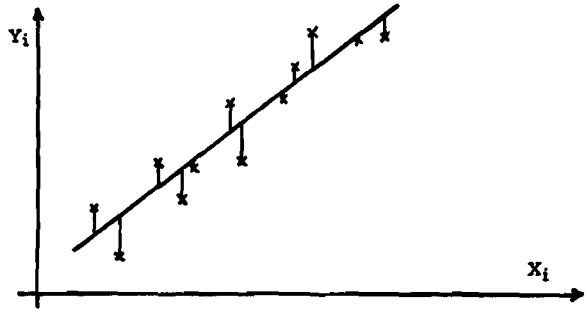


La varianza no explicada la determinan las desviaciones de los valores observados respecto de los valores calculados.

$$S^2_{Ys} = \frac{\Sigma\ (Y_i - Y_c)^2}{n}$$

Gráficamente:

GRÁFICA 21



Evidentemente:

$$S_Y^2 = S_{Yc}^2 + S_{Ys}^2$$

ya que:

$$\Sigma\ (Y_i - \overline{Y})^2 = \Sigma\ (Y_c - \overline{Y})^2 + \Sigma\ (Y_i - Y_c)^2$$

$$\Sigma Y_i^2 - 2\,\overline{Y}\,\Sigma\,Y_i + n\,\overline{Y}^2 = \Sigma\,Y_c^2 - 2\,\overline{Y}\,\Sigma\,Y_c + n\,\overline{Y}^2 + \Sigma\,Y_i^2 - 2\,\Sigma\,Y_i\,Y_c + \Sigma\,Y_c^2$$

Por otra parte $\Sigma\,Y_i = \Sigma\,Y_c$ ya que:

$\Sigma\,Y_i = a\,\Sigma\,X_i + n\,b$ 1ª ecuación normal.

$Y_c = a\,X_i + b$ Ecuación de regresión rectilínea, que al aplicarle el operador sumatoria, se transforma en:

$$\Sigma\,Y_c = a\,\Sigma\,X_i + n\,b$$

Luego

$$\Sigma\,Y_c = \Sigma\,Y_i$$

La relación original en consecuencia puede simplificarse:

$$2\,\Sigma\,Y_i\,Y_c - 2\,\overline{Y}\,\Sigma\,Y_i = 2\,\Sigma\,Y_c^2 - 2\,\overline{Y}\,\Sigma\,Y_c$$

pero, $\Sigma\,Y_i = \Sigma\,Y_c$

y queda: $\Sigma\, Y_i\, Y_c = \Sigma\, Y_c^2$

Pero: $Y_c = a\, X_i + b$

$$Y_c^2 = a^2\, X_i^2 + 2ab\, X_i + b^2$$

$$Y_c^2 = a^2\, X_i^2 + ab\, X_i + ab\, X_i + b^2$$

$$Y_c^2 = a\, (a\, X_i^2 + b\, X_i) + b\, (a\, X_i + b)$$

$$\Sigma\, Y_c^2 = a\, (a\, \Sigma\, X_i^2 + b\, \Sigma\, X_i) + b\, (a\, \Sigma\, X_i + n\, b)$$

Las expresiones entre parántesis son las ecuaciones normales de una recta, luego

$$\Sigma\, Y_c^2 = a\, \Sigma\, X_i\, Y_i + b\, \Sigma\, Y_i$$

Por otra parte

$$\Sigma\, Y_i\, Y_c = \Sigma\, Y_i\, (a\, X_i + b)$$

ya que

$$Y_c = a\, X_i + b$$

$$\Sigma\, Y_i\, Y_c = a\, \Sigma\, X_i\, Y_i + b\, \Sigma\, Y_i$$

Luego

$$\Sigma\, Y_i^2 = \Sigma\, Y_i\, Y_c$$

y por consiguiente

$$S_Y^2 = S_{Yc}^2 + S_{Ys}^2$$

De esto se deduce que el valor numérico del coeficiente de correlación, o de su cuadrado que se denomina coeficiente de determinación, fluctúa entre 0 y 1.

$$0 \leqslant r^2 \leqslant 1$$

$$0 \leqslant r \leqslant 1$$

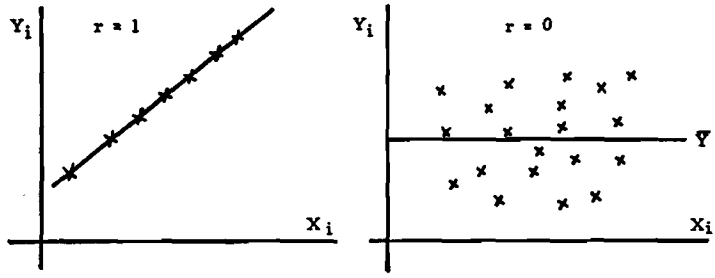
Los límites planteados son casos generales. Cuando no se elige en forma adecuada la ecuación de regresión, es posible encontrar coeficientes de correlación que adoptan valores no comprendidos entre los límites establecidos.

En correlación rectilínea se asigna el signo positivo de la raíz cuando se trata de correlación directa y el signo negativo si la correlación es inversa. En este caso, entonces los límites serán:

$$-1 \leqslant r \leqslant 1$$

El coeficiente de correlación tomará un valor igual a la unidad cuando *todos los puntos* observados estén situados sobre la ecuación de regresión, y tomará el valor cero cuando la ecuación de regresión *coincida con una paralela* al eje de las abscisas a la altura de la media aritmética.

GRÁFICA 22



2] Método abreviado de cálculo

El cálculo del coeficiente de correlación basado en las varianzas, es decir, en la definición, implica cuantificar los valores calculados por la ecuación de regresión, lo que por sí mismo representa un trabajo bastante laborioso. El método que a continuación se expone, aprovecha los cálculos que se debieron realizar para determinar los parámetros de la ecuación de regresión. Si se recuerda la definición:

$$r^2 = \frac{S^2_{Yc}}{S^2_Y} = \frac{\Sigma\,(Y_c - \overline{Y})^2}{\Sigma\,(Y_i - \overline{Y})^2}$$

$$= \frac{\Sigma Y_c^2 - 2\bar{Y}\Sigma Y_c + n\bar{Y}^2}{\Sigma Y_i^2 - 2\bar{Y}\Sigma Y_i + n\bar{Y}^2}$$

Obsérvese que $\Sigma Y_c = \Sigma Y_i = n\bar{Y}$

$$= \frac{\Sigma Y_c^2 - 2n\bar{Y}^2 + n\bar{Y}^2}{\Sigma Y_i^2 - 2n\bar{Y}^2 + n\bar{Y}^2} = \frac{\Sigma Y_c^2 - n\bar{Y}^2}{\Sigma Y_i^2 - n\bar{Y}^2}$$

Será necesario encontrar una expresión para ΣY_c^2. En el punto anterior se demostró que

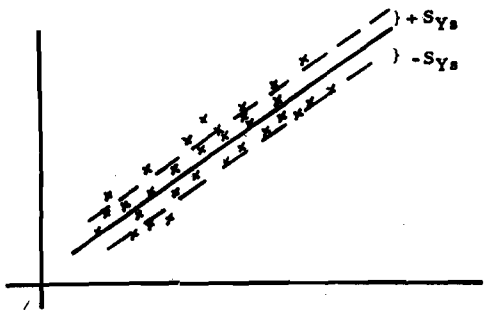
$$\Sigma Y_c^2 = a \Sigma X_i Y_i + b \Sigma Y_i$$

Remplazando esta nueva expresión en la última fórmula de r^2 se tiene:

$$r^2 = \frac{a\Sigma X_i Y_i + b\Sigma Y_i - n\bar{Y}^2}{\Sigma Y_i^2 - n\bar{Y}^2}$$

La fórmula anterior, como se dijo, posee la ventaja de utilizar cómputos que debieron hacerse para el ajuste por el método de los mínimos cuadrados, con excepción de ΣY_i^2. El numerador de esta fórmula equivale a n veces la varianza explicada, y el denominador equivale a n veces la varianza total. Por consiguiente, para obtener el cuadrado del error de proyección (varianza no explicada), será necesario restar el numerador del denominador y dividir la diferencia por n, antes de realizar simplificaciones numéricas. La utilización del error de proyección para predecir o estimar intervalos, tiene la siguiente interpretación gráfica:

GRÁFICA 23



Detrás de esta interpretación está el supuesto que las diferencias entre valores observados y valores calculados tienen una distribución de probabilidad normal. Por ese hecho pueden establecerse *niveles de confianza o probabilidades de acierto* en las estimaciones. Si se suma y resta una vez el error de proyección, el intervalo resultante implica un nivel de confianza de 68 por ciento; si se suma y resta dos veces el error de proyección, el nivel de confianza será de 95 por ciento; si se suma y resta tres veces el error de proyección, el nivel de confianza será de 99 por ciento, etcétera.

3] Otras fórmulas de cálculo

Existen otras fórmulas para cuantificar el grado de asociación entre ellas la fórmula llamada *momento-producto* que conduce a su vez a expresar el coeficiente de correlación como la *media geométrica* de los coeficientes de regresión angulares.

Dada la ecuación de regresión: $Y_c = a\,X_i + b$ aplicando el operador media aritmética, se tiene: $\overline{Y} = a\,\overline{X} + b$ de donde: $b = \overline{Y} - a\,\overline{X}$.

Por otra parte, las ecuaciones normales para la recta son:

i) $\Sigma\, Y_i = a\,\Sigma\, X_i + nb$

ii) $\Sigma\, X_i\, Y_i = a\,\Sigma\, X_i^2 + b\,\Sigma\, X_i$

Dividiendo la segunda ecuación por n, queda

$$\frac{\Sigma\, X_i\, Y_i}{n} = a\,\frac{\Sigma\, X_i^2}{n} + b\,\overline{X}$$

Remplazando el valor de $b = \overline{Y} - a\,\overline{X}$, se tiene:

$$\frac{\Sigma\, X_i\, Y_i}{n} = a\,\frac{\Sigma\, X_i^2}{n} + (\overline{Y} - a\,\overline{X})\,\overline{X}$$

$$= a\,\frac{\Sigma\, X_i^2}{n} + \overline{X}\,\overline{Y} - a\,\overline{X}^2$$

$$\frac{\Sigma\, X_i\, Y_i}{n} - \overline{X}\,\overline{Y} = a\left(\frac{\Sigma\, X_i^2}{n} - \overline{X}^2\right)$$

Observando estas expresiones, se concluye que el primer miembro no es otra cosa que la covarianza de las variables Y_i y X_i, y la

expresión dentro del paréntesis es la varianza de la variable independiente, es decir:

$$C\,[X_i\ Y_i] = a\ V\,[X_i]$$

Nótese que esta expresión corresponde a una ecuación de regresión de Y en X, es decir, donde X_i es la variable predictor y Y_i es la variable predictando. Para especificar la fórmula en este sentido el coeficiente angular "a" tendrá la siguiente expresión:

$$a_{YX} = \frac{C\,[X_i\ Y_i]}{V\,[X_i]}$$

Por analogía, si la ecuación de regresión fuera de X en Y se tendría:

$$X_c = a\ Y_i + b$$

Dado que

$$C\ (X_i\ Y_i) = C\ (Y_i\ X_i) = \frac{\Sigma\, X_i\, Y_i}{n} - \overline{Y}\ \overline{X}$$

donde el orden de los factores no altera el producto numérico; se tiene:

$$a_{XY} = \frac{C\,[X_i\ Y_i]}{V\,[Y_i]}$$

En resumen, hasta ahora se dispone de fórmulas para los coeficientes de regresión en términos de varianzas, covarianzas y medias aritméticas, que son útiles para obtener valores numéricos y para las demostraciones que a continuación se presentan.

La fórmula abreviada del coeficiente de correlación es

$$r^2 = \frac{a\,\Sigma\, X_i\, Y_i + b\,\Sigma\, Y_i - n\,\overline{Y}^2}{\Sigma\, Y_i^2 - n\,\overline{Y}^2}$$

dado que

$$b_{YX} = \overline{Y} - a_{YX}\ \overline{X}$$

remplazando se tiene:

$$r^2 = \frac{a_{YX} \Sigma X_i Y_i + (\overline{Y} - a_{YX} \overline{X}) \Sigma Y_i - n \overline{Y}^2}{\Sigma Y_i^2 - n \overline{Y}^2}$$

$$= \frac{a_{YX} \Sigma X_i Y_i + n \overline{Y}^2 - n a_{YX} \overline{X} \overline{Y} - n \overline{Y}^2}{\Sigma Y_i^2 - n \overline{Y}^2}$$

$$= \frac{a_{YX} \{\Sigma X_i Y_i - n \overline{X} \overline{Y}\}}{\Sigma Y_i^2 - n \overline{Y}^2}$$

dividiendo numerador y denominador por n se tiene:

$$r^2 = \frac{a_{YX} \{\frac{\Sigma X_i Y_i}{n} - \overline{X} \overline{Y}\}}{\frac{\Sigma Y_i^2}{n} - \overline{Y}^2} = a_{YX} \frac{C[X_i Y_i]}{V[Y_i]}$$

Luego $r^2 = a_{YX} a_{XY}$

$$r = \pm \sqrt{a_{YX} a_{XY}}$$

De otra manera

$$r^2 = \frac{(C[X_i Y_i])^2}{V[X_i] V[Y_i]} \therefore r = \frac{C[X_i Y_i]}{S_{Xi} S_{Yi}}$$

De esta manera se han deducido dos fórmulas adicionales para **el coeficiente de correlación**. La primera está dada por la **media geométrica** de los coeficientes angulares de regresión, y la **segunda tiene como** numerador a la covarianza, que es un *momento* **de orden** uno-uno respecto de las medias aritméticas, y **como denominador** al *producto* de las desviaciones típicas de las **variables; de donde** el nombre de momento-producto que se le da **a esta fórmula.**

Se ha hecho hincapié sobre la necesidad de distinguir el sentido **de la regresión** y correlación, es decir, si se trata de "Y sobre X" **o de "X sobre Y"** por el hecho que hay análisis donde puede pre**sentarse cierta** reversibilidad en la causalidad.

4] Correlación por rangos

Un caso particular de la correlación rectilínea es la llamada correlación por rangos u ordenamientos. Hay una cantidad de variables no susceptibles de medición exacta y, sin embargo, susceptibles de *ordenarse o jerarquizarse cualitativamente*; por ejemplo, una selección de candidatos a un cargo, basada en entrevistas personales, puede conducir a ordenamientos de los candidatos (por ejemplo, de mejor a peor) por parte de cada uno de los entrevistadores. El análisis de correlación por rangos determinará si estos ordenamientos son coincidentes o dispares, y cuál es la magnitud de la coincidencia o disparidad. El problema consiste en asignar a cada candidato un número de orden y determinar el grado de asociación entre dos ordenamientos, y el mismo puede ser enfrentado recurriendo a fórmulas generales ya vistas para la correlación rectilínea. Sin embargo, en este caso se consigue alguna ventaja de cálculo por el hecho que las variables tomarán *valores enteros equidistanciados*. Siguiendo el ejemplo, si dos supervisores hubieran ordenado a ocho postulantes en la siguiente forma:

Postulantes	*Ordenamiento del entrevistador 1* u_{1i}	*Ordenamiento del entrevistador 2* u_{2i}
A	4°	3°
B	2°	4°
C	7°	8°
D	6°	5°
E	3°	2°
F	8°	7°
G	5°	6°
H	1°	1°

El análisis de la correlación por rangos proporcionará un indicador cuantitativo acerca de la disparidad o coincidencia de los ordenamientos. Obsérvese que las variables son los números naturales en ambos tipos de ordenamientos; este hecho permite concluir que:

i) las medias aritméticas de los dos ordenamientos serán iguales, es decir: $M[u_{1i}] = M[u_{2i}]$.

ii) las varianzas de ambos ordenamientos serán iguales, es decir:

$$V[u_{1i}] = V[u_{2i}]$$

Una de las fórmulas generales para el coeficiente de correlación rectilínea era la siguiente:

$$r = \frac{C[X_i\, Y_i]}{S_{Xi}\, S_{Yi}} \therefore r^2 = \frac{(C[X_i\, Y_i])^2}{V[X_i]\, V[Y_i]}$$

Para las deducciones posteriores es importante establecer previamente cuál es la varianza de una suma de variables:

$$V[X_i + Y_i] = \frac{\Sigma\,(X_i + Y_i - \overline{X} - \overline{Y})^2}{n} = \frac{\Sigma\,[(X_i - \overline{X}) + (Y_i - \overline{Y})]^2}{n}$$

$$V[X_i + Y_i] = \frac{\Sigma\,(X_i - \overline{X})^2}{n} + \frac{\Sigma\,(Y_i - \overline{Y})^2}{n} + \frac{2\Sigma\,(X_i - \overline{X})\,(Y_i - \overline{Y})}{n}$$

$$V[X_i + Y_i] = V[X_i] + V[Y_i] + 2\,C[X_i\, Y_i]$$

ya que $\dfrac{2\Sigma\,(X_i - \overline{X})\,(Y_i - \overline{Y})}{n} = \dfrac{2\Sigma\,(X_i\, Y_i - \overline{X}\, Y_i - \overline{Y}\, X_i + \overline{Y}\,\overline{X})}{n}$

$$= 2\left[\frac{\Sigma\, X_i\, Y_i}{n} - \overline{X}\,\frac{\Sigma\, X_i}{n} - \overline{Y}\,\frac{\Sigma\, X_i}{n} + \overline{X}\,\overline{Y}\right]$$

$$= 2\left[\frac{\Sigma\, X_i\, Y_i}{n} - \overline{X}\,\overline{Y} - \overline{Y}\,\overline{X} + \overline{X}\,\overline{Y}\right]$$

$$= 2\left[\frac{\Sigma\, X_i\, Y_i}{n} - \overline{X}\,\overline{Y}\right] = 2\,C[X_i\, Y_i]$$

Por analogía la varianza de la diferencia de dos variables será

$$V[X_i - Y_i] = V[X_i] + V[Y_i] - 2\,C[X_i\, Y_i]$$

Dadas las variables ordenamiento u_{1i}, u_{2i}, se puede establecer una relación de diferencia entre ellas:

$$d_i = u_{1i} - u_{2i}$$

$$V[d_i] = V[u_{1i}] + V[u_{2i}] - 2\,C[u_{1i}\, u_{2i}]$$

El coeficiente de correlación en términos de estas variables será:

$$r = \frac{C\,[u_{1i}\, u_{2i}]}{\sqrt{V\,[u_{1i}]\;\; V\,[u_{2i}]}}$$

$$r = \frac{C\,[u_{1i}\, u_{i2}]}{V\,[u_{1i}]} \qquad \text{ya que} \qquad V\,[u_{1i}] = V\,[u_{2i}]$$

Despejando la relación $V\,[d_i]$, se tiene que:

$$C\,[u_{1i}\, u_{2i}] = (V[u_{1i}] + V\,[u_{2i}] - V\,[d_i]\,½$$

$$= (2\,V\,[u_{1i}] - V\,[d_i])\,½$$

Por otra parte

$$d_i = u_{1i} - u_{2i}$$

$$M\,[d_i] = M\,[u_{1i}] - M\,[u_{2i}] = 0$$

luego,

$$V\,[d_i] = M\,[d_i^2\,] - (M\,[d_i])^2$$

$$= \frac{\Sigma\, d_i^2}{n}$$

Entonces,

$$r = \frac{C\,[u_{1i}\, u_{2i}]}{V\,[u_{1i}]} = \frac{V\,[u_{1i}] - V\,[d_i]\,½}{V\,[u_{1i}]}$$

$$= 1 - \frac{V\,[d_i]}{\Sigma\, V\,[u_{1i}]}$$

Pero la varianza de u_{1i} $[V\,(u_{1i})]$ es la varianza de los n primeros números naturales.

$$V\,[u_{1i}] = \frac{\Sigma\, u_{1i}^2}{n} - \bar{u}_1^2$$

$$= \frac{n\ (n+1)\ (2n+1)}{6n} - \left(\frac{n\ (n+1)}{2n}\right)^2$$

$$= \frac{(n+1)\ (2n+1)}{6} - \left(\frac{n+1}{2}\right)^2$$

$$= \frac{(n+1)\ (2n+1)}{6} - \frac{(n+1)\ (n+1)}{4}$$

$$= (n+1)\left[\frac{2n+1}{6} - \frac{n+1}{4}\right]$$

$$= (n+1)\left[\frac{4n+2-3n-3}{12}\right]$$

$$= (n+1)\left(\frac{n-1}{12}\right) = \frac{n^2-1}{12}$$

Luego,

$$r = 1 - \frac{V\ [d_i]}{2\left(\frac{n^2-1}{12}\right)}$$

$$r = 1 - \frac{6 \Sigma d_i^2}{n\ (n^2-1)}$$

Para el ejemplo propuesto en páginas anteriores el cálculo se haría de la siguiente manera:

Postulantes	u_{1i}	u_{2i}	d_i	d_i^2
A	4	3	1	1
B	2	4	–2	4
C	7	8	–1	1
D	6	5	1	1
E	3	2	1	1
F	8	7	1	1
G	5	6	–1	1
H	1	1	0	0
	36	36	0	10

Aplicando la fórmula anterior resulta

$$r = 1 - \frac{6\ (10)}{8\ (63)} = 0.88$$

El resultado del indicador muestra que los dos ordenamientos están bastante asociados, sin que tengan, en general, discrepancias significativas.

A veces puede utilizarse con ventaja este tipo de indicador, aun en casos de variable cuantificable. Puede ocurrir que si el número de datos es muy grande y las variables toman valores que dificultan el cálculo numérico, sea conveniente ordenar las observaciones de acuerdo a sus valores numéricos y establecer la correlación entre los ordenamientos. Evidentemente esta simplificación implica rigidez por la introducción de supuestos adicionales, tales como la correlación rectilínea, y la necesidad que los ordenamientos reflejen adecuadamente la distribución de las variables originales. Sin embargo, muchas veces basta con saber si existe o no asociación sin que interese mucho refinar el análisis; para este tipo de estudios puede prestarse esta conversión arbitraria de variables.

La facilidad de cálculo del coeficiente de correlación por rangos tiene, como contrapartida, una *seria limitación.* Se supone que las distancias o diferencia de atributos es *constante* entre los casos considerados. En el ejemplo visto, esto quiere decir que la diferencia entre el postulante H y el postulante B es la misma que la que existe entre el postulante B y el postulante E, etc., para cada uno de los ordenamientos. En la práctica, difícilmente se cumplirá este supuesto, pero como se ha dicho, este tratamiento es adecuado para variables de atributos donde la jerarquización u ordenamiento constituyen la única forma de discriminación, y donde se observa con menos rigurosidad las limitaciones aludidas.

F. CORRELACIÓN NO RECTILÍNEA

Sin desconocer que el caso particular de la correlación rectilínea es útil para presentar los conceptos del análisis de regresión y correlación, su aplicación práctica es algo restringida por el hecho de que en los estudios socioeconómicos las relaciones entre las variables adquieren formas que, en general, difícilmente pueden ser representadas en forma adecuada por una línea recta.

De la misma manera que se distinguían diversas relaciones no rectilíneas en el capítulo de regresión, aquí desde ese mismo punto

de vista se expondrán los coeficientes de correlación respectivos. En correlación no rectilínea carece de utilidad distinguir entre directa e inversa, por el hecho que pueden haber tramos donde la relación sea directa y otros donde sea inversa.

1] Si la función es del tipo

$$Y_c = a \log X + b$$

es decir, si cambios relativos de X determinan cambios absolutos de Y, el coeficiente de determinación tendrá la expresión general

$$r^2 = \frac{S^2_{Y_c}}{S^2_Y} = \frac{\Sigma (Y_c - \overline{Y})^2}{\Sigma (Y_i - \overline{Y})^2} = \frac{\Sigma Y_c^2 - n \overline{Y}^2}{\Sigma Y_i^2 - n \overline{Y}^2}$$

pero,

$$Y_c^2 = a^2 (\log X_i)^2 + 2ab \log X_i + b^2$$

$$Y_c^2 = a\{a (\log X_i)^2 + b \log X_i\} + b \{a \log X_i + b\}$$

$$\Sigma Y_c^2 = a\{\underbrace{a \Sigma (\log X_i)^2 + b \Sigma \log X_i}_{\text{Ecuación normal}}\} + b \{\underbrace{a \Sigma \log X_i + nb}_{\text{Ecuación normal}}\}$$

$$\Sigma Y_c^2 = a \Sigma Y_i \log X_i + b \Sigma Y_i$$

$$r^2 = \frac{a \Sigma Y_i \log X_i + b \Sigma Y_i - n \overline{Y}^2}{\Sigma Y_i^2 - n \overline{Y}^2}$$

Nótese que este coeficiente de determinación resulta de la relación entre Y_i y $\log X_i$, que será distinto del que resulta de la relación entre Y_i y X_i (siendo X_i el antilogaritmo de $\log X_i$).

Para el cálculo del error de proyección se procede de manera similar al caso de correlación rectilínea, es decir, basta dividir por n la diferencia entre el numerador de la fórmula del coeficiente de determinación (n veces la varianza explicada) y el denominador (n veces la varianza total); la raíz cuadrada de esta diferencia dividida por n será el error de proyección. Nuevamente, el error de proyección está dado teniendo en cuenta la proyección de Y_i, en términos de la variable independiente $\log X_i$, que será distinto al error que se dé al proyectar Y_i en términos de X_i.

