Jesús Amón
CATEDRÁTICO DE PSICOLOGÍA MATEMÁTICA
DE LA FACULTAD DE PSICOLOGÍA DE LA UNIVERSIDAD COMPLUTENSE DE MADRID

ESTADÍSTICA PARA PSICÓLOGOS I

Estadística descriptiva

EDICIONES PIRÁMIDE

COLECCIÓN "PSICOLOGÍA"

Diseño de cubierta: C. Carabina

Decimoquinta edición, 1993
Reimpresión, 2003

Juan Ignacio Luca de Tena, 15. 28027 Madrid
Teléfono: 91 393 89 89
www.edicionespiramide.es
Depósito legal: M. 12.380-2003
ISBN: 84-368-0081-8 (Obra completa)
ISBN: 84-368-0082-6 (Tomo I)
Printed in Spain
Impreso en Lavel, S. A.
Polígono Industrial Los Llanos. Gran Canaria, 12
Humanes de Madrid (Madrid)

PRÓLOGO

La Estadística es hoy un instrumento profusamente utilizado por investigadores de áreas científicas muy diversas. En particular, su necesidad e importancia han ido creciendo durante estos últimos años dentro de las Ciencias de la Conducta y, más concretamente, dentro de la Psicología. Para convencerse de ello basta con leer las publicaciones contemporáneas sobre Psicología experimental, Psicología del aprendizaje, Psicología educacional, Psicología social, Psicofísica, etc. Aun la misma Psicología clínica exige ya un dominio no pequeño de las técnicas estadísticas. El psicólogo, por tanto, debe conocer dichas técnicas con cierta seriedad. No basta con que sepa aplicar unas cuantas fórmulas de modo más o menos mecánico. Es necesario que conozca el fundamento y la deducción de las mismas, así como las condiciones exigidas por cada técnica estadística para que su utilización resulte válida. Sólo así podrá aplicar la más apropiada en cada caso concreto y, a la vez, será capaz de llevar a cabo su tarea específica de psicólogo que es ofrecer una interpretación psicológica adecuada de los resultados numéricos obtenidos en sus investigaciones. Los meros cálculos no son de su incumbencia, sino propios de las máquinas que, por cierto, operan con mayor rapidez y seguridad que el hombre en la realización de los cálculos numéricos.

De acuerdo con estos criterios, hemos procurado definir claramente los conceptos y fórmulas, exponiendo, además, los fundamentos en que se basan y las propiedades que de ellos se derivan. También hemos intentado hacer ver las relaciones de unos con otros, cómo conceptos aparentemente nuevos no son más que una mera deducción de otros previamente expuestos o, dicho de modo distinto, cómo conceptos dispares a primera vista no son más que traducciones diversas de un mismo concepto ya conocido. Creemos que este planteamiento no sólo economiza energía mental sino que, además, ayuda al lector a estructurar racionalmente los conocimientos que va adquiriendo.

Según lo que diremos en el capítulo 3, la Estadística quedará dividida en dos partes fundamentales: Estadística Descriptiva y Estadística Inferencial. Allí explicaremos qué significa cada una de ellas y cómo están relacionadas entre sí. Por ahora nos contentaremos con indicar que en este volumen trataremos únicamente de la Estadística Descriptiva. En un segundo volumen expondremos la Estadística Inferencial tras unos capítulos previos dedicados a la Probabilidad. Creemos

que el contenido de este primer volumen puede ser expuesto en un semestre con tres clases semanales teóricas y tres clases prácticas.

Nos mantendremos a un nivel matemático moderadamente elemental, aunque suficiente para poder legitimar de modo razonable las fórmulas que vayamos proponiendo. En cualquier caso, el nivel exigido en este volumen dedicado a la Estadística Descriptiva será inferior al requerido en el segundo dedicado a la Estadística Inferencial. Esperamos que los actuales y los antiguos alumnos de Psicología podrán superar este nivel, aunque dicha superación exija, quizá, de alguno de ellos, un pequeño esfuerzo. Asimismo, dada la importancia de las Matemáticas en los nuevos planes de estudios pre-universitarios, esperamos que, dentro de dos años, los alumnos que lleguen a la Universidad española, vengan con una preparación matemática bastante superior a la que poseen la mayoría de los actuales universitarios de Psicología. Con ello, es previsible que los futuros psicólogos españoles serán capaces de manejar los instrumentos estadísticos con la cautela de quien conoce de cerca sus limitaciones y, a la vez, con la eficacia de quien sabe sacar de ellos el máximo partido posible en la solución de los problemas psicológicos.

Hemos introducido numerosos ejemplos que ayuden a comprender mejor las fórmulas teóricas. Algunos de ellos están tomados de investigaciones llevadas a cabo realmente en campos psicológicos. Otros han sido creados artificialmente. Los primeros tienen la ventaja de motivar al lector en el estudio de la Estadística, viendo cómo las fórmulas que estudia tienen aplicación en la vida real, pero tienen el inconveniente de exigir cálculos laboriosos y de no estar preparados expresamente para aclarar la naturaleza y propiedades de las fórmulas matemáticas. Los segundos tienen la ventaja de exigir cálculos más sencillos y de estar elaborados directamente para que el lector conozca mejor las proposiciones teóricas que se le presentan, aunque carecen del aliciente motivacional ofrecido por los primeros.

Al fin de cada capítulo el lector encontrará resumidas las definiciones y fórmulas en él propuestas y una serie de ejercicios cuya solución se halla en un apéndice, al final del libro.

Tras unos capítulos introductorios, la obra queda dividida en tres partes: *a*) Estudio de una sola variable, *b*) Estudio conjunto de dos variables, y *c*) Estudio conjunto de tres variables. En los capítulos introductorios presentamos ciertas relaciones entre Matemáticas y Psicología, exponemos el concepto de medida y de escala de medida (con mayor detalle, tal vez, que el usual en libros de nuestro nivel) y concluimos ofreciendo la definición y división de la Estadística. En la primera parte, tras unas definiciones previas, proponemos el modo de organizar unos datos numéricos y de calcular unos índices que nos indiquen el valor medio o central de todos ellos, su dispersión o variabilidad y su asimetría y apuntamiento. En la segunda parte estudiamos la relación entre dos variables (proponiendo diversos índices de correlación) y el modo de pronosticar las puntuaciones en una de ellas, conocidas las puntuaciones en la otra. En la tercera parte estudiamos la correlación entre dos de las tres variables (eliminando el influjo de la tercera), la correlación de una de ellas con las otras dos tomadas conjuntamente, y el modo de pronosticar las puntuaciones en una de ellas, conocidas las puntuaciones en las otras dos.

El doctor Luis Jáñez ha entresacado ejemplos apropiados de la literatura psicológica, ha colaborado en la corrección de pruebas y, sobre todo, ha ayudado con sus oportunas e inteligentes observaciones a definir con mayor precisión algunos conceptos. En la búsqueda de ejemplos ha trabajado también Vicente Sierra. Vicente Ponsoda ha ayudado a que mejorara la redacción primitiva del texto, ha colaborado en la corrección de pruebas y, gracias a su implacable caza del error, ha librado al texto de no pocos errores. A todos ellos agradecemos su inapreciable colaboración. De modo especial agradecemos el apoyo del doctor Rafael San Martín que en su docencia universitaria ha venido usando las primeras redacciones de este libro. A su claro juicio y a sus reconocidas dotes pedagógicas se deben valiosas mejoras en la redacción definitiva de la obra hoy editada.

Nuestro agradecimiento, finalmente, a Ediciones Pirámide por su entusiasta colaboración en todo momento.

JESÚS AMÓN

ÍNDICE GENERAL

II. ESTUDIO DE UNA SOLA VARIABLE

III. ESTUDIO CONJUNTO DE DOS VARIABLES

V. APÉNDICES

Introducción

1

Matemáticas en Psicología

1.1. Lenguaje matemático en las ciencias

Por regla general, las ciencias se han mantenido a un nivel meramente cualitativo en su infancia y han ido ascendiendo a niveles superiores cuantitativos al ir alcanzando su edad adulta. Los investigadores se han esforzado en ir traduciendo a lenguaje matemático la formulación verbal primitiva de las ideas científicas y este esfuerzo ha resultado fecundo para la ciencia por un doble motivo. En primer lugar, el intento de expresar matemáticamente las hipótesis científicas ha obligado a los investigadores a clarificar más sus conceptos y a perfilar mejor esas hipótesis antes de plasmarlas definitivamente. En segundo lugar, toda hipótesis científica es susceptible de una comprobación experimental más satisfactoria cuando viene formulada matemáticamente que cuando viene propuesta en forma meramente verbal.

Son múltiples las citas aducibles según las cuales van unidos estrechamente el progreso de una ciencia y el grado de su desarrollo matemático. Sólo voy a presentar el testimonio de dos psicólogos. Para Stevens (1951, pág. 1), «la importancia de una ciencia es medida comúnmente por el grado según el cual hace uso de las matemáticas». A juicio de Atkinson et al., (1965, pág. 2), «es un hecho histórico familiar que a medida que la ciencia progresa, sus teorías se van haciendo más y más matemáticas en la forma».

Parece, pues, clara la tendencia general de las ciencias modernas a expresar sus conceptos matemáticamente.

1.2. Lenguaje matemático en Psicología

La Psicología moderna, casi desde sus comienzos, ha intentado proponer sus leyes (o, al menos, algunas) bajo fórmulas matemáticas. A este respecto puede ser consultado Miller (1964). Thurstone (1959, pág. 9) cree que la Psicología seguirá el camino recorrido por otras ciencias, haciéndose cada vez más y más ma-

temática a medida que vaya formulando más rigurosamente sus ideas fundamentales. Horst (1966, pág. 2) piensa que la Psicología ha tardado tanto tiempo en ocupar su lugar apropiado entre las ciencias aplicadas, quizá por no haber reconocido la importancia de la medida en la investigación psicológica. Según Nunnally (1967, pág. 6) todas las teorías psicológicas, con el tiempo, irán siendo propuestas en forma matemática. Bailey (1967, IX) admite que la Biomatemática ayudará a la Biología y a la Medicina, como la Física matemática ha ayudado a la Física. Existe, pues, un convencimiento casi general de que el lenguaje cuantitativo irá asumiendo cada vez mayor relieve en las ciencias de la conducta y, en particular, en la Psicología. De hecho, los modelos matemáticos juegan hoy un papel importante en muchas áreas psicológicas: Psicología experimental, diferencial, social, industrial, pedagógica, ...y aun clínica. La lectura de bastantes artículos y libros relacionados con estas disciplinas exige una preparación matemática no ligera y, en ocasiones, profunda. Desde luego, sin esta última es imposible leer revistas de especialización psico o bio-matemática («Psychometrika», «Biometrica», «Biometrics», «Journal of Mathematical Psychology», etc.) o entender libros, cada vez más numerosos, que abordan los problemas psicológicos de modo rigurosamente métrico. Igualmente, es necesaria una sólida preparación matemática para comprender diversas leyes o teorías psicológicas, tal como hoy son propuestas. Así, por ejemplo, leyes psicofísicas y psicométricas, teorías sobre el aprendizaje, teoría de la decisión y teoría de la información aplicadas a la Psicología, teoría de tests, etcétera. En conclusión, las matemáticas se van haciendo cada vez más indispensables dentro de las ciencias humanas y, en particular, dentro de la Psicología. Los múltiples libros de Matemáticas para psicólogos, sociólogos, biólogos, etc., que vienen apareciendo durante estos últimos años, son un índice claro de que estos investigadores necesitan y piden una fundamentación matemática cada vez más seria para poder abordar adecuadamente muchos problemas de sus correspondientes especialidades.

Sin negar, ni mucho menos, la importancia del lenguaje matemático en Psicología, conviene, no obstante, advertir que la tarea fundamental de un psicólogo es llegar a consecuencias psicológicas a partir de premisas, también, psicológicas. Los instrumentos matemáticos serán útiles en Psicología en cuanto nos ayuden a alcanzar conclusiones psicológicas. Sin duda alguna, es mucho más estimable una afirmación con contenido psicológico, expuesta en términos verbales, que una afirmación carente de significado psicológico, por más elegante que sea su formulación matemática y por más riguroso que sea el proceso matemático que nos ha llevado a ella. Por tanto, como psicólogos, deberemos abordar matemáticamente un problema psicológico siempre y sólo cuando esta táctica nos ayude a encontrar una solución psicológica apropiada. Si un problema psicológico no admite un enfoque matemático, no nos empeñaremos en introducir modelos matemáticos que, en el caso más favorable, serán perfectamente inútiles. Con todo, si un problema psicológico admite un enfoque matemático y otro no matemático, será muy interesante acometer el problema bajo los dos puntos de vista. Ambos enfoques, lejos de ser antitéticos, se complementarán mutuamente. El resultado conjunto de ambos puede ofrecernos una solución más rica que la que nos hubiera

ofrecido cada uno de ellos por separado. En algunas ocasiones, la discordancia entre los resultados de uno y otro enfoque puede sugerirnos un planteamiento nuevo mucho más acertado que el concebido previamente.

1.3. Matemáticas y complejidad de las manifestaciones psicológicas

De lo dicho se infiere que hoy las técnicas matemáticas son utilizadas en Psicología con gran profusión y no poco éxito. Por ello consideramos superfluo ponernos a refutar ciertas dificultades clásicas contra la aplicabilidad de los métodos matemáticos en Psicología. Sin embargo, queremos responder a una pregunta que, obviamente, pueden hacer muchos lectores. ¿Es posible acercarse a los problemas psicológicos, de innegable complejidad, con instrumentos de naturaleza simple y elemental?

Comencemos admitiendo que tanto en Psicología humana, como, sobre todo, en Psicología animal, existen procesos bastante rudimentarios encajables fácilmente dentro de esquemas matemáticos. En estos casos no sólo es posible, sino, también, muy útil la introducción de instrumentos matemáticos. No obstante, debemos aceptar que son escasos los fenómenos psicológicos de gran simplicidad y que, sobre todo, los fenómenos psíquicos de máximo interés son precisamente los de contextura más compleja. ¿Qué hacer? En primer lugar, podemos considerar aspectos parciales del fenómeno complejo, limitando nuestra investigación a ellos y, por supuesto, restringiendo, después, nuestras conclusiones a esos únicos aspectos. De esta manera podremos utilizar técnicas matemáticas relativamente simples. En segundo lugar, conviene advertir que existen modelos matemáticos muy complejos y a los que podemos acudir en muchos casos. Desde luego, su manejo exige conocimientos matemáticos no asequibles a todos los psicólogos. Es una dificultad innegable, pero no invencible. Es extrínseca y superable con una preparación matemática adecuada. En conclusión, parece posible acercarse a muchos problemas psicológicos con instrumentos matemáticos con tal que éstos sean suficientemente apropiados. Aclaremos lo dicho con un ejemplo.

Dos personas, Pablo y Santiago, consideran una frase sobre cierto problema social. Tienen que decir si están de acuerdo o en desacuerdo con la misma. Pues bien, la situación siguiente es muy posible. Pablo responde que está de acuerdo y Santiago que está en desacuerdo. Cuando Santiago escucha a Pablo razonar su respuesta afirmativa, nos dice: Yo también habría estado de acuerdo si hubiera considerado la frase desde el punto de vista bajo el cual él la ha considerado. A su vez, cuando Pablo escucha a Santiago razonar su respuesta negativa, nos dice: Yo también habría estado en desacuerdo si hubiera considerado la frase desde el punto de vista bajo el cual él la ha considerado. Sin embargo, es muy previsible que no preguntemos a ninguno de los dos la razón de sus respuestas y nos contentemos con atribuir a Pablo un 1 (por estar de acuerdo) y a Santiago un 0 (por estar en desacuerdo). Si se trata de un cuestionario compuesto de diversas

frases, nos limitaremos a atribuir a cada persona una puntuación igual al número de frases con las que ha estado de acuerdo. Este tratamiento matemático es muy asequible. De modo muy elemental dispondremos enseguida de una puntuación para cada persona. Pero, ¿es esto legítimo? Probablemente, no. En efecto, la táctica anterior supone implícitamente que a alta puntuación en el cuestionario corresponde estar de acuerdo con el tema social del mismo. Ahora bien, tal suposición sólo sería defendible si dicho tema social fuera unidimensional, admitiese un único enfoque, y éste hubiera sido asumido por todas las personas que responden al cuestionario. Más aún, deberíamos estar seguros de que la única dimensión considerada por ellas ha sido entendida tal como lo pretendía el que aplica el cuestionario. De no cumplirse estas condiciones, el número atribuido a cada persona es susceptible de múltiples y aun equívocas interpretaciones psicosociales. Esto quiere decir que el modelo matemático simple aplicado en esta ocasión no es el más apropiado, pero ello no significa que sea rechazable todo tratamiento matemático. Podemos valernos de un modelo que nos descubra las dimensiones fundamentales simples del tema social complejo y que nos permita determinar la situación de cada persona respecto a cada una de esas dimensiones o facetas. Este modelo matemático trataría de considerar los diversos aspectos considerados por las personas que responden al cuestionario, evaluando a las personas según cada uno de estos aspectos, sin contentarse con clasificarlas en las dos únicas categorías «a favor» o «en contra». Es posible que muchos temas sociales, simples y unidimensionales en apariencia, sean complejos y multidimensionales*. Pero ello no implica que sea inviable todo enfoque matemático. Sólo quiere decir que los problemas psicosociales multidimensionales deben ser afrontados con modelos multidimensionales. Admitimos ciertas dificultades de orden práctico en su aplicación, pero negamos la imposibilidad de acometer dichos problemas complejos con instrumentos matemáticos.

1.4. Comentario sobre los modelos matemáticos complejos en Psicología

a) Ante todo, queremos reafirmar que estos modelos complejos no sólo son posibles, sino que aún son los únicos realistas en muchos casos. Es erróneo creer que las técnicas matemáticas algo complicadas son construcciones puramente ideales, muy propias para discusiones teóricas, pero sin relación alguna con los problemas de la vida real. Con gran frecuencia los instrumentos sencillos y de fácil aplicación suelen ser menos realistas y menos útiles que los modelos matemáticos complejos a la hora de tomar decisiones psicológicas de importancia en situaciones prácticas complejas.

* Por poner un ejemplo, Amón (1969) comprobó cómo una faceta social tan aparentemente unidimensional como la religiosidad utilitaria, medida con un cuestionario sumamente purificado, se manifestó como pluridimensional, siendo, además, bastante independientes entre sí las dimensiones en las que se descomponía.

b) Es verdad que los modelos matemáticos complejos suelen ser los únicamente válidos en situaciones psicológicas complejas y que el dominio de tales modelos exige un nivel matemático algo más que mediano. Sin embargo, conviene hacer algunas puntualizaciones. En primer lugar, según ya hemos indicado, no todos los problemas psicológicos son extremadamente complejos. Más adelante nos encontraremos, de hecho, con situaciones bastante simples. Comprobaremos cómo en ellas son aplicables legítimamente técnicas matemáticas sencillas y cómo sus resultados numéricos son susceptibles de una interpretación psicológica muy satisfactoria. En segundo lugar, la aplicación de modelos matemáticos complejos no implica necesariamente una preparación matemática extremadamente especializada. Son deseables unos conocimientos matemáticos serios, pero no es necesario que el investigador psicomatemático sea especialista en áreas matemáticas concretas. Lo verdaderamente importante es que sepa acercarse con mentalidad matemática a los problemas que se le presenten. Es decir, que se esfuerce en asimilar el proceso lógico subyacente al razonamiento matemático, que logre captar la estructura formal del modelo matemático de que se trate, que conozca las condiciones que lo hacen posible y, consiguientemente, las condiciones que éste exige de la realidad concreta para que sea legítima su aplicación a la misma. Sólo así, sabrá elegir el modelo matemático más apropiado en cada situación práctica de la vida real. Por otra parte, asimilada esta mentalidad matemática, no sólo podrá manejar con suficiente pericia instrumentos matemáticos bastante sofisticados, sino que podrá entablar diálogo con los especialistas en Estadística matemática para que le asesoren en el planteamiento métrico de algún problema psicológico y con los especialistas de un Centro de Cálculo para que le ayuden en la elección o creación del programa más adecuado con el que pueda resolver su problema psicológico mediante un ordenador electrónico.

c) Los modelos matemáticos, aun los más complejos, son todavía bastante elementales respecto a la realidad psicológica concreta. Las relaciones previstas por el modelo se verifican exactamente en una realidad simplificada, ideal, y sólo aproximadamente en la realidad compleja, existente. El grado de aproximación será tanto mayor, cuanto menor haya sido el proceso simplificador que ha permitido crear el modelo. Cuanto más mutilemos la realidad concreta, más sencillo y más manejable será el modelo creado a partir de esa realidad cercenada, pero más problemática será su aplicación a la realidad concreta, no mutilada. Cuanto menos mutilemos la realidad concreta, más complicado será el modelo, pero más fiable será su aplicación a dicha realidad. Tendremos que llegar a un compromiso: simplificar la realidad concreta lo menos posible, pero, a la vez lo suficiente de modo que el modelo creado a partir de ella sea fácilmente manejable desde el punto de vista instrumental matemático. En Psicología los casos prácticos que se nos presentan en la vida real suelen ser más complejos que los que aparecen en otras ciencias de la Naturaleza. Por ello, si deseamos modelos sencillos, deberemos simplificar la realidad concreta de manera notable. Aun quedándonos con modelos algo complejos, deberemos imponer a la realidad mutilaciones de cierta importancia. Ello quiere decir que en Psicología debemos ser prudentemente cautos en la aplicación de los modelos a la realidad concreta y en la traduc-

ción a lenguaje psicológico de los resultados numéricos obtenidos mediante el modelo.

Conviene advertir que los investigadores psicomatemáticos con sólida preparación matemática suelen ser muy cautos en la interpretación psicológica de los resultados numéricos. La razón es que ellos conocen muy bien las limitaciones de los modelos matemáticos aun de aquellos que pueden aparecer bastante complejos. Son mucho más audaces en dichas interpretaciones psicológicas los que carecen de adecuada preparación matemática, precisamente por carecer de ella, por desconocer la limitación de los modelos que usan. Es necesaria una buena preparación matemática para conocer la potencia y la debilidad de las técnicas estadísticas y, consiguientemente, para saber usarlas con eficacia y, a la vez, con prudencia.

2

Medida en Psicología

2.1. Introducción

Hemos visto que es posible enfocar matemáticamente los problemas psicológicos. Este enfoque implica atribuir números a las manifestaciones psicológicas, someter estos números a ciertas técnicas matemáticas de modo que lleguemos a un resultado numérico final e interpretar psicológicamente este resultado. En realidad, el estadio estrictamente matemático es el segundo que comienza con datos numéricos y concluye con resultados, también, numéricos. Pero éste es inconcebible sin una previa atribución de números a las manifestaciones psicológicas. Por ello, vamos ahora a referirnos brevemente a dicha atribución numérica o, lo que es equivalente, a la definición de medida.

2.2. Características y modalidades

Los objetos manifiestan características según diversas modalidades. Así, por ejemplo, las personas manifiestan la característica «sexo» según dos modalidades: varón y mujer; la característica «religión» según muchas modalidades: católico, protestante, mahometano, budista, ateo, etc.; la característica «peso» según infinitas modalidades, pues entre dos modalidades, por próximas que se encuentren, son siempre posibles otras modalidades intermedias.

2.3. Definición de medida

Atribución de números a los objetos según ciertas reglas*.

* La definición anterior está tomada de Stevens. En su primera publicación sobre la medida (1946, pág. 2), decía: «Parafraseando a N.R. Campbell, ... , podemos decir que medida, en su sentido más amplio, es definida como la atribución de numerales a objetos o sucesos de acuerdo con reglas».

Tenemos por una parte números y, por otra, objetos con sus correspondientes modalidades. Entre los números existen ciertas relaciones que son válidas siempre dentro del mundo aritmético, ideal. Entre las modalidades existen ciertas relaciones (en unos casos, pocas y simples; en otros, bastantes y complejas) que son verificables en el mundo empírico, real. Pues bien, la atribución de números a los objetos no va a ser arbitraria, sino de acuerdo con esta regla general: aceptar sólo como relaciones válidas entre los números aquellas que sean verificables empíricamente entre las correspondientes modalidades. Esta regla general se concretará en unas u otras reglas particulares, según sea mayor o menor el número de relaciones verificables empíricamente.

Exigimos, por tanto, un cierto paralelismo o isomorfismo entre las relaciones aceptadas como válidas entre los números y las relaciones verificables entre las correspondientes modalidades (en nuestro caso, psicológicas). Consiguientemente, con los números atribuidos a las modalidades, sólo podremos realizar aquellas operaciones que estén de acuerdo con las relaciones aceptadas como válidas entre los mismos. Por otra parte, estas relaciones son precisamente las verificables entre las modalidades empíricas, psicológicas. En consecuencia, parece que el resultado numérico final, obtenido operando de este modo con los números atribuidos a las modalidades psicológicas, admitirá una interpretación psicológica razonable.

Ordinariamente, la expresión «atribución de números a las modalidades» será una simplificación de «atribución de números a los objetos según las modalidades bajo las cuales manifiestan cierta característica». Pasamos por alto las ventajas e inconvenientes que esta equivalencia lleva consigo. Tampoco nos detenemos a discutir si los números deben ser atribuidos a los objetos o a sus características, por creer dicha discusión ajena al fin de este texto. Únicamente, nos contentamos con proponer la anterior equivalencia, usando preferentemente la expresión primera por meras razones de sencillez.

2.4. Definición de escala de medida

La palabra «escala» es usada profusamente en la literatura psicomatemática, pero raramente definida. Ya Suppes y Zinnes se quejaban (1968, pág. 10) de que «es raro encontrar en la literatura sobre la medida una definición exacta de escalas». Ni el mismo Stevens ha sido lo suficientemente explícito. Suppes y Zinnes han presentado una definición clara de «escala». Algo parecido ocurre con Coombs, Dawes y Tverski (1970) y con Pfanzagl (1971).

Veamos qué entendemos aquí por escala de medida. Comenzaremos definiéndola en algunos casos particulares, para concluir ofreciendo una definición general.

Consideremos la característica peso. Aceptemos como unidad empírica de medida la modalidad presentada por un cuerpo elegido arbitrariamente. Mediante una balanza podemos elegir otro cuerpo que manifieste la característica peso bajo la misma modalidad que el primero (es decir, que pese lo mismo). Mediante estos dos cuerpos podemos elegir otro nuevo cuya modalidad sea igual a la suma de las modalidades de los dos cuerpos anteriores (es decir, que pese lo mismo que los dos primeros juntos). Para ello, basta con colocar estos dos cuerpos en un platillo de la balanza y en el otro un tercer cuerpo tal que la balanza quede equilibrada. Esta operación la podemos ir repitiendo, poniendo tres cuerpos de peso unidad en un platillo y en el otro un cuerpo tal que la balanza permanezca en equilibrio. Así, llegaremos a obtener un conjunto de cuerpos materiales, uno con la modalidad unitaria, otro con una modalidad dos veces mayor que la unitaria, etc. Podemos ahora atribuir números a estas modalidades (o a estos cuerpos manifestando la característica, según estas modalidades). Una atribución obvia (no la única, como luego veremos) es asignar el 1 a la primera modalidad, el 2 a la modalidad empírica doble, el 3 a la triple, etc. Esta atribución es obvia, en cuanto que la primera modalidad ha sido elegida arbitrariamente como modalidad empírica unitaria.

Pues bien, llamaremos escala de medida a este conjunto de modalidades empíricas distintas y de números distintos, puestos en correspondencia biunívoca (a cada modalidad le corresponde un solo número y a cada número una sola modalidad). O, lo que es igual, a este conjunto de cuerpos materiales, con pesos distintos, y de los números distintos atribuidos a dichos cuerpos. Mediante esta escala podemos atribuir números de modo coherente a otro cuerpo cualquiera, comparando su modalidad peso con las modalidades de la escala tipo acabada de construir y atribuyéndole el número de la escala asociado a la modalidad que coincide empíricamente con la modalidad del cuerpo en cuestión.

Veamos otro ejemplo algo más complejo. Se trata de la característica temperatura. Consideremos una vasija llena de agua y situada sobre una fuente calorífica. Introduzcamos dentro del agua un tubo estrecho de vidrio en uno de cuyos extremos lleva un pequeño depósito conteniendo mercurio. Al aumentar la temperatura el mercurio se dilata y va avanzando por el tubo de vidrio que supondremos en posición vertical y con el depósito de mercurio en su extremo inferior. En un momento arbitrario la vasija manifestará la característica temperatura según cierta modalidad determinada y la columna de mercurio habrá alcanzado una altura determinada. Señalemos una raya en el tubo de vidrio junto al nivel alcanzado por el mercurio en ese momento. Vayamos aumentando la temperatura del agua y en otro momento, también arbitrario, veamos hasta dónde ha ascendido la columna de mercurio marcando otra raya en el tubo de vidrio junto al nivel alcanzado por el mercurio en este segundo momento. Sigamos calentando el agua hasta el preciso momento en que la diferencia entre esta altura y la obtenida en el momento segundo sea igual que la diferencia entre las alturas alcanzadas en los momentos segundo y primero respectivamente y hagamos otra raya en el tubo de vidrio junto al nivel alcanzado por el mercurio en este tercer momento. Si verificamos una ope-

En artículos posteriores (1951, 1959, 1966, 1968a, 1975, por ejemplo) volverá a proponer idénticas o parecidas formulaciones sobre la medida. Ordinariamente usa la palabra «numeral», pero, también, usa la palabra «número» (véase, por ejemplo, 1968b, pág. 850). Según algunos autores, hay que diferenciar cuidadosamente los dos vocablos «numeral y «número». Así, entre otros, Lorge (1967, pág. 44), Sender (1958, pág. 50). Según ellos, «numeral» es un concepto más amplio que «número». Con todo, Stevens (1959, pág. 19) no parece dar mayor importancia a esta distinción.

ración semejante en repetidas ocasiones, el tubo de vidrio habrá quedado marcado con una sucesión de rayas equidistantes, correspondientes a las modalidades de la característica temperatura manifestadas por el agua de la vasija en el primer momento, en el segundo, en el tercero, etc. Podemos, ahora, atribuir números a esas rayas. A una cualquiera de ellas le atribuimos el valor 0. A cada una de las rayas superiores (correspondientes a modalidades más cálidas) les vamos atribuyendo los números 1, 2, 3, ... A su vez, a cada una de las rayas inferiores (correspondientes a modalidades más frías) les vamos atribuyendo los números -1, -2, -3, ...

Pues bien, llamaremos escala de medida a este conjunto de rayas y de números relacionados biunívocamente. Mediante esta escala podemos atribuir números de modo coherente a un cuerpo cualquiera. Para ello, le ponemos en contacto con el tubo de vidrio tipo (o termómetro), acabado de construir, y observamos hasta qué raya de éste, r_k, ha ascendido la columna de mercurio. Hecho esto, atribuiremos al cuerpo en cuestión el número correspondiente a r_k.

Por supuesto, la sucesión de números (..., -3, -2, -1, 0, 1, 2, 3, ...), aunque obvia, no es la única, según veremos más adelante. Además, es claro que ahora la medida es indirecta, en cuanto que lo que medimos directamente es la altura alcanzada por el mercurio e indirectamente la temperatura causante de esa altura. Antes, en cambio, medíamos directamente la característica peso.

Consideremos, finalmente, la característica provincia de origen. Empíricamente podemos distinguir las modalidades Soria y Segovia, por ejemplo. Tendremos tantas modalidades distintas como provincias españolas. Podemos atribuir a cada provincia un número distinto. Tendremos tantos números distintos como provincias distintas.

Pues bien, llamaremos escala de medida a este conjunto de modalidades (tantas como provincias españolas) y de números relacionados biunívocamente. Mediante esta escala, podemos atribuir números de modo coherente a una persona cualquiera. Le atribuiremos, sencillamente, el número de la escala correspondiente a la modalidad bajo la cual manifiesta esa persona su provincia de origen. Por supuesto existen infinitas sucesiones de números distintos que puedan ser atribuidos a las modalidades anteriores, según veremos luego.

En conclusión, podemos ofrecer la siguiente definición general de escala: conjunto de modalidades (distintas) y conjunto de números (distintos) relacionados biunívocamente. Es decir, a cada modalidad le corresponde un solo número y a cada número una sola modalidad.

La escala, así definida, es un instrumento natural de medida. Con ella podemos atribuir números a cualquier objeto. Para ello, basta con observar qué modalidad, m_e, de la escala coincide empíricamente con la modalidad del objeto en cuestión, atribuyendo a éste el número correspondiente a m_e en la escala.

La definición anterior de escala está de acuerdo con la idea que tenemos de uno de los instrumentos de medida más usuales, la regla. Una regla no es más que un conjunto de rayas equidistantes marcadas sobre un listón de madera (o de otro material) y un conjunto de números, en correspondencia biunívoca: a cada raya le corresponde un solo número y a cada número una sola raya.

Nótese que hablamos de modalidades y números relacionados biunívocamente. Ordinariamente, tendremos conjuntos de objetos tales que, dentro de cada conjunto, todos sus objetos no manifiesten la característica de que se trate bajo modalidades distintas. Es decir, tendremos más objetos que modalidades. En este caso, a cada objeto le corresponde un solo número; pero un mismo número puede corresponder a varios objetos (todos los que manifiesten la característica según la misma modalidad). En el caso general, tendremos n objetos y m modalidades. Si definimos la escala como conjunto de objetos y de números, deberemos distinguir dos casos. En el primero, $m = n$; habrá correspondencia biunívoca entre objetos y números (isomorfismo). En el segundo, $m < n$; no habrá correspondencia biunívoca, será sólo unívoca: a cada objeto un solo número (homomorfismo). Para evitar estas distinciones, hemos definido la escala como un conjunto de modalidades (distintas) y de números (distintos) relacionados, siempre, biunívocamente.

2.5. Tipos de escalas de medida

Al medir, o sea, al atribuir números a los objetos, decíamos que sólo aceptábamos como válidas entre los números aquellas relaciones que fueran verificables empíricamente entre las correspondientes modalidades. Ahora bien, estas relaciones son muy simples en algunos casos y complejas en otros. Por consiguiente, en unos casos sólo aceptaremos como válidas entre los números, relaciones muy sencillas; en cambio, en los otros daremos validez a relaciones aritméticas más complejas.

Pues bien, diremos que nos encontramos a bajo o alto nivel de medida, según sea menor o mayor la complejidad de las relaciones que podamos verificar empíricamente entre las modalidades. Esta terminología parece la más oportuna. Sin embargo, en vez de niveles se suele hablar preferentemente de escalas de medida. Nosotros hablaremos a veces de niveles de medida; pero, por ahora, aceptaremos la expresión «escalas de medida», dada su aceptación casi universal. Esta última expresión tiene, también, un sentido razonable. La escala es el conjunto de modalidades empíricas y de números relacionados biunívocamente. Pues bien, según que entre las modalidades de la escala sea verificable uno u otro tipo de relaciones, tendremos uno u otro tipo de escala. Naturalmente, dentro de cada uno de estos tipos sólo serán válidas entre los números aquellas relaciones que sean verificables entre las correspondientes modalidades.

Distinguiremos cuatro tipos de escalas, siguiendo el esquema tradicional propuesto ya por Stevens en su primera publicación sobre teoría de la medida (1946).