2] Si la función es del tipo

$$Y = f\, d^X \qquad \text{si } \log f = b;\ \log d = a$$

$$\log Y = aX + b$$

es decir, una función de las llamadas exponenciales, el procedimiento para encontrar las fórmulas del coeficiente de correlación y del error de proyección es similar al caso anterior. Obsérvese que en esta función, variaciones absolutas de la variable X determinan variaciones relativas de Y. La fórmula general del coeficiente de determinación es la siguiente:

$$r^2 = \frac{\Sigma\ (Y_c - \overline{Y})^2}{\Sigma\ (Y_i - \overline{Y})^2} = \frac{\Sigma\ Y_c^2 - n\,\overline{Y}^2}{\Sigma\ Y_i^2 - n\,\overline{Y}^2}$$

Obsérvese que en la función aparece el logaritmo de Y_i. Por este hecho la fórmula particular será:

$$r^2 = \frac{\Sigma\ (\log Y_c)^2 - n\,\overline{\log Y}^2}{\Sigma\ (\log Y_i)^2 - n\,\overline{\log Y}^2}$$

donde,

$$\overline{\log Y} = M\ [\log Y_i] = \frac{\Sigma \log Y_i}{n}$$

Dado que $\log Y_c = aX_i + b$

$$\Sigma\ (\log Y_c)^2 = a^2\,\Sigma\ X_i^2 + ab\,\Sigma\ X_i + ab\,\Sigma\ X_i + nb^2$$

$$= a\ \{a\,\Sigma\ X_i^2 + b\,\Sigma\ X_i\} + b\ \{a\,\Sigma\ X_i + nb\}$$

$$= a\,\Sigma\ (\log Y_i)\ X_i + b\,\Sigma \log Y_i$$

$$r^2 = \frac{a\,\Sigma\ X_i \log Y_i + b\,\Sigma \log Y_i - n\,\overline{\log Y}^2}{\Sigma\ (\log Y_i)^2 - n\,\overline{\log Y}^2}$$

Cabe destacar nuevamente que el coeficiente de correlación calculado según la última expresión diferirá del calculado según la ex-

presión general. En la fórmula general se establece la asociación entre Y_i y X_i; en cambio, en el caso particular se establece la correlación entre el logaritmo de Y_i y la variable X_i. El error de proyección en uno y otro caso se obtiene calculando la raíz de la varianza no explicada que es la ene-ava parte de la diferencia entre el numerador y el denominador de la fórmula del coeficiente de determinación.

3] En una función potencial del tipo

$$Y_c = b\, X^a$$

cuya expresión logarítmica es

$$\log Y_c = \log b + a \log X_i$$

donde variaciones relativas de la variable independiente determinan variaciones también relativas de la variable dependiente; pueden así establecerse dos fórmulas para el coeficiente de correlación que conducirán a resultados distintos:

$$r^2 = \frac{\Sigma\ (\log Y_c - \overline{\log Y})^2}{\Sigma\ (\log Y_i - \overline{\log Y})^2}$$

y la otra con los antilogaritmos respectivos.

$$r^2 = \frac{\Sigma\ (Y_c - \overline{Y})^2}{\Sigma\ (Y_i - \overline{Y})^2}$$

Para el caso de la correlación logarítmica la fórmula abreviada de cálculo se obtiene de la misma forma que las anteriores.

$$r^2 = \frac{a \Sigma \log X_i \log Y_i + \log b \Sigma \log Y_i - n\, \overline{\log Y^2}}{\Sigma\ (\log Y_i)^2 - n \log Y^2}$$

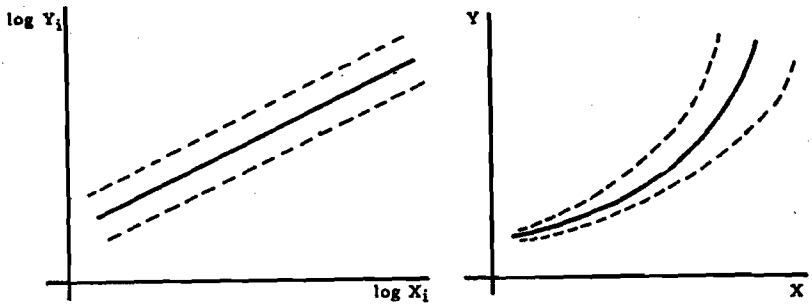
De esta fórmula también puede obtenerse el error de proyección logarítmico con el procedimiento ya conocido.

En el caso de correlación logarítmica la estimación de intervalos se hace en la misma forma que en el caso rectilíneo, pero teniendo en cuenta que para los valores reales el desvío contempla-

do representa una proporción constante de los valores dados por la ecuación de regresión.

La interpretación gráfica es la siguiente:

GRÁFICA 24



En escalas logarítmicas aparece una diferencia constante. Los antilogaritmos de estos valores conducen a la representación en escala natural donde puede observarse que el intervalo es cada vez mayor, lo que corresponde a una proporción constante del semiancho del intervalo respecto de los valores dados por la ecuación de regresión.

4] Correlación parabólica

Dada la función

$$Y_c = a\,X^2 + b\,X + c$$

El coeficiente de determinación está dado por

$$r^2 = \frac{\Sigma\,(Y_c - \bar{Y})^2}{\Sigma\,(Y_i - \bar{Y})^2} = \frac{\Sigma\,Y_c - n\,\bar{Y}^2}{\Sigma\,Y_i^2 - n\,\bar{Y}^2}$$

$$Y_c^2 = a^2\,X^4 + b^2\,X^2 + c^2 + 2\,a\,b\,X^3 + 2\,a\,c\,X^2 + 2\,b\,c\,X$$

$$Y_c^2 = a^2\,X^4 + a\,b\,X^3 + a\,b\,X^3 + b^2\,X^2 + a\,c\,X^2 + a\,c\,X^2 + c^2 + b\,c\,X + b\,c\,X$$

$$Y_c^2 = a\{a\,X_i^4 + b\,X_i^3 + c\,X_i^2\} + b\{a\,X_i^3 + b\,X_i^2 + c\,X_i\} + c\,\{a\,X^2 + b\,X + c\}$$

Aplicando el operador sumatoria

$$\Sigma Y_c^2 = a\{a\Sigma X_i^4 + b\Sigma X_i^3 + c\Sigma X_i\} + b\{a\Sigma X_i^3 + b\Sigma X_i^2 + c\Sigma X_i\} + c\{a\Sigma X^2 + b\Sigma X + nc\}$$

Las expresiones dentro de los paréntesis corresponden con las ecuaciones normales de la parábola, es decir:

$$\Sigma Y_c^2 = a\{\Sigma Y_i X_i^2\} + b\{\Sigma Y_i X_i\} + c\{\Sigma Y_i\}$$

La fórmula abreviada de cálculo para el coeficiente de correlación será:

$$r^2 = \frac{a\Sigma Y_i X_i^2 + b\Sigma Y_i X_i + c\Sigma Y_i - n\bar{Y}^2}{\Sigma Y_i^2 - n\bar{Y}^2}$$

El error de proyección se calcula por el método expuesto para los casos anteriores.

G. CORRELACIÓN MÚLTIPLE

En el caso de que existan dos o más variables independientes, la necesidad de disponer de indicaciones acerca de la asociación que simultáneamente tiene la variable dependiente con las variables independientes, conduce a la obtención de coeficientes de correlación múltiples; si bien son necesarios para el análisis los coeficientes de correlación simples, es preciso complementar este conjunto de indicadores con un estadígrafo que resuma simultáneamente los grados de asociación simples.

1] Correlación en un plano de regresión.
Si se tienen dos variables independientes, la ecuación de regresión es de la forma

$$X_{1c} = a_{1.23} + b_{12.3} X_2 + b_{13.2} X_3$$

El coeficiente de correlación se obtiene siempre a partir de la fórmula general que ahora tendrá la siguiente simbología:

$$R_{1.23}^2 = \frac{S_{Xc\ 1.23}^2}{S_{Xi}^2} = \frac{\Sigma(X_{1c} - \bar{X})^2}{\Sigma(X_{1i} - \bar{X})^2} = \frac{\Sigma X_{1c}^2 - n\bar{X}_1^2}{\Sigma X_{1i}^2 - n\bar{X}_1^2}$$

Donde $S^2_{Xc\ 1.23}$ es la varianza explicada por las variables X_2 y X_3 S^2_{X1} es la varianza total de la variable dependiente.

$$X^2_{1c} = a^2_{1.23} + b^2_{12.3}\ X^2_2 + b^2_{13.2}\ X^2_3 + a_{1.23}\ b_{12.3}\ X_2 +$$

$$+ a_{1.23}\ b_{12.3}\ X_2 + a_{1.23}\ b_{13.2}\ X_3 + a_{1.23}\ b_{13.2}\ X_3 +$$

$$+ b_{1.23}\ b_{13.2}\ X_2X_3 + b_{12.3}\ b_{13.2}\ X_2X_3$$

$$X^2_{1c} = a_{1.23}\ \{a_{1.23} + b_{12.3}\ X_2 + b_{13.2}\ X_3\}$$

$$+ b_{12.3}\ \{b_{12.3}\ X^2_2 + a_{1.23}\ X_2 + b_{13.2}\ X_2X_3\}$$

$$+ b_{13.2}\ \{b_{13.2}\ X^2_3 + a_{1.23}\ X_3 + b_{12.3}\ X_2X_3\}$$

Aplicando sumatoria se tiene:

$$\Sigma\ X^2_{1c} = a_{1.23}\ \{n\ a_{1.23} + b_{12.3}\ \Sigma\ X_2 + b_{13.2}\ \Sigma\ X_3\}$$

$$+ b_{12.3}\ \{b_{12.3}\ \Sigma\ X^2_2 + a_{1.23}\ \Sigma\ X_2 + b_{13.2}\ \Sigma\ X_2X_3\}$$

$$+ b_{13.2}\ \{b_{13.2}\ \Sigma\ X^2_3 + a_{1.23}\ \Sigma\ X_3 + b_{12.3}\ \Sigma\ X_2X_3\}$$

Las expresiones dentro de los paréntesis corresponden con las ecuaciones normales de un plano de regresión. En efecto,

$$\Sigma\ X^2_{1c} = a_{1.23}\ \Sigma\ X_1 + b_{12.3}\ \Sigma\ X_1X_2 + b_{13.2}\ \Sigma\ X_1X_3$$

La fórmula abreviada del coeficiente de determinación queda en consecuencia,

$$R^2_{1.23} = \frac{a_{1.23}\ \Sigma\ X_1 + b_{12.3}\ \Sigma\ X_1X_2 + b_{13.2}\ \Sigma\ X_1X_3 - n\ \overline{X}^2_1}{\Sigma\ X^2_1 - n\ \overline{X}^2_1}$$

En correlación múltiple no tiene sentido el signo de R, ya que puede haber variables que influyan positiva o negativamente en la variable dependiente.

En cuanto a la forma de cálculo del error de proyección, no difiere de los vistos anteriormente; la diferencia entre numerador y denominador es n veces la varianza no explicada.

En el caso de correlación múltiple se presenta un problema particular originado por la necesidad de disponer de indicaciones so-

bre la *asociación neta* existente entre la variable dependiente y cada una de las variables independientes. El coeficiente de correlación múltiple indica el grado de asociación que simultáneamente se presenta entre la variable dependiente y las variables independientes. Un coeficiente de correlación simple indica el grado de asociación entre dos variables: dependiente e independiente, pero sin eliminar o depurar estadísticamente la asociación entre ambas variables de la influencia de otras que actúan a través de la variable independiente. Por ejemplo, puede haber una alta correlación entre la cantidad vendida de un artículo y su precio; pero esta asociación puede disminuir en forma sustancial al eliminar explícitamente la influencia de la variable precio de un sustituto.

Este concepto de asociación neta o depurada se cuantifica a través del coeficiente de correlación parcial que, en el caso de tres variables, se define de la siguiente manera:

$$r_{12.3} = \left(\frac{S^2_{Xc\ 1.23} - S^2_{Xc\ 1.3}}{S^2_{Xs\ 1.3}}\right)^{1/2}$$

$r_{1.23}$ representa la asociación entre las variables X_1 y X_2, eliminando estadísticamente la influencia de la variable X_3. En efecto, si se observa el numerador, se concluye que representa el incremento en la varianza explicada al incluir la variable X_2. Este incremento se compara con la varianza que dejaba sin explicar la variable X_3. Sustituyendo el numerador por varianzas totales y no explicadas se tiene:

$$r^2_{12.3} = \frac{S^2_{X1} - S^2_{Xs\ 1.23} - S^2_{X1} + S^2_{Xs\ 1.3}}{S^2_{Xs\ 1.3}}$$

$$= 1 - \frac{S^2_{Xs\ 1.23}}{S^2_{Xs\ 1.3}}$$

El otro coeficiente de correlación parcial se define

$$r^2_{13.2} = \frac{S^2_{Xc\ 1.23} - S^2_{Xc\ 1.2}}{S^2_{Xs\ 1.2}} = 1 - \frac{S^2_{Xs\ 1.23}}{S^2_{Xs\ 1.2}}$$

Con todos estos estadígrafos, en el caso de tres variables se tiene un conjunto de indicadores complementarios que permiten obtener conclusiones objetivas. Por una parte, se dispone de tres coefi-

cientes de correlación simple: r_{12}, r_{13}, r_{23}; además dos coeficientes de correlación parcial: $r_{12.3}$ y $r_{13.2}$; por último un coeficiente de correlación múltiple: $R_{1.23}$. Por otra parte, se dispone de todos los errores de proyección correspondientes que permitirán obtener intervalos para las proyecciones.

Las relaciones que se plantean entre estos coeficientes de correlación permiten realizar análisis de consistencia. Cualquier coeficiente de correlación parcial es menor, y a lo sumo igual, que un coeficiente de correlación simple, por la eliminación explícita de la influencia de otras variables:

$$r_{ij \cdot k} \leqslant r_{ij}$$

Un coeficiente de correlación múltiple será siempre mayor, o por lo menos igual, que un coeficiente de correlación simple, por el hecho que aquél toma en cuenta un mayor número de variables independientes que explican la variabilidad de la variable dependiente.

$$R_{i\,jk} \geqslant r_{ij}$$

$$R_{i\,jk} \geqslant r_{ik}$$

2] Correlación en un hiperplano de regresión.

Cuando se tienen más de dos variables independientes, se presenta el caso general de la correlación múltiple; los conceptos analizados para el caso de tres variables son también aplicables al caso general. Lo que ocurre es que si se consideran muchas variables independientes se dificulta un tanto el análisis, y el cálculo de estadígrafos, cuando no se dispone de computadores, resulta en extremo laborioso. En todo caso, a continuación se presentan las fórmulas de los estradígrafos más importantes para una ecuación lineal que considera 3 variables independientes, es decir:

$$X_{1c} = a_{1.234} + b_{12.34}\, X_2 + b_{13.24}\, X_3 + b_{14.23}\, X_4$$

$$R^2_{1.234} = \frac{a_{1.234}\,\Sigma X_1 + b_{12.34}\,\Sigma X_1X_2 + b_{13.24}\,\Sigma X_1X_3}{\Sigma X_1^2 - n\,\bar{X}_1^2} + \frac{b_{14.23}\,\Sigma X_1X_4 - n\,\bar{X}_1^2}{\Sigma X_1^2 - n\,\bar{X}_1^2}$$

El error de proyección se calcula recordando que la diferencia

entre numerador y denominador de la fórmula anterior es n veces la varianza no explicada.

Los coeficientes de correlación parcial, aislando el efecto de dos variables, son los siguientes:

$$r^2_{12.34} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.34}}{S^2_{Xs\ 1.34}}$$

$$r^2_{13.24} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.24}}{S^2_{Xs\ 1.24}}$$

$$r^2_{14.23} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.23}}{S^2_{Xs\ 1.23}}$$

Puede definirse otro tipo de coeficientes de correlación parcial donde se aísla el efecto de una variable:

$$r^2_{122.4} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.4}}{S^2_{Xs\ 1.4}}$$

$$r^2_{134.2} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.2}}{S^2_{Xs\ 1.2}}$$

$$r^2_{124.3} = \frac{S^2_{Xc\ 1.234} - S^2_{Xc\ 1.3}}{S^2_{Xs\ 1.3}}$$

3] Correlación múltiple logarítmica

La linealidad de los planos de regresión ya vistos puede no ser adecuada en muchos problemas de estimación; en esos casos es conveniente probar con otro tipo de funciones, como por ejemplo:

$$X_{1c} = a_{1.23} + b_{12.3}\, X_2^2 + b_{13.2}\, X_3^{1/2}$$

La metodología que permite obtener ecuaciones normales para determinar los parámetros de regresión y la deducción de las fórmulas de coeficientes de correlación múltiples y parciales, es la misma que se presentó en páginas anteriores. Con ecuaciones del tipo que ahora se muestra puede representarse adecuadamente la relación entre las variables.

Sin embargo, una función que es muy utilizada en los problemas de proyección es la llamada función logarítmica:

$$X_{1c} = \alpha\, X_2^{\beta}\, X_3^{\gamma}$$

Para poder aplicar la metodología de los mínimos cuadrados, es necesario previamente "linealizar" esta función, aplicando logaritmos.

$$\log X_{1c} = \log \alpha + \beta \log X_2 + \gamma \log X_3$$

La función así linealizada es similar al caso de correlación múltiple lineal; la única diferencia radica en el hecho de que en los cómputos deben tomarse los logaritmos de las variables. El coeficiente de correlación múltiple logarítmico, por analogía con el caso de correlación múltiple lineal (siendo $\log \alpha = \alpha^*$) es:

$$R^2_{\log(1.23)} = \frac{\alpha^* \Sigma \log X_1 + \beta \Sigma \log X_1 \log X_2}{\Sigma (\log X_1)^2 - n\,\overline{\log X_1^2}}$$

$$+ \frac{\gamma \Sigma \log X_1 \log X_3 - n\,\overline{\log X_1^2}}{\Sigma (\log X_1)^2 - n\,\overline{\log X_1^2}}$$

La diferencia entre numerador y denominador resulta ser n veces la varianza no explicada. Esta función tiene amplias posibilidades de ser aplicada por el hecho que los parámetros β y γ son coeficientes de elasticidad entre las variables X_1 y X_2 y las variables X_1 y X_3 respectivamente. Esta ecuación será tratada con más detalle, en el capítulo correspondiente a las proyecciones por coeficientes de elasticidad.

En cuanto a los coeficientes de correlación parcial, tampoco aparecen diferencias, ya que la metodología de deducción y cálculo no varía, mas lo que no debe descuidarse en el trabajo con los logaritmos de las variables es que deben tener una precisión equivalente a los 6 decimales.

H. ETAPAS DE LA CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN Y CORRELACIÓN

Es necesario diferenciar dos tipos de moledos de regresión en lo tocante al objetivo que persiguen: los *modelos de análisis,* utilizados para cuantificar relaciones y explicar adecuadamente qué sucedió con una variable en términos de otras variables que tienen influencia sobre aquélla, y los *modelos predictivos* que además de

ser útiles en el análisis están diseñados para "predecir" o estimar valores de la variable dependiente en términos de las variables independientes en el supuesto de que se conoce su comportamiento. Además, es conveniente distinguir entre modelos *temporales y atemporales*. Los primeros son aquellos que analizan y estiman valores en el tiempo, por ejemplo, estimación de los precios agrícolas del próximo año en función de las siembras y la política de importaciones. Los modelos atemporales, en cambio, no toman en cuenta explícita ni implícitamente la variable tiempo: son cortes transversales; sería éste el caso de la estimación de los consumos familiares en función de la variable ingreso, pero teniendo como datos los consumos e ingresos de una muestra en un momento o período dado.

La metodología que a continuación se presenta tiene aplicación general; con todo, se aclararán aquellos puntos que son más críticos en uno y otro tipo de modelo.

1] El primer punto, obviamente es la *determinación clara y precisa del objetivo del estudio*. Es necesario especificar los objetivos de la investigación general y los objetivos del análisis de regresión y correlación en particular. En esencia es necesario responder a las interrogantes ¿en qué se utilizará el modelo?, ¿qué se pretende demostrar por medio de la regresión y correlación?

2] Una vez aclarado el primer punto básico, se hace necesaria una *evaluación lógica* para determinar qué variables deben incorporarse al análisis. En principio deben tomarse en cuenta todas las que razonablemente pueden estar asociadas a la variable que se estudia.

3] A continuación se procede a recopilar las estadísticas, ya sea históricas, cuando se trata de modelos temporales, o las estadísticas pertinentes si se trata de un modelo atemporal.

4] Siempre es indispensable un análisis de la *calidad* de los datos recolectados. Ya aquí quedan eliminadas algunas de las variables que, en principio, fueron seleccionadas, por el hecho que sus valores pueden no ser confiables; por otra parte, pueden haber algunas variables que, pese a ser confiables, no pueden tomarse en cuenta porque constituyen muy *pocas observaciones*. Sobre este punto hay que decidir cuál es el número mínimo de datos y observaciones que puede considerarse satisfactorio. Recuérdese que tamaños de muestra insuficientes conducen a resultados erróneos. En los modelos temporales no puede pensarse en un número inferior a las 10 o 12 observaciones (puntos en el tiempo). Además, en los modelos predictivos, el número de variables independientes

está condicionado por la posibilidad de disponer con cierta confianza de valores futuros de tales variables.

5] Las variables restantes deben ser *depuradas* de otras variables que actúan a través de éstas. Como se apuntó antes, es indispensable trabajar con series que representen valor real o "quantum". La consideración de valores nominales exagera la correlación por el hecho de que la variable inflación o alzas de precios puede actuar sobre la variable dependiente y, simultáneamente, sobre las variables independientes. Es conveniente también, en lo posible, representar las series en términos por habitante, si no hubiera un propósito específico para hacerlo de otra manera.

6] Una vez que se dispone de las estadísticas de las principales variables depuradas, se hace necesario determinar la forma y cuantificar el grado de la asociación siemple que cada una de estas variables tenga con la variable dependiente estudiada. También puede ser conveniente calcular los coeficientes de correlación simple entre las variables independientes para advertir las posibles dependencias que existan entre ellas. A esta altura del análisis ya se tiene bastante definido el campo de la posible metodología que finalmente se utilizará; por lo menos, se habrá decidido si se trata de correlación simple o múltiple.

7] Otro punto de gran importancia es la determinación de la forma general de la función. Si se trata de correlación simple, será útil la representación gráfica, es decir, con la ayuda del diagrama de dispersión puede solucionarse adecuadamente este problema. Si se trata, en cambio, de correlación múltiple, hay que considerar principalmente los coeficientes cuantificados en el punto 5 y las formas particulares de relación entre las variables. A veces se dispone de *modelos teóricos* ya probados, donde sólo se requiere comprobar si tal teoría corresponde al caso que se estudia; por ejemplo, la función consumo de Friedman, donde ya se tienen especificadas las variables independientes y la forma de la función, y sólo resta calcular el valor de los parámetros. El caso más corriente es determinar la función (formulación de la teoría), primero en *términos conceptuales* y, segundo, *cuantificando resultados*. En los modelos temporales un punto delicado es la especificación de las asincronías entre las variables. Por ejemplo, la producción del período t podría depender de la inversión del período "t — a", donde "a" indicaría el tiempo de maduración de la inversión. La representación gráfica por parejas de variables (dependiente o independiente) puede ayudar a la especificación mencionada.

8] El paso siguiente es la cuantificación de estadígrafos: medias, varianzas, coeficientes de correlación simples, múltiples, parciales,

errores de proyección y, por último, *la estimación* en los modelos predictivos y *el análisis* en los modelos descriptivos. Es conveniente también calcular, por medio de la ecuación de regresión, los valores de la variable dependiente en términos de los valores conocidos en la variable independiente, para compararlos con valores observados y analizar la bondad del ajuste. Las formulaciones de pruebas de consistencia entre los estadígrafos calculados constituyen, tal vez, los puntos más descuidados en los análisis de regresión y correlación. Por otra parte, es aquí donde cabe calificar el análisis a la luz de las *cuantificaciones apropiadas*. Es conveniente comparar la magnitud de los errores con los valores calculados, estableciendo porcentualmente la cuantía de los probables desvíos.

9] Finalmente, en la presentación de los resultados es imprescindible destacar:

a) Clara definición de las variables;
b) Tamaño de muestra y tipo de modelos;
c) Forma de la función;
d) Estadígrafos pertinentes.

No se debe dejar de señalar las limitaciones particulares del método, los supuestos utilizados y las fuentes de obtención de informaciones.

1. MÉTODO DE ESTIMACIÓN POR MEDIO DEL COEFICIENTE DE ELASTICIDAD

1] *Presentación conceptual*

Un método muy utilizado en proyecciones de variables socioeconómicas es el que utiliza el coeficiente de elasticidad entre las variables. El coeficiente de elasticidad se define como

$$E = \frac{\dfrac{dY}{Y}}{\dfrac{dX}{X}} = \frac{dY}{dX} \cdot \frac{X}{Y} = Y' \frac{X}{Y}$$

Como puede observarse, el coeficiente de elasticidad es *una medida de cambios porcentuales* experimentados por una variable Y (dependiente) ante cambios porcentuales de una variable X (independiente).

En la definición está implícita la función que relaciona ambas

variables. Desde un punto de vista estricto, se trata de un cociente entre cambios porcentuales infinitesimales; cuando se trata de estimar valores de una variable, no interesan los cambios demasiado pequeños, sino los cambios significativos.

El objetivo inmediato será, entonces, encontrar funciones donde el coeficiente de elasticidad sea constante en cualquier punto de la función. Solamente tal tipo de funciones podrán ser utilizadas en la proyección, ya que de otra manera el coeficiente de elasticidad variará para cada punto de la función haciendo impracticable la proyección.

Si la función es una recta, el coeficiente de elasticidad no es constante, como se ve a continuación.

$$Y = a\,X + b$$

$$\frac{dY}{dX} = a$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = a \cdot \frac{X}{Y}$$

pero $Y = a\,X + b$ ∴

$$E = a\,\frac{X}{a\,X + b}$$

Como puede observarse, el coeficiente de elasticidad E está en función de X, y tomará un valor distinto para cada valor de X. Si la función es una hipérbola equilátera, se tiene:

$$Y = \frac{a}{X} = a\,X^{-1}$$

$$\frac{dY}{dX} = -\,a\,X^{-2}$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = -\,a\,X^{-2} \cdot \frac{X}{\frac{a}{X}}$$

$$E = -\,a\,X^{2} \cdot \frac{X^{2}}{a} = -\,1$$

El resultado se interpreta de modo tal que aumentos porcentuales en la variable independiente determinen disminuciones de igual magnitud porcentual en la variable dependiente. En este resultado, en consecuencia, puede estimarse cualquier valor de la variable dependiente, si se supone conocido un valor de la variable independiente. En la función potencial general también puede verificarse la constancia del coeficiente de elasticidad. En efecto,

$$Y = b\ X^a$$

$$\frac{dY}{dX} = ab\ X^{a-1}$$

$$E = \frac{dY}{dX} \cdot \frac{X}{Y} = ab\ X^{a-1} \cdot \frac{X}{Y}$$

pero $Y = b\ X^a$ ∴

$$E = ab\ X^{a-1} \frac{X}{b\ X^a} = a$$

El hecho de que el coeficiente de elasticidad sea constante en esta función hace que se la utilice periódicamente en las proyecciones. La proyección se basa en lo siguiente:

Dada la función: $Y = b\ X^a$

Aplicando logaritmos: $\log Y = \log b + a \log X$

Las relaciones correspondientes al año 0 (base de proyección) y al año n (período para el que se quiere estimar la variable dependiente), son las siguientes:

$$\log Y_0 = \log b + a \log X_0$$

$$\log Y_n = \log b + a \log X_n$$

Restando la primera de la segunda se tiene:

$$\log Y_n - \log Y_0 = a\ (\log X_n - \log X_0)$$

El antilogaritmo de la relación anterior

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0}\right)^a$$

Observando la fórmula puede concluirse que los cambios porcentuales en la variable dependiente son equivalentes a los cambios porcentuales en la variable independiente, elevados a la potencia a. A veces, suele interpretarse erróneamente la relación potencial; por ejemplo, si a = 2, se dice que un cambio de 0 por ciento en X determinará un cambio de 100 por ciento en Y. Evidentemente, la conclusión es falsa, porque ella supone una relación de linealidad entre las variables que está lejos de presentarse en este caso.