2.5.1. Escala nominal

Supongamos que, dadas dos o más modalidades, sólo podemos comprobar empíricamente si ellas son iguales o distintas. Consiguientemente, entre los números atribuidos a las mismas sólo aceptaremos como válida la relación igualdad-des-

igualdad. Si, por ejemplo, se trata de la característica «provincia de origen» y atribuimos el 1 a la modalidad Soria, el 2 a la modalidad Segovia y el 3 a la modalidad Avila, entre los números 1, 2 y 3 sólo aceptaremos como válida la relación igualdad-desigualdad. Es decir, el símbolo 1 será considerado como algo distinto de los símbolos 2 y 3, y éstos, también, como distintos entre sí, del mismo modo que son distintas entre sí las tres modalidades correspondientes; pero, el 2 no será considerado como mayor que el 1, ni el 3 como mayor que el 2 y el 1, del mismo modo que la modalidad Segovia no es una manifestación mayor de la característica «provincia de origen» que la modalidad Soria, sino, simplemente, una manifestación distinta. A este nivel, los números atribuidos son puros «nombres» (de aquí el apelativo nominal) que podían ser sustituidos por cualesquiera símbolos no numéricos: letras, colores, figuras geométricas, etc. Por consiguiente, los números no gozan aquí de ninguna de las propiedades aritméticas. No tiene ningún sentido aceptar a este nivel que $3 = 2 + 1$, pues ello implicaría que la modalidad Avila era el resultado de unir las modalidades Soria y Segovia. En otras palabras, que el resultado de unir una persona con la modalidad Soria y otra persona con la modalidad Segovia daría como resultado una nueva persona con la modalidad Avila.

Evidentemente, la escala nominal permanece invariante frente a cualquier transformación que a números distintos haga corresponder números distintos. Es decir, seguiremos teniendo una misma escala nominal, cuando, permaneciendo las mismas modalidades tipo, los números atribuidos a las mismas sean transformados en otros con la única limitación de que a números distintos primitivos, correspondan, también, números distintos nuevos. Si, por ejemplo, a tres modalidades distintas les hemos atribuido los números 5, 7, 10, podemos atribuirles con igual derecho cualesquiera otra terna compuesta de números distintos como 6, 9, 2; 7, 3, 1; etc.

2.5.2. Escala ordinal

Supongamos que, dadas dos o más modalidades, no sólo podemos comprobar si son iguales o distintas, sino, siendo distintas, cuál de cada dos es la mayor. Es decir, dados dos objetos, podemos comprobar empíricamente si ambos manifiestan una característica según la misma o distinta modalidad y, supuesto que la manifiesten según distinta, podemos comprobar cuál de los dos la manifiesta según una modalidad mayor. Consiguientemente, entre los números atribuidos a las modalidades admitiremos como válidas las relaciones igualdad-desigualdad y orden.

Si, por ejemplo, se trata de la característica dureza, diremos que A es más duro que B si A raya a B y no es rayado por éste, al frotarlos entre sí. Es una definición operativa de dureza, es decir, verificable empíricamente. Vamos a construir una escala de dureza, o sea, una escala para medir la característica dureza. Elijamos diez cuerpos de naturaleza física distinta y ordenémoslos empíricamente de acuerdo con la definición anterior. Pongamos en primer lugar aquel cuerpo que es rayado por todos y no raya a ninguno. Pongamos en segundo lugar aquel que raya al primero (sin ser rayado por éste) y no raya a ninguno de los restantes (siendo rayado por ellos). Sigamos así hasta poner en último lugar aquel que raya a los nueve restantes y no es rayado por ninguno de ellos.

Tenemos, por tanto, diez modalidades de la característica dureza ordenadas empíricamente desde la más blanda hasta la más dura. Atribuyamos a la primera modalidad el número 1, a la segunda el número 2, . . . , a la última el número 10. Ahora aceptaremos no sólo que los números 1, 2, . . . , 10 son símbolos distintos, sino que, además, el 2 es mayor que el 1, el 3 es mayor que el 2 y el 1, . . . , el 10 es mayor que los nueve primeros números enteros positivos. Pero, a este nivel no tiene sentido admitir como válida entre los números una igualdad del tipo $7 - 6 = 3 - 2$, pues no podemos comprobar empíricamente si la diferencia de dureza entre la modalidad a la que he atribuido el 7 y la modalidad a la que he atribuido el 6 es igual que la diferencia de dureza entre la modalidad a la que he atribuido el 3 y la modalidad a la que he atribuido el 2.

Evidentemente, la escala ordinal permanece invariante frente a cualquier transformación monótona creciente. En otras palabras, seguiremos teniendo una misma escala ordinal, cuando, permaneciendo las mismas modalidades tipo, los números atribuidos a las mismas sean sometidos a una transformación monótona creciente, es decir, que haga corresponder a una sucesión ordenada de números otra sucesión de números ordenados del mismo modo que los primeros. Si, por ejemplo, a tres modalidades ordenadas de menor a mayor les hemos atribuido los números 3, 5 y 8, podemos atribuirles, con igual derecho, otras ternas del tipo: 6, 70, 95; 82, 195, 981; etc.

2.5.3. Escala de intervalos

Supongamos que, dadas dos o más modalidades, no sólo podemos comprobar empíricamente la igualdad-desigualdad y el orden, sino que, también, podemos establecer una unidad empírica de medida y observar cuántas veces se encuentra contenida dentro de la diferencia entre dos modalidades cualesquiera. Consiguientemente, dadas tres modalidades a, b y c, podemos comprobar empíricamente cuántas veces la diferencia entre a y b es mayor (o menor) que la diferencia entre b y c, suponiendo que las dos diferencias son distintas. Consiguientemente, entre los números atribuidos a las modalidades admitiremos como válidas las relaciones igualdad-desigualdad y orden, y las operaciones suma y resta entre ellos. Además, podemos admitir como válidas la multiplicación y división entre las diferencias obtenidas a partir de dichos números (no entre los mismos números).

Supongamos que se trata de la característica temperatura. Elegimos tres cuerpos y los ponemos en contacto con un tubo de vidrio en uno de cuyos extremos lleva un pequeño depósito con mercurio. Observamos ahora el nivel alcanzado por el mercurio en cada uno de los tres casos. Tendremos tres niveles termométricos n_1, n_2, n_3. Elegimos arbitrariamente una unidad empírica de medida, es decir, una distancia arbitraria sobre el tubo de vidrio. Por sencillez, supongamos que las diferencias entre n_1, n_2 y n_3 contienen esa unidad un número entero de veces. Para ser más concretos, supongamos que la unidad empírica queda comprendida dos

veces entre n_1 y n_2, y ocho veces entre n_2 y n_3, es decir, que la diferencia entre n_2 y n_3 es cuatro veces mayor que la diferencia entre n_1 y n_2. (No se olvide que todas estas relaciones las estamos constatando empíricamente.)

Pues bien, una atribución obvia de números a las tres modalidades anteriores puede ser: 4, 6, 14. Entre estos tres números aceptaremos como válidas las relaciones: $4 \neq 6 \neq 14$, $4 < 6 < 14$, $14 - 6 = 4(6 - 4)$.

Evidentemente, la escala de intervalos permanece invariante frente a cualquier transformación de la forma $y = ax + b$, donde a y b son dos constantes arbitrarias. En otras palabras, seguiremos teniendo una misma escala de intervalos, cuando, permaneciendo las mismas modalidades tipo, los números atribuidos a ellas son sometidos a una transformación de la forma $y = ax + b$. Esto es debido a que son arbitrarios tanto el origen, como la unidad de medida. Por consiguiente, tan legítima es la terna primitiva 4, 6, 14, como la terna $(4a + b)$, $(6a + b)$, $(14a + b)$, donde a y b son dos constantes arbitrarias. Así, por ejemplo, para $a = 2$ y $b = -5$, tendremos la terna 3, 7, 23. Para esta nueva terna siguen siendo válidas las tres relaciones que fueron válidas para la terna primitiva, a saber, $3 \neq 7 \neq 23$, $3 < 7 < 23$, $23 - 7 = 4(7 - 3)$.

Nótese que la introducción de la unidad empírica de medida legitima la suma y resta entre los números atribuidos a las modalidades y la multiplicación y la división entre las diferencias obtenidas a partir de dichos números. Pero no legitima la multiplicación y división entre los números mismos. La legitimidad de estas últimas operaciones sólo es posible cuando contemos con un origen empírico absoluto y no con un origen meramente arbitrario. Ahora bien, esta arbitrariedad en el origen es propia de las escalas de intervalos. Así, por ejemplo, el origen empírico de temperaturas en la graduación centígrada no corresponde a la temperatura nula, a la carencia total de calor, es decir, no es absoluto. Ese origen corresponde a la modalidad o grado de temperatura a la cual se funde el hielo, es decir, es arbitrario. Con el mismo derecho podríamos haber elegido como modalidad origen cualquier otra temperatura inferior o superior a la que se funde el hielo. Así se hace, por ejemplo, en las escalas Fahrenheit y Reaumur.

2.5.4. Escala de razón

Supongamos que, dadas dos o más modalidades, no sólo podemos comprobar empíricamente la igualdad-desigualdad, el orden y cuántas veces la diferencia entre dos modalidades es mayor que la diferencia entre otras dos, sino, además, cuántas veces una modalidad es mayor que la otra. Por tanto, entre dos números atribuidos a las modalidades admitiremos como válidas las relaciones igualdad-desigualdad y orden, y las operaciones suma, resta, multiplicación y división.

Supongamos que se trata de la característica longitud. Elegimos tres varillas metálicas que manifiesten la característica longitud según tres modalidades distintas. (Basta con escoger tres varillas tales que, al compararlas simultáneamente, coincidan las tres en uno de sus extremos y difieran en el otro.) Determinamos arbitrariamente una unidad empírica de medida, es decir, un trocito de varilla

arbitrario que llamaremos v_u. Aplicamos v_u a las tres varillas A, B y C y contamos el número de veces que v_u cabe en A, en B y en C. Supongamos que v_u cabe tres veces en A, seis veces en B y 24 veces en C. Tenemos, pues, tres objetos manifestando la característica longitud según tres modalidades distintas y equivalentes respectivamente a tres, seis y veinticuatro veces la modalidad unitaria.

Una atribución obvia de números a las tres modalidades anteriores puede ser: 3, 6, 24. Entre estos tres números son válidas las relaciones siguientes: $3 \neq 6 \neq 24$, $3 < 6 < 24$, $24 - 6 = 6(6 - 3)$, y, además, $6/3 = 2$, $24/3 = 8$, $24/6 = 4$, dado que empíricamente podemos comprobar que la modalidad «longitud» de B es doble de la de A, la de C es ocho veces la de A y la de C es cuatro veces la de B.

Evidentemente, la escala de razón es invariante frente a cualquier transformación de la forma $y = ax$, donde a es una constante arbitraria. En otras palabras, seguiremos teniendo la misma escala de razón, cuando, permaneciendo, las mismas modalidades tipo, los números atribuidos a ellas son sometidos a una transformación de la forma $y = ax$. Esto es debido a ser arbitraria la unidad de medida, pero no el origen. Ahora el origen empírico corresponde siempre a la carencia total de la característica, a la modalidad nula. No lo podemos elegir arbitrariamente donde nos parezca.

Por consiguiente, tan legítima es la terna primitiva: 3, 6, 24, como la terna: $3a$, $6a$, $24a$, donde a es una constante arbitraria. Así, por ejemplo, para $a = 5$, tendremos la terna: 15, 30, 120. Para esta nueva terna siguen siendo válidas las mismas relaciones que fueron válidas para la terna primitiva: $15 \neq 30 \neq 120$, $15 < 30 < 120$, $120 - 30 = 6(30 - 15)$, $30/15 = 2$, $120/15 = 8$, $120/30 = 4$.

2.6. Comentario sobre las escalas de medida

En primer lugar, la aceptación de cuatro tipos de escalas es tan arbitraria como lo hubiera sido aceptar dos, veinticinco u otro número cualquiera. El mismo Stevens, propugnador de las cuatro escalas, acepta la posibilidad de esquemas no cuatripartitos. «Las anteriores escalas representan los cuatro tipos de uso común. Otros tipos son posibles» (1968b, pág. 850). De hecho, otros autores han aceptado otros tipos de escalas o han aceptado el cuatripartito de Stevens, pero modificado.

En segundo lugar, conviene distinguir entre medida y estadística. Esta es una ciencia matemática que comienza y concluye con números, sin atender al origen extramatemático de los mismos y sin pretender interpretaciones ultramatemáticas (psicológicas, por ejemplo). La medida, en cambio, como atribución de números a modalidades empíricas (psicológicas, en particular), es el eslabón que une las modalidades empíricas con los números y, gracias a él, podemos interpretar empíricamente (psicológicamente, en particular) los resultados numéricos finales que nos ofrece la estadística. Los puros estadísticos prescinden de las escalas de medida. Así, por ejemplo, Savage (1957, pág. 331). Los psicólogos, en cambio, no pueden adoptar una postura asépticamente matemática. Necesitan recoger datos psicológicos, atribuirles números, operar con estos números e interpretar psicológicamente los resultados

finales. Si los números con los que empezamos una investigación estadística están desligados de las realidades psicológicas, no será fácil dar una interpretación psicológica al resultado obtenido a partir de aquellos números iniciales.

Está claro, pues, que como psicólogos no podemos prescindir de la interconexión entre modalidades psicológicas y números, ni del paralelismo entre la contextura relacional de las modalidades psicológicas y la contextura relacional de los números. Sin embargo, tampoco debemos exagerar la necesidad de este paralelismo. En efecto, si lleváramos a sus últimas consecuencias este isomorfismo entre modalidades psicológicas y números, nos sería imposible aplicar muchas de las técnicas estadísticas que muy frecuente y razonablemente usamos en Psicología. Supongamos, por ejemplo, que tres personas *A*, *B* y *C* hacen un examen de Geografía. Aunque con las naturales reservas, podemos aceptar que la lectura imparcial de los tres exámenes nos permitirá ordenar empíricamente a las tres personas según las modalidades bajo las cuales manifiesta cada una de ellas su ciencia geográfica. Pero, ¿nos atreveríamos a determinar empíricamente cuántas veces la diferencia de ciencia geográfica entre *A* y *B* es mayor o menor que la diferencia de ciencia geográfica entre *B* y *C*? Esta determinación no parece muy viable. Ahora bien, sólo bajo esta condición los números atribuidos a las tres modalidades pueden ser considerados como auténticos números, es decir, susceptibles de ser sometidos a operaciones aritméticas como la suma y la resta. En otras palabras, la atribución de un 10, un 9 y un 7 (con propiedades de números estrictamente dichos) a las tres modalidades, sólo sería legítima si pudiéramos comprobar empíricamente que la diferencia entre la ciencia de *A* y la de *B* era la mitad que la diferencia entre la ciencia de *B* y la de *C*. Pero, como ya hemos dicho, esta comprobación es muy difícil en una gran mayoría de los casos que se presentan en Psicología.

¿Qué decisión tomar? Llegar a un compromiso razonable guiados por el sentido común. Usar aquellas técnicas estadísticas que creamos más apropiadas en cada caso, en cuanto usándolas esperamos llegar, como psicólogos, a conclusiones psicológicas razonables. La experiencia parece confirmar que podemos llegar a resultados numéricos interpretables psicológicamente, aunque las técnicas estadísticas utilizadas no hayan sido las más oportunas, teniendo en cuenta el nivel o escala de medida requeridos por los datos. El mismo Stevens (1968) reconoce este hecho. Por su parte, Amón (1968) pudo comprobar cómo eran susceptibles de una misma interpretación psicológica los resultados obtenidos a partir de unos datos, valiéndose de unas técnicas que implicaban mero nivel nominal, de otras que implicaban nivel ordinal y, finalmente, de otras que requerían nivel de intervalos.

De todo lo dicho se desprende que no aceptamos la postura de textos como Sender (1958), Freeman (1968), el mismo Siegel (1958), etc., que encuadran algo rígidamente las técnicas estadísticas dentro del esquema cuatripartito de Stevens. Si con alguna frecuencia acudimos nosotros, también, a dicho esquema, será con gran flexibilidad y valiéndonos de él con fines preferentemente didácticos. En ningún caso condenaremos a los que, por ejemplo, calculen la media aritmética de las puntuaciones dadas por un profesor a sus alumnos en un examen o use otras técnicas estadísticas sin encontrarse a su estricto nivel de medida. Conviene advertir, además, que muchos de los que anatematizan a quienes usan instrumentos estadísticos sin encontrarse al nivel estricto de medida requerido por estos últimos, luego los utilizan ellos mismos sin graves escrúpulos en circunstancias idénticas o muy parecidas.

2.7. Resumen: Definiciones

Supuesto que los objetos manifestaban ciertas características según diversas modalidades, proponíamos las siguientes definiciones:

Medida: atribución de números a los objetos según ciertas reglas. Estas reglas se resumen en la siguiente: aceptar sólo como válidas entre los números aquellas relaciones que sean verificables empíricamente entre las correspondientes modalidades.

Escala de medida: conjunto de modalidades (distintas) y de números (distintos) relacionados biunívocamente. Es decir, a cada modalidad le corresponde un solo número y a cada número una sola modalidad. Tendremos uno u otro tipo de escala, según que sean verificables empíricamente más o menos relaciones entre las modalidades que forman parte de la escala. De acuerdo con este criterio, hemos distinguido cuatro tipos de escalas.

a) *Nominal*: sólo es verificable empíricamente la igualdad-desigualdad.

b) *Ordinal*: son verificables empíricamente igualdad-desigualdad y orden.

c) *De intervalos*: son verificables empíricamente igualdad-desigualdad y orden. Podemos, además, comprobar cuántas veces queda contenida una unidad empírica, elegida arbitrariamente, dentro de la diferencia entre dos modalidades.

d) *De razón*: son verificables empíricamente igualdad-desigualdad y orden. Además de poder comprobar empíricamente cuántas veces queda contenida una unidad empírica, elegida arbitrariamente, dentro de la diferencia entre dos modalidades, podemos, también, comprobar cuántas veces una modalidad cualquiera contiene dicha unidad empírica.

3

¿Qué es la Estadística?