Los datos necesarios para proyectar mediante este método son: disponer del coeficiente de elasticidad (a), conocer el valor base y dado de la variable independiente (X_0 y X_n), o por lo menos su variación porcentual. Con estos datos puede aplicarse la fórmula antes citada:

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0}\right)^a$$

Por ejemplo, si a = 2

$$Y_0 = 100$$

$$X_0 = 200$$

$$X_n = 300$$

Remplazando:

$$\frac{Y_n}{100} = \left(\frac{300}{200}\right)^2 = 2.25$$

$$Y_n = 225$$

2] *Tipos de elasticidad*

Es necesario distinguir el tipo de elasticidad según las variables consideradas. De este modo, si la variable Y representa consumos y la variable X representa ingresos, se habla de elasticidad ingreso del consumo o de la demanda. Por otra parte, si la variable X representa consumo y la variable Y representa consumo específico de un bien o conjunto de bienes similares, se habla de elasticidad gasto del consumo específico. Estos dos son los conceptos más conocidos y utilizados. Sin embargo, según las denominaciones de las variables puede hablarse de otros tipos distintos de elasticidad, como

por ejemplo, elasticidad de la tributación al ingreso, elasticidad del ahorro al producto, de las importaciones al tipo de cambio, etc.

3] *Métodos de cálculo*

A continuación se presentarán las formas de cálculo del coeficiente de elasticidad. Cuando se está utilizando la forma general de proyección,

$$\frac{Y_n}{Y_o} = \left(\frac{X_n}{X_o}\right)^a$$

implícitamente se están aceptando dos supuestos: a) que las variables están relacionadas mediante la función potencial; b) que el coeficiente de correlación entre los logaritmos de X y de Y sea *significativo*. El cumplimiento de estos supuestos garantiza una buena proyección.

De lo anterior se deduce que una forma de obtener el coeficiente de elasticidad consiste en ajustar la función potencial, por el método de mínimos cuadrados, a los datos retrospectivos de que se disponga. Es decir, dada

$$Y = b\ X^a\ *$$

el ajuste a una nube conocida de puntos permitirá calcular los parámetros.

El valor de "a" corresponde, como se demostró, al coeficiente de elasticidad. Existe una forma aproximada de estimar este coeficiente de elasticidad por el método gráfico. En la expresión:

$$\log Y = \log b + a \log X$$

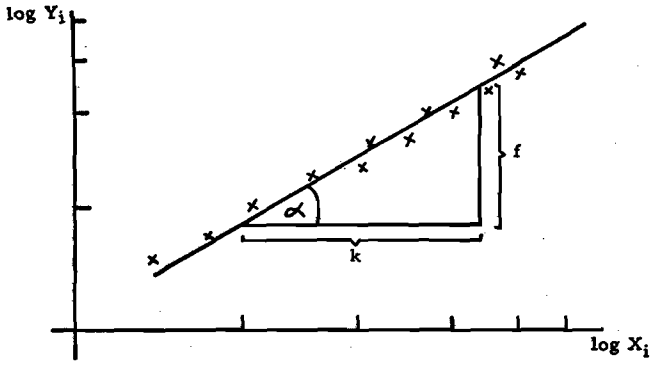
el coeficiente de elasticidad es el coeficiente angular de la recta logarítmica.

Si los puntos retrospectivos de que se dispone se representan en escalas logarítmicas, es posible a simple vista ajustar la recta tratando de aproximarse a la recta minimocuadrática.

---

* A la parte final de este capítulo se agrega una exposición sobre las diferencias de las proyecciones a través de ecuaciones de regresión y de coeficientes de elasticidad.

GRÁFICA 25



La recta "ajustada a ojo" puede entregar estimaciones muy cercanas al valor efectivo con un *ahorro* de tiempo considerable. La manera de obtener el valor de "a" es la siguiente:

$$a = T\,g\,\alpha = \frac{f}{k}$$

donde f es el cateto opuesto al ángulo α en el triángulo del gráfico *medido en centímetros* u otras unidades de longitud, y k es el cateto adyacente al ángulo α, también medido en las mismas unidades de f. El cociente dará la inclinación de la recta logarítmica y, por consiguiente, el coeficiente de elasticidad. La bondad de esta estimación depende de la *habilidad y cuidado* que se tenga para hacer pasar la recta por entre los puntos, de manera que se minimice el cuadrado de las diferencias.

Cualquiera de los dos métodos anteriores supone disponer de *estadísticas retrospectivas*. Ocurre con frecuencia que es necesario proyectar variables para las cuales no es posible recopilar suficientes antecedentes que permitan garantizar una cierta representatividad del coeficiente de elasticidad.

En estos casos es corriente utilizar *comparaciones internacionales*, eligiendo países que tengan similitudes marcadas con el país cuya proyección se necesita hacer. Por ejemplo, es posible utilizar el coeficiente de elasticidad gasto del consumo de artefactos eléctricos en Colombia, al realizar una primera estimación para Chile; entre ambos países existen características comunes en cuanto a población y su concentración, nivel de ingreso, grado de industria-

lización, etc. Otra manera sería seleccionar un conjunto de países dentro de un rango de nivel de ingreso comparable al del país considerado y calcular el coeficiente de elasticidad por los métodos anteriores, contando con informaciones de estos países en vez de las estadísticas retrospectivas mencionadas. Este método es conocido con el nombre de estimación del coeficiente de elasticidad por medio de datos internacionales; el último método tiene por objetivo generalmente la estimación de coeficientes de elasticidad ingreso de la demanda. Por último, otra manera de cuantificar coeficientes de elasticidad, principalmente elasticidad gasto, es la realización de muestras en un período dado, con las cuales se averigua los valores que toman las variables que interesa analizar y proyectar. Si bien el hecho de calcular un coeficiente de corte transversal en el tiempo y utilizarlo en proyecciones hacia el futuro, *tiene limitaciones,* hay que reconocer que si se tiene buen cuidado de definir las unidades muestrales de manera tal que reflejen internamente las condiciones de una cierta dinámica, las proyecciones no serán distorsionadas seriamente. Con anterioridad deben realizarse pruebas sobre la racionalidad y consistencia de los resultados alcanzables.

4] *La ecuación de regresión y el coeficiente de elasticidad como instrumentos de proyección.**

Como se ha visto, la ecuación de regresión puede utilizarse directamente como instrumento para estimar valores futuros de la variable dependiente, una vez planteadas determinadas hipótesis sobre el comportamiento de la variable independiente. Cabría discutir entonces qué diferencia existiría entre utilizar la ecuación de regresión o el coeficiente de elasticidad como instrumento de proyección, por ejemplo, de la demanda de un bien en función del ingreso. En primer término, la ecuación de regresión es de aplicabilidad mucho más general, ya que podría utilizarse cualquiera que fuese la forma de relación que se admita entre las dos variables; desde este punto de vista, el concepto de elasticidad no sería sino un caso particular donde, como se ha visto, se admite una relación logarítmica.

Podría suceder, sin embargo, que en determinados casos prácticos no pudiera disponerse de la ecuación de regresión correspondiente; en cambio, sí utilizar una estimación del coeficiente de elasticidad. Si sólo se tienen las cifras globales de consumo e ingreso para un único período reciente, podrían por ejemplo utili-

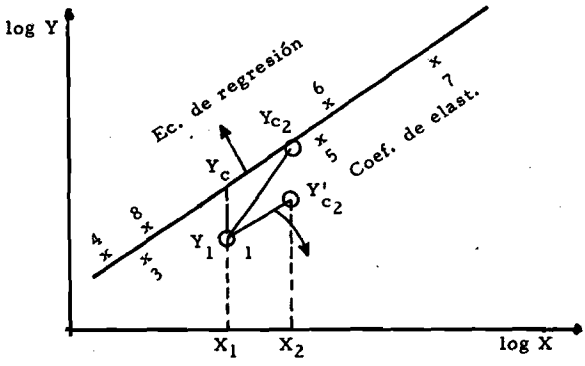
* El texto del punto 4 es del profesor Pedro Vuskovic.

zarse coeficientes de elasticidad deducidos de las experiencias de otros países de condiciones similares.

Aun cuando se dispusiera de los dos intrumentos —ecuación de regresión y coeficiente de elasticidad-ingreso— y se admitiera una relación logarítmica, los resultados de las proyecciones a que conducirían uno y otro serían diferentes en la generalidad de los casos. Supóngase, por ejemplo, que las comparaciones correspondientes se hayan referido al consumo de determinado bien en países con distinto nivel de ingreso (países 1, 2, 3, ... en la gráfica siguiente; interesa para la proyección, el país 1).

y: consumo por habitante
x: ingreso por habitante

GRÁFICA 26



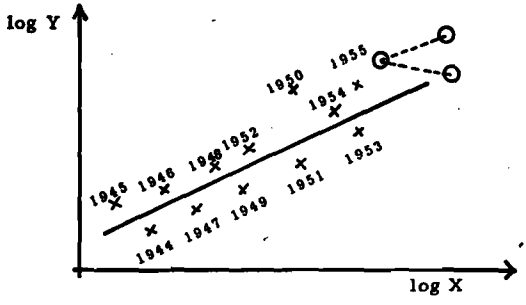
En la medida en que la relación entre las dos variables esté más alejada de la línea de regresión en el país que interesa para las proyecciones, mayor sería la diferencia a que se llegue utilizando la ecuación de regresión y el coeficiente de elasticidad como instrumento de proyección. Si, como en la gráfica anterior, la relación está en ese país por debajo de la línea de regresión, ello significaría que existe allí un consumo relativamente bajo (en comparación con el nivel de ingreso) del bien considerado; una estimación del consumo futuro basado sobre la ecuación de regresión supondría que tal situación se eliminaría, y el consumo tendería a aumentar no sólo por efecto del incremento del ingreso, sino también para superar ese retraso relativo; la utilización del coeficiente de elasticidad, en cambio, equivaldría a admitir que el consumo au-

mentará sólo por el efecto ingreso, pero que continuará registrándose un consumo relativamente bajo (en comparación con el nuevo nivel de ingreso).

En otras palabras, al aumentar el ingreso de X_1 a X_2 en la ecuación de regresión, se admite un aumento del consumo de Y_1 a Y_{c2}. En el primer caso, se supone que el nuevo nivel del consumo corresponderá exactamente al valor dado por la ecuación de regresión; en el segundo, se admite que perdurará una discrepancia entre el valor teórico (dado por la ecuación de regresión y el valor efectivo proporcionalmente igual al que existirá en el período base).

Es difícil juzgar en términos generales cuál de los dos métodos podría ser más adecuado ante una situación de esta índole. Si el nivel relativamente bajo del consumo en el país considerado es atribuible a limitaciones de la oferta, u otros factores de carácter temporal, podría ser más adecuada la proyección basada sobre la ecuación de regresión; si se debe, en cambio, a diferencias en materia de hábitos de los consumidores, a factores climáticos u otros de carácter relativamente permanente, sería más adecuada la proyección basada en el coeficiente de elasticidad. Aun en el primer caso sería necesario tener en cuenta si el período al que se refiere la proyección es lo suficientemente prolongado como para que lleguen a eliminarse los efectos adversos de los factores temporales. Las diferencias anotadas entre los dos métodos de proyección podrían presentarse también en el caso que todo el análisis se hubiera basado en series cronológicas correspondientes a un mismo país, puesto que las cifras correspondientes al período tomado como base seguramente serán diferentes de los valores teóricos dados por la ecuación de regresión.

GRÁFICA 27



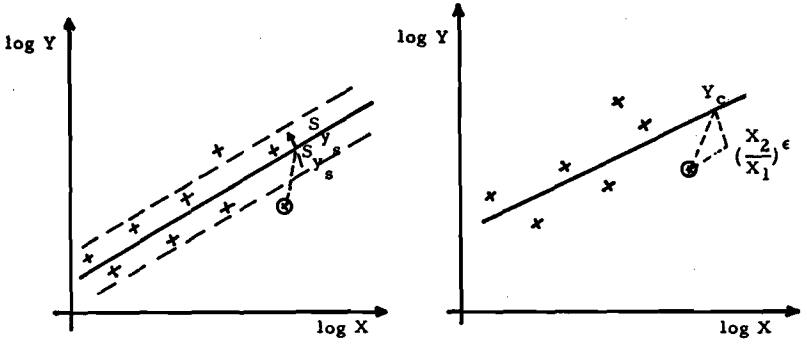
La interpretación sería, sin embargo, algo diferente, si se trata de un mismo país; el consumo relativamente bajo o relativamente

elevado) registrado en el período base sería, con mayor probabilidad, atribuible a factores de carácter temporal. En consecuencia, podría considerarse como más adecuada la proyección basada sobre la ecuación de regresión.

De modo que, en términos generales, la utilización de la ecuación de regresión y del coeficiente de elasticidad conducirá en la mayor parte de los casos a proyecciones diferentes, sin que resulte posible precisar cuál de las dos tendría que considerarse más adecuada. Esto puede llevar a la proyección de un intervalo probable para la variable dependiente, basado no en la magnitud del error estándar de estimación, sino en la diferencia entre la proyección obtenida con la ecuación de regresión y la proyección a que conduce la aplicación del coeficiente de elasticidad.

La gráfica anterior y la siguiente ilustran esta alternativa; en el primer caso se utilizan la ecuación de regresión y el error estándar de estimación para proyectar el intervalo correspondiente. Al utilizar $Y_c \pm C_{ys}$ se está admitiendo no sólo que se elimina el bajo

GRÁFICA 28



consumo relativo registrado durante el período base o en el país correspondiente (si se trata de una comparación internacional), sino que además se estima como probable que llegue a registrarse un consumo relativamente elevado durante el período abarcado por la proyección. Es evidente que las posibilidades prácticas de que esto ocurra son muy limitadas.

En el segundo caso, se utilizan la ecuación de regresión y el coeficiente de elasticidad, y se proyecta un intervalo delimitado por estos dos valores. De este modo se estima un intervalo más amplio por debajo de la línea de regresión y ninguno por encima de ésta,

lo que parecería más lógico en una situación como la supuesta en las gráficas.

5] *Relaciones y propiedades del coeficiente de elasticidad*

a) El coeficiente de elasticidad ingreso del consumo está relacionado con las propensiones media y marginal al consumo de la siguiente manera:

$$E = \frac{dY}{dX} \cdot \frac{X}{Y}$$

Si Y = consumo X = ingreso

$$\frac{dY}{dX} = \text{Propensión marginal a consumir}$$

$$\frac{X}{Y} = \text{Inverso de la propensión media al consumo}$$

Luego

$$E = \text{Propensión marginal} \cdot \frac{1}{\text{Propensión media}}$$

b) Si el consumo total X se divide en consumos parciales u_1, u_2, u_3 ... u_k, de manera que,

$$\sum_{i=1}^{k} u_i = X,$$

la media aritmética ponderada de las elasticidades gasto respectivas será igual a la unidad.

La definición de la elasticidad gasto en estos términos es la siguiente:

$$E_{gi} = \frac{du_i}{dX} \cdot \frac{X}{u_i}$$

donde i = 1, 2, ... k son los componentes del consumo total.

$$M\,[E_{gi}] = \Sigma\, E_{gi}\, Wi = 1$$

donde $Wi = \frac{u_i}{X}$ (participación porcentual del consumo específico del consumo total).

Remplazando E_{gi} y Wi por las definiciones, se tiene:

$$\Sigma \frac{du_i}{dX} \cdot \frac{X}{u_i} \left(\frac{u_i}{X}\right) = 1$$

Σ d $(u_i) =$ d x, dado que la suma de las diferencias es equivalente a la diferencia de la suma:

$$d [\Sigma u_i] = d X$$

recordando que $\Sigma u_i = X$

se tiene:

$$d X = d X$$

con lo que se comprueba la proposición enunciada.

6] *Elasticidad en regresión múltiple*

Parece oportuno presentar aquí el caso del cálculo de elasticidades simultáneas para más de una variable independiente. Es muy frecuente tratar con funciones potenciales múltiples cuando se deben encarar problemas de análisis económico; determinar, por ejemplo, en forma simultánea cómo juegan las elasticidades con respecto al precio y al ingreso en sus relaciones con la cantidad vendida; o cuál es la elasticidad de la tributación respecto a variaciones en las tasas y variaciones en el ingreso. El tratamiento simultáneo implica evitar la superposición que podría presentarse cuando se efectúan cálculos parciales por separado.

Sea la función:

$$Y = \alpha X^{\beta} W^{\gamma}$$

La elasticidad X de Y se encontrará derivando parcialmente la función respecto de X y multiplicando por la relación $\frac{X}{Y}$; en otras palabras, aplicando la definición de elasticidad.

$$E = \frac{dY}{dX} \cdot \frac{X}{Y}$$

En la función que se presenta habrá dos elasticidades: una que relaciona X con Y y otra que relaciona W con Y.

Para el primer caso es tiene:

$$\frac{dY}{dX} = \alpha\, W^{\gamma}\, \beta\, X^{\beta-1}$$

$$E_X = \alpha\, W^{\gamma}\, \beta\, X^{\beta-1} \cdot \frac{X}{Y}$$

pero $Y = \alpha\, X^{\beta}\, W^{\gamma}$

Luego,

$$E_X = \alpha\, W^{\gamma}\, \beta\, X^{\beta-1} \cdot \frac{X}{\alpha\, X^{\beta}\, W^{\gamma}}$$

$$E_X = \beta$$

Para el segundo caso se tiene:

$$E_W = \alpha\, X^{\beta}\, \gamma\, W^{\gamma-1} \cdot \frac{W}{\alpha\, X^{\beta}\, W^{\gamma}}$$

$$E_W = \gamma$$

Los exponentes de la función potencial múltiple corresponden a los conceptos de elasticidad respectivos.

Para determinar la magnitud de los parámetros α, β, γ, donde los dos últimos son las elasticidades mencionadas, se sigue el método tradicional del ajuste por mínimos cuadrados. Previamente será necesario "linealizar" la función aplicando los logaritmos, es decir:

$$\log Y = \log \alpha + \beta \log X + \gamma \log W$$

Interesa minimizar la expresión:

$$Z = \Sigma\ (\log Y_i - \log Y_c)^2$$

$$Z = \Sigma\ (\log Y_i - \log \alpha - \beta \log X - \gamma \log W)^2$$

y haciendo

$$\frac{\delta Z}{\delta \log \alpha} = 0$$

$$\frac{\delta Z}{\delta \beta} = 0$$

$$\frac{\delta Z}{\delta \gamma} = 0$$

se tienen las tres ecuaciones normales que permiten determinar los parámetros. Estas ecuaciones normales son:

$$\Sigma \log Y_i = n \log \alpha + \beta \Sigma \log X_i + \gamma \Sigma \log W_i$$

$$\Sigma (\log Y_i) \log X_i = \log \alpha \Sigma \log X_i + \beta \Sigma (\log X_i)^2 + \gamma \Sigma \log W_i \log X_i$$

$$\Sigma \log Y_i \log W_i = \log \alpha \Sigma \log W_i + \beta \Sigma \log X_i \log W_i + \gamma \Sigma (\log W_i)^2$$

donde Y_i, W_i, X_i son los valores observados de las tres variables, ya sea que correspondan a valores en el tiempo (temporal) o en el espacio (atemporal). Los límites de las sumatorias corresponden al total de observaciones que se disponga simultáneamente sobre las tres variables.

Con una función ajustada de esa manera, pueden realizarse proyecciones y análisis entre las variables. Para las proyecciones, como en el caso de regresión simple, queda la alternativa de hacerlo a través de la ecuación de regresión o a través de los coeficientes de elasticidad.

Para proyectar por medio de la ecuación de regresión, bastará con fijar exógenamente el comportamiento de las variables independientes y remplazar tales valores en la función. Si se desea proyectar a través de los coeficientes de elasticidad se tiene para el período 0:

$$\log Y_0 = \log \alpha + \beta \log X_0 + \gamma \log W_0$$

para el período n:

$$\log Y_n = \log \alpha + \beta \log X_n + \gamma \log W_n$$

restando ambas ecuaciones:

$$\log Y_n - \log Y_0 = \beta(\log X_n - \log X_0) + \gamma[\log W_n - \log W_0]$$

El antilogaritmo de la anterior relación conduce a

$$\frac{Y_n}{Y_0} = \left(\frac{X_n}{X_0}\right)^{\beta}\left(\frac{W_n}{W_n}\right)^{\gamma}$$

que es la fórmula básica de proyección a través de coeficientes de elasticidad en el caso de más de una variable independiente. Como ejemplo de una proyección de este tipo, admítanse los siguientes casos: Y variable que representa la recaudación efectiva tributaria; X variable que representa la tasa tributaria promedio, y W el Producto Geográfico real.

Si se tienen estimaciones que el producto crecerá en los próximos cinco años en 20 por ciento, siendo la elasticidad producto de la tributación unitaria, y se desea aumentar la tasa promedio en 44 por ciento, siendo la elasticidad tasa de la tributación equivalente a 0.5, el incremento porcentual de la recaudación tributaria será:

$$\frac{Y_n}{Y_0} - 1 = (1.44)^{0.5} \cdot (1.20)^1 - 1$$

$$\frac{Y_n}{Y_0} - 1 = (1.2)\ (1.2) - 1 = 1.44 - 1 = 0.44$$

Evidentemente que el tratamiento puede extenderse a más de dos variables independientes. La metodología presentada puede fácilmente ampliarse a tales casos.

7] *Limitaciones en la utilización de coeficientes de elasticidad*

Pese a la profusa, y quizá hasta exagerada utilización de coeficientes de elasticidad en las proyecciones económicas, cabe reconocer que su aplicación, en países donde se registran cambios político-económicos con inusitada frecuencia, tiene serias limitaciones. Se tratará el caso principal de los coeficientes de elasticidad ingreso. Un coeficiente de este tipo calculado, por ejemplo, con estadísticas retrospectivas, *lleva implícita la distribución de ingresos* y estructura de preferencias de los consumidores durante el período que

comprenden las estadísticas retrospectivas. Surge una objeción inmediata cuando se piensa en proyecciones hacia el futuro cuyo objetivo puede ser precisamente modificar esa distribución de ingresos. Igual objeción puede hacerse a las proyecciones por medio de coeficientes de elasticidad calculados a partir de muestras de corte transversal en el tiempo. De igual manera los coeficientes de elasticidad resultantes de comparaciones internacionales llevan implícita la distribución de ingresos de los países considerados.

El problema de las proyecciones supone un método de aproximaciones sucesivas; los tres métodos planteados para calcular coeficientes de elasticidad no son métodos alternativos sino complementarios. Un coeficiente calculado a través de estadísticas retrospectivas puede ser corregido por muestras sucesivas que registren los cambios en la distribución de ingresos. Por otra parte, la utilización de coeficientes "internacionales" supone seguir las distribuciones de ingresos de los países considerados y, en determinados casos, esa tendencia puede no estar demasiado apartada de los planes que sobre esta materia se formulen. En general, puede llegarse a proyecciones razonables considerando el problema como un método iterativo sujeto a revisiones periódicas. En los modelos temporales, *el transcurso del tiempo proporciona nuevas informaciones,* que, al tomarse en cuenta, modifican las proyecciones a medio y a largo plazo. Ése debería ser el verdadero sentido de las proyecciones y no la estimación esporádica. En las instituciones más avanzadas existen equipos de técnicos que están dedicados permanentemente a preparar, corregir, revisar e integrar modelos predictivos.

Como regla general, para calificar una proyección, debería establecerse, por una parte, un mínima razonable de confiabilidad en la proyección, porque malas proyecciones en general pueden ocasionar serios perjuicios; por otra parte, aunque se carezca de estimaciones certeras, disponer de aproximaciones, aunque burdas, siempre será mejor que el desconocer o ignorar las características de un fenómeno en el futuro.

TEMAS DE DISCUSIÓN

Indique si las siguientes afirmaciones son ciertas o falsas, y justifique su opinión.

1] Si la covarianza toma un valor negativo, el coeficiente de correlación es imaginario.

2] Los siguientes datos son consistentes en correlación rectilínea:

$$a_{YX} = 10 \qquad a_{XY} = 0.08$$

$$S^2_{Xc} = 16 \qquad S^2_{Xs} = 4$$

3] El error de proyección puede tomar valores superiores a la unidad.
4] La flexibilidad de la función potencial hace que los coeficientes de correlación sean siempre mayores que si el ajuste hubiese sido a una recta.
5] Los siguientes datos son consistentes:

$$Y_c = 0.2\ X_i - 4 \qquad V\ [X_i] = 50 \qquad \overline{X} = 30$$

$$\Sigma Y_i X_i = 50\ n$$

6] En la siguiente función:

$$X_1 = 4 + 2\ X_2 + 0.1\ X_3$$

el coeficiente de correlación simple entre X_1 y X_2 será mayor que el coeficiente de correlación simple entre X_1 y X_3.
7] Los siguientes datos son consistentes:

$$Y_c = 0.4\ X_i + 4 \qquad \text{(Y en X)}$$

$$X_c = 2.4\ Y_i - 3 \qquad \text{(X en Y)}$$

$$V\ (Y_i) = 20 \qquad \text{Error de proyección: 2}$$

8] Los siguientes datos son consistentes:

$$V\ [X_i] = 16 \qquad r = 0.3 \qquad S_{Ys} = 6 \qquad Y_c = 2\ X_i + 5$$

9] La ecuación normal de la función $Y_c = a\ X_i$ es $\Sigma\ Y_i = a\ \Sigma\ X_i$
10] Los siguientes datos son consistentes:

$$Y_c = 0.4\ X_i + 2$$

$$\overline{X} = \overline{Y}$$

$$\frac{\Sigma\ X_i\ Y_i}{n} = 20$$

$$V\ [X_i] = 50$$

11] Si las varianzas explicada y no explicada son iguales, el coeficiente de correlación será superior a 0.5.