3.1. Conceptos previos

3.1.1. Población

Conjunto de objetos (realmente existentes o posibles) que verifican una definición bien determinada. Por objeto entendemos cualquier persona, animal, cosa, operación, familia, institución, etc. Así, por ejemplo, constituirán una población los universitarios españoles, las familias europeas, los coches fabricados al año por cierta empresa automovilística, los posibles lanzamientos de un dado.

3.1.2. Muestra

Cualquier subconjunto de una población. La muestra hace siempre referencia a una población de la cual es parte. Así, por ejemplo, constituirán una muestra de las anteriores poblaciones: 300 universitarios españoles, 1.200 familias europeas, 213 coches, 80 lanzamientos de un dado.

Supongamos que observamos una característica de los objetos de una población. Por ejemplo, consideremos la altura de los universitarios españoles (población). Tendremos una población de observaciones y una población de números. Paralelamente, observando la altura de una muestra de universitarios españoles, tendremos una muestra de observaciones y una muestra de números.

Es claro que, dada una misma población de objetos, podemos tener diversas poblaciones de observaciones y, consiguientemente, diversas poblaciones de números, según que estudiemos una u otra característica. Así, con los mismos universitarios españoles, podíamos haber considerado su peso, su capacidad intelectual, su actitud frente a la guerra, etc.

3.1.3. Parámetro

Toda función definida sobre los valores numéricos de una población. Así, por ejemplo, será parámetro la media aritmética de las alturas de todos los universitarios españoles, pues dicha media aritmética no es más que la suma de las alturas de la población de universitarios españoles dividida por el número de éstos, es decir, es una función definida sobre los valores numéricos de la población.

Conviene distinguir entre la función, como tal, y el resultado numérico obtenido mediante la misma en cada caso concreto. La función es idéntica en todos los casos particulares. Por el contrario, el resultado numérico varía, en general, de caso a caso.

3.1.4. Estadístico

Toda función definida sobre los valores numéricos de una muestra. Así, por ejemplo, será estadístico la media aritmética de las alturas de una muestra de 300 universitarios españoles.

Supongamos una población constituida por diez personas a quienes hemos aplicado una prueba objetiva. Tenemos diez observaciones y diez valores numéricos. Supongamos que éstos son: 3, 12, 14, 8, 7, 7, 3, 10, 6, 8. Las puntuaciones 8 y 10, por ejemplo, constituirán una muestra de esa población. Será un parámetro la media aritmética de las diez puntuaciones (población), es decir, $(3 + 12 + 14 + 8 + 7 + 7 + 3 + 10 + 6 + 8)/10 = 7{,}8$. Será un estadístico la media aritmética de las dos puntuaciones 8 y 10 (muestra), es decir $(8 + 10)/2 = 9$.

Consideremos ahora el sexo de las diez personas. Atribuyamos un 1 a los varones y un 0 a las mujeres. Tenemos, también, diez observaciones y diez valores numéricos. Supongamos que éstos son: 0, 1, 1, 0, 1, 0, 1, 1, 1, 1. Ellos constituyen la población. Los tres primeros, por ejemplo, constituirán una muestra de esa población. Será un parámetro la proporción de «unos» (varones) en la población de las diez personas, es decir, $7/10 = 0{,}70$. Será un estadístico la proporción de «unos» (varones) en la muestra de las tres primeras personas, es decir, $2/3 = 0{,}67$.

3.2. Definición de Estadística

Ciencia que recoge, ordena y analiza los datos de una muestra, extraída de cierta población, y que, a partir de esa muestra, valiéndose del Cálculo de Probabilidades, se encarga de hacer inferencias acerca de la población.

Ordinariamente, las inferencias versarán sobre los parámetros de la población a partir de los estadísticos de la muestra. Pero, también, haremos inferencias acerca de la forma de la distribución* de la población, a partir de la forma de la distribución de la muestra. En cualquier caso las inferencias estarán basadas únicamente en la información objetiva contenida en la muestra. La información será exclusivamente objetiva, no subjetiva; contenida en la muestra y no en otras fuentes extrañas a la misma. Esta postura es la llamada «clásica» cuyo exponente máximo ha sido Ronald Aymer Fisher (1890-1962). Nos limitamos a este punto de vista

* En el capítulo 4 trataremos sobre distribución de frecuencias.

«clásico» por una doble razón. En primer lugar, sólo con una base sólida en estadística clásica es posible acceder a otros puntos de vista como el bayesiano o el de la teoría de la decisión. En segundo lugar, ni la estadística enfocada bayesianamente, ni la teoría de la decisión pueden presentar hoy un cuerpo de doctrina tan estructurado como el que presenta la estadística clásica. Además, el enfoque clásico hoy por hoy es mucho más útil en la aplicación a los casos prácticos psicológicos que los otros dos enfoques.

Conviene distinguir entre Estadística, estadísticas y estadístico (o estadísticos).

a) Estadística es la ciencia acabada de definir.

b) Estadísticas son los resultados numéricos obtenidos mediante la Estadística: número de accidentes de tráfico durante un mes, proporción de alcohólicos en diversas naciones, consumo medio semanal de leche por familia, etc.

c) Estadístico es todo valor numérico obtenido a partir de los valores presentados por una muestra, según lo dicho anteriormente.

Por supuesto, estadístico como sustantivo es, también, usado para denominar a la persona dedicada a la Estadística. Como adjetivo es utilizado para calificar personas y cosas relacionadas con la Estadística.

3.3. División de la Estadística

Según la definición acabada de dar en el párrafo anterior, la Estadística consta de dos partes fundamentales:

a) Recogida, ordenación y análisis de los datos de una muestra.

b) Verificación de inferencias acerca de la población (de sus parámetros, de la forma de su distribución), a partir de la muestra (de sus estadísticos, de la forma de su distribución).

La Probabilidad es el puente que nos permite pasar válidamente de la muestra a la población, que legitima el salto desde las características (conocidas) de la muestra hasta las características (desconocidas) de la población.

La primera parte constituye la Estadística Descriptiva, cuyo cometido es describir una muestra. La segunda parte constituye la Estadística Inferencial, cuyo cometido es hacer inferencias sobre la población, a partir de la muestra. En este primer tomo nos limitaremos a la Estadística Descriptiva. En el segundo tomo, tras unos capítulos sobre Probabilidad, estudiaremos la Estadística Inferencial. Allí discutiremos la diferencia entre Probabilidad y Estadística y veremos el papel que juega la Probabilidad en la fundamentación de las inferencias estadísticas.

3.4. Tareas de la Estadística Descriptiva

3.4.1. Recogida de datos

Posponemos la consideración de este apartado, dejándola para el tomo II. Es muy difícil hablar de recogida de datos sin haber tratado sobre el muestreo.

A su vez, es prácticamente imposible presentar el muestreo sin haber expuesto algunas nociones previas de Probabilidad.

3.4.2. Organización de los datos

Supongamos una muestra de 300 niños a quienes aplicamos una prueba de inteligencia. Estos niños manifestarán dicha característica según diversas modalidades. Si la prueba consta de 20 preguntas, un conjunto posible de modalidades sería: «ninguna pregunta bien respondida», «una pregunta bien respondida», . . . , «veinte preguntas bien respondidas». Atribuyamos números a esas modalidades. Una atribución razonable (no la única) puede ser la siguiente: un 0 a la primera modalidad, un 1 a la segunda, . . . , un 20 a la vigésimo primera. (Recuérdese que esta atribución de números, en rigor, no es propia de la Estadística que comienza a actuar sobre unos números ya atribuidos previamente.)

Tenemos, por tanto, 300 números (varios de ellos necesariamente iguales entre sí) correspondientes a los 300 niños. Este conjunto desordenado de números nos ofrece una información muy pobre sobre la inteligencia de la muestra. Una ordenación razonable consiste en colocar los 21 números posibles (0, 1, 2, . . . , 20) de menor a mayor. Pues bien, al número de ceros le llamaremos frecuencia correspondiente a la primera modalidad (mejor aún, correspondiente al 0, atribuido a la primera modalidad), al número de unos le llamaremos frecuencia correspondiente a la segunda modalidad (mejor aún, correspondiente al 1, atribuido a la segunda modalidad), . . . , al número de veintes le llamaremos frecuencia correspondiente a la vigésimo primera modalidad (mejor aún, correspondiente al 20, atribuido a la vigésimo primera modalidad). Por fin, llamaremos distribución de frecuencias al conjunto de todas las modalidades (mejor aún, de todos los números atribuidos a dichas modalidades) y de sus correspondientes frecuencias. De esta manera obtendremos una información más clara sobre la inteligencia de la muestra.

Es posible que sean muchas veintiuna modalidades. Por esta razón, podíamos reducirlas, por ejemplo, a siete clases, cada una de ellas con tres modalidades. La primera compuesta de las modalidades: «ninguna pregunta bien respondida», «una pregunta bien respondida», «dos preguntas bien respondidas». Y así, sucesivamente. En este caso la distribución de frecuencias quedaría constituida por el conjunto de las siete clases (mejor aún, por el conjunto de los siete números atribuidos a las siete clases) y de sus correspondientes frecuencias. Por regla general, este agrupamiento en pocas clases nos ofrece una información más asequible que la ofrecida por las 21 modalidades.

En el caso de características no cuantificables, por ejemplo, «provincia de origen», podemos seguir haciendo algo análogo a lo acabado de realizar. Supongamos una muestra de 1.000 españoles que manifiestan la característica «provincia de origen» según diversas modalidades. En principio el número de modalidades posibles es 50. Podemos atribuir los números 1, 2, . . . , 50 a estas modalidades. Tendremos 1.000 números (varios de ellos necesariamente iguales entre sí), corres-

pondientes a los 1.000 españoles. Podemos seguir un esquema análogo al anterior hasta llegar a una distribución de frecuencias. De igual modo podemos reducir las 50 modalidades (provincias) a un número menor de clases (por ejemplo, haciendo de cada región una clase). Es claro que, bajo estas condiciones, las diversas clases no estarán constituidas por el mismo número de modalidades.

3.4.3. Análisis de los datos

a) *Supuesta una sola característica*

En el caso de características cuantificables nos será muy útil obtener un solo número, promedio de todos los números de la muestra, y que, como tal, los represente a todos ellos y nos indique su posición. También, nos será útil obtener un valor numérico que nos diga si los números de la muestra se encuentran muy próximos entre sí (y respecto del promedio de todos ellos) o muy distantes o dispersos unos de otros. En resumen, calcularemos estadísticos de tendencia central o de posición y estadísticos de variabilidad o de dispersión.

b) *Supuestas dos o más características*

Comenzaremos estudiando conjuntamente dos características. Así, por ejemplo, podemos considerar la inteligencia espacial y la habilidad mecánica de 368 adultos. Tendremos 368 pares de números (cada persona tiene dos puntuaciones, una en inteligencia espacial y otra en habilidad mecánica). A partir de ellos construiremos diversos índices que nos manifiesten el grado de relación existente entre esas dos características y que nos permitan pronosticar, del mejor modo posible, la puntuación de una persona en una de las dos características, conociendo la que ha obtenido en la otra. Veremos más adelante cómo nos será posible elaborar índices de correlación e instrumentos de pronóstico, tanto en el caso de variables cuantificables como en el caso de no cuantificables, aunque sean distintos los modos de alcanzarlos en uno y otro caso.

Estudiado el caso más simple de solas dos variables, trataremos el caso de tres variables.

3.5. Resumen: Definiciones

Población: conjunto de objetos (actuales o posibles) que verifican una definición bien determinada.

Muestra: cualquier subconjunto de una población.

Parámetro: toda función definida sobre los valores numéricos de una población.

Estadístico: toda función definida sobre los valores numéricos de una muestra.

Estadística: ciencia que recoge, ordena y analiza los datos de una muestra, extraída de cierta población, y que, a partir de esa muestra, valiéndose del Cálculo de Probabilidades, se encarga de hacer inferencias acerca de la población.

Estadística Descriptiva: parte de la Estadística que se limita a recoger, ordenar y analizar los datos de una muestra. Es decir, se limita a *describir* la muestra.

Estadística Inferencial: parte de la Estadística que se encarga de hacer *inferencias* acerca de la población a partir de una muestra extraída de la misma.

II

Estudio de una sola variable

4

Organización de datos

4.1. Definiciones previas

4.1.1. Constante

Característica que sólo puede manifestarse bajo una única modalidad. Por ejemplo, la longitud de todas las circunferencias con el mismo radio.

4.1.2. Variable

Característica que puede manifestarse según dos o más modalidades. Por ejemplo, el peso, la inteligencia, la edad, la agudeza visual, etc. Cuando una característica, en sí misma variable, sólo puede manifestarse bajo una modalidad, será considerada como constante. Por ejemplo, si estudiamos la extroversión en un grupo de varones, diremos que la característica sexo se mantiene constante en dicho grupo.

Variable cualitativa

Característica que sólo puede ser considerada a nivel meramente nominal: sexo, profesión, lugar de origen, etc. Los números atribuidos a sus modalidades solamente gozan de la relación igualdad-desigualdad.

Variable cuasi cuantitativa

Característica que puede ser considerada, como máximo, a nivel ordinal: dureza de los cuerpos, responsabilidad de un grupo de operarios estimada por su capataz, etc. Los números atribuidos a sus modalidades sólo gozan de las relaciones igualdad-desigualdad y orden.

Variable cuantitativa

Característica que puede ser considerada, al menos, a nivel de intervalos: peso, inteligencia, fuerza física, número de hijos, etc. Con los números atribuidos a las mismas podemos realizar operaciones aritméticas.

Variable cuantitativa discreta

Característica que no admite siempre una modalidad intermedia entre dos cualesquiera de sus modalidades: número de hijos, número de coches vendidos al año, número de caras al lanzar diez monedas al aire, etc. Una familia puede tener, por ejemplo, cuatro o cinco hijos, pero no cuatro y medio. Esta modalidad es imposible.

Variable cuantitativa continua

Característica que admite siempre una modalidad intermedia entre dos cualesquiera de sus modalidades: fuerza física, longitud, inteligencia, etc.

4.1.3. Modalidades y clases

Como ya hemos indicado, frecuentemente es muy grande el número de modalidades bajo las cuales puede manifestarse una característica. Conviene reducir estas múltiples modalidades a un número menor de clases. Estas clases deben estar bien definidas (es decir, debemos saber claramente qué modalidades incluye cada una de ellas dentro de sí), deben ser mutuamente exclusivas (es decir, ninguna modalidad puede pertenecer simultáneamente a dos o más clases distintas), deben ser exhaustivas (es decir, toda modalidad debe pertenecer necesariamente a alguna de las clases).

4.1.4. Frecuencia, proporción, porcentaje

Frecuencia (o, frecuencia absoluta) de una clase es el número de observaciones contenidas dentro de ella.

Proporción (o, frecuencia relativa) de una clase es el cociente entre la frecuencia absoluta de dicha clase y el número total de observaciones (en todas las clases).

Porcentaje de una clase es igual a la proporción multiplicada por 100.

Distribución de frecuencias: conjunto de las clases y de las frecuencias (proporciones o porcentajes) correspondientes a cada una de aquellas. O, mejor aún, conjunto de los números atribuidos a las clases y de las frecuencias (proporciones o porcentajes) correspondientes a cada una de aquellas.

4.2. Organización de datos

4.2.1. Variables cualitativas

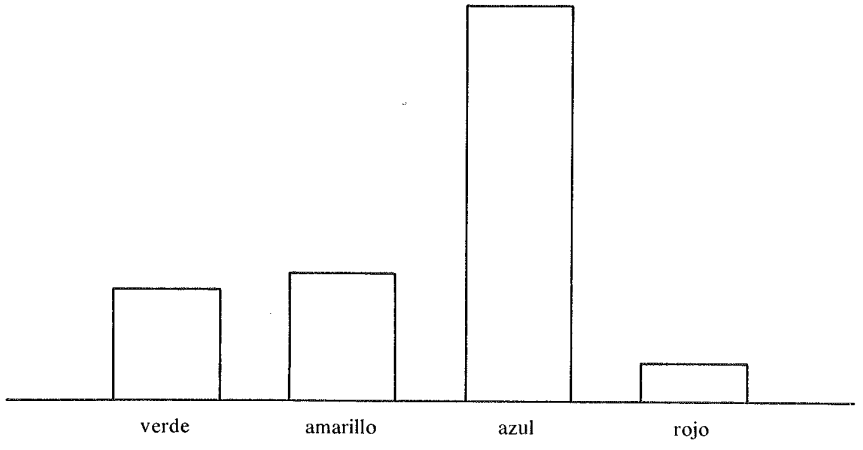
Warren (1974) investigó la característica «tipo de color asociado a la palabra *paz*» (entre otras). Para ello, la presentó a un grupo de cien personas, pidiendo que cada una escogiese entre 4 colores (rojo, azul, amarillo y verde) el que mejor se ajustase a dicha palabra.

La distribución de frecuencias fue la siguiente:

Distribución de frecuencias

Color	Frecuencia	Proporción	Porcentaje
rojo	6	0,06	6
azul	58	0,58	58
amarillo	19	0,19	19
verde	17	0,17	17
	100	1,00	100

Representación gráfica



Los cuatro rectángulos anteriores tienen la misma base y sus alturas (y áreas) son proporcionales a las frecuencias (proporciones y porcentajes) correspondientes.

Otras representaciones gráficas son posibles: ciclogramas, pictogramas, etc. El fin de todas ellas es representar de modo intuitivo las frecuencias de cada una de las modalidades (o clases de modalidades).

En el diagrama de barras las modalidades (o clases de modalidades) pueden ser colocadas en cualquier orden, pues representan distintos aspectos, no ordenados, de una característica.

En el ejemplo anterior las modalidades eran sólo cuatro y no parecía razonable agruparlas en clases. Pero pueden darse otros casos en los que el agrupamiento en clases sea muy conveniente. Así, por ejemplo, supongamos un colegio mayor con 80 universitarios. Estudiemos la característica «carrera universitaria» y supongamos que tenemos 15 modalidades distintas: Filología Clásica, Arte, Física, Ingeniería Naval, Química, etc. Hagamos con estas 15 modalidades cuatro clases que las engloben a todas ellas, de acuerdo con el siguiente esquema:

Distribución de frecuencias

Carrera univ.	Frecuencia	Proporción	Porcentaje
Letras	24	0,30	30
Ciencias	28	0,35	35
Ingeniería	8	0,10	10
Derecho	20	0,25	25
	80	1,00	100

Con esta distribución de frecuencias tendríamos una representación gráfica semejante a la anterior, con la única diferencia que ahora cada barra representa a una clase de modalidades y no a una sola modalidad.

4.2.2. Variables cuasi-cuantitativas

Con objeto de investigar la eficacia diagnóstica y terapéutica de algunas técnicas clínicas, Harrower (1965) recopiló los datos que exponemos a continuación sobre la mejoría de 622 pacientes.

Mejoría		Frec.	Prop.	Porc.	Fr. ac.	Prop. ac.	Porc. ac.
Máxima	(4)	134	0,21543	21,543	622	1,0000	100,00
Moderada	(3)	212	0,34084	34,084	488	0,7846	78,46
Leve	(2)	129	0,20740	20,740	276	0,4437	44,37
Nula	(1)	147	0,23633	23,633	147	0,2363	23,63
		622	1,00000	100,000			

A nivel ordinal tiene sentido hablar de frecuencias, proporciones y porcentajes acumulados (Fr. ac., Prop. ac., Porc. ac.). Ordinariamente, se suele comenzar la acumulación a partir de la clase inferior. Así lo hemos hecho en el cuadro adjunto. La primera frecuencia acumulada es la frecuencia de la clase inferior. La segunda frecuencia acumulada es la suma de las frecuencias de las dos clases inferiores. La tercera frecuencia acumulada es la suma de las frecuencias de las tres clases inferiores o, lo que es equivalente, la suma de la segunda frecuencia acumulada más la de la tercera no acumulada. Y así sucesivamente. Por supuesto la última frecuencia acumulada será igual, siempre, a la frecuencia total. En nuestro ejemplo,

Primera frecuencia acumulada: 147
Segunda frecuencia acumulada: $147 + 129 = 276$
Tercera frecuencia acumulada: $147 + 129 + 212 = 276 + 212 = 488$
Cuarta frecuencia acumulada: $147 + 129 + 212 + 134 = 488 + 134 = 622$

De modo análogo se obtienen las proporciones y los porcentajes acumulados a partir de las proporciones y porcentajes sin acumular. Naturalmente, las proporciones acumuladas pueden, también, ser obtenidas dividiendo cada frecuencia acumulada por el total de las observaciones. Así, en nuestro caso, $147/622 = 0{,}2363$, $276/622 = 0{,}4437$, $488/622 = 0{,}7846$, $622/622 = 1{,}0000$. Multiplicando por 100 estas proporciones acumuladas, obtendremos los correspondientes porcentajes acumulados.

Representación gráfica

Usaremos el diagrama de barras, como antes. Sin embargo, ahora las clases deben ser colocadas según un orden bien determinado, pues representan aspectos ordenados de una característica.

La figura 4.1 y la figura 4.2 han sido construidas con distinta unidad de medida. La primera barra de la izquierda de la figura 4.1 y la primera barra de la izquierda de la figura 4.2 representan lo mismo, a saber, la frecuencia (proporción o porcentaje) correspondiente a la modalidad «mejoría nula». La razón de elegir una distinta unidad de medida ha sido meramente práctica. Si hubiéramos elegido la misma unidad, o la figura 4.1 habría quedado excesivamente reducida, o la figura 4.2 habría quedado exageradamente grande.

4.2.3. Variables cuantitativas discretas

Consideremos cierta situación experimental en que una persona debe aprender una lista de pares de palabras, de manera que al presentarle una palabra de cada par sepa decir cuál es la otra que forma parte del mismo. Tomaremos como índice de dificultad de la tarea, el número de ensayos necesarios para asociar cada palabra con la correspondiente de su par. A continuación proponemos la distribución

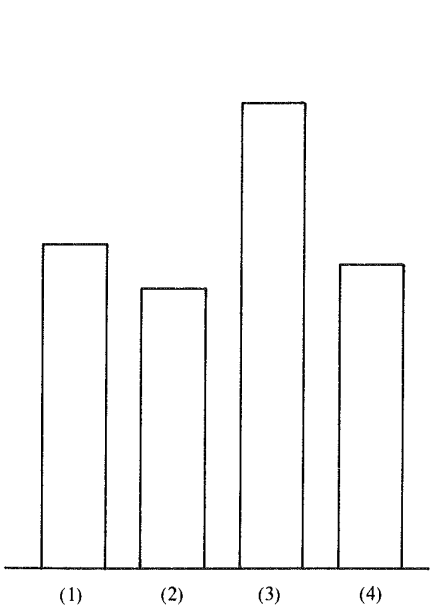


Fig. 4.1 (Sin acumular)

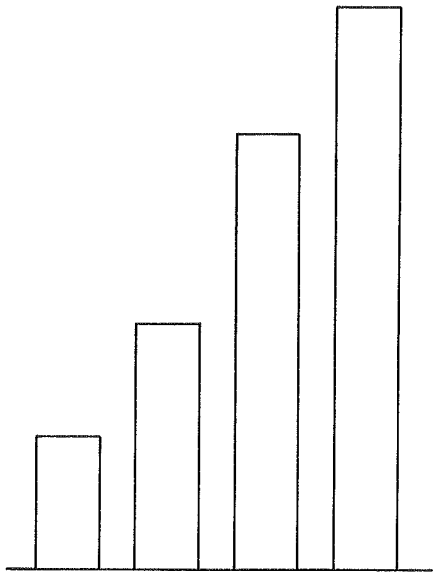


Fig. 4.2 (Acumulando)

de frecuencias y la representación gráfica acerca del número de ensayos necesitados por un grupo de 59 estudiantes de Psicología para aprender una lista de dificultad media formada por seis pares de palabras (Jáñez, 1976).

Distribución de frecuencias

Número de ensayos	Frec.	Prop.	Porc.	Frec. ac.	Prop. ac.	Porc. ac.
13	2	0,0339	3,39	59	1,0000	100,00
12	2	0,0339	3,39	57	0,9661	96,61
11	3	0,0508	5,08	55	0,9322	93,22
10	6	0,1017	10,17	52	0,8814	88,14
9	10	0,1695	16,95	46	0,7797	77,97
8	8	0,1356	13,56	36	0,6102	61,02
7	7	0,1186	11,86	28	0,4746	47,46
6	6	0,1017	10,17	21	0,3559	35,59
5	10	0,1695	16,95	15	0,2542	25,42
4	5	0,0847	8,47	5	0,0847	8,47
	59	0,9999	99,99			

Representación gráfica

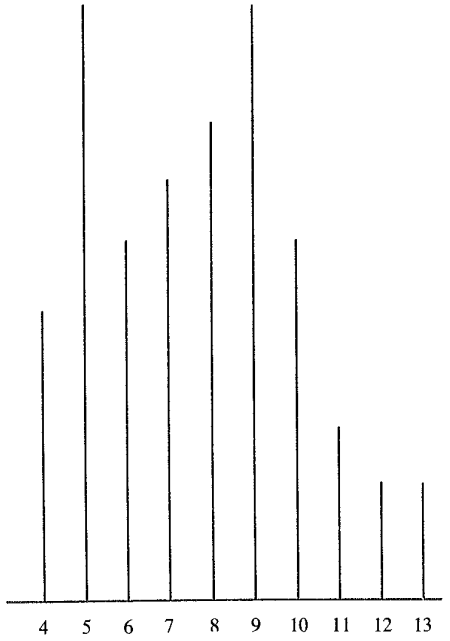


Fig. 4.3 (Sin acumular)

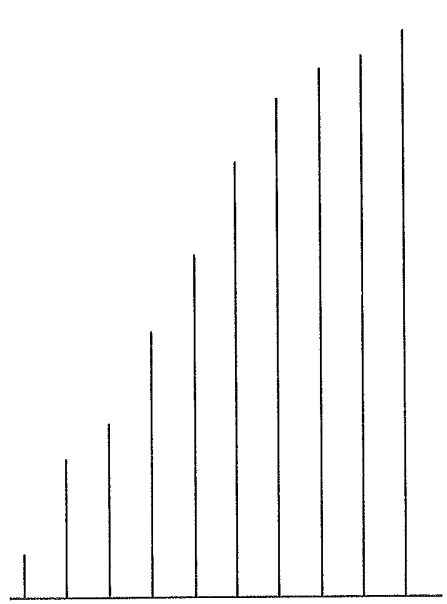
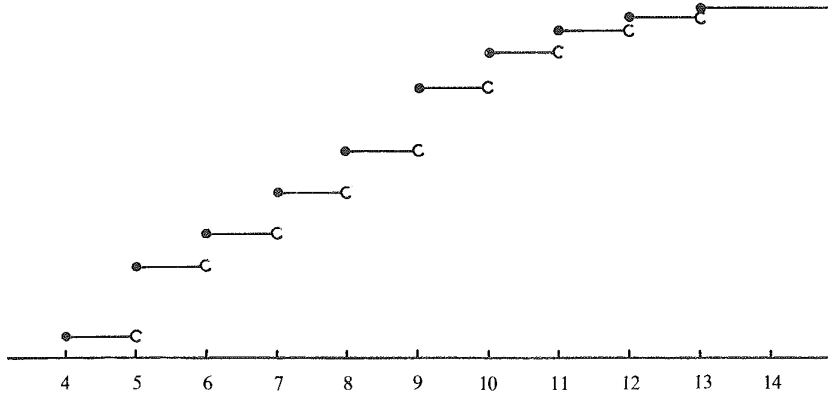


Fig. 4.4 (Acumulando)

Respecto a la unidad de medida usada en la figura 4.3 y la usada en la figura 4.4 vale lo dicho para las figuras 4.1 y 4.2.

Otra manera de representar gráficamente las frecuencias acumuladas, en el caso de variables discretas, es la siguiente:



Este gráfico es el que usaremos en Probabilidad para representar la función de distribución (concepto muy parecido al de distribución de frecuencias acumuladas) en el caso de variables aleatorias discretas.

Nótese que los «saltos» vienen dados en los puntos 4, 5, 6, ... Por ejemplo, al punto 6 (necesitar 6 o menos ensayos) le corresponde exactamente la proporción acumulada 0,3559. Al punto 7 le corresponde 0,4746. A todos los puntos entre 6 y 7 (excluido el 7) les hacemos corresponder la proporción 0,3559, pues necesitar 6,45; 6,78; ... (cualquier valor entre 6 y 7, excluidos ambos) equivale a necesitar seis ensayos, ya que es imposible necesitar seis ensayos y una fracción de ensayo. Por eso, a todos los valores entre 6 y 7 (excluidos ambos) les hacemos corresponder la misma proporción que corresponde al punto 6 (necesitar seis ensayos).

4.2.4. Variables cuantitativas continuas

4.2.4.1. Interpretación continua de los valores discretos

Consideremos, por ejemplo, la longitud. Entre dos modalidades cualesquiera existe un número infinito de modalidades posibles. Sin embargo, de hecho, sólo somos capaces de detectar un número finito de ellas, debido a la imperfección del instrumento de medida, en este caso, la regla. Cuanto más fina sea dicha regla, es decir, cuanto mayor número de subdivisiones contenga, tanto mayor será el número de modalidades que podremos detectar. Pero, en todo caso, ese número será finito, por culpa del instrumento de medida. En conclusión, la variable que en sí misma es continua se manifiesta, de hecho, como discreta. El número de modalidades discernibles es finito y, por tanto, será finito el número de valores atribuibles a dichas modalidades. Contemplemos algo más despacio este número finito de valores discretos.

Supongamos que nuestra regla no discierne más allá de los centímetros. Ello nos permitirá atribuir valores tales como 1,87, 1,88, 1,89, por ejemplo, pero no valores intermedios. Ahora bien, esta limitación, según lo visto, es debida a la imperfección del instrumento material de medida, no a que sean imposibles esos valores intermedios. Por ello, para salvar la continuidad, vamos a admitir que cada valor discreto representa a todos los infinitos valores situados media unidad de medida (medio centímetro) a su izquierda y media unidad a su derecha. En nuestro caso, 1,68 representa a los infinitos valores que van desde 1,675 hasta 1,685 (incluido el mismo 1,68), el valor 1,69 representa a todos los valores que van desde 1,685 hasta 1,695, etc. Es decir, 1,68 representa a una clase con infinitas modalidades. Llamaremos «intervalo elemental» a cada una de estas clases. Diremos que 1,675 es el límite exacto inferior del intervalo representado por 1,68, y 1,685 su límite exacto superior. Diremos que 1,685 es el límite exacto inferior del intervalo representado por 1,69, y 1,695 es su límite exacto superior. Admitiremos, por tanto, que 1,685 es, a la vez, límite exacto superior de un intervalo y límite exacto inferior del intervalo siguiente.

Habría que distinguir entre intervalos abiertos y cerrados, abiertos por la derecha (izquierda) y cerrados por la izquierda (derecha). En rigor, la amplitud de cada intervalo elemental valdrá la unidad de medida utilizada si nos valemos de intervalos semiabiertos. Sin embargo, estas distinciones, aunque importantes a nivel matemático, tienen poca importancia a nivel psicológico-estadístico. Por ello, las pasaremos por alto y aceptaremos un mismo valor como límite exacto común de dos intervalos consecutivos, admitiendo que la amplitud de cada intervalo elemental vale la unidad de medida.

4.2.4.2. Intervalos elementales y compuestos

Recordemos que cada valor discreto representa a todos los valores situados media unidad a su izquierda y media unidad a su derecha. Es decir, con cada valor discreto va asociado un intervalo de amplitud unidad que hemos llamado elemental. Llamaremos intervalo compuesto (o, simplemente, intervalo) al conjunto de varios intervalos elementales consecutivos. Por regla general, todos los intervalos compuestos (para un conjunto de datos) contendrán cada uno de ellos el mismo número de intervalos elementales.

4.2.4.3. Límites exactos y límites aparentes

Supongamos que en una investigación el valor discreto mínimo obtenido es 8 y el máximo es 19. Los valores discretos posibles (incluyendo el 8 y el 19) serán: 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19. Cada uno de ellos representa un intervalo elemental unitario. Formemos un intervalo compuesto con los intervalos elementales representados por 8, 9 y 10. Dado que es 7,5 el límite exacto inferior del intervalo elemental representado por el 8 y es 10,5 el límite exacto superior del intervalo elemental representado por el 10, aceptaremos 7,5 como límite exacto inferior del intervalo compuesto y 10,5 como límite exacto superior del mismo. Por consiguiente, tendremos 4 intervalos compuestos cuyos límites exactos serán: 7,5-10,5; 10,5-13,5; 13,5-16,5; 16,5-19,5. Llamaremos límites aparentes de estos cuatro intervalos a: 8 — 10, 11 — 13, 14 — 16, 17 — 19. Nótese que estos valores aparentes son los únicos que, de hecho, pueden aparecer.

4.2.4.4. Amplitud y punto medio de los intervalos. Amplitud total

La amplitud de un intervalo es la diferencia entre su límite exacto superior y su límite exacto inferior. En el ejemplo anterior la amplitud de los cuatro intervalos será: $10,5 - 7,5 = 13,5 - 10,5 = 16,5 - 13,5 = 19,5 - 16,5 = 3$. Aceptamos como punto medio de cada intervalo la media aritmética de sus dos límites exactos. Así, los puntos medios de los intervalos anteriores serán: $(7,5 + 10,5)/2 = 9$, $(10,5 + 13,5)/2 = 12$, $(13,5 + 16,5)/2 = 15$, $(16,5 + 19,5)/2 = 18$. A los mismos

resultados habríamos llegado calculando la media aritmética de los dos límites aparentes de cada intervalo. En el ejemplo anterior: $(8 + 10)/2 = 9$, $(11 + 13)/2 = 12$, $(14 + 16)/2 = 15$, $(17 + 19)/2 = 18$.

Llamaremos amplitud total de una serie de valores numéricos a la diferencia entre el límite exacto superior del intervalo máximo y el límite exacto inferior del intervalo mínimo. En nuestro caso, $19{,}5 - 7{,}5 = 12$.

4.2.4.5. Número y amplitud de los intervalos

Para una misma amplitud total, si aumenta el número de intervalos, tanto menor será la amplitud de cada uno de ellos. Se recomienda que, con 100 o más observaciones, el número de intervalos no sea menor que 12, ni mayor que 18. Según otros, ni menor que 10 ni mayor que 20. Sin embargo, ninguna de estas reglas es inflexible. Elegiremos en cada caso la regla que juzguemos más oportuna.

Ordinariamente, comenzamos fijando el número de intervalos en función del número total de observaciones. La amplitud de cada uno de los intervalos, dependerá de la amplitud total, una vez fijado su número. Por ejemplo, supongamos que 8, 8, 10, 11, 11, 12, 14, 15, 15, 15, 17, 19, 20, 21 son las puntuaciones obtenidas en una prueba por 14 personas. La amplitud total es $21{,}5 - 7{,}5 = 14$. Supongamos que decidimos hacer tres intervalos. La amplitud de cada uno de ellos tiene que ser 5 por lo menos. Si fuera sólo 4, alguna puntuación quedaría no contenida dentro de esos tres intervalos. Así, aceptando como intervalo ínfimo el 8-11, el segundo sería el 12-15 y el tercero el 16-19. No quedarían incluidas dentro de ellos las puntuaciones 20 y 21. La amplitud 5 sería suficiente, pues dentro de tres intervalos de amplitud 5 pueden caber todas las puntuaciones. (Adviértase que, también, podrían caber dentro de tres intervalos de amplitud 6, 7, 8, 9, 10, 11 y aun 12.) Lo ordinario es elegir la mínima entre todas las posibles. En el caso anterior elegiríamos la amplitud 5. Con esta amplitud serían posibles dos ternas de intervalos: (8-12, 13-17, 18-22) y (7-11, 12-16, 17-21).