12] Dada la función $Y = \frac{a}{X}$, la ecuación normal correspondiente será:

$$\Sigma Y_i = a \Sigma \frac{1}{X_i}$$

13] Los siguientes datos son perfectamente consistentes:

$$S^2_{Xc\ 1.234} = 200 \qquad S^2_{Xs\ 1.23} = 80$$

$$S^2_{X\ 1.23} = 260 \qquad S^2_{Xs\ 1.234} = 50$$

$$R^2_{1.234} = 0.80 \qquad r^2_{14.23} = 0.25$$

14] Mientras mayor número de parámetros tenga la ecuación de regresión, más alto será el valor del coeficiente de correlación.

15] Un coeficiente de correlación de 0.95, indica siempre una asociación significativa entre las variables.

16] Si el coeficiente de elasticidad ingreso de la demanda de papel es inferior a la unidad, quiere decir que el consumo de papel en períodos futuros será cada vez menor en cantidades absolutas.

17] Si el gasto total de una comunidad se divide entre consumo esencial y no esencial, es razonable admitir que, si los coeficientes de elasticidad gasto respectivos son 1.4 y 0.6 el volumen del gasto se distribuye entre ambos tipos de consumo, en partes iguales.

18] Si el producto interno bruto crece a una tasa de 8% simple anual y el valor agregado del sector construcción lo hace a una tasa acumulativa de 4% anual, quiere decir que la participación del sector construcción, dentro del producto será cada vez mayor.

PROBLEMAS PROPUESTOS

1] La ecuación de regresión entre el consumo total y el ingreso total de una región, es la siguiente:

$$C = 40 + 0.6\ Y$$

En 1967 la propensión media al consumo es de 80%; el 20% del consumo total está constituido por productos importados, la elasticidad gasto del consumo de productos nacionales es de 0.8. Si se estima que el ingreso en 1972 será un 40% mayor que en 1967, ¿cuál será el valor de las importaciones de bienes de consumo en 1972?

2] El ajuste entre consumo (C) y consumo de artefactos (CA) proporcionó la siguiente función:

$$CA = 0.4\ C^{1.5}$$

Si el consumo de artefactos representa el 50% del total, ¿qué porcentaje del consumo representará cuando éste se duplique?

3] Sobre el consumo de acero en una región, se tienen los siguientes antecedentes:

Años	*Consumo de acero (miles de ton)*	*Valor del consumo de acero (mill de u.m. corrientes)*
1960	200	10
1961	250	15
1962	320	20
1963	400	28

Se pide calcular la tendencia rectilínea del índice de precios del acero (el tiempo como variable independiente) y estimar el valor probable del índice en 1965 con base en 1963.

4] Se desea estimar el monto de gastos en consumo de alimentos para 1967. Para ello se dispone de una ecuación de regresión entre el ingreso real y el consumo total real del siguiente tipo:

$$C = 0.9\ Y^{0.7}$$

Por otra parte se sabe que la elasticidad gasto de alimentos es de 0.8. Realice la proyección deseada si además se sabe que el consumo en alimentos en el año 1964 fue de 2 000 millones de unidades monetarias, siendo el siguiente el índice de base variable del ingreso real según el Plan de Desarrollo:

1964	*1965*	*1966*	*1967*
—	105	106	106

EJERCICIOS

1] Ajuste los siguientes datos:

Años	*Ingreso* (Y) *(u.m. constantes)*	*Consumo* (C)
1957	2.0	1.6
1958	2.1	1.7
1959	2.4	2.0
1960	2.4	2.1
1961	2.5	2.2
1962	2.8	2.5
1963	3.0	2.6

a) a una recta: $C = a\ Y + b$
b) a una potencial: $C = b\ Y^a$
c) calcule los coeficientes de correlación respectivos, comentando los resultados;
d) calcule la tasa de crecimiento del consumo mediante la función: $C = a\ b^t$
e) estime la propensión media al consumo en 1965 si admite que el ingreso será en ese período de 3.7.

2] Los siguientes datos corresponden a precios y cantidades transadas de cierto artículo en un mercado:

Precio (p) *(u.m.)*	*Cantidad vendida* (q) *(miles de u.m.)*
10	2.00
12	1.60
15	1.50
18	1.20
20	1.00
25	0.80
30	0.30

Se le pide:
a) Determine la función de demanda mediante:

i) $q = \frac{a}{p}$

ii) $q = \frac{a}{p} + b$

b) Calcule los coeficientes de correlación respectivos;
c) Calcule los errores de proyección;
d) Estime la cantidad vendida a un precio de 40, por medio de las dos funciones;
e) Estime el precio que garantizaría una venta de 3.00 miles de unidades.

3] Con el objeto de estudiar la relación entre las variables consumo de energía eléctrica (X_i) y volumen de producción en las empresas industriales (Y_i), se tomó una muestra de 20 empresas, para las cuales se computaron los siguientes valores:

$$\Sigma X_i = 11.34 \qquad \Sigma Y_i = 20.72 \qquad \Sigma X_i^2 = 12.16$$

$$\Sigma Y_i^2 = 84.96 \qquad \Sigma X_i Y_i = 22.13$$

Se le pide:
a) Calcule las ecuaciones de regresión de "Y en X" y de "X en Y";
b) Calcule el coeficiente de correlación rectilíneo;
c) Calcule el error de proyección de la regresión de "Y en X".

4] En el diseño de un modelo de simulación se necesitaba disponer de una función consumo de bienes de origen industrial; para lograrlo se tienen los siguientes datos:

Años	*Consumo de bienes industriales* *Unidades*	*Ingreso disponible* *monetarias*	*Importaciones bienes de consumo* *constantes*
1960	45	52	10
1961	42	58	13
1962	48	58	10
1963	55	60	14
1964	53	65	16
1965	65	70	18

Se le pide:
a) Ajuste una función del tipo: $C = \alpha Y + \beta M + \gamma$
donde: α, β, γ, parámetros de regresión
C: consumo de bienes industriales
Y: ingreso disponible
M: importaciones de bienes de consumo
b) Calcule el coeficiente de correlación múltiple y el error de proyección;
c) Calcule el coeficiente de correlación parcial entre consumo e ingreso, aislando el efecto de las importaciones.

5] Interesa disponer de estimaciones de las variaciones en los pre-

cios de bienes agrícolas de consumo esencial; para lograrlo, después de algunos estudios, se concluyó que una metodología posible podría ser el ajuste de una ecuación de regresión a los siguientes datos:

Período	*Precio de bienes agrícolas* P	*Costo unitario de producción* C	*Expectativas de inflación* V
1	10	6	6%
2	12	7	8%
3	15	12	4%
4	17	13	7%
5	20	16	12%
6	30	25	12%
7	35	25	8%

Se le pide:

a) Ajuste una función del tipo: $P\ (t + 1) = C\ (t)\ [1 + a + b\ V\ (t)]$

donde, a: margen de ganancia de empresarios agrícolas

b: coeficiente de realimentación de inflación

b) Estime el precio cuando el costo sea de 40 y hayan expectativas de estabilidad en la actividad económica.

6] De acuerdo al plan de desarrollo de un país, se comprueba que el producto real del sector agrícola muestra los siguientes crecimientos porcentuales anuales en el período 1956/1962:

1957	*1958*	*1959*	*1960*	*1961*	*1962*
2.0	3.0	4.0	3.5	3.0	3.3

a) Determine la tasa anual promedio de crecimiento para el período considerado utilizando la ecuación de regresión adecuada.*

b) El año 1962, el sector agrícola aportaba el 30% del producto interno bruto (PIB). Si se emplea la ecuación de regresión encontrada y se supone que el PIB crecerá el 5% anual acumula-

* Por razones implícitas en el modelo de proyección, la recta queda excluida como posible solución.

tivo durante el período 1962/1970, estime la participación que tendrá el sector agrícola en el PIB, el año 1970.

7] Los siguientes datos corresponden a ingresos y consumos de un conjunto de profesionales.

Período (t)	*Ingreso* (Y)	*Consumo* (C)
0	100	60
1	120	65
2	120	66
3	130	70
4	140	72
5	160	75
6	160	80

Se le pide calcular el valor de los parámetros en la función consumo de Friedman:

$$C\ (t) = \alpha + \beta\ Y\ (t) + \gamma\ C\ (t - 1)$$

8] Para una región de cierto país, se tienen los siguientes antecedentes:

Años	*Ingreso por habitante*	*Consumo por habitante*	*Consumo de alimentos por habitante*
	Unidades monetarias constantes		
1958	200	180	120
1959	220	210	130
1960	245	230	150
1961	270	250	170
1962	300	280	200
1963	340	320	220

Se le pide:

a) Calcule la elasticidad ingreso del consumo mediante la función: $Y_c = b\ X_i^a$;

b) Estime, utilizando el método gráfico, el coeficiente de elasticidad gasto del consumo de alimentos;

c) Teniendo en cuenta las dos elasticidades calculadas anteriormente, estime el consumo de alimentos en 1970, admitiendo que el ingreso por habitante crecerá a partir de 1963, a una tasa de 3% acumulativo anual.

8] La distribución del gasto en 1960 es la siguiente:

Alimentos	100
Otros productos manufacturados	60
Servicios	40
Consumo total	200

Si además se sabe que la elasticidad gasto de alimentos es 0.8 y la de productos manufacturados 1.2, se le pide calcule:

a) El coeficiente de elasticidad gasto de servicios;

b) La distribución del gasto en 1966, considerando que el ingreso crecerá al 4% anual, si se mantiene constante la propensión media al consumo, al nivel de 0.85.

10] En 1955 el consumo de papel para cierto país se calculó en 4.5 kg por habitante. Con el propósito de programar el desarrollo de la industria de la celulosa, se necesita estimar el consumo para 1967, teniendo en cuenta los siguientes antecedentes:

a) Elasticidad gasto de la demanda *per capita* de papel: 1.5;

b) Crecimiento del ingreso: 4% acumulativo anual;

c) Ecuación del gasto total en relación al ingreso total: $G = 0.9\ Y$.

Agregue, si necesita, la información adicional que estime útil para efectuar la proyección.

SOLUCIÓN DE EJERCICIOS

1] *a)* Las ecuaciones normales son:

$$\Sigma\ C = a\ \Sigma\ Y + nb$$

$$\Sigma\ CY = a\ \Sigma\ Y^2 + b\ \Sigma\ Y$$

Y	C	YC	C^2	Y^2
2.0	1.6	3.20	2.56	4.00
2.1	1.7	3.57	2.89	4.41
2.4	2.0	4.80	4.00	5.76
2.4	2.1	5.04	4.41	5.76
2.5	2.2	5.50	4.84	6.25
2.8	2.5	7.00	6.25	7.84
3.0	2.6	7.80	6.76	9.00
17.2	14.7	36.91	31.71	43.02

Si se remplazan estos valores en las ecuaciones normales:

$$14.7 = 17.2\ a + 7b$$

$$36.91 = 43.02\ a + 17.2\ b$$

Si se resuelve el sistema: $a = 1.043$; $b = -0.463$
la función queda: $C = 1.043\ Y - 0.463$

b) Para el ajuste de la potencial, será necesario aplicar logaritmos, para utilizar las ecuaciones normales

$$C = b\ Y^a$$

$$\log C = \log b + a \log Y$$

Las ecuaciones normales son:

$$\Sigma \log C = n \log b + a \Sigma \log Y$$

$$\Sigma \log C \log Y = \log b \Sigma \log Y + a \Sigma (\log Y)^2$$

	log C	*log Y*	*log C log Y*	*(log C)²*	*(logY)²*
	0.20412	0.30103	0.061446	0.041665	0.090619
	0.23045	0.32222	0.074256	0.053107	0.103826
	0.30103	0.38021	0.114455	0.090619	0.144560
	0.32222	0.38021	0.122511	0.103826	0.144560
	0.34242	0.39794	0.136263	0.117251	0.158256
	0.39794	0.44716	0.177943	0.158356	0.199952
	0.41497	0.47712	0.197990	0.172200	0.227643
Total	2.21315	2.70589	0.884864	0.737024	1.069516

Si se remplazan estos valores en las ecuaciones normales

$$2.213150 = 7 \log b + 2.70589\ a$$

$$0.884864 = 2.705890 \log b + 1.069516\ a$$

Si se resuelve el sistema:

$$a = 1.247$$

$$\log b = -0.166$$

La función quedará:

$$\log C = 1.247 \log Y - 0.166$$

$$C = \frac{1}{1.306} Y^{1.247}$$

c) Los coeficientes de correlación serán:
para la recta:

$$r^2 = \frac{a \Sigma Y_i C_i + b \Sigma C_i - \overline{C} \Sigma C_i}{\Sigma C_i^2 - \overline{C} \Sigma C_i}$$

$$= \frac{1.043\ (36.91) - 0.463\ (14.7) - 2.1\ (14.7)}{31.71 - 2.1\ (14.7)}$$

$$= \frac{38.4971 - 6.8061 - 30.87}{31.71 - 30.87} = \frac{0.82}{0.84} = 0.976$$

$$r = + \sqrt{0.976} = 0.987$$

para la potencial:

$$r^2 \log C \log Y = \frac{a \Sigma \log Y_i \log C_i + \log b \Sigma \log C_i - \overline{\log C} \Sigma \log C_i}{\Sigma (\log C_i)^2 - \overline{\log C} \Sigma \log C_i}$$

$$r^2 \log C \log Y = \frac{1.247\ (0.884864) - 0.166\ (2.21315) - 0.31616\ (2.21315)}{0.737024 - 0.31616\ (2.21315)}$$

$$r^2 \log C \log Y = \frac{1.10343 - 0.36738 - 0.69971}{0.737024 - 0.69971} = \frac{0.03634}{0.03731} = 0.974$$

$r \log C \log Y = 0.987$

d) Las ecuaciones normales después de aplicar logaritmos son:

$$\Sigma \log C_i = n \log a + \log b \Sigma t_i$$

$$\Sigma (\log C_i)\ t_i = \log a \Sigma t_i + \log b \Sigma t_i^2$$

Los datos se disponen en la siguiente forma:

t_i	C_i	$\log C_i$	$t_i \log C_i$	t_i^2
−3	1.6	0.20412	−0.61236	9
−2	1.7	0.23045	−0.46090	4
−1	2.0	0.30103	−0.30103	1
0	2.1	0.32222	0	0
1	2.2	0.34242	0.34242	1
2	2.5	0.39794	0.79588	4
3	2.6	0.41497	1.24491	9
0		2.21315	1.00892	28

Remplazando estos valores en las ecuaciones normales:

$$2.21315 = 7 \log a + 0$$

$$1.00892 = 0 + 28 \log b$$

$$\log a = 0.3162 \therefore a = 2.071$$

$$\log b = 0.0360 \therefore b = 1.0865$$

En la función se tenía $C = a\, b^t$, donde $b = 1 + i$ en que "i" es la tasa de crecimiento acumulativo. En el problema $i = 0.0865 = 8.65\%$ acumulativa anual.

e) Para encontrar la propensión media al consumo en 1965, es necesario disponer de estimaciones de consumo e ingreso para ese año; el consumo fácilmente puede estimarse a través de la función encontrada en d.

$$C_{1965} = 2.071\ (1.0865)^5 \qquad t = 5 \text{ para } 1965$$

$$C_{1965} = 2.071\ (1.7313) = 3.5855$$

El ingreso está dado; luego,

$$\text{Propensión media el consumo} = \frac{3.5855}{3.7000} = 0.969$$

2] *a)* Para responder i e ii, será necesario determinar sus ecuaciones normales.

i) $\Sigma\, q/p = a\, \Sigma\, 1/p^2$

ii) $\Sigma q = a \Sigma \frac{1}{p} + nb$

$\Sigma q/p = a \Sigma 1/p^2 + b \Sigma 1/q$

Se calcula a continuación los datos para remplazar en estas ecuaciones:

q	p	q/p	$1/p$	$1/p^2$
2.00	10	0.2000	0.1000	0.010000
1.60	12	0.1333	0.0833	0.006941
1.50	15	0.1000	0.0667	0.004444
1.20	18	0.0667	0.0556	0.003309
1.00	20	0.0500	0.0500	0.002500
0.80	25	0.0320	0.0400	0.001600
0.30	30	0.0100	0.0333	0.001111
8.40		0.5920	0.4288	0.029688

Remplazando valores;

i) $0.5920 = 0.2968\ a$

$a = 19.946$

La función queda:

$q = 19.946/p$

ii) $8.400 = 0.42880\ a + 7\ b$

$0.5920 = 0.02968\ a + 0.4288\ b$

$a = 22.6966$

$b = -0.1903$

La función queda: $q = \frac{22.6966}{p} - 0.1903$

b) Para calcular los coeficientes de correlación, se utilizará la fórmula la general:

$$r = \left(\frac{\Sigma (Y_c - \bar{Y})^2}{\Sigma (Y_i - \bar{Y})^2} \right)^{1/2}$$

Caso i		*Caso ii*		*Ambos casos*	
q_c	$(q_c - \bar{q})^2$	q_c	$(q_c - \bar{q})^2$	q_i	$(q_i - \bar{q})^2$
1.995	0.632	2.079	0.773	2.0	0.64
1.662	0.213	1.700	0.250	1.6	0.16
1.330	0.017	1.324	0.015	1.5	0.09
1.108	0.008	1.072	0.016	1.2	—
0.997	0.041	1.025	0.031	1.0	0.04
0.798	0.162	0.718	0.232	0.8	0.16
0.665	0.286	0.565	0.403	0.3	0.81
	1.359		1.720	8.4	1.90

$$\bar{q} = \frac{\Sigma q_i}{n} = \frac{8.4}{7} = 1.20$$

Caso i)

$$r = \left(\frac{1.359}{1.900}\right)^{1/2} = (0.715)^{1/2} = 0.845$$

Caso ii)

$$r = \left(\frac{1.720}{1.900}\right)^{1/2} = (0.905)^{1/2} = 0.952$$

c) Para el cálculo de los errores de proyección recuérdese que

$$S_Y^2 = S_{Yc}^2 + S_{Ys}^2$$

$$\text{Error} = \sqrt{S_{Ys}^2} = S_{Ys}$$

Caso i)

$$S_Y^2 = \frac{1.90}{7} = 0.27$$

$$S_{Yc}^2 = \frac{1.359}{7} = 0.19$$

$$S_{Yc}^2 = 0.08$$

$$\text{Error} = 0.283$$

Caso ii)

$$S_Y^2 = \frac{1.90}{7} = 0.27$$

$$S_{Yc}^2 = \frac{172}{7} = 0.25$$

$$S_{Yc}^2 = 0.02$$

$$\text{Error} = 0.142$$

d) Basta remplazar $p = 40$ en ambas funciones:

Caso i)

$$q = \frac{19.946}{p} = \frac{19.946}{40} = 0.499$$

Caso ii)

$$q = \frac{22.6966}{p} - 0.1903$$

$$= \frac{22.6966}{40} - 0.1903 = 0.377$$

e) Basta remplazar $q = 3.00$

Caso i)

$$3 = \frac{19.946}{p} \therefore p = 6.65$$

Caso ii)

$$3 = \frac{22.6966}{p} - 0.1903$$

$$p = 7.11$$

3] *a)* Para el caso de la regresión de "Y en X" se sabe que

$$a_{YX} = \frac{C[X_i Y_i]}{V[X_i]} = \frac{\dfrac{\Sigma X_i Y_i}{n} - \overline{X}\,\overline{Y}}{\dfrac{\Sigma X_i^2}{n} - \overline{X}^2}$$

Remplazando valores se tiene

$$a_{YX} = \frac{\dfrac{22.13}{20} - \dfrac{11.34}{20} \cdot \dfrac{20.72}{20}}{\dfrac{12.16}{20} - \left(\dfrac{11.34}{20}\right)^2} = \frac{1.1065 - 0.567 \cdot 1.036}{0.608 - 0.321} =$$

$$= \frac{0.519}{0.287} = 1.808$$

Para determinar el coeficiente de posición se tiene

$$\overline{Y} = a_{YX}\,\overline{X} + b_{YX}$$

$$b_{YX} = \frac{20.72}{20} - 1.808\,\frac{11.34}{20} = 1.036 - 1.025 = 0.009$$

La ecuación de "Y en X" queda: $Y_c = 1.808\ X_i + 0.09$

$$a_{XY} = \frac{C[X_i Y_i]}{V[Y_i]} = \frac{0.519}{\dfrac{84.96}{20} - \left(\dfrac{20.72}{20}\right)^2} = \frac{0.519}{4.248 - 1.073} = 0.163$$

Para determinar el coeficiente de posición, se tiene

$$\overline{X} = a_{XY}\,\overline{Y} + b_{XY}$$

$$b_{XY} = \frac{11.34}{20} - 0.163\,\frac{20.72}{20} = 0.567 - 0.169 = 0.396$$

La ecuación de "X en Y" queda:

$$X_c = 0.163\ Y_i + 0.396$$

b) Para el cálculo del coeficiente de correlación rectilíneo, se tiene que

$$r^2 = a_{XY}\, a_{YX} = (0.163)\ \ (1.808) = 0.295$$

$$r = + \sqrt{0.295} = 0.543$$

c) Para el cálculo del error de proyección recuérdese que

$$r = \frac{S^2_{Ys}}{S_Y{}^2} = 1 - \frac{S^2_{Ye}}{S_Y{}^2}$$

$$0.295 = 1 - \frac{S^2_{Ys}}{3.175} \text{ ya que } S^2_Y = V\,[Y_i] = 3.175$$

$$S^2_{Ys} = (0.705)\ \ (3.175) = 2.238$$

$$\text{Error de proyección } S_{Ys} = \sqrt{S^2_{Ys}} = \sqrt{2.238} = 1.5$$

4] Si se llama X_1 al consumo, X_2 al ingreso, X_3 a las importaciones, las ecuaciones normales son:

a)

$$\Sigma\, X_1 = \alpha\, \Sigma\, X_2 + \beta\, \Sigma\, X_3 + n\, \gamma$$

$$\Sigma\, X_1X_2 = \alpha\, \Sigma\, X_2^2 + \beta\, \Sigma\, X_2X_3 + \gamma\, \Sigma\, X_2$$

$$\Sigma\, X_1X_3 = \alpha\, \Sigma\, X_2X_3 + \beta\, \Sigma\, X_3^2 + \gamma\, \Sigma\, X_3$$

Los datos se disponen así:

X_1	X_2	X_3	X_1X_2	X_2^2	X_2X_3
45	52	10	2 340	2 704	520
42	58	13	2 436	3 364	754
48	58	10	2 784	3 364	580
55	60	14	3 300	3 600	840
53	65	16	3 445	4 225	1 040
65	70	18	4 550	4 900	1 260
308	363	81	18 855	22 157	4 994

X_1X_3	X_3^2	X_1^2
450	100	2 025
546	169	1 764
480	100	2 304
770	196	3 025
848	256	2 809
1 170	324	4 225
4 264	1 145	16 152

El sistema queda:

$$308 = 363\ \alpha + 81\ \beta + 6\ \gamma$$

$$18.855 = 22157\ \alpha + 4994\ \beta + 363\ \gamma$$

$$4.264 = 4994\ \alpha + 1145\ \beta + 81\ \gamma$$

Cuya resolución es

$$\alpha = b_{12.3} = 1.114$$

$$\beta = b_{13.2} = 0.040$$

$$\gamma = a_{1.23} = -16.59$$

b) La fórmula del coeficiente de correlación múltiple que se utiliza es la siguiente:

$$R_{1.23} = \left(\frac{a_{1.23} \Sigma X_1 + b_{12.3} \Sigma X_1X_2 + b_{13.2} \Sigma X_1X_3 - n \bar{X}_1^2}{\Sigma X_1^2 - n \bar{X}_1^2} \right)^{1/2}$$

Si se remplazan valores:

$$R_{1.23} = \left(\frac{-16.59\ (308) + 1.114\ (18\,855) + 0.04\ (4\,264) - 6\ (2\,635)}{16\,152 - 6\ (2\,635)} \right)^{1/2}$$

$$R_{1.23} = \left(\frac{-5\,109.7 + 21\,004.5 + 170.6 - 15\,810.0}{16\,152 - 15\,810} \right)^{1/2}$$

$$R_{1.23} = \left(\frac{255.4}{342.0}\right)^{1/2} = (0.747)^{1/2} = 0.865$$

Para el cálculo del error de proyección, recuérdese que

$$R^2_{1.23} = \frac{n\, S^2_{Xc\ 1.23}}{n\, S^2_{X\ 1.23}}$$

De los datos anteriores se tiene

$$n\, S^2_{Xc\ 1.23} = 255.4 \quad \therefore \quad S^2_{Xc\ 1.23} = 42.6$$

$$n\, S^2_{X\ 1.23} = 342 \quad \therefore \quad S^2_{X\ 1.23} = 57$$

Luego

$$S^2_{Xs\ 1.23} = \frac{86.6}{6} = 14.4$$

$$\text{Error} = \sqrt{14.4} = 3.8$$

c) Se trata de calcular:

$$r_{12.3} = \left(\frac{S^2_{Xc\ 1.23}\ S^2_{Xc\ 1.3}}{S^2_{Xs\ 1.3}}\right)^{1/2}$$

En esta fórmula se desconocen los valores de las varianzas explicada y no explicada entre las variables X_1 y X_3; en cambio el valor de $S^2_{Xc\ 1.23}$ se obtiene del cálculo anterior

$$S^2_{Xc\ 1.23} = \frac{255.4}{6} = 42.6$$

Para las otras varianzas se recurre a las siguientes fórmulas:

$$n\, S^2_{Xc\ 1.3} = a_{1.3} \Sigma X_1 + b_{1.3} \Sigma X_1 X_3 - n\, \overline{X^2_1}$$

La ecuación de ajuste rectilíneo es

$$X_{1.3} = a_{1.3} + b_{1.3} X_3$$

Sus ecuaciones normales:

$$\Sigma X_1 = n\, a_{1.3} + b_{1.3} \Sigma X_3$$

$$\Sigma X_1X_3 = a_{1.3} \Sigma X_3 + b_{1.3} \Sigma X_3^2$$

Todas las sumatorias aparecen tabuladas en la primera hoja. Remplazando valores se tiene:

$$308 = 6\ a_{1.3} + 81\ b_{1.3}$$

$$4261 = 81\ a_{1.3} + 1145\ b_{1.3}$$

$$a_{1.3} = 23.52$$

$$b_{1.3} = 2.06$$

Remplazando valores en la fórmula de la varianza explicada, se tiene

$$S^2_{Xc\ 1.3} = 23.52\ (308) + 2.06\ (4\,264) - 6\ (2\,635)$$

$$S^2_{Xc\ 1.3} = \frac{7\,244.2 + 8\,783.8 - 15\,810}{6} = \frac{218}{6}$$

$$S^2_{Xc\ 1.3} = 36.3$$

Recuérdese que

$$S^2_{X\ 1.3} = S^2_{Xc\ 1.3} + S^2_{Xs\ 1.3}$$

La varianza total de la variable dependiente es única, cualquiera sea el número de variables independientes que se tomen; en consecuencia, este valor es el mismo que aparece en el denominador del coeficiente de correlación múltiple.