Según algunos autores, es preferible elegir amplitudes iguales a uno de los valores siguientes: 1, 2, 3, 5, 10 ó 20. Estos números y sus múltiplos son fácilmente manejables. Sin embargo, este criterio es arbitrario y puede ser rechazado siempre que sea conveniente.

4.2.4.6. Distribución de frecuencias

Cincuenta estudiantes han obtenido en una prueba de inteligencia las siguientes puntuaciones:

8	11	11	8	9	10	16	6	12	19	13	14	9	13	15	9
12	16	8	7	14	11	15	6	14	14	17	11	6	9	10	19
12	11	12	6	15	16	16	12	13	12	12	8	17	13	7	12
14	12														

Ordenemos estas puntuaciones:

6	6	6	6	7	7	8	8	8	8	9	9	9	9	10	10
11	11	11	11	11	12	12	12	12	12	12	12	12	12	13	13
13	13	14	14	14	14	14	15	15	15	16	16	16	16	17	17
19	19														

Decidimos elegir cinco intervalos. La amplitud total vale $19{,}5 - 5{,}5 = 14$. Calculamos $14/5 = 2{,}8$. La amplitud mínima posible de cada intervalo será 3. Y, por ello, la elegimos como amplitud común a los cinco intervalos.

Ordinariamente, se suele comenzar por fijar el límite inferior del intervalo mínimo. Hay varias reglas convencionales en la elección de dicho límite. Según unos, debe ser la puntuación mínima (en nuestro caso 6). Según otros, debe ser múltiplo de la amplitud elegida (en nuestro caso, ó 3, ó 6). De acuerdo con algunos autores, es mejor comenzar fijando el límite superior del intervalo máximo. Desde luego, estas reglas son convencionales y no tenemos por qué acomodarnos a ellas.

La distribución de frecuencias en nuestro caso puede ser la siguiente:

		Frec.	Prop.	Porc.	Fr. ac.	Prop. ac.	Porc. ac.
17-19	////	4	0,08	8	50	1,00	100
14-16	/|| /|| //	12	0,24	24	46	0,92	92
11-13	/|| /|| /|| ///	18	0,36	36	34	0,68	68
8-10	/|| /||	10	0,20	20	16	0,32	32
5-7	/|| /	6	0,12	12	6	0,12	12
		50	1,00	100			

Desde luego, podíamos haber elegido 14 intervalos de amplitud unidad y cuyos puntos medios fueran 6, 7, . . . , 19. Es claro que, bajo esta condición, las puntuaciones originales coincidirán con los puntos medios de los intervalos unitarios.

4.2.4.7. Representación gráfica

a) *Histograma y polígono de frecuencias no acumuladas*

Comenzamos aceptando que todos los intervalos tienen la misma amplitud. Esto supuesto, sobre cada uno de ellos, como base, levantamos un rectángulo cuya altura (y área) sea proporcional a la frecuencia (proporción o porcentaje) no acumulados de dicho intervalo. Llamamos histograma de frecuencias no acumuladas a este conjunto de rectángulos consecutivos.

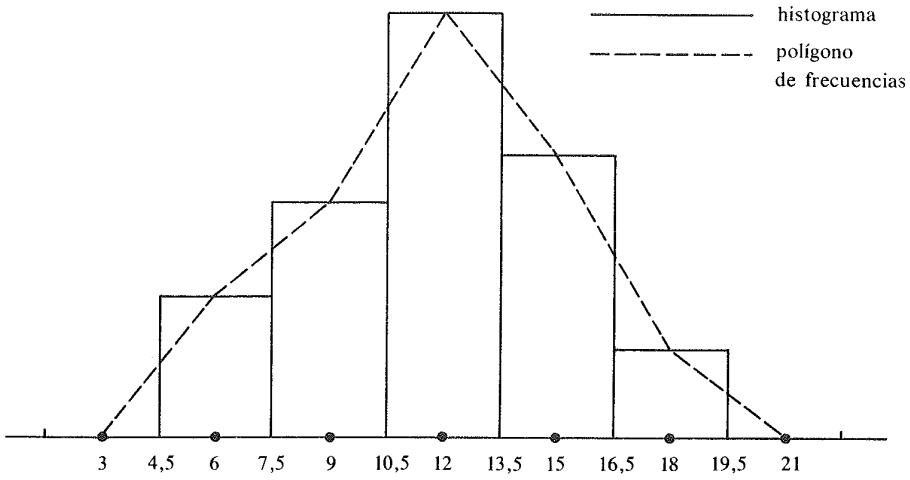


Fig. 4.5

Si los rectángulos no tuvieran todos la misma amplitud, el área de cada rectángulo seguirá siendo proporcional a la correspondiente frecuencia, pero no lo serán las alturas. Es claro en la figura 4.6 que la frecuencia (y área) del intervalo (I) es mayor que la del intervalo (II), pero la altura del primero es menor que la del segundo. No nos detenemos en discutir este caso, pues usaremos siempre intervalos con la misma amplitud.

Sobre el punto medio del lado superior (el opuesto a la base) de cada rectángulo del histograma dibujamos un punto. Unimos, después, cada dos puntos consecutivos mediante un segmento rectilíneo. Pues bien, llamaremos polígono de frecuencias no acumuladas a la línea originada por este conjunto de segmentos rectilíneos. Suelen ser añadidos dos intervalos, uno a la izquierda del inferior y

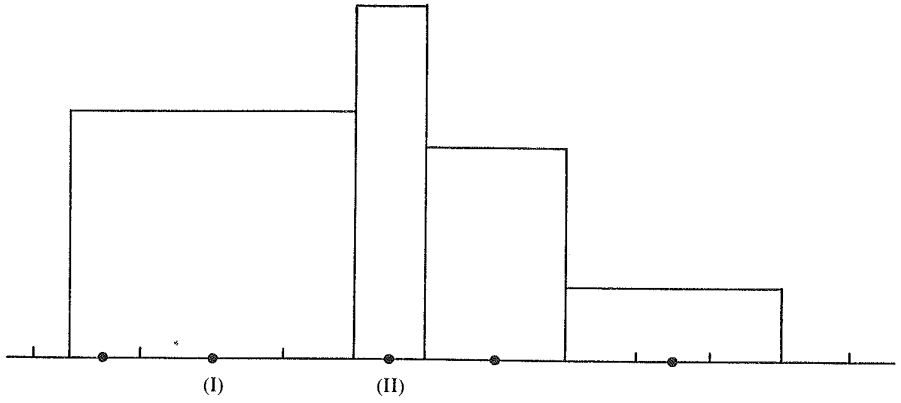
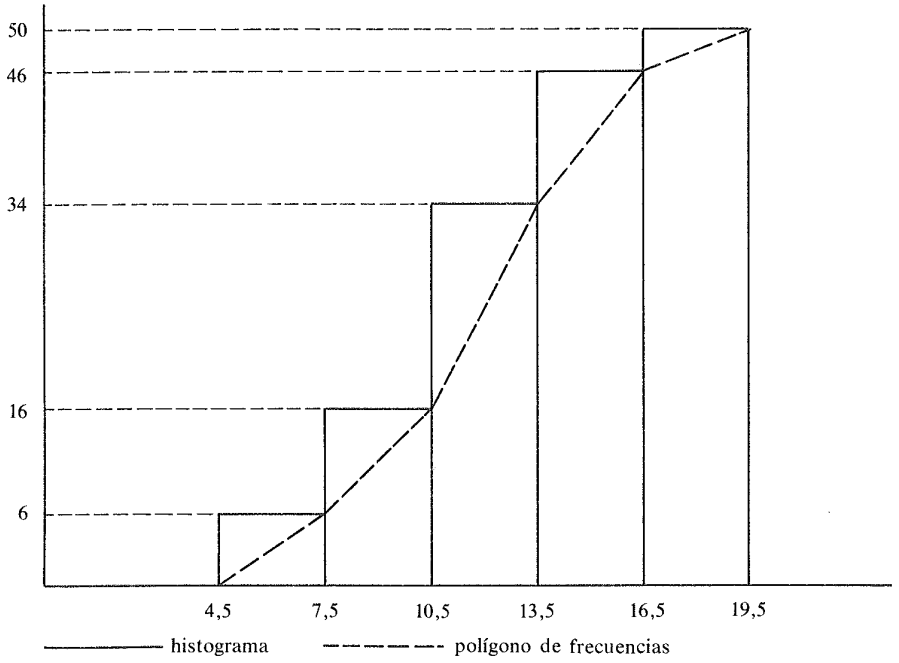


Fig. 4.6

otro a la derecha del superior, ambos con frecuencia nula. Así, el polígono de frecuencias comienza a partir del eje de abscisas y termina en el mismo. Así, además, el área contenida dentro del histograma es igual a la que queda debajo del polígono de frecuencias (véase Fig. 4.5).

b) *Histograma y polígono de frecuencias acumuladas*

Supuestos intervalos de igual amplitud, sobre cada uno de ellos, como base, levantamos un rectángulo cuya altura (y área) sea proporcional a la frecuencia (proporción o porcentaje) acumulados de dicho intervalo. Llamaremos histograma de frecuencias acumuladas a este conjunto de rectángulos consecutivos.



Unamos, mediante un segmento rectilíneo, el vértice inferior izquierdo del primer rectángulo (el situado a la izquierda de todos) con su vértice superior derecho. Este punto con el vértice superior derecho del siguiente rectángulo. Este punto con el vértice superior derecho del tercer rectángulo. Y así, sucesivamente. Pues bien llamaremos polígono de frecuencias acumuladas a la línea originada por este conjunto de segmentos rectilíneos. En otras palabras, esta línea une los puntos situados en las verticales levantadas sobre los límites exactos superiores de cada uno de los intervalos y a una altura proporcional a la frecuencia, proporción o porcentaje acumulados de dicho intervalo.

4.2.4.8. Normas prácticas para las representaciones gráficas

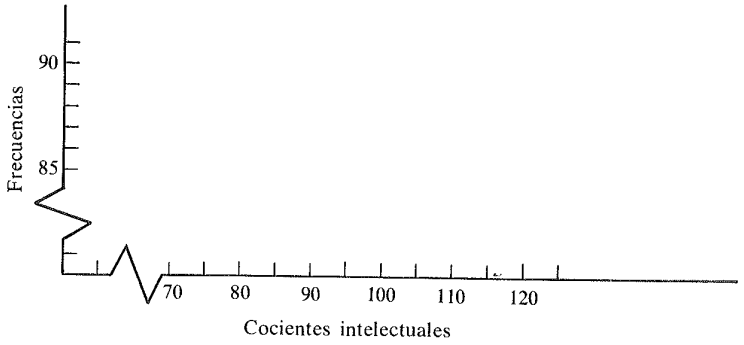
a) El eje de abscisas (horizontal) representará las puntuaciones de la variable de que se trate, y el eje de ordenadas (eje vertical) representará las frecuencias, proporciones o porcentajes.

b) En el eje de abscisas pondremos las puntuaciones menores a la izquierda y las mayores a la derecha. En el eje de ordenadas pondremos las frecuencias menores abajo y las mayores arriba.

c) La unidad de medida elegida en cada uno de los ejes será tal que el gráfico construido tenga una altura y una anchura cuya relación aproximada sea igual a 3/5. Es decir, si tiene una altura de 9 cm, por ejemplo, deberá tener una anchura aproximada de 15 cm.

d) La intersección de los dos ejes será tomada como origen de puntuaciones en el eje de abscisas y como origen de frecuencias, proporciones o porcentajes en el eje de ordenadas.

e) Si la puntuación mínima de que se trate es alta y la frecuencia mínima en cuestión es, también, alta, tanto en el eje de abscisas como en el de ordenadas se suelen hacer dos cortes, según la figura adjunta.



f) Conviene indicar explícitamente qué representa el gráfico en general y qué representa cada uno de los ejes, siempre que sea necesario.

Por supuesto, las anteriores normas son convencionales y sólo las seguiremos en tanto que nos sean útiles.

4.2.4.9. Polígonos de frecuencias de varios grupos considerados conjuntamente

Veamos la distribución de edades de 157 adolescentes (87 varones y 70 mujeres) con defectos auditivos y en los cuales Balow, Fulton y Peploe (1971) estudiaron algunas implicaciones educativas de la sordera.

TABLA 4.1

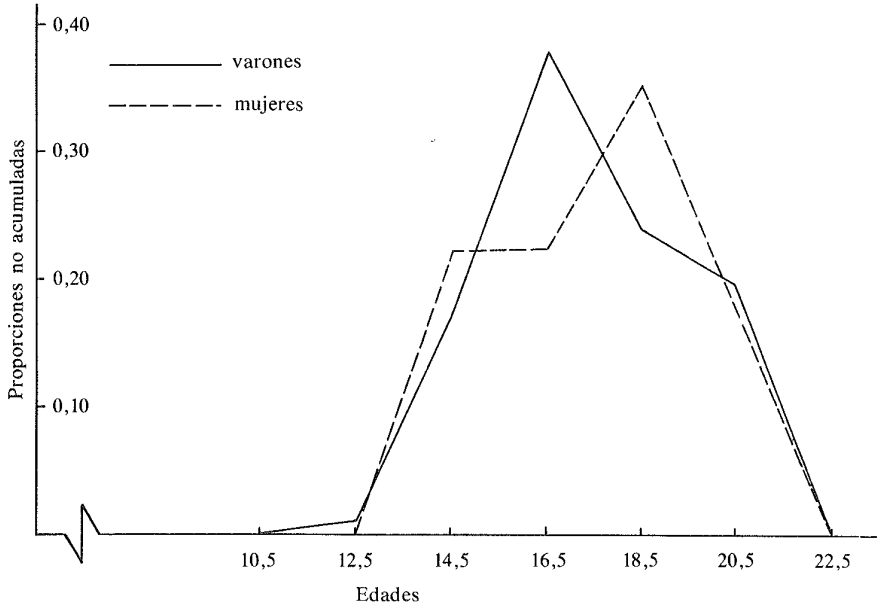
Distribución de edades correspondientes a los varones

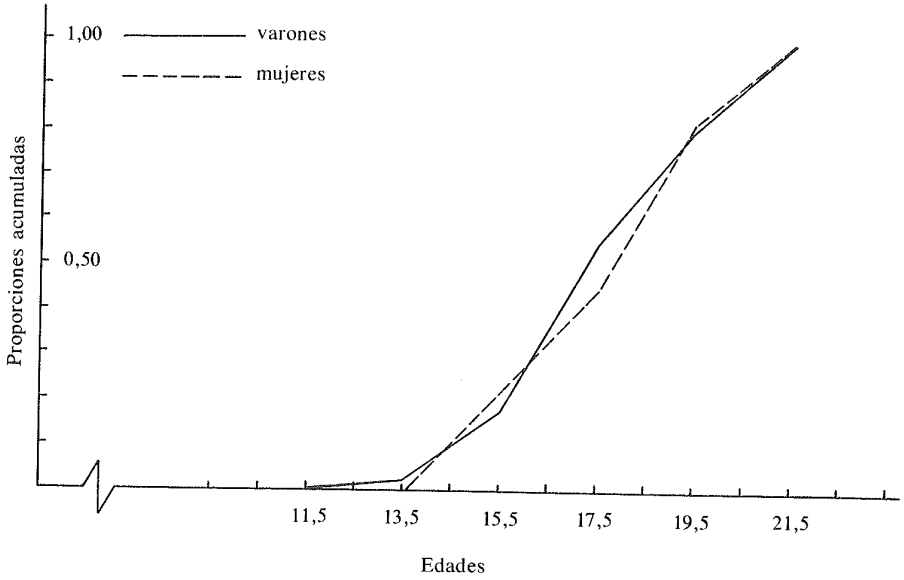
Edad	Frec.	Prop.	Porc.	Frec. ac.	Prop. ac.	Porc. ac.
20-21	17	0,1954	19,54	87	1,0000	100,00
18-19	21	0,2414	24,14	70	0,8046	80,46
16-17	33	0,3793	37,93	49	0,5632	56,32
14-15	15	0,1724	17,24	16	0,1839	18,39
12-13	1	0,0115	1,15	1	0,0115	1,15
	87	1,0000	100,00			

TABLA 4.2

Distribución de edades correspondientes a las mujeres

Edad	Frec.	Prop.	Porc.	Frec. ac.	Prop. ac.	Porc. ac.
20-21	13	0,1857	18,57	70	1,0000	100,00
18-19	25	0,3571	35,71	57	0,8143	81,43
16-17	16	0,2286	22,86	32	0,4571	45,71
14-15	16	0,2286	22,86	16	0,2286	22,86
12-13	0	0,0000	00,00	0	0,0000	00,00
	70	1,0000	100,00			





Valiéndonos de proporciones no acumuladas y de proporciones acumuladas, tendremos las dos representaciones gráficas anteriores.

4.2.4.10. Datos sin agrupar y agrupados

Agrupar ciertas puntuaciones en intervalos implica hacerlas equivalentes a los puntos medios de los intervalos dentro de los cuales se encuentra cada una de ellas. Así, por ejemplo, al agrupar en cinco intervalos las cincuenta puntuaciones propuestas en el apartado 4.2.4.6, nos quedaremos con sólo cinco valores distintos, los puntos medios de los cinco intervalos. Es decir,

6, 6, 6, 6, 7, 7 (dentro del intervalo 5-7) equivaldrán a 6.
8, 8, 8, 8, 9, 9, 9, 9, 10, 10 (dentro del intervalo 8-10) equivaldrán a 9.
11, 11, 11, 11, 11, 12, 12, 12, 12, 12, 12, 12, 12, 12, 13, 13, 13, 13 (dentro del intervalo 11-13) equivaldrán a 12.
14, 14, 14, 14, 14, 15, 15, 15, 16, 16, 16, 16 (dentro del intervalo 14-16) equivaldrán a 15.
17, 17, 19, 19 (dentro del intervalo 17-19) equivaldrán a 18.