$$S^2_{Xs\ 1.3} = 57 - 36.3 = 20.7$$

Luego,

$$r_{12.3} = \left(\frac{42.6 - 36.3}{20.7}\right)^{1/2} = \sqrt{0.304} = 0.551$$

5) *a)* La ecuación de ajuste puede representarse así:

$$P\ (t + 1) = (1 + a)\ C\ (t) + b\ C\ (t)\ V\ (t)$$

Llamando

$$P = X_1$$

$$C\ (t) = X_2$$

$$C\ (t)\ V\ (t) = X_3$$

$$1 + a = d$$

La ecuación queda:

$$X_1 = d\ X_2 + b\ X_3$$

Las ecuaciones normales son las siguientes:

$$\Sigma\ X_1X_2 = d\ \Sigma\ X_2^2 + b\ \Sigma\ X_3X_2$$

$$\Sigma\ X_1X_3 = d\ \Sigma\ X_3X_2 + b\ \Sigma\ X_3^2$$

La tabulación de valores se dispone así:

P X_1	C X_2	CV X_3	PC X_1X_2	C^2 X_2^2	VC^2 X_3X_2	PCV X_1X_3	$(CV)^2$ X_3^2
12	6	0.36	72	36	2.16	4.32	0.1296
15	7	0.56	105	49	3.92	8.40	0.3136
17	12	0.48	204	144	5.76	8.16	0.2304
20	13	0.91	260	169	11.83	18.20	0.8281
30	16	1.92	480	256	30.72	57.90	3.6864
35	25	3.00	875	625	75.00	105.00	9.0000
129	79	7.23	1 996	1 279	129.39	201.98	14.1881

El sistema queda

$$1996.00 = 1279.00\ d + 129.39\ b$$

$$201.98 = 129.39\ d + 14.1881\ b$$

Resolviendo para d y b

$$d = 1.555$$

$$b = 0.054$$

Luego la ecuación ajustada queda

$$P\ (t + 1) = C\ (t)\ [1.555 + 0.054\ V\ (t)]$$

b) Bastará remplazar C (t) por 40 y V por 0, para estimar P (t + 1), en tales circunstancias

$$P\ (t + 1) = 40\ [1.555 + 0] = 62.22$$

6] En vista de que hay variaciones porcentuales anuales, se calculará un índice de base fija, y luego mediante la función

$$Y = a\ b^t$$

se calculará la tasa pedida.

a)

Índice del prod. real agrícola	*Tiempo*	
Y_i	t_i	años
100.0	−3	1956
102.0	−2	1957
105.1	−1	1958
109.3	0	1959
113.1	1	1960
116.5	2	1961
120.3	3	1962
	0	

Las ecuaciones normales son:

$$\Sigma \log Y_i = n \log a + \log b\ \Sigma\ t_i$$

$$\Sigma\ t_i \log Y_i = \log a\ \Sigma\ t_i + \log b\ \Sigma\ t_i^2$$

$\log Y_i$	t_i^2	$(\log Y_i)\ t_i$
2.00000	9	−6.00000
2.00860	4	−4.01720
2.02160	1	−2.02160
2.03862	0	—
2.05346	1	2.05346
2.06633	4	4.12516
2.08027	9	6.23754
14.26888	28	0.37736

Si se remplazan estos valores en las ecuaciones normales

$$14.26888 = 7 \log a \qquad (\text{Ya que } \Sigma t_i = 0)$$

$$0.37736 = 28 \log b$$

$$\log a = \frac{14.26888}{7} = 2.03841 \quad \therefore \quad a = 109.24$$

$$\log b = \frac{0.37736}{28} = 0.01348 \quad \therefore \quad b = 1.0315$$

La función queda

$$Y_c = 109.24 \ (1.0315)^t$$

La tasa de crecimiento es 3.15% acumulativa anual.

b) Para encontrar la participación relativa, se puede trabajar con los mismos porcentajes dados en los datos iniciales.

	1962	*1970*
PIB	100	$100\ (1 + 0.05)^8$
P. agríc.	30	$30\ (1 + 0.0315)^8$

$$\text{PIB}_{1970} = 100 \ (1.477) = 147.7$$

$$\text{P. agríc.}_{1970} = 30 \ (1.282) = 38.5$$

La participación relativa del sector agrícola en 1970 será

$$\frac{38.5}{147.7} = 26.07\%$$

7] Los datos se disponen de la siguiente manera (se introduce el desajuste del consumo)

t	$Y(t)$	$C(t)$	$C(t-1)$
1	120	65	60
2	120	66	65
3	130	70	66
4	140	72	70
5	160	75	72
6	160	80	75
	830	428	408

Las ecuaciones normales son las siguientes:

$$\Sigma\, C\ (t) = n\ \alpha + \beta\ \Sigma\, Y\ (t) + \gamma\ \Sigma\, C\ (t-1)$$

$$\Sigma\, C\ (t)\ Y\ (t) = \alpha\ \Sigma\, Y\ (t) + \beta\ \Sigma\, [Y\ (t)]^2 + \gamma\ \Sigma\, C\ (t-1)\ Y\ (t)$$

$$\Sigma\, C\ (t)\ C\ (t-1) = \alpha\ \Sigma\, C\ (t-1) + \beta\ \Sigma\, Y\ (t)\ C\ (t-1) + \gamma\ \Sigma\, [C\ (t-1)]^2$$

Es necesario tabular además las siguientes sumatorias.

$C(t)\,Y(t)$	$[Y(t)]^2$	$C(t-1)\ Y(t)$	$C(t)\,C(t-1)$	$[C(t-1)]^2$
7 800	14 400	7 200	3 900	3 600
7 920	14 400	7 800	4 290	4 225
9 100	16 900	8 580	4 620	4 356
10 080	19 600	9 800	5 040	4 900
12 000	25 600	11 520	5 400	5 184
12 800	25 600	12 000	6 000	5 625
59 700	116 500	56 900	29 250	27 890

El sistema de ecuaciones normales queda:

$$428 = 6\ \alpha + 830\ \beta + 408\ \gamma$$

$$59\,700 = 830\ \alpha + 116\,500\ \beta + 56\,900\ \gamma$$

$$29\,250 = 408\ \alpha + 56\,900\ \beta + 27\,890\ \gamma$$

Si se resuelve el sistema se llega a

$$\alpha = 38.017$$

$$\beta = 0.1333$$

$$\gamma = 0.2187$$

La función consumo queda

$$C\ (t) = 38.017 + 0.1333\ Y\ (t) + 0.2187\ C\ (t-1)$$

8] *a)* Las ecuaciones normales de la función son:

$$\Sigma \log Y_i = n \log b + a\ \Sigma \log X_i$$

$$\Sigma \log Y_i \log X_i = \log b \, \Sigma \log X_i + a \, \Sigma (\log X_i)^2$$

Siendo X_i = Ingreso; Y_i = Consumo

X_i	Y_i	$\log X_i$	$\log Y_i$	$\log X_i \log Y_i$	$\log X_i^2$
200	180	2.30103	2.25527	5.18944	5.29474
220	210	2.34242	2.32222	5.43961	5.48693
245	230	2.38917	2.36173	5.64257	5.70813
270	250	2.43136	2.39794	5.83026	5.91151
300	280	2.47712	2.44716	6.06191	6.13612
340	320	2.53148	2.50515	6.34174	6.40839
		14.47258	14.28947	34.50553	34.94582

Si se remplazan estos valores:

$$14.28947 = 6 \log + 14.47258 \, a$$

$$34.50553 = 14.47258 \log b + 34.94582 \, a$$

$$\log b = - 0.12203$$

$$a = 1.038$$

La función queda:

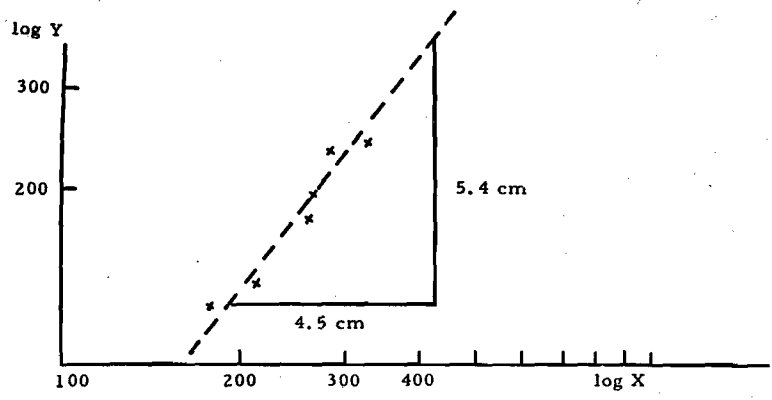
$$Y_c = \frac{1}{1.324} X_i^{1.038}$$

Luego el coeficiente de elasticidad ingreso del consumo es:

$$a = 1.038$$

b) Utilizando el método gráfico, se tiene:

GRÁFICA 29



$$\text{Elasticidad gasto} = \frac{5.4}{4.5} = 1.2$$

c) Si las elasticidades son las siguientes:

Elasticidad ingreso del consumo: $E_y = 1.038$

Elasticidad gasto del consumo de alimentos: $E_G = 1.20$

Para realizar la estimación pedida, se hará primero una estimación del consumo en 1970 mediante la relación

$$\left(\frac{Y_{70}}{Y_{63}}\right)^{E_y} = \frac{C_{70}}{C_{63}}$$

El valor del ingreso en 1970 está dado por

$$Y_{70} = Y_{63}\ (1 + 0.03)^{7} = 340\ (1.23) = 418.2$$

Luego,

$$\left(\frac{418.2}{340.0}\right)^{1.038} = \frac{C_{70}}{320}$$

$$(1.23)^{1.038} = \frac{C_{70}}{320}$$

$$C_{70} = 320\ (1.23)^{1.038} = 320\ (1.24) = 396.8$$

Para la proyección del consumo de alimentos:

$$\left(\frac{C_{70}}{C_{63}}\right)^{E_G} = \frac{\text{C. alim. 70}}{\text{C. alim. 63}}$$

Si se remplazan valores se tiene:

$$\left(\frac{396.8}{320.0}\right)^{1.2} = \frac{\text{C. alim. 70}}{220}$$

$$\text{C. alim.}_{70} = 220\ (1.24)^{1.2} = 220\ (1.295) = 284.9$$

9] Para calcular la elasticidad gasto de los servicios se recurre a la propiedad:

a)

$$E_S W_S + E_A W_A + E_M W_M = 1$$

$$W_S = \frac{\text{gasto en servicios}}{\text{gasto total}} = \frac{40}{200} = 0.20$$

$$W_A = \frac{\text{gasto en alimentos}}{\text{gasto total}} = \frac{100}{200} = 0.50$$

$$W_M = \frac{\text{gasto en prod. manufac.}}{\text{gasto total}} = \frac{60}{200} = 0.30$$

Luego,

$$E = \frac{1 - 0.8\ (0.50) - 1.2\ (0.30)}{0.20} = \frac{0.24}{0.20} = 1.20$$

b) Si la propensión media al consumo se mantiene constante, el consumo crecerá a la misma tasa que el ingreso.
El consumo en 1966 será:

$$C_{1966} = C_{1960}\ (1 + i)^n$$

$$C_{1966} = 200\ (1.04)^6 = 200\ (1.265) = 253$$

Para proyectar los consumos específicos se tiene:

i) Servicios:

$$\left(\frac{C_{1966}}{C_{1960}}\right)^{E_S} = \frac{\text{C. serv. 1966}}{\text{C. serv. 1960}};\ \text{dado que}\ \frac{253}{200} = 1.265$$

$$(1.265)^{1.2} = \frac{\text{C. serv.1966}}{40}$$

$$\text{C. serv. 1966} = 40\ (1.265)^{1.2} = 40\ (1.326) = 53.04$$

ii) Otros productos manufacturados:

$$\left(\frac{C_{1966}}{C_{1960}}\right)^{E_M} = \frac{\text{C. prod. man. 1966}}{\text{C. prod. man. 1960}}$$

$$(1.265)^{1.2} = \frac{\text{C. prod. man. 1966}}{60}$$

C. prod. man. 1966 $= 60\ (1.265)^{1.2} = 60\ (1.326) = 79.56$

iii) Alimentos:

$$\left(\frac{C_{1966}}{C_{1960}}\right)^{E_A} = \frac{\text{C. alim. 1966}}{\text{C. alim. 1960}}$$

$$(1.265)^{0.8} = \frac{\text{C. alim. 1966}}{100}$$

C. alim. 1966 $= 100\ (1.265)^{0.8} = 100\ (1.207) = 120.7$

La distribución del gasto en 1960 y 1966 es la siguiente:

	1960	%	*1966*	%
Alimentos	100	50	120.7	47.7
Otros productos manufac.	60	30	79.6	31.4
Servicios	40	20	53.0	20.9
Total	200	100	253.3	100.0

10] Los datos se resumen así:
Consumo de papel por habitante en 1955: $CP_{1955} = 4.5$
Consumo de papel por habitante en 1967: $CP_{1967} = ?$
Elasticidad gasto de la demanda de papel por habitante: $EG = 1.5$
Tasa de crecimiento del ingreso: $i = 4\%$ ac. anual
Relación gasto ingreso: $G = 0.8\ Y$

La fórmula para la proyección será la siguiente:

$$\left(\frac{\text{Consumo por habitante 1967}}{\text{Consumo por habitante 1955}}\right)^{EG}$$

$$= \frac{\text{Consumo de papel por habitante 1967}}{\text{Consumo de papel por habitante 1955}}$$

Es necesario estimar previamente los consumos por habitante en 1955 y 1967. Para ello es necesario

$$G_{1955} = 0.9\ Y_{1955}$$

$$G_{1955}\ \text{(por habit.)} = \frac{0.9\ Y_{1955}}{P_{1955}}$$

donde P = Población

$$G_{1967}\ \text{(por habit.)} = \frac{0.9\ Y_{1955}\ (1 + 0.4)^{12}}{P_{1955}\ (1.025)^{12}}$$

$$\frac{G_{1967}}{G_{1955}}\ \text{(por habit.)} = \frac{0.9\ Y_{1955}\ (1.04)^{12}}{P_{1955}\ (1.025)^{12}} \cdot \frac{P_{1955}}{0.9\ Y_{1955}}$$

$$= \left(\frac{1.040}{1.025}\right)^{12}$$

Remplazando en la fórmula de la proyección se tiene:

$$\left[\left(\frac{1.040}{1.025}\right)^{12}\right]^{1.5} = \frac{\text{Consumo de papel por habitante 1967}}{4.5}$$

$$\text{Consumo de papel por habit. 1967} = 4.5 \left(\frac{1.040}{1.025}\right)^{18}$$

$$= 4.5\ (1.0146)^{18} = 4.5\ (1.308) = 5.89$$

Luego, se estima para 1967 un consumo de 5.89 kg de papel por habitante.

TERCERA PARTE

ELEMENTOS DE MUESTREO

I

ASPECTOS TEÓRICOS

A. INTRODUCCIÓN

Cuando se acepta la idea de planificar, implícitamente se admite la necesidad de contar con más, mejores y periódicas informaciones básicas. En el proceso de planificación, desde el diagnóstico hasta la reformulación de los planes, es indispensable contar con magnitudes e indicadores que muestren lo que en verdad está ocurriendo en una actividad económica. Ahora bien, hay varias posibilidades de captación de información: los censos, el análisis de casos típicos y las muestras. Al considerar el *tipo* y *fidelidad* de las informaciones necesarias, *el costo* de obtenerlas y *el tiempo* que demandará su recolección, surgen nítidas las ventajas del muestreo. No debe interpretarse con esto, que el muestreo sea un sustituto del censo. Informaciones básicas para el total de una población o universo siempre serán necesarias; de otra manera no será posible diseñar una muestra con reales ventajas y suficiente garantía. Lo que ocurre es que mediante una muestra pueden obtenerse informaciones que resultarían poco menos que imposibles con una enumeración total de la población: justamente por problemas de *costo* y *tiempo*. Por otra parte los censos se realizan cada 5 o más años y las muestras pueden ser utilizadas para estimar parámetros en períodos intermedios. Conocer el tamaño de una población y otras características básicas es indispensable para extraer una buena muestra representativa. Se concluye pues, que censo y muestra son dos métodos complementarios y no excluyentes en el proceso de captación de información.

El objeto de la primera parte de este trabajo es mostrar, a los profesionales planificadores, las posibilidades y limitaciones de esta técnica, por una parte, y plantear los principales conceptos envueltos en el muestreo. La segunda parte está destinada a analizar los principales problemas que se presentan en las encuestas industriales y señalar las posibles soluciones.

A esta altura es indispensable advertir que existe una gran profusión bibliográfica sobre el tema, pero es difícil encontrar un

trabajo que deje de lado todo el tratamiento matemático y se centre en la parte práctica. El planificador debe conocer los elementos básicos del muestreo, sus limitaciones y alcances, de modo que pueda determinar qué tipo de investigaciones son susceptibles de enfrentarse por muestreo y pueda decidir con base fundada sobre las posibilidades que le ofrezca el muestrista o especialista en muestreo. El experto en muestreo puede presentar posibilidades para una investigación dada, que fácilmente pueden implicar costos que van desde los 5 o 10 mil dólares hasta los 50 mil o más, dependiendo del grado de precisión, de los niveles de confianza o probabilidad de acierto, del tipo de diseño muestral, etc. Sobre esas posibilidades el planificador debe estar en condiciones de tomar una decisión racional.

Se intentará cumplir con los objetivos señalados en líneas anteriores, prescindiendo hasta donde sea posible del aparato matemático. El trabajo está diseñado para ser interpretado contando con conocimientos de estadística descriptiva, elementos de probabilidades y álgebra superior.

En los anexos finales se presentan las principales demostraciones matemáticas y formularios completos de los principales diseños muestrales.

Se piensa que con estos antecedentes, el planificador estará en condiciones de interpretar en su justa dimensión las reales posibilidades de las técnicas muestrales en general, de calificar las estimaciones resultantes y de tener una posición realista en los problemas en que deberá tomar decisiones.

B. FUNDAMENTOS TEÓRICOS DEL MUESTREO [1]

Para presentar los elementos básicos del muestreo, se tomará como punto de partida el diseño muestral aleatorio simple por dos razones principales: se trata de un diseño sencillo, de fácil interpretación, y constituye la base de los diseños muestrales aleatorios.

Por población, universo o marco muestral se entenderá el conjunto de elementos de los que se extraerá información y sobre los que se podrá generalizar la conclusión obtenida a partir de la muestra. La población podrá estar formada por personas, empresas, familias, áreas, objetos, etc. El número de elementos que componen la población se designará por N.

[1] Se utilizará la simbología de W. C. Cochran, para facilitar al lector interesado la consulta del libro *Sampling techniques* del autor citado.

Por muestra se entenderá una *parte representativa* de esa población y se designará por n.

Será necesario definir la variable que se investigará: ingresos, valor agregado, productividad, etc., y las unidades en que se expresará.

Por y_i, se denominará el valor de la variable de la i-ésima unidad de la población o de la muestra ($i = 1, 2, 3 \ldots n \ldots N$).

El número posible de muestras de composición distinta que se podrá obtener estará dado por el número combinatorio C_n^N. Se insiste que se trata de muestras de composición distinta, en cuanto a sus elementos, aunque muchas de estas muestras pueden tener la misma media aritmética.

Los estadígrafos que interesa definir son los siguientes:

$\bar{Y}$: media aritmética de la población

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

$\bar{y}$: media aritmética de la muestra

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

σ^2: varianza de la población

$$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N}$$

S^2: varianza de Cochran de la población

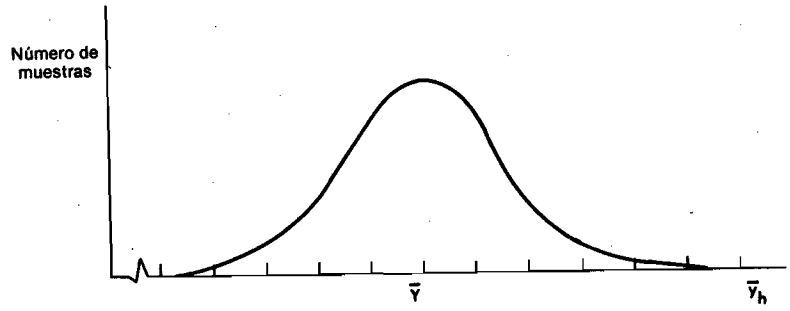
$$S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N - 1}$$

s^2: varianza de Cochran de la muestra

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

Ahora bien, si el número de muestras distintas que se puede obtener es $\binom{N}{n}$, habrá igual número de medias aritméticas muestrales: $\bar{y}_h$, donde $h = 1, 2, 3 \ldots \binom{N}{n}$.

Si se extraen todas esas muestras computándose sus medias aritméticas y clasificándolas en una tabla de distribución de frecuencias, se comprobará que en general estas medias aritméticas tienen distribución muy aproximada a la distribución normal. Se recalca en general, porque esto no se cumple cuando la muestra es muy pequeña (menos de 30-40 elementos),[2] o la población es muy reducida o con características poco frecuentes. En el siguiente gráfico se ilustra este comentario:



Si se promedian todas estas medias aritméticas (E: esperanza matemática) se encuentra que el promedio es exactamente igual a la media aritmética poblacional, es decir (demostración en el apéndice):

[2] El desconocimiento de la varianza poblacional (S^2) es otro aspecto a considerar. Para investigaciones en el campo industrial, en donde generalmente las muestras superarán los 30 o 40 elementos, no es muy peligroso aceptar esa simplificación teórica.

$$E(\bar{y}_h) = \frac{\sum_{h=1}^{\binom{N}{n}} \bar{y}_h}{\binom{N}{n}} = \bar{Y}$$

Por eso a la media aritmética muestral ($\bar{y}_h$) se le llama *estimador insesgado* de la media aritmética poblacional.

De la misma manera, a todas las posibles muestras se les puede computar una varianza (s^2_h). Si se promedian estas varianzas, el resultado es la varianza de Cochran para la población:

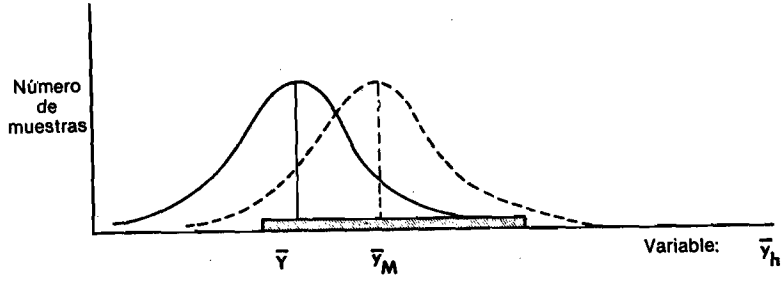
$$E(s^2_h) = \frac{\sum_{h=1}^{\binom{N}{n}} s^2_h}{\binom{N}{n}} = S^2$$

Por ello a la varianza de Cochran para la muestra se le llama estimador insesgado de la varianza de Cochran para la población. Esta igualdad sólo es válida para el caso en que las varianzas hayan sido calculadas con denominadores N — 1 (poblacional) y n — 1 (muestral) y ésa es la razón para restarle una unidad al denominador de las varianzas.

En la práctica sólo se extrae *una muestra* y sobre la base de ella se hacen las estimaciones para la población. La selección de esa única muestra se hace, en este diseño muestral, en forma aleatoria y sin reposición, es decir, todas y cada una de las posibles muestras tienen la misma probabilidad de ser elegidas, y la muestra con la que se trabaja estará formada por elementos distintos. Por ejemplo, en una encuesta industrial no habrá empresas repetidas en la muestra. Cada empresa aparecerá una vez y sólo una vez en la muestra. En resumen, este diseño muestral es, pues, equiprobabilístico y sin reposición.

Ahora bien, con el cómputo de la media aritmética de la muestra se tiene una estimación de la media aritmética poblacional, es decir, una estimación puntual. Es prácticamente imposible que la media aritmética muestral coincida exactamente con la media aritmética poblacional. Es necesario determinar un rango o inter-

valo a partir de la media aritmética muestral, donde se espera esté comprendida la media aritmética poblacional. A ese intervalo específicamente se le denominará intervalo de confianza y a su mitad se simbolizará por d, que corrientemente se le denomina semiancho del intervalo de confianza. Lo anterior puede interpretarse gráficamente de la siguiente manera:



La curva continua representa la distribución de las medias muestrales en torno a la media poblacional. Si se ha seleccionado una muestra que tiene por media aritmética $\bar{y}_M$, se traslada la distribución con eje en $\bar{y}_M$, que en el gráfico está representada por la línea punteada. A partir de $\bar{y}_M$ se determina el intervalo de confianza (zona sombreada sobre el eje de las abscisas).

El intervalo de confianza tiene dos componentes:

El coeficiente de confianza que está determinado por la probabilidad de acierto de la estimación, y tiene su origen en la estructura de una distribución normal. En una distribución normal, en el intervalo comprendido entre $\bar{Y} \pm \sigma$ se encuentra el 68 por ciento de los casos, entre $\bar{Y} \pm 2\sigma$ se encuentra el 96 por ciento de los casos.[3] El coeficiente de confianza es el número de veces que se debe tomar σ a derecha e izquierda de $\bar{Y}$ para agrupar, entre medio, cierto porcentaje de casos. Asociado al concepto de coeficiente de confianza está el nivel de confianza o probabilidad de acierto (porcentaje de casos comprendidos en cierto intervalo central). A continuación se dan los coeficientes de confianza más frecuentemente utilizados en una distribución normal, los que se simbolizarán por t.

Nivel de confianza	98%	95%	90%	80%	68%
Coef. de confianza (t)	2.33	1.96	1.65	1.28	1.0

[3] Estas proporciones deben interpretarse respecto del número posible de muestras a obtener $\binom{N}{n}$

El otro componente del intervalo es el que representa la dispersión de la variable. En una distribución normal corriente, por ejemplo la distribución de una población, ese componente está representado por σ. La distribución normal que ahora se está tratando es bastante particular: en primer lugar la variable está dada por medias aritméticas que provienen de muestras, en segundo lugar se trata de una distribución teórica ya que se suponen computadas todas las posibles medias aritméticas. El estadígrafo que indica la dispersión de esta particular variable estará dado por:

$$V(\bar{y}) = \frac{\sum_{h=1}^{\binom{N}{n}} (\bar{y}_h - \bar{Y})^2}{\binom{N}{n}}$$

Nótese que se trata de la fórmula de una varianza cualquiera, pero donde la variable está dada por todas las posibles medias aritméticas muestrales.

Se puede matemáticamente demostrar que

$$V(\bar{y}) = \frac{\sum_{h=1}^{\binom{N}{n}} (\bar{y}_h - \bar{Y})^2}{\binom{N}{n}} = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

que constituye la fórmula de cálculo práctico de la *varianza verdadera del estimador de la media aritmética.* Este estadígrafo se conoce también con el nombre de cuadrado del "error estándar verdadero". En general, cuando se realizan estimaciones sobre media aritmética, no se dispone del valor de la varianza poblacional de Cochran. En esos casos se utiliza la varianza de la muestra s^2, en virtud de que es un estimador insesgado de S^2. Cuando ocurre tal cosa, es decir cuando se utiliza s^2 en vez de S^2, el estadígrafo toma el nombre de varianza estimada del estimador de la media aritmética, y se simboliza así:

$$v(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

Intervalos de confianza:

Con los elementos mencionados, ya se puede cuantificar el intervalo de confianza para la media aritmética. Este intervalo estará dado por:

$$\bar{y} \pm d$$

donde d es la magnitud del semiancho del intervalo de confianza y está dado por:

$$d = t \sqrt{V(\bar{y})}$$

Como se recordará, t es el coeficiente de confianza que corresponde a un nivel de confianza. En general el nivel de confianza fluctúa entre 80 y 95 por ciento, en la práctica generalmente se toma un 90 o 95 por ciento de confianza, lo que en otros términos indicaría que en 95 de cada 100 muestras la media aritmética poblacional estará dentro del intervalo de confianza. $V(\bar{y})$ indicaba la dispersión de las medias muestrales en torno a la media poblacional. Con los elementos vistos ya se está en condiciones de estimar una media aritmética poblacional:

Ejemplo: se ha tomado una muestra de 30 de las 108 empresas pesqueras de un país. Se pretende estimar el número promedio de obreros por empresa. Para las 30 empresas de muestra se calculó una media aritmética de 46 obreros por empresa y una varianza de 18. El número promedio de obreros por empresa en toda la población estará comprendido entre:

$$\bar{y} \pm d$$

es decir

$$\bar{y} \pm t \sqrt{V(y)}$$

Los datos que se conocen son:

$$N = 108;\ n = 30;\ \bar{y} = 46;\ s^2 = 18.$$

$$V(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{18}{30}\left(1 - \frac{30}{108}\right) = 0.433$$

Si se toma como nivel de confianza un 90 por ciento, el correspondiente valor de t será de 1.65. El intervalo de confianza estará dado por:

$$46 \pm 1.65 \sqrt{0.43}$$

lo que equivale a decir que la media aritmética poblacional estará comprendida entre 44.92 y 47.08, con 90 por ciento de probabilidad de que tal estimación sea cierta.

Para la estimación de un total, es decir, de la suma de los valores de la variable de cada uno de los elementos de la población, se procede a multiplicar el intervalo para la media aritmética por N; en virtud de que:

El total (Y) estará dado por:

$$Y = N \bar{Y}$$

y el total estimadó ($\hat{Y}$):

$$\hat{Y} = N \bar{y}$$

Si en el ejemplo anterior se deseara encontrar el intervalo para el total de obreros ocupados en las 108 empresas, con el mismo nivel de confianza se tendrá:

$$N \bar{y} \pm N t \sqrt{v(\bar{y})}$$

que equivale a:

$$(108)(46) \pm (108)(1.65)(0.656)$$

El total de la población estará comprendido entre: 4851 y 5085, con 90 por ciento de probabilidad de que tal cosa sea cierta.

C. DETERMINACIÓN DEL TAMAÑO DE MUESTRA

Hasta ahora se ha supuesto un tamaño de muestra dado, interesa pues analizar brevemente cuáles son los elementos que determinan la magnitud de n.

Fundamentalmente hay cuatro elementos técnicos que condicionan el tamaño de una muestra. Por otra parte hay un quinto ele-

mento de extraordinaria importancia práctica: el monto de recursos financieros y elementos humanos y materiales sin cuyo concurso no es posible garantizar estimaciones confiables.

1] *Homogeneidad de la población*

Es fácil interpretar que el grado de homogeneidad indicado por la magnitud de S^2 condicionará el tamaño de muestra. Donde la variable en la población se distribuya uniformemente, sólo serán necesarios unos pocos elementos de muestra para tener una idea bastante precisa de lo que ocurre en la población para la variable que se investiga. En cambio en poblaciones muy heterogéneas para la variable investigada, será necesario un tamaño de muestra bastante grande para poder realizar estimaciones sin riesgo de grandes errores. Hay, pues, una relación directa entre n y S^2.

2] *Precisión de la estimación*

Nuevamente el título sugiere algo obvio: mientras más precisa sea una estimación (menor tamaño de d), más grande deberá ser el tamaño de muestra. Se había adelantado que la precisión estaba dada por la magnitud de d. Es el investigador: planificador, sociólogo, economista, etc., que investiga el comportamiento de una variable, quien tiene que decidir sobre la magnitud del máximo de desviación respecto de la media aritmética verdadera. Se concluye pues que hay una relación inversa entre tamaño de muestra y precisión.

3] *Nivel de confianza*

El nivel de confianza que representaba la probabilidad de que la estimación fuera verdadera, también tiene una relación directa con el tamaño de muestra, a través del coeficiente de confianza t. Mientras mayor probabilidad de acierto se desea tener, más grande deberá ser el tamaño de muestra.

4] *Tamaño de la población*

Otro elemento que es indispensable analizar es la relación que existe entre los tamaños de muestra y población. De poblaciones numerosas, cabrá esperar muestras grandes y viceversa.

Del equilibrio de todas estas condicionantes se determina la magnitud del tamaño de una muestra, como se verá en páginas posteriores.

5] *Recursos*

Este elemento, si bien no es tomado en cuenta en la determinación "técnica" del tamaño de muestra, desempeña, en nuestros países, un papel de primera línea. Existe toda una problemática entre la compatibilización del tamaño de muestra determinado técnicamente, y el tamaño de muestra que resultaría del monto de recursos, principalmente financieros, asignados para la investigación.

En general el monto de recursos determina una muestra menor que la que resultaría de la aplicación de las fórmulas. La muestra menor determinará una menor precisión; será necesario analizar si esa precisión resultante garantiza la obtención de estimaciones adecuadas. Es en este punto donde hay que decidir: o se renuncia a la precisión especificada técnicamente y se acepta la resultante de la limitación de recursos, o se destinan mayores recursos financieros. Las dos posibilidades restantes son: una combinación de los dos procesos anteriores, o la decisión de no llevar adelante la investigación por estos métodos.[4]

Las fórmulas del tamaño de muestra se deducen fácilmente a partir del intervalo de confianza:

Recuérdese la fórmula del intervalo de confianza para la media aritmética:

$$d = t \frac{S}{\sqrt{n}} \left(1 - \frac{n}{N}\right)^{1/2}$$

Después de elevar al cuadrado, se despeja n y se tiene:

$$n = \frac{\left(\frac{t\,S}{d}\right)^2}{1 + \left(\frac{t\,S}{d}\right)^2 \cdot \frac{1}{N}}$$

Ésta es la fórmula del tamaño de una muestra aleatoria simple, que garantiza la estimación de una media aritmética con los requisitos de precisión y confianza o probabilidad de acierto especificados.

Es necesario prestar atención a estos conceptos mencionados:

[4] Véase, del autor, "El costo en las investigaciones muestrales", en la revista *Economía*, núm. 83, Facultad de Ciencias Económicas de la Universidad de Chile.

t, como se había adelantado, es el coeficiente de confianza que provenía de la distribución normal estandarizada (media = 0 y varianza = 1) y estaba automáticamente determinado al elegir el nivel de confianza. Una interpretación para este concepto, en este caso, sería: la proporción dentro de todas las muestras posibles, que entregan resultados para la media aritmética que no difieran respecto de la media aritmética verdadera en más del valor d especificado.

En cuanto a la precisión representada por la magnitud de d, se había adelantado que quien está en mejores condiciones para decidir sobre el desvío máximo que se puede tolerar es el investigador, en este caso el planificador, porque sabe cuáles serán los fines de la estimación. No está de más señalar que una alta precisión (d pequeño) sólo se podrá conseguir a expensas de una muestra grande, y que muestras pequeñas sólo pueden entregar resultados poco precisos. Los anteriores juicios son válidos cuando el grado de heterogeneidad en la población (magnitud de S^2) es apreciable. En poblaciones homogéneas, una muestra pequeña puede ser suficiente para lograr una aceptable precisión. Todo lo anterior exige una decisión del planificador, que para este fin deberá agregar dos elementos importantes: el costo de la investigación y el tiempo que demandará.

El muestrista se limitará a presentarle alternativas, el planificador deberá elegir la más conveniente.

Observando la fórmula de tamaño de muestra, se puede visualizar que cuando la población es bastante grande, el denominador tiende a la unidad, y el numerador puede aceptarse como una adecuada aproximación del tamaño de muestra.

$$n_0 = \left(\frac{t\ S}{d}\right)^2$$

El lector debe haberse planteado una interrogante: en las fórmulas para tamaño de muestra se supone conocida la varianza. ¿Qué sentido tiene determinar una muestra para estimar una media aritmética si se supone conocida la varianza y esto implica conocer la media aritmética? Lo que sucede es que el supuesto no es hipotético y el propósito no carece de sentido. Hay muchos estudios que exigen investigaciones muestrales periódicas. Con estas investigaciones se pueden tener buenas estimaciones de S^2 y porque en muchos casos la varianza puede ser poco cambiante, lo que no siempre ocurre con la media (recuérdese que $V(y_i + K) = V(y_i)$ y que $M(y_i + K) = K + M(y_i)$).

Sin embargo, el caso más frecuente es que se desconozca el valor de S^2. El muestrista dispone de métodos que le permiten estimar la magnitud de S^2, en este caso el planificador deberá tomar nota de que se trata de una estimación que puede constituir una fuente de posibles errores. Es necesario que, una vez realizada la muestra, se compulse el valor estimado de S^2 y la varianza de la muestra. Si el valor estimado de S^2 es similar o mayor a la varianza de la muestra (s^2) no hay motivo de preocupación, excepto que se haya tomado tal vez una muestra mayor que la necesaria y esto puede estar compensado en parte por la mayor precisión que se consigue. Pero si la varianza de la muestra es apreciablemente mayor que el valor estimado de S^2 para fines de cálculo del tamaño de muestra, la situación deberá ser motivo de mayor investigación. Habrá que recalcular el tamaño de muestra en función de s^2. En esta etapa será necesaria otra decisión: se toma una muestra complementaria o se renuncia a la precisión especificada primitivamente.

Por último, en muchas investigaciones resulta que los datos censales disponibles escasamente proporcionan un valor de la magnitud de la población. En estos casos, a veces como primera aproximación, se utilizan varianzas de la variable que se investiga provenientes de otras poblaciones, otros países u otras áreas dentro del mismo país. Nótese que tal procedimiento puede contener serios errores y deberá ser analizado concienzudamente. Además es necesario recalcar que estas suposiciones son con el solo objeto de tener una primera aproximación del tamaño de la muestra. Posteriormente, si se realiza la encuesta, deberán compararse las magnitudes de las varianzas y proceder en la forma indicada en páginas anteriores.

Cuando no es posible tener estimaciones previas de S^2, a veces se opta por una muestra piloto o de iluminación, cuyo tamaño generalmente está condicionado por el costo que ésta implica. Dicha muestra de iluminación puede dar una idea del posible tamaño definitivo de muestra. Con todos los conceptos que se han señalado se pretende sistematizar en un cuadro las alternativas que se le presentarán al planificador para que decida sobre el tamaño de muestra definitivo, a la luz de los objetivos que tiene planteados. Sólo se incluyen los principales elementos de decisión; así, se omite presentar posibilidades en cuanto al nivel de confianza porque éste generalmente tiene poca variación (entre 90 y 95 por ciento), y su omisión no distorsionaría la decisión.

Tamaño de muestra	*Desvío máximo*	*Costo (miles de dólares)*	*Tiempo (días)*
150	0.10	4 — 5	30 — 35
600	0.05	8 — 9	60 — 70
938	0.04	9 — 10	70 — 85
1666	0.03	15 — 17	120 — 140
3750	0.02	22 — 25	210 — 240
6000[a]	0.00	50 — 56	400 — 500

[a] Censo.

El cuadro anterior puede corresponder por ejemplo a posibilidades de tamaño de muestra para una investigación en el sector industrial, donde la variable estratégica sea la estimación del coeficiente producto-capital. Se ha supuesto, para fines de ilustrar el ejemplo, que la varianza de los coeficientes producto-capital es 0.375, y en todos los casos se ha considerado una probabilidad de acierto de 96 por ciento. He aquí un caso de ocurrencia periódica en la práctica.

¿Puede un investigador decidir racionalmente, seleccionando la mejor posibilidad sin contar con ideas fundamentales de los conceptos que entran en juego? Evidentemente que decisiones realistas y racionales sólo provienen de investigadores que conocen los elementos básicos de la teoría de muestras.

D. LA ESTIMACIÓN DE PROPORCIONES

Dentro de los diagnósticos y en la preparación de los programas hay cierto tipo de variables cuyo tratamiento puede ser indispensable. Es el caso de las variables cualitativas. Por ejemplo, la estimación de la proporción de empresas que son sociedades anónimas, la proporción de empresas que cuentan con más de 100 obreros, el número de obreros que tienen salarios superiores al promedio del país, etc. El lector puede imaginarse una serie de casos donde la estimación de la proporción puede ser útil. No es otra cosa que un caso particular del diseño anterior; aquí la variable puede estar clasificada en dos categorías: posee la característica que se investiga, o no la posee. Por convención se asigna el valor 1 a aquellas unidades en la población que tienen la característica y 0 si no la tienen. Con esa convención, la proporción en la población será:

$$P = \frac{\sum_{i=1}^{N} y_i}{N}$$

La proporción de los elementos que no poseen la característica que se investiga será:

$$Q = 1 - P$$

En la muestra este estadígrafo estará dado por:

$$P = \frac{\sum_{i=1}^{n} y_i}{n} \quad y \quad q = 1 - p$$

Si se observan las fórmulas en ambos casos, corresponden a las conocidas fórmulas de una media aritmética. En este caso también las proporciones muestrales se distribuyen en general en forma aproximadamente normal, en torno a la proporción poblacional. Donde hay que poner un poco de atención es en la varianza de la proporción. Como se verá en seguida, en este caso la varianza tiene dos límites: sólo puede tomar valores comprendidos entre 0 y 0.25.[5]

Cuando la variable era cuantitativa, la varianza sólo tenía límite inferior: no podía ser negativa pero podía tomar un valor tan grande como se quisiera.

Es fácil deducir la fórmula de la varianza de la proporción, partiendo de la fórmula conocida para la varianza de Cochran:

$$S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N - 1}$$

Desarrollando el cuadrado y reduciendo términos semejantes, se tiene:

$$S^2 = \frac{\sum_{i=1}^{N} y_i^2}{N - 1} - \frac{N \bar{Y}^2}{N - 1}$$

[5] Para poblaciones no demasiado pequeñas. La varianza máxima se tiene cuando $P = Q$.

Se trata de remplazar los valores de esta expresión, por los que resultan de considerar la *proporción* en el caso en que la variable toma valores cero o uno.

Si

$$P = \frac{\sum_{i=1}^{N} y_i}{N},$$

puede asociarse con $\bar{Y}$.

Además:

$$\sum_{i=1}^{N} y_i = N\ P$$

y

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} y_i^2$$

porque la variable sólo toma valores cuyos cuadrados son los mismos valores. Remplazando en la fórmula de Cochran para la varianza se tiene:

$$S^2 = \frac{N\ P\ -\ N\ P^2}{N\ -\ 1} = \frac{N\ P\ Q}{N-1}$$

Por analogía, la varianza de la muestra estará dada por:

$$s^2 = \frac{n\ p\ q}{n-1}$$

Para la determinación de los intervalos de confianza, se utilizan las mismas fórmulas, teniendo cuidado de remplazar el valor de

$$S^2 \text{ por } \frac{N\ P\ Q}{N-1} \quad \text{y} \quad s^2 \text{ por } \frac{n\ p\ q}{n-1}$$

En todo caso, en el formulario anexo de este trabajo se incluyen fórmulas detalladas, cuya consulta puede ser beneficiosa.

Este diseño muestral no ofrece complicación alguna. Es simplemente la combinación de varios diseños aleatorios simples.

Se trata de estratificar la población, es decir dividirla en partes excluyentes, de acuerdo con la característica que se investiga. Así, por ejemplo, si se quiere estimar el valor agregado del sector industrial, sería conveniente dividir la población de industrias en 4 o 5 estratos de manera que en cada estrato las unidades tengan características homogéneas en cuanto a valor agregado, es decir, en cada estrato los valores agregados de cada industria deberían estar dentro de cierto rango. De esta manera se minimiza la magnitud de la varianza [6] en cada estrato y, por consiguiente, se trata de disminuir el error de muestreo. Es importante insistir que el criterio de estratificación debería estar en directa relación con el objetivo de la encuesta. Por ejemplo, si se quiere estimar un coeficiente producto-capital, un posible criterio de estratificación podría ser según el número de turnos diarios; un primer estrato comprendería a las industrias que trabajan un turno diario, el segundo estrato a las que lo hacen dos turnos diarios, el tercer estrato las de tres turnos y un último estrato que comprendería a aquellas industrias que no tienen un número de turnos definido.

La principal ventaja de este diseño muestral es su eficiencia comparada con otros diseños muestrales: para un mismo tamaño de muestra el error de muestreo es en general menor cuando se estratifica la población. La otra ventaja importante es que permite realizar estimaciones no sólo para la población total sino para cada uno de los estratos, lo que sin duda, especialmente en el campo industrial, permite afinar conclusiones. La desventaja es la necesidad de tener un conocimiento anticipado de las características generales del universo que se investigará; de la calidad de las informaciones básicas dependerá la bondad del criterio de estratificación.

En cuanto al número de estratos, hay que considerar que, mientras mayor sea, más precisas serán las estimaciones y una mayor desagregación en los resultados permitirá obtener conclusiones específicas y detalladas. Por otra parte, es difícil disponer de antecedentes que permitan clasificar adecuadamente (estratificar) la po-

[6] Recuérdese que los componentes de la varianza son la intervarianza y la intravarianza. Dado que para los errores de muestreo se promedian las varianzas de cada estrato, en el hecho sólo se está tomando en cuenta la intravarianza.

blación. En la práctica, con los antecedentes que en general se dispo.en, lo más que puede hacerse es dividir la población en gruesas categorías. Incluso si se pudiera estratificar en un número grande de estratos, la complejidad administrativa y la organización del trabajo de campo puede significar trabajo adicional considerable.

En las investigaciones industriales, normalmente se toman entre 4 y 7 estratos.

1] *Estimadores e intervalos de confianza*

En cada estrato se podrán computar los siguientes estadígrafos:

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$ Media aritmética del estrato h

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1}$$ Varianza de Cochran del estrato h

$$Y_h = N_h \bar{Y}_h$$ Total del estrato h

La media aritmética de la población (con L = número de estratos) estará dada por

$$\bar{Y} = \frac{\sum_{h=1}^{L} \bar{Y}_h N_h}{N} \quad \text{donde} \quad \sum_{h=1}^{L} N_h = N$$

y la varianza de la población

$$S^2 = \frac{\sum_{h=1}^{L} (\bar{Y}_h - \bar{Y})^2 N_h}{N - 1} + \frac{\sum_{h=1}^{L} S_h^2 \cdot (N_h - 1)}{N - 1}$$

Para la muestra se tendrían los siguientes estadígrafos:

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$ media aritmética de la muestra del estrato h

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}$$ varianza del estrato h

Ahora bien, dado que en cada estrato se extrae una muestra aleatoria simple, tanto la media de la muestra $(\bar{y}_h)$ como la varianza son respectivamente estimadores insesgados de la media aritmética $(\bar{Y}_h)$ y de la varianza (S_h^2) del estrato:

$$E(\bar{y}_h) = \bar{Y}_h$$

$$E(s_h^2) = S_h^2$$

Si ahora se combinan todos los estratos,[7] la media de la muestra será:

$$\bar{y}_n = \frac{\sum_{h=1}^{L} \bar{y}_h \, n_h}{n}$$

donde

$$\sum_{h=1}^{L} n_h = n$$

y el estimador insesgado de la media aritmética poblacional será:

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} \bar{y}_h \cdot N_h}{N}$$

[7] En general $\bar{y}_n$ no es estimador insesgado de $\bar{Y}$, sólo lo es cuando $\frac{n_h}{n} = \frac{N_h}{N}$.

Es fácil verificar que:

$$E(\bar{y}_{st}) = \bar{Y}$$

lo que justifica que $\bar{y}_{st}$, sea estimador insesgado de $\bar{Y}$.

En la misma forma se pueden combinar las varianzas del estimador de la media. Recuérdese que en muestreo aleatorio simple se tenía:

$$V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

Este estadígrafo para un estrato cualquiera será:

$$V(\bar{y}_h) = \frac{S_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right)$$

Promediando las $V(\bar{y}_h)$ para todos los estratos se tendrá la varianza del estimador de la media poblacional:

$$V(\bar{y}_{st}) = \frac{\sum_{h=1}^{L} N_h^2\, V(\bar{y}_h)}{N^2} = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\,\frac{S_h^2}{n_h}$$ [8]

Cuando no se dispone de los valores de las varianzas de los estratos (S_h^2), se utilizan los resultantes de las muestras de cada estrato (s_h^2) y se tiene:

$$v(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h\,(N_h - n_h)\,\frac{s_h^2}{n_h}$$

que es el estimador de la varianza del estimador de la media aritmética poblacional.

En igual forma que en el muestreo aleatorio simple, el intervalo de confianza para la media estará dado por:

$$\bar{y}_{st} \pm d$$

[8] Fórmula general de la varianza del estimador de la media aritmética poblacional.

donde

$$d = t \sqrt{V(\bar{y}_{st})}$$

El intervalo de confianza para el total estará dado por

$$N \bar{y}_{st} + d$$

donde

$$d = t\ N \sqrt{V(\bar{y}_{st})}$$

2] *Afijaciones y tamaños de muestra*

Es importante considerar lo que en muestreo se denomina afijación; el significado de esta expresión es el de distribución y asignación de la muestra total en cada uno de los estratos. Por afijación, entonces, se entenderá el proceso que permite distribuir un tamaño de muestra dado (n) entre los estratos, de manera de tener una muestra en cada estrato (n_h). Hay distintas maneras de "afijar" una muestra; a continuación se expondrán las principales.

i] Afijación proporcional. Se trata de distribuir una muestra dada de tamaño n entre los estratos, en forma proporcional al tamaño de cada estrato. El tamaño de muestra en cada estrato estará dado por:

$$n_h = \frac{N_h}{N} \cdot n$$

La justificación de este método radica en el hecho de que de mayores estratos se extraerán mayores tamaños de muestra. Cuando se sigue una afijación proporcional, la fórmula general de la varianza del estimador $V(\bar{y}_{st})$ puede simplificarse remplazando n_h por $\frac{N_h}{N} n$ en la forma:

$$V(\bar{y}_{st})_{A.P.} = \frac{1 - f}{n} \sum_{h=1}^{L} W_h S_h^2$$

donde

$$f = \frac{n}{N} \quad y \quad W_h = \frac{N_h}{N}$$

De la fórmula anterior puede despejarse n y se tendrá la fórmula del tamaño de muestra cuando se decide previamente seguir un diseño muestral estratificado por afijación proporcional:

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

donde

$$n_o = \sum_{h=1}^{L} \frac{W_h \, S_h^2}{V(\bar{y}_{st})}$$

Es necesario aclarar que, cuando se trata de determinar un tamaño de muestra, es necesario tener estimaciones de la varianza de la variable que se desea investigar (S_h^2), ya sea por antecedentes de otras encuestas, por comparaciones con otras variables de distribución similar, por muestras de iluminación o por otros métodos estadísticos de estimación. En la fórmula aparece $V(\bar{y}_{st})$ que es el cuadrado del error de muestreo que se puede tolerar. La manera de especificar este valor es a través del intervalo de confianza.

$$d = t \sqrt{V(\bar{y}_{st})}$$

de donde

$$V(\bar{y}_{st}) = \frac{d^2}{t^2}$$

Recuérdese que d es el desvío máximo que se puede tolerar, y está expresado en las mismas unidades de la variable que se investiga; este valor lo determina el planificador. t es el coeficiente de confianza que está dado por el nivel de confianza o probabilidad de acierto en la estimación. El nivel de confianza también es materia de decisión del investigador. Será necesario plantear

alternativas para diferentes magnitudes de d. Con un cuadro similar al presentado en el anterior diseño muestral, se puede decidir en forma objetiva.

ii] Afijación óptima. La distribución de una muestra en forma proporcional al tamaño del estrato puede adolecer de ciertos defectos. Es posible que existan estratos muy grandes pero bastante homogéneos, y al contrario, puede ocurrir que haya pequeños estratos sumamente heterogéneos, o ambas cosas a la vez. Si en estos casos se sugiere una afijación proporcional, sucedería que de los grandes estratos homogéneos se extraería una muestra más que suficiente con el correspondiente desperdicio de recursos, y en los pequeños estratos altamente heterogéneos se extraerían muestras de tamaño insuficiente. Para evitar tales desajustes, la afijación óptima distribuye la muestra total (n) entre los estratos tomando simultáneamente el tamaño y el grado de heterogeneidad del estrato.[9] La fórmula para afijar óptimamente una muestra es:[10]

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \cdot n$$

Observando la fórmula se comprueba que la distribución de n entre los estratos es proporcional simultáneamente al tamaño y al grado de variabilidad de la variable en el estrato.