Si el número de datos fuera mayor (por ejemplo, 350), haciendo diez intervalos nos quedaríamos sólo con diez puntuaciones distintas. Evidentemente, es más manejable el nuevo conjunto de diez datos distintos que el conjunto de las puntuaciones originales y tanto más manejable, cuanto mayor sea el número de puntuaciones originales distintas. El agrupamiento tiene la ventaja de hacer más manejable la masa informe de datos primitivos. Sin embargo, tiene el inconveniente de falsear más o menos, de cercenar en parte, la información contenida en los datos originales. En resumen, la información ofrecida por los datos no agrupados es la verdadera e íntegra, pero es menos manejable. La información ofrecida por los datos agrupados es más manejable, pero es menos íntegra y verdadera.

4.3. Resumen: Definiciones

Constante: característica que sólo puede manifestarse bajo una única modalidad.

Variable: característica que puede manifestarse bajo dos o más modalidades.

Variable cualitativa: característica que sólo puede ser considerada a nivel meramente nominal.

Variable cuasi cuantitativa: característica que sólo puede ser considerada a nivel máximo ordinal.

Variable cuantitativa: característica que puede ser considerada, al menos, a nivel de intervalos.

Variable cuantitativa discreta: característica que no admite siempre una modalidad intermedia entre dos cualesquiera de sus modalidades.

Variable cuantitativa continua: característica que admite siempre una modalidad intermedia entre dos cualesquiera de sus modalidades.

Frecuencia de una clase: número de observaciones contenidas dentro de ella.

Proporción de una clase: cociente entre la frecuencia de una clase y el número total de observaciones.

Porcentaje de una clase: proporción de la misma multiplicada por 100.

Distribución de frecuencias: conjunto de números, atribuidos a las modalidades o clases, y de las frecuencias (proporciones o porcentajes) correspondientes a cada una de aquellas.

Intervalo: tratándose de variables continuas, clase compuesta de infinitas modalidades o, mejor, compuesta de los infinitos números atribuibles a dichas modalidades y situados entre dos valores numéricos que llamaremos sus *límites exactos*. Llamaremos *amplitud* del intervalo a la diferencia entre esos dos límites exactos y *punto medio* a la semisuma de dichos límites. Diremos que un intervalo es elemental si su amplitud es la unidad de medida utilizada y diremos que es compuesto si su amplitud es mayor que dicha unidad.

Histograma: supuesto que todos los intervalos tienen la misma amplitud, levantamos sobre cada intervalo, como base, un rectángulo cuya altura (y área) sea proporcional a la frecuencia (proporción o porcentaje) de dicho intervalo, llamando histograma al conjunto de estos rectángulos consecutivos.

Polígono de frecuencias: dibujado un punto sobre la mitad del lado superior (el opuesto a la base) de cada rectángulo, unimos cada dos puntos consecutivos mediante un segmento rectilíneo, llamando polígono de frecuencias a la línea originada por este conjunto de segmentos rectilíneos.

EJERCICIOS

4.1. ¿Cuáles son los límites exactos de los siguientes intervalos?

a) 15 – 24 , *b*) 62,5 – 68,5 , *c*) 20,0 – 20,8 , *d*) 44,35 – 54,35

4.2. ¿Cuál es la amplitud de los intervalos anteriores?

4.3. ¿Cuál es su punto medio?

4.4. En un examen de Estadística los alumnos han obtenido las siguientes puntuaciones: 16, 18, 26, 15, 17, 21, 27, 21, 21, 26, 14, 20, 23, 16, 19, 24, 22, 23, 20, 26, 18, 20, 14, 17, 21, 17, 24, 27, 18, 17, 25, 19, 22, 21, 21, 15, 24, 22, 15, 18.
Preparar una distribución de frecuencias, sin acumular y acumuladas, introduciendo intervalos de amplitud 3. Calcular las proporciones y porcentajes sin acumular y acumulados.

4.5. Dibujar el correspondiente histograma y el polígono de frecuencias sin acumular y acumuladas a partir de los datos del ejercicio anterior.

4.6. A partir de la siguiente distribución de frecuencias, calcular las frecuencias acumuladas y las proporciones y porcentajes sin acumular y acumulados. Dibujar, además, el histograma y el polígono de frecuencias, considerando éstas sin acumular y acumuladas.

X	n_j
17-19	8
14-16	9
11-13	12
8-10	10
5-7	7
2-4	4

4.7. Durante el curso 1971-1972 en las Facultades universitarias estatales españolas estaban matriculados los siguientes alumnos:

Facultad de Ciencias	42.572
Fac. de Cienc. Polít. Económ. y Comerc.	25.683
Fac. de Derecho	22.665
Fac. de Farmacia	8.083
Fac. de Filosofía y Letras	49.049
Fac. de Medicina	37.578
Fac. de Veterinaria	2.166
	187.796

FUENTE: *Comentario sociológico, Estructura social de España, 1973-74.* Confederación Española de Cajas de Ahorro, año II, núms. 4-5.

Calcular las proporciones y porcentajes correspondientes a cada una de las categorías y dibujar el correspondiente diagrama de barras.

4.8. A partir de la siguiente distribución de frecuencias, calcular las frecuencias acumuladas y las proporciones y porcentajes sin acumular y acumulados. Dibujar, además, el histograma y el polígono de frecuencias, considerando éstas sin acumular y acumuladas.

X	n_j
20-24	12
15-19	18
10-14	24
5-9	16
0-4	10

4.9. Las puntuaciones en una prueba de inteligencia abstracta han sido las siguientes:

91, 92, 83, 81, 88, 94, 91, 87, 90, 94, 85, 85, 93, 90, 89, 86, 87, 89, 85, 89

Preparar una distribución de frecuencias, sin acumular y acumuladas, introduciendo intervalos de amplitud 4. Calcular las proporciones y porcentajes sin acumular y acumulados.

4.10. Dibujar el correspondiente histograma y el polígono de frecuencias sin acumular y acumuladas a partir de los datos del ejercicio anterior.

5

Estadísticos de posición o tendencia central

5.1. Introducción

Supongamos que deseamos comparar el aprovechamiento en Estadística de una muestra de 200 varones con el de otra muestra de 250 mujeres. Esta comparación será muy difícil si hemos de tener en cuenta todas las puntuaciones de ambos grupos. Lo que solemos hacer es comparar el promedio de la primera muestra con el promedio de la segunda. En otras palabras, lo que hacemos es determinar un estadístico (función de las puntuaciones de la muestra) que nos ofrezca la posición de una y otra muestra en la variable aprovechamiento. En general, este tipo de estadísticos se utiliza para darnos la posición de cada una de las muestras a las que va representando y, por esta razón, deberá ir tomando siempre un valor situado hacia el centro de las puntuaciones de cada una de dichas muestras. Debido a esta circunstancia, suelen ser llamados de posición o tendencia central.

Antes de comenzar a estudiar estos estadísticos, el lector debe consultar el Apéndice A que trata sobre el signo de sumar Σ. Este signo será utilizado muy profusamente de ahora en adelante y, por ello, conviene que el lector se familiarice con su uso.

5.2. Media aritmética*

5.2.1. Definición

Dados n valores, $X_1, X_2, \ldots, X_n$, su media aritmética, $\bar{X}$, viene definida por

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\Sigma X_i}{n} \tag{5.1}$$

* Siempre que usemos la expresión «media», nos referimos a la «media aritmética».

Es decir, la media aritmética de n valores no es más que su suma dividida por el número de ellos.

5.2.2. Cálculo

a) *Datos no agrupados*

Aplicación directa de la fórmula (5.1) a los datos originales, es decir, sumando una a una las n puntuaciones y dividiendo el resultado por n.

EJEMPLO 5.1. Para obtener un índice de la carga de significación que posee la palabra GOLA, Jáñez (1976) la presentó a 23 personas con objeto de observar el número de palabras que GOLA suscitaba en cada persona durante cuarenta y cinco segundos y obtuvo los siguientes resultados:

10, 5, 2, 7, 9, 5, 7, 6, 5, 9, 12, 2, 6, 6, 9, 12, 6, 6, 6, 4, 9, 7, 12

La media aritmética fue tomada como índice de la carga de significación de GOLA y valió:

$$\bar{X} = \frac{10 + 5 + 2 + \cdots + 7 + 12}{23} = 7{,}04$$

b) *Datos agrupados*

Supongamos n observaciones agrupadas en r intervalos, todos ellos de igual amplitud. Sea X_1 el punto medio del intervalo primero y n_1 el número de observaciones dentro del mismo. Sea X_2 el punto medio del intervalo segundo y n_2 el número de observaciones dentro del mismo. Sea X_r el punto medio del intervalo r y n_r el número de observaciones dentro del mismo. Según sabemos, al agrupar las puntuaciones en intervalos, atribuimos a cada una de ellas (como puntuación) el punto medio dentro del intervalo dentro del cual se encuentra. Por consiguiente, dentro del intervalo primero tendremos n_1 puntuaciones iguales a X_1 y su suma valdrá n_1X_1. Dentro del intervalo segundo tendremos n_2 puntuaciones iguales a X_2 y su suma valdrá n_2X_2. Dentro del intervalo r-simo tendremos n_r puntuaciones iguales a X_r y su suma valdrá n_rX_r. En conclusión, la suma total de las $n_1 + n_2 + \cdots + n_r = n$ puntuaciones valdrá $n_1X_1 + n_2X_2 + \cdots + n_rX_r$ y su media aritmética valdrá

$$\bar{X} = \frac{n_1X_1 + n_2X_2 + \cdots + n_rX_r}{n_1 + n_2 + \cdots + n_r} = \frac{\Sigma n_jX_j}{\Sigma n_j} = \frac{\Sigma n_jX_j}{n} \tag{5.2}$$

EJEMPLO 5.2. Valiéndonos de los datos del ejemplo 5.1, agrupémoslos en cuatro intervalos de acuerdo con el cuadro siguiente y calculemos su media aritmética.

X	n_j	X_j	n_jX_j
10-12	4	11	44
7-9	7	8	56
4-6	10	5	50
1-3	2	2	4
	23		154

$$\bar{X} = \frac{154}{23} = 6{,}70$$

Nótese que la definición de media es la misma: sumar n puntuaciones y dividir esa suma por n. Las que no suelen ser exactamente iguales son las puntuaciones antes y después del agrupamiento en intervalos. Así, comparando el ejemplo 5.1 con el ejemplo 5.2, tenemos:

EJEMPLO 5.1: 2, 2, 4, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 9, 9, 9, 9, 10, 12, 12, 12
EJEMPLO 5.2: 2, 2, 5, 5, 5, 5, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8, 8, 8, 11, 11, 11, 11

Consiguientemente, el valor de la media aritmética obtenida a partir de ciertos datos no agrupados en intervalos diferirá, en general, del valor de la media aritmética obtenida a partir de estos mismos datos, pero agrupados en intervalos. Además, este último valor será en general, uno u otro según que los mismos datos sean agrupados de una u otra manera.

Por supuesto, podíamos haber elegido intervalos de amplitud unidad, cuyos puntos medios coincidan con números enteros consecutivos y tales que dentro de la sucesión de éstos se encuentren las puntuaciones originales. Naturalmente, bajo estas condiciones coincidirán la media aritmética obtenida mediante datos no agrupados y la obtenida mediante datos agrupados en intervalos (unitarios).

EJEMPLO 5.3. Valiéndonos de los datos del ejemplo 5.1 calculemos la media aritmética agrupando los datos en intervalos unitarios, según lo acabado de exponer.

X_j	n_j	n_jX_j
12	3	36
11	0	0
10	1	10
9	4	36
8	0	0
7	3	21
6	6	36
5	3	15
4	1	4
3	0	0
2	2	4
	23	162

$$\bar{X} = \frac{\Sigma\, n_jX_j}{n} = \frac{162}{23} = 7{,}04$$

5.2.3. Propiedades

a) Si $k = \bar{X}$, la suma de las diferencias de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a k vale cero. Es decir, la suma de las diferencias de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a su media vale cero.

En efecto, si $k = \bar{X}$ tendremos

$$\Sigma(X_i - k) = \Sigma(X_i - \bar{X}) = \Sigma X_i - n\bar{X} = \Sigma X_i - n\frac{\Sigma X_i}{n} = \Sigma X_i - \Sigma X_i = 0$$

a') Si la suma de las diferencias de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a k vale cero, $k = \bar{X}$.

En efecto, si se verifica que $\Sigma(X_i - k) = 0$, entonces $\Sigma X_i - nk = 0$. Es decir, $\Sigma X_i = nk$. Por tanto,

$$k = \frac{\Sigma X_i}{n} = \bar{X}$$

EJEMPLO 5.4. La media de 1, 2, 5, 8 vale $\frac{16}{4} = 4$. Pues bien, $(1 - 4) + (2 - 4) + (5 - 4) + (8 - 4) = (-3) + (-2) + (1) + (4) = 0$.

b) Si $k = \bar{X}$, la suma de las diferencias cuadráticas de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a k es mínima. Es decir, la suma de las diferencias cuadráticas de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a su media es menor que la suma de las diferencias cuadráticas respecto a cualquier otro valor distinto de la media.

Intentamos probar que $\Sigma(X_i - k)^2 > \Sigma(X_i - \bar{X})^2$ si $k \neq \bar{X}$.

En efecto, si $k \neq \bar{X}$, $k = \bar{X} + c$ (con $c \neq 0$, bien positiva, bien negativa). Por tanto,

$\Sigma(X_i - k)^2 = \Sigma[X_i - (\bar{X} + c)]^2 = \Sigma[(X_i - \bar{X}) - c]^2 = \Sigma(X_i - \bar{X})^2 - 2c\,\Sigma(X_i - \bar{X}) + nc^2 = \Sigma(X_i - \bar{X})^2 + nc^2$, pues $\Sigma(X_i - \bar{X}) = 0$, (según *a*).

Ahora bien, nc^2 es un valor esencialmente positivo, pues tanto n como c^2 son esencialmente positivos. Por otra parte, $\Sigma(X_i - \bar{X})^2$ es esencialmente no negativa (será siempre positiva, salvo el caso extremo en que $X_1 = X_2 = \cdots = X_n$ y en el cual valdrá cero). Por tanto, dado que $\Sigma(X_i - k)^2 = \Sigma(X_i - \bar{X})^2 + nc^2$, necesariamente $\Sigma(X_i - k)^2 > \Sigma(X_i - \bar{X})^2$.

EJEMPLO 5.5. Restemos 1, 2, 5, 8 de su media 4, elevemos al cuadrado estas diferencias y sumémoslas. Hagamos lo mismo con las diferencias cuadráticas respecto a dos valores distintos de 4, el 5 (mayor que 4) y el 3 (menor que 4).

$$(1 - 4)^2 + (2 - 4)^2 + (5 - 4)^2 + (8 - 4)^2 = 9 + 4 + 1 + 16 = 30$$
$$(1 - 5)^2 + (2 - 5)^2 + (5 - 5)^2 + (8 - 5)^2 = 16 + 9 + 0 + 9 = 34$$
$$(1 - 3)^2 + (2 - 3)^2 + (5 - 3)^2 + (8 - 3)^2 = 4 + 1 + 4 + 25 = 34$$

Vemos cómo la suma de las diferencias cuadráticas respecto a 4 (media) es menor que la suma de las diferencias cuadráticas respecto a 5 y a 3 (distintos de la media).

b') Si la suma de las diferencias cuadráticas de n puntuaciones $X_1, X_2, \ldots, X_n$ respecto a k es mínima, $k = \bar{X}$.

En efecto, si k fuera distinta de la media, la suma de las diferencias cuadráticas no podría ser mínima (según *b*), contra lo supuesto.

c) La media de $Y_1 = AX_1 + B$, $Y_2 = AX_2 + B$, ..., $Y_n = AX_n + B$, siendo A y B dos constantes arbitrarias, es $\bar{Y} = A\bar{X} + B$.

En efecto,

$$\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{\Sigma (AX_i + B)}{n} = \frac{A \Sigma X_i + nB}{n} = A\frac{\Sigma X_i}{n} + \frac{nB}{n} = A\bar{X} + B$$

EJEMPLO 5.6. A partir de las puntuaciones 1, 2, 6 (con media 3) formemos las nuevas puntuaciones $Y_i = 2X_i + 4$, es decir, las puntuaciones $Y_1 = 2X_1 + 4 = (2)(1) + 4 = 6$, $Y_2 = 2X_2 + 4 = (2)(2) + 4 = 8$, $Y_3 = 2X_3 + 4 = (2)(6) + 4 = 16$ y veamos cómo la media de las nuevas puntuaciones debe valer $\bar{Y} = 2\bar{X} + 4 = (2)(3) + 4 = 10$. En efecto,

$$\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{6 + 8 + 16}{3} = \frac{30}{3} = 10$$

Si $A = 1$ y $B \neq 0$, $\bar{Y} = (1)(\bar{X}) + B = \bar{X} + B$. Es decir, si sumamos a todas las puntuaciones una constante no nula, la media de las nuevas puntuaciones es igual a la media de las antiguas más esa constante B.

Si $A \neq 0$ y $B = 0$, $\bar{Y} = A\bar{X} + 0 = A\bar{X}$. Es decir, si multiplicamos todas las puntuaciones por una constante A, la media de las nuevas puntuaciones es igual a la media de las antiguas multiplicada por esa constante A.

d) La media es sensible a la variación de cada una de las puntuaciones. Basta con que varíe una sola puntuación, para que varíe la media. La media es función de todas y cada una de las puntuaciones y variará con que varíe una sola de ellas.