Remplazando este valor de n_h por la expresión recién comentada, en la fórmula general de la varianza del estimador para la media aritmética, se tiene:

$$V(\bar{y}_{st})_{A.O.} = \frac{\left(\sum_{h=1}^{L} W_h S_h\right)^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

Si no se dispone de los valores de S_h^2, se podrá introducir el corres-

[9] El calificativo de óptima tiene su origen en el hecho de que mediante esta afijación se consigue el menor error de muestreo.

[10] Existe una demostración que justifica esta fórmula, basándose en la minimización de la varianza del estimador.

pondiente s_h^2 de la muestra, con lo que se tendrá el estimador de la varianza del estimador de la media:

$$v(\bar{y}_{st})_{A.O.} = \frac{\sum_{h=1}^{L} W_h \, s_h^2}{n} - \frac{\sum_{h=1}^{L} W_h \, s_h^2}{N}$$

Si de la fórmula de la varianza del estimador se despeja n, se tiene la fórmula del tamaño de muestra cuando se decide aplicar un diseño muestral estratificado por afijación óptima:

$$n = \frac{\left(\sum_{h=1}^{L} W_h \, S_h \right)^2}{V(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^{L} W_h \, S_h^2}$$

Los elementos que conforman esta fórmula ya han sido tratados. Se insiste que $V(\bar{y}_{st})$ es el cuadrado del error de muestreo que se toleraría. Para especificar su valor, habrá que plantearse un desvío máximo d y una probabilidad de acierto que determinará el valor de t.

$$V(\bar{y}_{st}) = \frac{d^2}{t^2}$$

iii] Afijación óptima económica. Aparte de considerar simultáneamente el tamaño y variabilidad del estrato (N_h y S_h^2), a veces es recomendable introducir un tercer elemento: el costo por unidad encuestada en cada estrato. Sucede que en algunas investigaciones hay diferencias sustanciales en cuanto a las facilidades de acceso a la información por colectar y puede ser justificado tomar este elemento que tiene su origen en limitaciones financieras. Cabe aclarar que tal procedimiento sólo se justifica en la medida que su introducción no signifique pérdida de representatividad en las muestras de los estratos. La forma de "afijar" una muestra dada obedece en este caso a la siguiente relación.

$$n_h = \frac{N_h \, S_h \, / \sqrt{C_h}}{\sum_{h=1}^{L} N_h \, S_h \, / \sqrt{C_h}} \cdot n$$

donde C_h es el costo de investigar una unidad en el estrato h.

iv] Afijación arbitraria. Finalmente, cuando se "decide" distribuir la muestra sobre la base del buen sentido, cuidando de no perder representatividad y tomando supuestos tentativos en cuanto a heterogeneidad de los estratos, es decir, dirigiendo la distribución sobre la base de un conocimiento ilustrado de la población, se dice que la afijación es arbitraria. Para la determinación de los intervalos de confianza se utiliza la fórmula general de la varianza del estimador.

Es necesario advertir que el diseño muestral estratificado va cobrando mayores ventajas, sobre todo en cuanto a eficiencia, a medida que se tiene un mayor conocimiento de la población. De la adecuada elección del criterio de estratificación depende en gran medida la fidelidad y precisión de los resultados. Dado que las encuestas industriales, en un proceso de planificación, deben tener cierta periodicidad, permitirá conseguir cada vez mejores informaciones con lo que se facilitará la tarea iterativa de acercarse, con los estimadores, a valores que estén muy próximos a los que se dan en la realidad.

II

ENCUESTAS INDUSTRIALES

En gran parte de la bibliografía disponible sobre muestreo aparece, en general, un esquema de las etapas que se deben cumplir en una investigación muestral. Pero puede observarse al mismo tiempo que obedecen a circunstancias que precisamente no parecen ser las que se dan en los países en desarrollo, específicamente los de América Latina, desde el punto de vista de la carencia de informaciones básicas. Si bien es cierto que las etapas son inevitables, cualesquiera sean las circunstancias que se contemplen, no lo es menos que hay necesidad de adaptar y jerarquizar cada punto, en función de las dificultades y posibilidades específicas de la planificación industrial en estos países.

A continuación se detallarán las diferentes etapas y facetas de una investigación del sector industrial, a través de técnicas muestrales. Se recalcará aquellos problemas sobre los que la experiencia en investigaciones del sector industrial aconseja poner el mayor cuidado, detallando hasta donde sea posible la compatibilización de los conceptos teóricos con las metodologías y sus alternativas prácticas.

A. DETERMINACIÓN CLARA Y PRECISA DE LOS OBJETIVOS

Es de fundamental importancia especificar en la mejor forma posible cuáles son los fines que persigue la investigación muestral. El diseño de la muestra obedecerá en cada caso a consideraciones distintas. Es diferente diseñar una muestra para averiguar la magnitud que de la capacidad instalada en el sector industrial se utiliza efectivamente, en comparación a indagar la estructura de insumos importados que tienen las industrias dinámicas, o a investigar la situación de este sector, identificar sus principales escollos y disponer de antecedentes para proyecciones en el nivel agregado. Desde otro punto de vista, una misma muestra no puede ser igualmente eficiente si se trata de una primera investigación sobre rela-

ciones interindustriales, que si se trata de corregir coeficientes técnicos cuando se tienen sospechas de que éstos puedan haber variado. Es necesario tomar conciencia de que, según lo que se desea averiguar y la ponderación que la investigación tenga dentro del análisis general, se podrá diseñar en cada caso una muestra adecuada. No hay pues diseños "matrices" en toda su extensión; lo más que puede hacerse en aras de una cierta orientación es fijar gruesas líneas dentro de las cuales hay una infinidad de alternativas. Más aún, otra condicionante es la situación del sector en cuanto al conocimiento que de él se tiene. En dos encuestas industriales en que se persiguen los mismos objetivos los diseños muestrales pueden ser muy diferentes si se considera la cantidad, calidad y antigüedad de las informaciones de que se disponga en uno y otro caso.

Sólo cuando haya completo acuerdo en el equipo investigador sobre lo que se desea averiguar se podrá iniciar el diseño muestral propiamente tal, siempre que éste sea factible. Recuérdese que el muestreo no es apto para toda averiguación; hay casos donde es saludable renunciar a utilizar este procedimiento, porque es preferible desconocer algo antes que tener una información apreciablemente errada.

Una clara delimitación de objetivos permitirá centrar la investigación en los puntos más importantes. Debe dejarse de lado la idea de aprovechar la encuesta para averiguar todo aquello que tiene y "puede llegar" a tener importancia en el futuro; ese frecuente criterio es altamente perjudicial, porque hace de la encuesta un conjunto de preguntas en las que se diluyen los objetivos centrales. Basta analizar los formularios de encuestas a la industria, y en no pocos casos se encontrará este tipo de defectos. Aquí el planificador debe tomar en consideración la puntos señalados compatibilizando y jerarquizando los objetivos.

B. DELIMITACIÓN DEL MARCO MUESTRAL

Las estimaciones que proporcione una muestra se generalizan para una población. El campo de donde proviene una muestra, es decir el marco muestral, exige una muy precisa definición. En el sector industrial es especialmente peligroso dejar de ser acucioso en esta etapa. Como se sabe es necesario definir lo que se entenderá por unidad industrial. Así, corrientemente se divide este sector en industria manufacturera, que puede estar constituida por

las empresas industriales que cuentan con cinco o más obreros, e industria artesanal, las que no alcanzan dicha cifra. Pero aquí el problema aún no estará resuelto, porque dentro de las empresas manufactureras puede haber unidades que se dediquen tanto a actividades industriales como a comerciales, agrícolas, etc. Será necesario fijar un criterio que permita clasificarlas nítidamente en una u otra categoría. Es indispensable contar con un listado, directorio o empadronamiento donde estén inscritas todas las unidades sobre las cuales se desea investigar. Normalmente los países disponen de censos industriales que son utilizados como marco de la muestra; sobre este punto es necesario destacar que dichos censos no pueden hacerse anualmente: cuando más se realizan cada cinco y más años y en consecuencia un trabajo adicional que no puede evitarse es la actualización de los directorios. Hay empresas que se forman después del censo, otras que desaparecen y finalmente las que cambian de giro y tamaño dentro de la misma industria; estos cambios deben ser identificados antes de la extracción de la muestra. Es frecuente que en las encuestas industriales se desee tener alguna clasificación, al menos sobre las principales variables, según las ramas o agrupaciones industriales. Sobre lo mencionado surge en algunos casos otro problema: el de la identificación de todas y cada una de las unidades que conforman la población, dentro de dichas categorías.

Un criterio de clasificación podría encontrarse en la escritura jurídica de formación de la empresa, en la que por regla general se destaca el giro o actividad, pero ocurre que en algunas legislaciones sobre el particular el cambio de giro exige una serie de trámites e impuestos que hacen que los industriales, al formar una empresa, aprovechen para especificar una serie de actividades o giros potenciales, para evitarse molestias y erogaciones en un eventual cambio de actividad. Este hecho obliga en muchos casos a prescindir de esa fuente de clasificación y puede no quedar otro camino que una inspección censal. Muchas veces, la misma muestra puede ser utilizada para "ajustar" tentativamente el directorio industrial, en cuanto a clasificación de industrias por giro o tamaño. No está de más señalar que en estos casos puede ser necesaria una muestra complementaria por razones obvias.

Cuando no se disponga de marcos muestrales ni siquiera medianamente aceptables, como ocurre con la industria artesanal, corrientemente se "estiman" los valores de sus correspondientes variables mediante asociación y ajustes correctivos con las peque-

ñas empresas manufactureras. Es probable que tales estimaciones no estén muy distantes de los valores efectivos. En todo caso, para establecer los coeficientes de ajuste es necesario averiguar las similitudes y diferencias, cualitativas y cuantitativas, entre la pequeña industria manufacturera y la industria artesanal. No puede prescindirse de un empadronamiento por áreas de este tipo de unidades industriales; lo que ocurrirá, en general, es que sólo será necesaria una pequeña muestra para tener las estimaciones necesarias para realizar los mencionados ajustes, o para cuantificar directamente los estadígrafos que interesen a este subsector.

Finalmente, en muchos casos no será posible determinar para la industria artesanal el marco de la muestra definitivo y tendrá que recurrirse a marcos muestrales más generales, siempre que tengan una relación razonable con la investigación.

C. ELECCIÓN DEL DISEÑO MUESTRAL

Dadas las características de los sectores industriales y pensando siempre en función de la recolección de informaciones para la planificación de la industria, no hay mucha posibilidad de elección en cuanto a los diseños muestrales. El muestreo estratificado calza perfectamente bien con los objetivos que se plantean en este tipo de estudios; tanto en lo que concierne a la eficiencia (menor error comparado con otros diseños muestrales para igual tamaño de muestra) como en lo que toca a la posibilidad de tener estimaciones para cada estrato o subestrato.

Ahora bien, en cuanto a las informaciones básicas de las que normalmente disponen los países a través de los censos industriales, cabe advertir que prácticamente hay un criterio obligado de estratificación: por tamaño, asimilando éste al número de obreros con que cuenta cada unidad industrial. Vale la pena advertir sobre este punto que no siempre se justifica este tipo de estratificación, a no ser que se piense en términos de factibilidad, es decir porque puede no ser posible estratificar de otra manera, aunque los conceptos teóricos así lo indiquen. Si se piensa que la finalidad principal que debe cumplir un criterio de estratificación es la de hacer lo más homogéneas posibles las unidades de cada estrato, de modo de disminuir la magnitud de la varianza, surge inmediatamente la pregunta: ¿Qué tipo de homogeneidad? Debe elegirse una variable estratégica o clave, para agrupar las unidades de la población según los valores que toman en cuanto a dicha

variable. Agrupar las unidades según el número de obreros implica admitir que esta variable es representativa, en cuanto a clasificación, de las otras variables que interesan, como el valor agregado, los insumos, la capacidad utilizada, etc. El avance tecnológico produce cierto escepticismo para pronunciarse sobre la validez de ese supuesto. La mecanización puede disminuir la ocupación al mismo tiempo que aumentar considerablemente, por ejemplo, el valor agregado. Según uno u otro criterio, una industria con esas características puede quedar clasificada en estratos muy diferentes. Si el objetivo central es estudiar la absorción de recursos, principalmente mano de obra, su productividad y composición, el primer criterio sería adecuado, pero si lo que realmente interesa es tener estimaciones sobre el producto industrial, su estructura y dinamismo, parecería preferible utilizar, siempre que fuera factible, el segundo criterio.

En general en una encuesta industrial se plantean una serie de objetivos; de su jerarquización dependerá la elección del criterio adecuado. Si hay más de un objetivo importante, como ocurre con frecuencia, habrá que buscar un criterio que compatibilice ambos; tal vez no se consiga hacerlo en forma ideal, pero al menos podrá lograrse una correspondencia aceptable.

Una vez que la población ha quedado estratificada de acuerdo con el criterio elegido, será indispensable establecer nuevas clasificaciones para propósitos de planificar el desenvolvimiento del sector. Dentro de cada estrato podrá requerirse subclasificaciones que conformarán subestratos, como por ejemplo en industrias tradicionales y dinámicas, según ramas industriales, según ubicación geográfica, muy importante dentro de planes de integración regional, etc. De esta manera se podrá disponer de una serie de clasificaciones cruzadas que permitirán al programador realizar análisis en un distinto nivel de agregación. Los diagnósticos serán más acertados, las metas serán más detalladas y concretas y el control se facilitará muchísimo. La reformulación de los planes se hará sobre la base de un sólido conocimiento de lo ocurrido.

D. CÁLCULO DEL TAMAÑO DE MUESTRA

De más está insistir sobre la trascendencia de esta fase. Prácticamente en este punto es donde se conjugan todos los elementos conceptuales vistos en la primera parte. De otro lado, es donde más cuenta respetar los principios teóricos, al confrontarse con una

realidad que no se compadece con los supuestos simplificadores que debieron hacerse en busca de organicidad y claridad de exposición.

Recuérdese que, principalmente desde un punto de vista puramente técnico, hay tres elementos fundamentales que determinan el tamaño de una muestra: la magnitud de la población, la variabilidad con que se distribuye la característica que se investiga y la precisión que se desea tener. Así presentado el asunto es extraordinariamente sencillo, pero analizado con un poco más de detenimiento se comprenderá la complejidad que representa. En lo que se refiere a tamaño de la población, el problema radica principalmente en la definición de la unidad a investigarse y en la delimitación del marco muestral, ambos aspectos tratados ya en el punto B.

El problema crucial dice relación con la variabilidad de la población, es decir con el estadígrafo S^2. Prescindiendo por un momento de discutir sobre si se cuenta o no con esa información, es interesante discutir sobre cuál será la variable para la que es necesario cuantificar o disponer de una varianza. Revisando cualquier encuesta industrial, se comprobará que se indaga sobre una cantidad grande de variables; pues bien, ¿la varianza de cuál de ellas se tomará? Habrá algunas variables cuya distribución es muy homogénea, en tanto que habrá otras que presentan enorme heterogeneidad. En rigor, habrá que elegir sólo una varianza. Puede pensarse en elegir aquella que presente el más alto valor, y es difícil que ello conduzca a resultados prácticos; normalmente en ese caso se calculará un tamaño de muestra excesivamente grande, sin posibilidades de financiamiento. Hay variables cuya varianza es tan grande que el cálculo del tamaño de muestra, con una precisión razonable, equivale en el hecho a casi "censar" la población. En todo caso, constituye la alternativa más defendible desde un punto de vista teórico. Las limitaciones que ofrece una investigación, en la realidad, obligan en general a desecharla. Deberá pensarse en elegir una variable estratégica o representativa del conjunto de variables. En primer lugar es sumamente difícil disponer de varianzas para todas las variables que interese investigar; además, como se vio, aun en el hipotético caso de que se dispusiera de tales estadígrafos, la muestra resultante podría exceder en demasía el presupuesto disponible. En la práctica es un poco más realista empezar determinando el tamaño de muestra en función de los recursos, y luego calcular "técnicamente" el ta-

maño de muestra en términos de la variable elegida como representativa, es decir, la que más interese en la investigación. De la comparación de los tamaños de muestra resultantes por esos dos métodos diferentes nacerá el ajuste que será necesario hacer en aras de la factibilidad de la encuesta. Habrá tres posibilidades: se aumenta la cantidad de recursos, se disminuye *razonablemente* la precisión, o ambas cosas en alguna medida.

De cualquier manera, la muestra que resulte sólo garantizará la estimación de la media aritmética, con la precisión y confianza especificadas (d y t) de aquella variable cuya varianza se ha tomado en cuenta para el cálculo de n. Para aquellas variables que tengan una varianza menor que la de la variable clave, el tamaño de muestra será más que suficiente, es decir, se conseguirá mayor precisión que la especificada; en cambio para aquellas variables con mayor varianza, el tamaño de muestra será insuficiente y, en consecuencia, para dichas variables se estará especificando implícitamente una precisión menor (mayor d). Una vez realizada la encuesta y en conocimiento de las varianzas muestrales de todas las variables, deberá procederse a calcular los intervalos de confianza. Es frecuente encontrarse con sorpresas desagradables; hay algunos intervalos, cuya magnitud es tan extraordinariamente grande que tomar la estimación puntual implica serios riesgos. El planificador deberá tomar conciencia de estos hechos, para que en todo momento contemple, en la utilización de tales estimadores, sus bondades y limitaciones. Es muy frecuente escuchar la frase de que basta tener un orden de magnitud para formarse una idea, pero ocurre que a veces estos "órdenes de magnitud" tienen un rango de variabilidad tan amplio que, situándose en uno u otro límite del intervalo, las conclusiones pueden ser del todo diferentes. Esta situación ocurre particularmente en las investigaciones industriales en niveles muy desagregados, donde las poblaciones son demasiado pequeñas.

La elección de la variable clave es de una trascendencia que es innecesario recalcar. En la encuesta sobre la industria en Centroamérica, para fines del cálculo del tamaño de muestra, se ha tomado como variable representativa el tamaño de las industrias a través del número de obreros que ocupan. No sería aventurado adelantar que habrá otras variables, como el valor agregado, los insumos, etc., que tendrán mayor variabilidad que la variable elegida. En esos casos, con seguridad que la precisión será mucho menor que la especificada para estimar el número promedio de

obreros por industria. Lo interesante será analizar cuál es la precisión efectivamente alcanzada y la confianza que ello merece.

En cuanto a la precisión, indicada por la magnitud del semiancho del intervalo de confianza (d), en otras palabras el desvío máximo tolerable, constituye un grado de libertad en el simplificado modelo del cálculo de una muestra estratificada. Puede teóricamente tomar cualquier valor positivo; lo interesante es determinar un valor, un desvío, de manera que no signifique por una parte un rango demasiado amplio que no permita obtener conclusiones generales, y por otra que no sea tan innecesariamente pequeño como para que exija un tamaño de muestra prohibitivo. En verdad, en una encuesta industrial sólo será necesario determinar la precisión deseada para la variable clave, ya que para el resto de las variables habrá dejado de ser un grado de libertad, ya que el tamaño de muestra es único para toda la investigación. La precisión alcanzada para las otras variables habrá que calcularla en función de ese único tamaño de muestra y de la desviación típica (s) de cada una de las variables. Como en general cada variable tiene distinto grado de heterogeneidad, tendrá también una diferente precisión. En la fijación del desvío máximo tolerable para la variable representativa, el planificador tendrá que tomar una decisión. Si por ejemplo se sabe que el valor agregado de la industria en Chile fue en 1963 de 811 millones de escudos de 1960, el promedio por industria será del orden de 115 000 escudos de 1960. Si además se supone que el valor agregado por industria crece al 5 por ciento anual, el promedio para 1964 será de 121 000 escudos de 1960 aproximadamente. La pregunta que debe plantearse el planificador es: si se quisiera estimar este promedio, ¿cuál sería el desvío máximo (d) que se podría tolerar: 10 000, 20 000 o 25 000 escudos? Compulsando los tamaños de muestra resultantes para cada alternativa, los objetivos de la investigación, y el presupuesto disponible, se tendrá que tomar esa importante decisión. Cuando se tiene conciencia de que el resto de las variables por investigar tiene una heterogeneidad mucho mayor que la de la variable clave, será necesario considerar en el tamaño de muestra un desvío menor al necesario, en la medida que no encarezca en exceso la encuesta industrial. Una vez calculado el tamaño definitivo de la muestra, el siguiente paso será la distribución de ella, entre los estratos y/o subestratos por la correspondiente afijación elegida.

E. SELECCIÓN DE LAS UNIDADES MUESTRALES

Se había advertido que el muestreo estratificado no era más que una combinación de diseños muestrales aleatorios simples; la selección de las unidades deberá ser hecha en forma aleatoria. Sin embargo, la extracción por medio de una tabla de números aleatorios puede significar un procedimiento muy engorroso, como primera desventaja, y luego que la concentración industrial, alrededor de las zonas urbanas, determina que una gran parte de la muestra provenga justamente de esas áreas. En el hecho lo mencionado en segundo término no constituye una desventaja, a no ser que se tome en cuenta que la muestra no entregará resultados por zonas geográficas, aspecto que debiera interesar al programador. En tales casos puede justificarse seleccionar muestras sistemáticas, es decir, una de cada k industrias de la población (siendo $k = \frac{N}{n}$). En el caso de la encuesta industrial de la Corporación de Fomento de Chile, se siguió una extracción sistemática de norte a sur del país, garantizando de esta manera cierta representatividad de zonas geográficas que de otra manera no la hubieran tenido. La extracción de una muestra sistemática puede asociarse con una muestra aleatoria, siempre que la población no presente un determinado ordenamiento coincidente con el intervalo de sistematización y que la iniciación de la selección sea aleatoria, es decir, que una de las k primeras unidades sea elegida al azar. En tal caso, podrán utilizarse los mismos estimadores de las muestras aleatorias o las aproximaciones correspondientes al muestreo sistemático.

F. ENTRENAMIENTO DE ENUMERADORES

El enumerador deberá ser una persona que, conociendo perfectamente los objetivos de la encuesta, tenga alguna experiencia en materia de industrias. En general, en dichas encuestas, se inquiere sobre una serie de datos técnicos, muchos de los cuales son dados directamente por el respondiente; el enumerador debe estar en condiciones de calificar estas respuestas. En las encuestas de opinión pública, en materia de política, religión, simpatías, actitudes, etc., es correcto que el enumerador sea un elemento en lo posible neutro hacia las respuestas del encuestado. En las encuestas industria-

les, sobre todo en las preguntas que dicen relación con magnitudes, el encuestador tendrá una participación activa y verificadora. Incluso para la realización de encuestas industriales donde se averigüe por datos técnicos, es conveniente que el encuestador sea un técnico entendido en la materia; de otro modo se corre el riesgo de obtener respuestas deficientes que pueden estropear el trabajo.

En todo caso será necesario un proceso de adiestramiento, probando en el terreno la eficiencia de cada enumerador sobre la base de indagaciones que exijan una cabal interpretación, teniendo previamente a disposición los datos que averiguará en el terreno cada encuestador.

G. COLABORACIÓN DE LA POBLACIÓN INFORMANTE

En las encuestas industriales, es de una significación muy grande conseguir una actitud positiva de la población que se investiga, imprescindible para garantizar una cierta fidelidad en la masa de informaciones que recolectar. Es frecuente que se advierta a la población la trascendencia de la muestra, y una manera de hacer más efectiva esa colaboración es realizar conjuntamente la investigación con asociaciones de industriales u organismos de probado prestigio, universidades, institutos, etc. De una u otra manera existe siempre un margen considerable de no respuesta. Sobre el particular hay que destacar que si la no respuesta tiene su origen en una desaparición o cambio de estrato de la unidad que se pretende investigar, no debe ser motivo de preocupación, ya que puede tomarse como una muestra representativa de una parte del universo que desaparece o cambia de estrato. El problema es muchísimo más serio cuando, existiendo la unidad, hay negativa de dar respuesta. Basar las estimaciones solamente en poblaciones de respondientes, puede ocasionar sesgos de alguna magnitud. En esos casos las estimaciones realizadas, podrán ser generalizadas sólo a la población de respondientes y no a la población total. Para tener informaciones sobre la población de no respondientes, habrá que disponer de encuestas de seguimiento u otros métodos indirectos. Sobre este problema de la no respuesta y, además de él, sobre la respuesta defectuosa, es necesario una consideración detenida. La necesidad de planificar y para ello de contar con informaciones serias, periódicas y oportunas, son cuestiones de aceptación general. Para poder contar con informaciones con tales

atributos, parece que no hay otra salida más que en las legislaciones de los países se contemple la obligación y las sanciones correspondientes, en forma similar a una declaración de impuestos.

No puede esperarse a que las poblaciones tomen conciencia del agudo problema. Esto implicaría la elaboración previa de planes nacionales de estadística, donde se coordinarían todas las necesidades de información de los diferentes organismos de planificación, ejecución y control.

H. ORGANIZACIÓN DEL TRABAJO EN EL TERRENO

La programación de la recolección de informaciones debe plantearse en términos de las peculiaridades geográficas, concentración industrial en la urbe, etc. Debiera haber un contacto permanente entre la oficina de diseño de la muestra y el equipo de encuestadores. Siempre se presentarán tipos de respuesta o características de la industria que no se han contemplado en el proceso de adiestramiento. En tales casos es preferible solucionar el problema consultando a los directores de la investigación. En encuestas industriales en el nivel nacional, es conveniente que en los lugares donde no hay oficinas a cargo de los que dirigen la encuesta la enumeración se realice por personas altamente calificadas para este tipo de trabajos. En la práctica, la recolección de datos en provincias la realiza personal de las juntas de planificación u organismos similares, dejando la investigación en la capital o sede de la encuesta a personal temporal contratado para este efecto.

Es indispensable que se implanten ciertos controles en forma simultánea a la recolección de informaciones. De esta manera se previene que se cometan errores voluntarios e involuntarios. Verificar una pequeña parte del número de encuestas que realiza cada enumerador puede ser un medio para disminuir un posible sesgo. Es importante que se realice esta forma de control simultáneamente al proceso de extracción de información, para que puedan tomarse medidas *oportunas* de corrección. Los controles *ex post* son en general tardíos y no se alcanza a remediar el problema, si no a expensas de muchos mayores gastos y demoras perjudiciales.

I. SISTEMATIZACIÓN DE DATOS

Toda la masa de informaciones extraídas deberá ser codificada,

tabulada, clasificada, interpretada y verificada. En esta altura es donde se deberá diseñar el nivel de desagregación y tipos de clasificaciones necesarias. Por una parte, las informaciones sumamente desagregadas no permiten obtener conclusiones; por la otra, estimaciones globales sólo dan una idea muy general de la situación de la industria en sus diversos aspectos. Es imprescindible pues combinar estimadores en distintos niveles de desagregación; unas y otras son necesarias cuando se pretende fijar una política y estrategia de desarrollo industrial. Conocer por ejemplo el coeficiente producto-capital para toda la industria manufacturera es un dato de mucha utilidad, pero además es necesario conocer sus componentes, ya sea por ramas industriales, por ubicación geográfica, por tamaño de establecimiento, etc. Con la utilización de computadores electrónicos, la sistematización de datos se hace mucho más amplia y rápida. Previamente a la introducción de los datos en los computadores, aquéllos debieron ser sometidos a un estricto control, máxime si fueron extraídos por enumeradores que no tenían un cabal conocimiento técnico-económico de la industria. Este control en general se realiza por comparación entre industrias similares, por antecedentes que se tienen acerca de los probables resultados que arrojaría la encuesta en cada caso y por revisión analítica en cuanto a lo razonables que pudieran ser esas informaciones.

J. DETERMINACIÓN DEL COSTO TOTAL

Para la realización de una investigación muestral se ha debido contar con un determinado presupuesto, el que será necesario comparar con el costo real de la investigación. En la misma forma en que se ha realizado el presupuesto, contemplando cada una de las principales etapas, se deberá realizar la comparación con la realidad. De esta manera se acumularán experiencias útiles para posteriores investigaciones, ya que en todas las diferencias que se produzcan habrá que analizar sus causas; así se tomará una mayor conciencia de la trascendencia de cada etapa, lo que puede redundar en mayor aprovechamiento de los recursos disponibles en futuras investigaciones.

K. PUBLICACIÓN DE RESULTADOS

En la publicación de los resultados obtenidos mediante muestras es fundamental indicar las principales características del diseño utilizado: criterio de estratificación, tamaño de muestra, tipo de afijación, estratos en que se hizo censo, precisión para cada una de las variables principales, probabilidad de acierto, formas de selección, etc. De esta manera los usuarios conocerán las bondades y limitaciones de que son objeto las estimaciones resultantes. Incluso es necesario presentar un anexo con una descripción minuciosa de toda la investigación, detallando los principales problemas y las formas de solución.

III

DEMOSTRACIONES MATEMÁTICAS DE MUESTREO ALEATORIO SIMPLE

1. La media de la muestra, $\bar{y}$, es un estimador insesgado de $\bar{Y}$, media de la población.

Demostración: se sabe que

$$\text{A]} \quad E(\bar{y}_h) = \frac{\sum_h \bar{y}_h}{\binom{N}{n}} = \frac{\sum_h (y_1 + y_2 + y_3 \cdots + y_n)/h}{n} \cdot \frac{1}{\frac{N!}{n!\,(N-n)!}}$$

donde $h = 1, 2, 3, \ldots\ldots \binom{N}{n}$

$i = 1, 2, 3, \ldots\ldots n \ldots N$

Para poder evaluar la suma que aparece en el numerador, es necesario averiguar en cuántas muestras aparece un valor específico cualquiera y_i. Aparte de ese valor y_i, habrá otras (N—1) unidades disponibles para el resto de la muestra y otros (n—1) lugares a ocupar en la muestra. Luego el número de muestras que contienen el valor y_i estará dado por:

$$C_{n-1}^{N-1} = \frac{(N-1)!}{(n-1)!\,(N-n)!}$$

Por lo tanto:

$$\text{B]} \quad \sum_{h=1}^{\binom{N}{n}} (y_1 + y_2 + \ldots\ldots + y_n)/h =$$

$$= \frac{(N-1)!}{(n-1)!\,(N-n)!}\,(y_1 + y_2 + y_3 + \ldots\ldots + y_N)$$

Nótese que, por el anterior artificio, la suma de los valores se extiende hasta y_N, porque el valor específico y_1, puede ser cualquiera de los valores poblacionales.

Remplazando la expresión obtenida en B en el numerador de la igualdad señalada con A, se tiene:

$$E(\bar{y}_h) = \frac{(N-1)!}{(n-1)!\,(N-n)!}(y_1 + y_2 + y_3 + \cdots + y_N)\,\frac{1}{n}\cdot\frac{1}{\frac{N!}{n!\,(N-n)!}}$$

$$E(\bar{y}_h) = \frac{(N-1)!}{(n-1)!\,(N-n)!}(y_1 + y_2 + y_3 + \cdots + y_N)\cdot\frac{1}{n}\,\frac{n!\,(N-n)!}{N!}$$

Simplificando factoriales:

$$E(\bar{y}_h) = \frac{(y_1 + y_2 + y_3 + \cdots + y_N)}{N} = \bar{Y}$$

II. Demostración

$$E[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2] = \frac{n\,(N-1)}{N}S^2$$

Donde $S^2 = \dfrac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$ (varianza de Cochran)

La esperanza propuesta es igual a:

$$C] \quad \frac{\sum_{h=1}^{\binom{N}{n}}(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2 / h}{\binom{N}{n}}$$

Nótese que los índices de la sumatoria se extienden a todas las posibles muestras.

Utilizando el mismo artificio que en la demostración 1, el numerador se puede evaluar de la siguiente manera:

$$\binom{N-1}{n-1} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots (y_N - \bar{Y})^2]$$

Remplazando esta expresión, en el numerador de C se tiene:

$$\frac{\binom{N-1}{n-1} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + (y_3 - \bar{Y})^2 + \cdots (y_N - \bar{Y})^2]}{\binom{N}{n}} =$$

$$= \frac{(N-1)!}{(n-1)!\,(N-n)!} \cdot \frac{n!\,(N-n)!}{N!} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_N - \bar{Y})^2]$$

Simplificando se obtiene:

$$\frac{n}{N} [(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_N - \bar{Y})^2]$$

Multiplicando y dividiendo por N — 1:

$$\frac{(N-1)n}{N} \left[\frac{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_N - \bar{Y})^2}{N-1} \right] =$$

$$= \frac{(N-1)n}{N} \left[\frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1} \right]$$

La expresión dentro del paréntesis es lo que se había llamado S^2 (varianza de Cochran).
Luego, queda demostrado que:

$$E[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2] = \frac{n\,(N-1)}{N} S^2$$

III. $E[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \cdots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})] = \frac{n(n-1)}{2N} S^2$

La esperanza indicada significa:

$$\frac{\sum_{h=1}^{\binom{N}{n}} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \cdots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})]}{\binom{N}{n}} /h$$

Para evaluar la suma del numerador, debe pensarse en el mismo artificio utilizado en las demostraciones I y II, teniendo en cuenta que, como ahora se trata de productos de dos desviaciones, habrá dos valores específicos en la población que podrán ocupar dos lugares en la muestra. La evaluación de la suma del numerador estará dada, pues, por:

$$\binom{N-2}{n-2} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_2 - \bar{Y}) + (y_3 - \bar{Y}) + \cdots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})]$$

$$\binom{N-2}{n-2} \frac{[(y_1 - \bar{Y})(y_2 - \bar{Y} + \cdots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})]}{\binom{N}{n}}$$

$$\frac{(N-2)!}{(n-2)!(N-n)!} \cdot \frac{(N-n)!\, n!}{N!} [(y_1 - \bar{Y})(y_2 - \bar{Y}) + \cdots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})]$$

Simplificando factoriales y multiplicando y dividiendo por 2 se tiene:

$$\frac{n(n-1)}{2N(N-1)} [2\{(y_1 - \bar{Y})(y_2 - \bar{Y}) + \cdots + (y_{N-1} - \bar{Y})(y_N - \bar{Y})\}]$$

A la expresión dentro del paréntesis cuadrado le sumaremos y restaremos los términos:

$$(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \ldots + (y_N - \bar{Y})^2$$

Luego

$$\frac{n\,(n-1)}{2N\,(N-1)}\,[2\{(y_1 - \bar{Y})\;(y_2 - \bar{Y}) + \ldots + (y_{N-1} - \bar{Y})\;(y_N - \bar{Y})\} +$$

$$+ \{(y_1 - \bar{Y})^2 + y_2 - \bar{Y})^2 + \ldots + (y_N - \bar{Y})^2\} -$$

$$- \{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \ldots + (y_N - \bar{Y})^2\}]$$

Observando los términos dentro del paréntesis cuadrado se concluye que la expresión que tiene signo negativo es el numerador de la varianza de Cochran, y los dos primeros que tienen signo positivo son el desarrollo del cuadrado de una suma de desviaciones, ya que aparecen las desviaciones al cuadrado y los dobles productos correspondientes, es decir

$$\sum_{i=1}^{N} (y_i - \bar{Y})^2 + 2 \sum_{i \neq j=1}^{N} (y_i - \bar{Y})\;(y_j - \bar{Y}) =$$

$$= \left[\sum_{i=1}^{N} (y_i - \bar{Y})\right]^2 = 0$$

por tratarse del cuadrado de la suma de las desviaciones respecto de la media aritmética (es decir, el cuadrado de cero). Con esta reducción, la expresión queda de la siguiente manera:

$$\frac{n\,(n-1)}{2N\,(N-1)}\,[-\{(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + (y_3 - \bar{Y})^2 + \ldots + (y_N - \bar{Y})^2\}] =$$

$$= -\frac{n\,(n-1)}{2N(N-1)}\left[\sum_{i=1}^{N} (y_i - \bar{Y})^2\right]$$

$$- \frac{n\ (n-1)}{2N} \left[\frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1} \right]$$

$$- \frac{n\ (n-1)}{2N} S^2$$

IV.

$$V(\bar{y}) = E(\bar{y}_h - \bar{Y})^2 = \frac{\sum_{h=1}^{\binom{N}{n}} (\bar{y}_h - \bar{Y})^2}{\binom{N}{n}} = \frac{S^2}{n} \left(1 - \frac{n}{N}\right)$$

Empezamos la demostración por esta identidad

$$n(\bar{y} - \bar{Y}) = n \cdot \frac{\sum_{i=1}^{n} y_i}{n} - n\,\bar{Y}$$

$$= y_1 + y_2 + y_3 + \cdots + y_n - n\,\bar{Y}$$

$$= y_1 - \bar{Y} + y_2 - \bar{Y} + y_3 - \bar{Y} + \cdots + y_n - Y$$

Elevando al cuadrado ambos miembros se tiene:

$$n^2 (\bar{y} - \bar{Y})^2 = [(y_1 - \bar{Y}) + (y_2 - \bar{Y}) + (y_3 - \bar{Y}) + \cdots + (y_n - \bar{Y})]^2$$

$$n^2 (\bar{y} - \bar{Y})^2 = (y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2 +$$

$$+ 2\,[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \cdots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})]$$

Aplicando el operador esperanza matemática:

$$n^2\, E(\bar{y} - \bar{Y})^2 = E[(y_1 - \bar{Y})^2 + (y_2 - \bar{Y})^2 + \cdots + (y_n - \bar{Y})^2] +$$

$$+ 2E[(y_1 - \bar{Y})(y_2 - \bar{Y}) + (y_1 - \bar{Y})(y_3 - \bar{Y}) + \cdots + (y_{n-1} - \bar{Y})(y_n - \bar{Y})]$$

Los dos términos de la derecha tienen expresiones conocidas, demostradas en los puntos II y III.

Remplazando tales expresiones se tiene:

$$n^2\ E(\bar{y} - \bar{Y})^2 = \frac{n(N-1)}{N} S^2 + 2\left[-\frac{n(n-1)}{2N} S^2\right]$$

$$n^2\ E(\bar{y} - \bar{Y})^2 = \frac{n(N-1)}{N} S^2 - \frac{n(n-1)}{N} S^2$$

$$n^2\ E(\bar{y} - \bar{Y})^2 = \frac{nS^2\ [(N-1) - (n-1)]}{N}$$

$$n^2\ E(\bar{y} - \bar{Y})^2 = nS^2\ \frac{(N-n)}{N}$$

$$E(\bar{y} - \bar{Y})^2 = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

Ésta es la fórmula de cálculo práctico de la varianza de la media de una muestra aleatoria simple. (Cuadrado del error estándar de estimación.)

v. Demostración de que la varianza de una muestra aleatoria simple es una estimación insesgada de la varianza de la población (ambas con denominadores restados en una unidad).

$$E(s_h^2) = E\left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}\right]$$

Sumando y restando la media poblacional:

$$(n - I)\ E(s_h^2) = E\left[\sum_{i=1}^{n} (y_i - \bar{y} \pm \bar{Y})^2\right]$$

$$(n - I)\ E(s_h^2) = E\left[\sum_{i=1}^{n} (y_1 - \bar{Y}) - (\bar{y} - \bar{Y})\right]^{2\,(*)}$$

Si hacemos:

$$y_i - \bar{Y} = W_i;\ \bar{y} - \bar{Y} = \bar{W}$$

tenemos:

$$\sum_{i=1}^{n} [(y_i - \bar{Y}) - (\bar{y} - \bar{Y})]^2 = \sum_{i=1}^{n} (W_i - \bar{W})^2$$

$$= \sum_{i=1}^{n} W_i^2 - 2\,\bar{W} \sum_{i=1}^{n} W_i + n\,\bar{W}^2$$

$$= \sum_{i=1}^{n} W_i^2 - 2\,\bar{W}\,(n\,\bar{W}) + n\,\bar{W}^2$$

$$= \sum_{i=1}^{n} W_i^2 - n\,\bar{W}^2$$

$$= \sum_{i=1}^{n} (y_i - \bar{Y})^2 - n(\bar{y} - \bar{Y})^2$$

Remplazando esta expresión en aquella marcada con el asterisco (p. 281) se tiene:

$$(n-1)\,E(s_h^2) = E\left[\sum_{i=1}^{n} (y_i - \bar{Y})^2 - n\,(\bar{y} - \bar{Y})^2\right]$$

$$= E\left[\sum_{i=1}^{n} (y_i - \bar{Y})^2\right] - n\,E\,(\bar{y} - \bar{Y})^2$$

Teniendo en cuenta las demostraciones II y IV y remplazando dichas expresiones en el miembro de la derecha de la igualdad anterior:

$$(n-1)\,E(s_h^2) = \frac{n(N-1)}{N} S^2 - n\,\frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

$$= \frac{n(N-1)}{N} S^2 - S^2\left(1 - \frac{n}{N}\right)$$

$$= S^2\,\frac{n(N-1) - (N-n)}{N}$$

$$= S^2\,\frac{n(N-1) - (N-n)}{N}$$

$$= S^2\,\frac{N(n-1)}{N}$$

Se obtiene finalmente:

$$E(s^2) = S^2$$

FORMULARIO SOBRE MUESTREO ALEATORIO SIMPLE

A. VARIABLE CUANTITATIVA

N: Número de elementos que componen la población.
n: Número de elementos que componen la muestra.
y_i: Valor de la variable de la i-ésima unidad de la población ($i = 1, 2, 3, \ldots N$).
y_i: Valor de la variable en la i-ésima unidad de la muestra ($i = 1, 2, 3, \ldots n$).
$\bar{Y}$: Media aritmética poblacional

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

$\bar{y}$: Media aritmética de la muestra

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \qquad \left\{\begin{array}{l}\text{Estimador insesgado de la} \\ \text{media aritmética poblacional}\end{array}\right\}$$

σ^2: Varianza de la población

$$\sigma^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N}$$

S^2: Varianza de Cochran para la población

$$S^2 = \frac{\sum_{i=1}^{N} (y_i - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^{N} y_i^2}{N-1} - \frac{N\bar{Y}^2}{N-1}$$

s^2: Varianza de Cochran para la muestra

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2}{n-1} - \frac{n\bar{y}^2}{n-1}$$

Estimador insesgado de la varianza de Cochran para la población.

Y: Total poblacional

$$Y = N\,\bar{Y} = \sum_{i=1}^{N} y_i$$

$\hat{Y}$: Estimador del total poblacional

$$\hat{Y} = N\,\bar{y}$$

$V(\bar{y})$: Varianza verdadera del estimador de la media o cuadrado de la desviación estándar verdadera

$$V(\bar{y}) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

$v(\bar{y})$: Estimador de la varianza del estimador de la media

$$v(\bar{y}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$$

$V(\hat{Y})$: Varianza verdadera del estimador del total

$$V(\hat{Y}) = N^2\,V(\bar{y})$$

$v(\hat{Y})$: Estimador de la varianza del estimador del total

$$v(\hat{Y}) = N^2\,v(\bar{y})$$

n_o: Tamaño de muestra (primera aproximación)

$$n_o = \frac{t^2\,S^2}{d^2}$$

en que t es el coeficiente de confianza y d es el semiancho del intervalo de confianza.

$$d = t\,\sqrt{V(\bar{y})}$$

n: Tamaño de la muestra (definitivo)

$$n = \frac{n_o}{1 + \frac{n_o}{N}}$$

B. VARIABLE CUALITATIVA (DOS CATEGORÍAS)

La variable solamente puede tomar dos valores: 1 si el elemento (en la población o en la muestra) posee la característica que se investiga, y 0 si no la posee.

P: Proporción en la población

$$P = \frac{\sum_{i=1}^{N} y_i}{N}$$

p: Proporción en la muestra

$$p = \frac{\sum_{i=1}^{n} y_i}{n}$$

Estimador insesgado de la proporción poblacional.

S^2: Varianza de Cochran para la población

$$S^2 = \frac{N}{N-1} PQ \qquad Q = 1 - P$$

s^2: Varianza de Cochran para la muestra

$$s^2 = \frac{n}{n-1} pq$$

Estimador insesgado de la varianza de Cochran para la población.

$$q = 1 - p$$

A: Número de elementos que poseen la característica en la población

$$A = NP = \sum_{i=1}^{N} y_i$$

Â: Estimador del número de elementos que poseen la característica en la población

$$A = Np$$

$V(p)$: **Varianza verdadera del estimador de la proporción o cuadrado de la desviación estándar verdadera**

$$V(p) = \frac{P\,Q}{n}\left(\frac{N - n}{N - 1}\right)$$

$v(p)$: **Estimador de la varianza del estimador de la proporción**

$$v(p) = \frac{p\,q}{n - 1}\left(\frac{N - n}{N}\right)$$

$V(\hat{A})$: **Varianza verdadera del estimador del número de elementos que poseen la característica que se investiga**

$$V(\hat{A}) = N^2\,V(p)$$

$v(\hat{A})$: **Estimador de la varianza del estimador del número de elementos que poseen la característica que se investiga**

$$v(\hat{A}) = N^2\,v(p)$$

n_o: **Tamaño de la muestra (primera aproximación)**

$$n_o = \frac{t^2\,P\,Q}{d^2}$$

en que: t es el coeficiente de confianza y d es el semiancho del intervalo de confianza.

$$d = t\sqrt{V(p)}$$

n: **Tamaño de muestra (definitivo)**

$$n = \frac{n_o}{1 + \dfrac{n_o - 1}{N}}$$

FORMULARIO SOBRE MUESTREO ALEATORIO ESTRATIFICADO

A. VARIABLES CUANTITATIVAS

N: Número de elementos que componen la población
n: Número de elementos que componen la muestra
N_h: Número de elementos del estrato h-ésimo
n_h: Número de elementos de la muestra del estrato h-ésimo
y_{hi}: Valor de la i-ésima unidad del estrato h-ésimo
$i = 1, 2, 3, \ldots N_h$; $h = 1, 2, 3, \ldots L$
y_{hi}: Valor de la i-ésima unidad de la muestra del estrato h-ésimo
$i = 1, 2, 3, \ldots n_h$; $h = 1, 2, 3, \ldots L$
$\bar{Y}_h$: Media aritmética del estrato h-ésimo

$$\bar{Y}_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h}$$

$\bar{Y}$: Media aritmética de la población

$$\bar{Y} = \frac{\sum_{h=1}^{L} \bar{Y}_h N_h}{N}$$

$\bar{y}_h$: Media aritmética de la muestra del estrato h-ésimo

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h}$$ Estimador insesgado de $\bar{Y}_h$

$\bar{y}$: Media aritmética de la muestra total

$$\bar{y} = \frac{\sum_{h=1}^{L} \bar{y}_h n_h}{n}$$

$\bar{y}_{st}$: Estimador insesgado de la media aritmética poblacional

$$\bar{y}_{st} = \frac{\sum_{h=1}^{L} \bar{y}_h \, N_h}{N}$$

S_h^2: Varianza de Cochran para el estrato h-ésimo

$$S_h^2 = \frac{\sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2}{N_h - 1} = \frac{\sum y_{hi}^2}{N_h - 1} - \frac{N_h \bar{Y}_h^2}{N_h - 1}$$

s_h^2: Varianza de Cochran para la muestra del estrato h-ésimo (estimador insesgado de S_h^2)

$$s_h^2 = \frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1} = \frac{\sum_{i=1}^{n_h} y_{hi}^2}{n_h - 1} - \frac{n_h \bar{y}_n^2}{n_h - 1}$$

Y: Total poblacional

$$Y = N \bar{Y} = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}$$

$\hat{Y}_{st}$: Estimador del total poblacional

$$\hat{Y}_{st} = N \bar{y}_{st}$$

$V(\bar{y}_{st})$: Varianza verdadera del estimador de la media

a] Fórmula general

$$V(\bar{y}_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

b] Para afijación proporcional

$$V(\bar{y}_{st}) = \frac{1 - f}{n} \sum_{h=1}^{L} W_h S_h^2 \quad \text{donde:}$$

$$f = \frac{n}{N} \quad \text{y} \quad W_h = \frac{N_h}{N}$$

c] Para afijación óptima

$$V(\bar{y}_{st}) = \frac{\left[\sum_{h=1}^{L} W_h S_h\right]^2}{n} - \frac{\sum_{h=1}^{L} W_h S_h^2}{N}$$

$v(y_{st})$: Varianza estimada del estimador de la media (las mismas fórmulas anteriores, pero utilizando s_h^2 en vez de S_h^2, por ser estimador insesgado)

n_h: Tamaño de la muestra en el estrato h-ésimo

a] Afijación proporcional

$$n_h = \frac{N_h}{N} \cdot n$$

b] Afijación óptima

$$n_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \cdot n$$

c] Afijación óptima económica

$$n_h = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}} \cdot n$$

n: Tamaño de la muestra; para estimar la media aritmética

a] Afijación proporcional

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde} \quad n_0 = \frac{\sum_{h=1}^{L} W_h S_h^2}{V(\bar{y}_{st})}$$

b] Afijación óptima

$$n = \frac{\left[\sum_{h=1}^{L} W_h S_h\right]^2}{V(\bar{y}_{st}) + \frac{1}{N} \sum_{h=1}^{L} W_h S_h^2}$$

B. VARIABLE CUALITATIVA (DOS CATEGORÍAS)

La variable solamente puede tomar dos valores: 1 si el elemento (en la población o en la muestra) posee la característica que se investiga, y 0 si no la posee.
(La nomenclatura es similar a la presentada en el punto A.)

P_h: Proporción en el estrato h-ésimo

$$P_h = \frac{\sum_{i=1}^{N_h} y_{hi}}{N_h} = \frac{A_h}{N_h}$$

P: Proporción en la población

$$P = \frac{\sum_{h=1}^{L} P_h N_h}{N}$$

p_h: Proporción en la muestra del estrato h-ésimo

$$p_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} = \frac{a_h}{n_h}$$ Estimador insesgado de P_h

p: Proporción de la muestra total

$$p = \frac{\sum_{h=1}^{L} p_h n_h}{n}$$

p_{st}: Estimador insesgado de la proporción poblacional

$$p_{st} = \frac{\sum_{h=1}^{L} p_h N_h}{N}$$

A: Número de elementos que poseen la característica en la población

$$A = N P$$

$\hat{A}$: Estimador del total poblacional (número de elementos)

$$\hat{A} = N p_{st}$$

S_h^2: Varianza del estrato h-ésimo

$$S_h^2 = \frac{N_h P_h Q_h}{N_h - 1}$$

$V(p_{st})$: Varianza verdadera del estimador de la proporción poblacional

a] Fórmula general

$$V(p_{st}) = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 (N_h - n_h)}{N_h - 1} \frac{P_h Q_h}{n_h}$$

b] Afijación proporcional

$$V(p_{st}) = \frac{N - n}{N} \frac{1}{n N} \sum_{h=1}^{L} \frac{N_h^2 P_h Q_h}{N_h - 1}$$

$$= \frac{1 - f}{n} \sum_{h=1}^{L} W_h P_h Q_h$$

c] Afijación óptima

$$V(p_{st}) = \frac{\left(\sum_{h=1}^{L} W_h \sqrt{\frac{N_h P_h Q_h}{N_h - 1}} \right)^2}{n} - \frac{\sum_{h=1}^{L} W_h \frac{N_h P_h Q_h}{N_h - 1}}{N}$$

$V(\hat{A})$: Varianza verdadera del estimador del total

$$V(\hat{A}) = N^2 V(p_{st})$$

$v(p_{st})$: Varianza estimada del estimador de la proporción. Se utilizan las mismas fórmulas anteriores, pero sustituyendo:

$\frac{N_h P_h Q_h}{N_h - 1}$ por $\frac{n_h p_h q_h}{n_h - 1}$, por ser estimador insesgado.

$v(\hat{A})$: Estimador de la varianza del estimador del total

$$v(\hat{A}) = N^2 \, v(p_{st})$$

n_h: Tamaño de la muestra en el estrato h-ésimo

a] Afijación proporcional

$$n_h = \frac{N_h}{N} \cdot n$$

b] Afijación óptima

$$n_h = \frac{N_h \, (P_h \, Q_h)^{1/2}}{\sum_{h=1}^{L} N_h \, (P_h \, Q_h)^{1/2}} \cdot n$$

c] Afijación óptima económica

$$n_h = \frac{N_h \, (P_h \, Q_h / c_h)^{1/2}}{\sum_{h=1}^{L} N_h \, (P_h \, Q_h / c_h)^{1/2}} \cdot n$$

n: Tamaño de la muestra para estimar una proporción

a] Afijación proporcional

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \quad \text{donde} \quad n_0 = \frac{\sum_{h=1}^{L} W_h \, P_h \, Q_h}{V(p_{st})}$$

b] Afijación óptima

$$n = \frac{n_0}{1 + \frac{1}{N \, V \, (p_{st})} \sum_{h=1}^{L} W_h \, P_h \, Q_h}$$

$$\text{donde } n_0 = \frac{\left[\sum_{h=1}^{L} W_h \sqrt{P_h \, Q_h} \right]^2}{V(p_{st})}$$