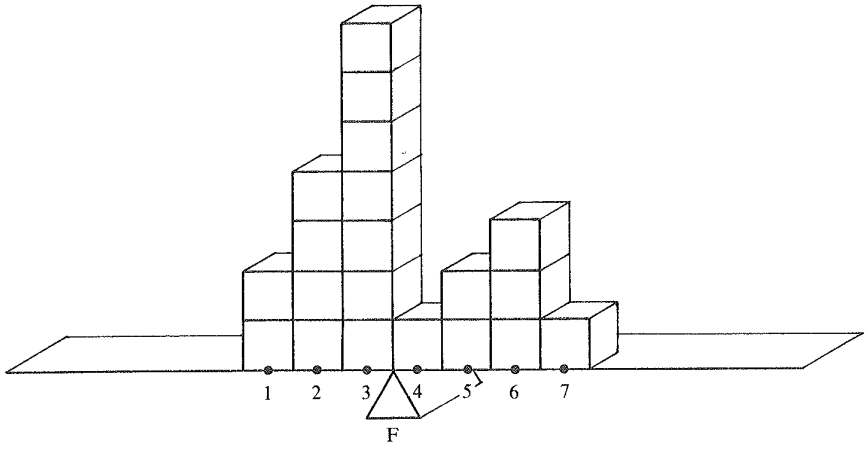
e) Es función de los intervalos elegidos (de su amplitud, de su número y de los límites de los mismos).

f) Es fundamento de muchas otras técnicas estadísticas.

g) No podrá ser calculada si el intervalo máximo no tiene límite superior y/o el intervalo mínimo no lo tiene inferior. Pues si no conocemos los límites extremos, no podremos calcular los puntos medios de los intervalos máximo y mínimo y, consiguientemente, no podremos calcular la media que exige conocer los puntos medios de todos los intervalos. Así, por ejemplo, no podremos calcular la media a partir de la siguiente distribución de frecuencias:

X	n_j
17 o más	9
14-16	15
11-13	22
8-10	13
7 o menos	8

h) La media es el «centro de gravedad» de la distribución de frecuencias. Consideremos un tablero ideal, sin peso, largo y estrecho. Representemos cada observación por un cubo de peso unidad. Todas las observaciones con la misma puntuación son colocadas una encima de otra sobre el punto del tablero que coincide con esa puntuación. Apoyemos el tablero, así cargado con las observaciones, sobre un fulcro F. Pues bien, solamente se mantendrá en equilibrio el tablero, cuando la media sea el punto de apoyo del mismo sobre el fulcro. Es decir, la media es el «centro de gravedad» del sistema. Por ejemplo, supuestas las puntuaciones 1, 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 5, 5, 6, 6, 6, 7, su media valdrá 70/20 = 3,5. Pues bien, el tablero cargado con las veinte puntuaciones se mantendrá en equilibrio si se apoya sobre el fulcro F en la puntuación 3,5 y perderá dicho equilibrio si se apoya sobre un punto a la derecha o a la izquierda de 3,5.



i) No es recomendable calcular la media cuando la distribución de frecuencias es muy asimétrica. Es decir, cuando existen una o muy pocas puntuaciones en uno de los dos extremos (o muy altas o muy bajas, respecto a las restantes que constituyen la mayoría).

j) Dados r grupos con $n_1, n_2, \ldots, n_r$ puntuaciones, respectivamente, y siendo $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_r$ sus correspondientes medias, la media de las $n_1 + n_2 + \cdots + n_r = n$ puntuaciones vale $\bar{X} = (n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_r\bar{X}_r)/n$.

En efecto, sean

$X_{11}, X_{21}, \ldots, X_{n_1 1}$ las n_1 puntuaciones del grupo primero
$X_{12}, X_{22}, \ldots, X_{n_2 2}$ las n_2 puntuaciones del grupo segundo
. . . . . .
$X_{1r}, X_{2r}, \ldots, X_{n_r r}$ las n_r puntuaciones del grupo r

Esto supuesto, tendremos

$$\begin{aligned}
&\bar{X}_1 = (X_{11} + X_{21} + \cdots + X_{n_1 1})/n_1; \quad \text{por tanto, } n_1\bar{X}_1 = X_{11} + X_{21} + \cdots + X_{n_1 1}\\
&\bar{X}_2 = (X_{12} + X_{22} + \cdots + X_{n_2 2})/n_2; \quad \text{por tanto, } n_2\bar{X}_2 = X_{12} + X_{22} + \cdots + X_{n_2 2}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (5.3)\\
&\bar{X}_r = (X_{1r} + X_{2r} + \cdots + X_{n_r r})/n_r; \quad \text{por tanto, } n_r\bar{X}_r = X_{1r} + X_{2r} + \cdots + X_{n_r r}
\end{aligned}$$

Ahora bien, por definición,

$$\bar{X} = \frac{(X_{11} + X_{21} + \cdots + X_{n_1 1}) + (X_{12} + X_{22} + \cdots + X_{n_2 2}) + \cdots + (X_{1r} + X_{2r} + \cdots + X_{n_r r})}{n_1 + n_2 + \cdots + n_r} \quad (5.4)$$

Por consiguiente, sustituyendo los r términos del numerador de (5.4) por sus equivalentes en (5.3), nos queda

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_r\bar{X}_r}{n_1 + n_2 + \cdots + n_r} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + \cdots + n_r\bar{X}_r}{n}$$

EJEMPLO 5.7.

Grupo primero	Grupo segundo	Grupo tercero
2 5 4 7 7 5	7 14 7 8	14 17 8 12 9
30 $n_1 = 6$ $\bar{X}_1 = 5$	36 $n_2 = 4$ $\bar{X}_2 = 9$	60 $n_3 = 5$ $\bar{X}_3 = 12$

$$\bar{X} = \frac{(6)(5) + (4)(9) + (5)(12)}{15} = \frac{126}{15} = 8{,}40$$

EJEMPLO 5.8. Las puntuaciones de la tabla adjunta representan el aumento en el cociente intelectual de un grupo de dos niños y de un grupo de cuatro niñas, todos de dos años, tras haber sido estimulada la interacción verbal con sus madres durante cuatro meses (Levenstein, 1968). Calculemos el aumento medio del grupo total, a partir del aumento medio de los niños y del de las niñas.

Niños	Niñas
13 29	16 16 2 6
42 $n_1 = 2$ $\bar{X}_1 = 21$	40 $n_2 = 4$ $\bar{X}_2 = 10$

$$\bar{X} = \frac{(2)(21) + (4)(10)}{6} = \frac{82}{6} = 13{,}67$$

En el caso particular en el que $n_1 = n_2 = \cdots = n_r = k$,

$$\bar{X} = \frac{k\bar{X}_1 + k\bar{X}_2 + \cdots + k\bar{X}_r}{kr} = \frac{\bar{X}_1 + \bar{X}_2 + \cdots + \bar{X}_r}{r}$$

k) Sean r grupos con $n_1, n_2, \ldots, n_r$ personas. Sean $p_1, p_2, \ldots, p_r$ las proporciones con cierta característica dentro de cada grupo. Por ejemplo, sea p_1 la proporción de varones en el grupo primero, p_2 la proporción de varones en el grupo segundo, . . . , p_r la proporción de varones en el grupo r. Esto supuesto, la proporción de varones en el grupo total vale

$$p = \frac{n_1p_1 + n_2p_2 + \cdots + n_rp_r}{n_1 + n_2 + \cdots + n_r}$$

En efecto, llamando $n_1', n_2', \ldots, n_r'$ al número de varones en el grupo primero, en el segundo, . . . , en el r, tendremos

$p_1 = \dfrac{n_1'}{n_1}$, de donde, $n_1' = n_1p_1$; $p_2 = \dfrac{n_2'}{n_2}$, de donde, $n_2' = n_2p_2$; . . . ; $p_r = \dfrac{n_r'}{n_r}$, de donde $n_r' = n_rp_r$.

Por definición, la proporción de varones dentro del grupo total valdrá,

$$p = \frac{n_1' + n_2' + \cdots + n_r'}{n_1 + n_2 + \cdots + n_r}$$

Sustituyendo n'_1 por n_1p_1, n'_2 por n_2p_2, ..., n'_r por n_rp_r, nos queda

$$p = \frac{n_1p_1 + n_2p_2 + \cdots + n_rp_r}{n_1 + n_2 + \cdots + n_r}$$

EJEMPLO 5.9. En tres grupos distintos, con 270, 180 y 300 personas, la proporción de demócratas es 0,70, 0,65 y 0,62, respectivamente. Esto supuesto, ¿cuál es la proporción de demócratas en el grupo total?

$$p = \frac{(270)(0,70) + (180)(0,65) + (300)(0,62)}{270 + 180 + 300} = \frac{492}{750} = 0,656$$

La proporción de demócratas en el grupo total es 0,656 o, dicho de otro modo, el 65,6 por 100 del grupo total son demócratas.

5.2.4. Método abreviado para el cálculo de la media

Supongamos n puntuaciones agrupadas en intervalos, todos ellos con amplitud I. Sea X_0 el punto medio de uno de ellos, elegido arbitrariamente, al que llamaremos intervalo origen. Hagamos

$$A = \frac{1}{I}, \qquad B = \frac{X_0}{I}$$

Según 5.2.3. *c*), las puntuaciones

$$x'_j = \frac{1}{I}X_j + \left(-\frac{X_0}{I}\right) = \frac{X_j - X_0}{I}$$

tendrán como media

$$\bar{x}' = \frac{1}{I}\bar{X} - \frac{X_0}{I} = \frac{\bar{X} - X_0}{I}$$

De donde, $\bar{X} = I\bar{x}' + X_0$. Esta última fórmula nos permite obtener $\bar{X}$ mediante $\bar{x}'$, cuyo cálculo suele ser mucho más sencillo y breve que el de $\bar{X}$.

Si X_0 es el punto medio del intervalo origen, los puntos medios de los intervalos superiores al origen serán $X_0 + I$, $X_0 + 2I$, $X_0 + 3I$, ... y los puntos medios de los intervalos sucesivos inferiores al origen serán $X_0 - I$, $X_0 - 2I$, $X_0 - 3I$, ..., ya que la diferencia entre dos puntos medios consecutivos es igual a la amplitud, I, del intervalo. Consiguientemente la transformación $x'_j = \frac{X_j - X_0}{I}$ establece la siguiente correspondencia:

X_j	$x'_j = \frac{X_j - X_0}{I}$
.	.
.	.
.	.
$X_0 + 3I$	$\frac{(X_0 + 3I) - X_0}{I} = 3$
$X_0 + 2I$	$\frac{(X_0 + 2I) - X_0}{I} = 2$
$X_0 + I$	$\frac{(X_0 + I) - X_0}{I} = 1$
$X_0 + 0$	$\frac{(X_0 + 0) - X_0}{I} = 0$
$X_0 - I$	$\frac{(X_0 - I) - X_0}{I} = -1$
$X_0 - 2I$	$\frac{(X_0 - 2I) - X_0}{I} = -2$
$X_0 - 3I$	$\frac{(X_0 - 3I) - X_0}{I} = -3$
.	.
.	.

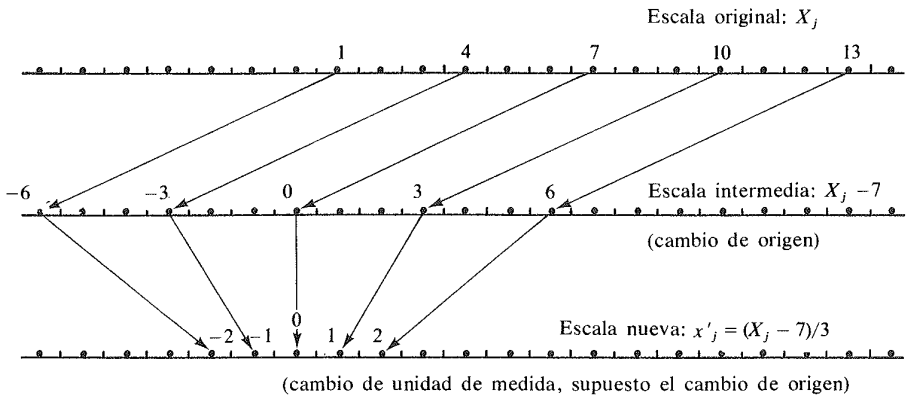
EJEMPLO 5.10. Apliquemos el método abreviado a la siguiente distribución de frecuencias, donde todos los intervalos tienen amplitud $I = 3$ y donde hemos elegido como origen el intervalo 6 – 8.

	n_j	x'_j	$n_jx'_j$
12-14	8	$\frac{13 - 7}{3} = 2$	16
9-11	12	$\frac{10 - 7}{3} = 1$	12
6-8	16	$\frac{7 - 7}{3} = 0$	0
3-5	10	$\frac{4 - 7}{3} = -1$	−10
0-2	4	$\frac{1 - 7}{3} = -2$	−8
	50		10

$$\bar{x}' = \frac{10}{50} = 0,2$$

$$\bar{X} = (3)(0,2) + 7 = 7,6$$

Veamos en qué ha consistido la transformación $x'_j = \dfrac{X_j - X_0}{I}$ en el ejemplo que acabamos de exponer.



(cambio de unidad de medida, supuesto el cambio de origen)

De lo dicho se infiere que es muy sencilla la manera práctica de transformar las X_j en las x'_j. En efecto, basta con elegir como origen un intervalo cualquiera, atribuyéndole como punto medio el valor 0. A continuación, atribuir un 1 al inmediatamente superior, un 2 al superior siguiente, etc.; atribuir un -1 al intervalo inmediatamente inferior, un -2 al inferior siguiente, etc. Es claro que en el ejemplo acabado de exponer no compensa utilizar el método abreviado puesto que el no abreviado es tan sencillo como el abreviado. Pero en otros casos no sucede esto. La diferencia suele ser de cierta importancia cuando las puntuaciones contienen cifras decimales, cuando son muchas las observaciones y grande el número de intervalos.

Desde luego, la introducción creciente de máquinas calculadoras y miniordenadores va haciendo cada vez menos necesario el uso de métodos abreviados. Igualmente, poseyendo dichos instrumentos de cálculo, deberemos obtener la media a partir de las puntuaciones originales, sin agruparlas en intervalos, siempre que las tengamos a nuestra disposición.

5.2.5. Media ponderada

La media de 2, 4 y 12 es 6. Pero, si por ejemplo, atribuimos a la primera puntuación un «peso» igual a 4,5, a la segunda un «peso» igual a 3,5 y a la tercera un «peso» igual a 2, tendremos la siguiente media «ponderada»

$$\frac{(4{,}5)(2) + (3{,}5)(4) + (2)(12)}{4{,}5 + 3{,}5 + 2} = \frac{47}{10} = 4{,}7$$

Es decir, la media ponderada de n puntuaciones es la media de esas puntuaciones multiplicadas o ponderadas por coeficientes o pesos apropiados. Nótese que atribuir los pesos 4,5; 3,5 y 2 no es más que hacer aparecer la primera puntuación 4,5 veces, la segunda 3,5 veces y la tercera 2 veces. Es, por tanto, como si tuviéramos $4{,}5 + 3{,}5 + 2 = 10$ puntuaciones. De aquí que el denominador de la media ponderada sea igual a la suma de los pesos.

En realidad, la media obtenida con datos agrupados en intervalos es una media ponderada. El peso asignado a cada puntuación (o punto medio del intervalo de que se trate) es igual al número de observaciones dentro de dicho intervalo.

$$\bar{X} = \frac{n_1 X_1 + n_2 X_2 + \cdots + n_r X_r}{n_1 + n_2 + \cdots + n_r}$$

5.2.6. Medias aritméticas generalizadas

a) *Media geométrica*

Llamamos media geométrica, $\bar{X}_g$, de n valores $X_1, X_2, \ldots, X_n$ a la raíz enésima del producto de esos n valores. Es decir, $\bar{X}_g = \sqrt[n]{(X_1)(X_2)\ldots(X_n)}$.

Mientras que la media aritmética se obtenía sumando n puntuaciones y dividiendo esa suma por n, la media geométrica se obtiene multiplicando las n puntuaciones y calculando la raíz enésima de ese producto. Es decir, lo que allí era suma, aquí es producto; lo que allí era división, aquí es radicación.

Usando logaritmos, tendremos

$$\log \bar{X}_g = \frac{\log X_1 + \log X_2 + \cdots + \log X_n}{n} = \frac{\Sigma \log X_i}{n}$$

En otras palabras, la media geométrica, $\bar{X}_g$, es un valor tal que su logaritmo es igual a la media aritmética de los logaritmos de los datos. De aquí que digamos que la media geométrica es una media aritmética generalizada.

EJEMPLO 5.11. En investigaciones sobre promedios de tiempos (latencias, tiempos de reacción, tiempo empleado en solución de problemas, ...) es frecuente la utilización de medias geométricas. Por tanto, si 22, 11, 7, 7, 28 son los tiempos (en segundos) empleados por una persona para la solución de cinco problemas (Sokolov, 1972), aceptaremos, como índice del tiempo medio consumido por esa persona en cada problema, la media geométrica de esos datos. Es decir,

$$\bar{X}_g = \sqrt[5]{(22)(11)(7)(7)(28)} = 12{,}71 \text{ segundos}$$

EJEMPLO 5.12. En Psicofísica suele preferirse la media geométrica a la media aritmética cuando se trata de promediar razones. Un caso típico es el siguiente.

A un sujeto se le impone la tarea de valorar las veces que cada uno de los estímulos supera a los restantes, según cierta característica, es decir, la tarea de ofrecer la la razón entre cada par de estímulos que él juzga más exacta. Con frecuencia, el método experimental utilizado permite establecer varias estimaciones de cada razón. Pues bien, para poder alcanzar una mejor estimación de ella, se suele calcular la media geométrica entre las distintas estimaciones de la misma. (Véase, por ejemplo, Torgerson, 1963.)

Sean 3; 2,5; 4; 3,5 y 3,8 cinco estimaciones de la razón entre los estímulos A y B. Es decir, A es juzgado como 3; 2,5; 4; 3,5 y 3,8 veces mayor que B, en cinco ocasiones distintas. Aceptaremos como estimación media de la razón entre A y B $\sqrt[5]{(3)(2,5)(4)(3,5)(3,8)} = 3,31$. En otras palabras, aceptaremos que el estímulo A es juzgado como 3,31 veces mayor que el estímulo B.

Si los datos estuvieran agrupados en intervalos,

$$X_g = \sqrt[n]{(X_1)^{n_1}(X_2)^{n_2} \ldots (X_r)^{n_r}}$$

donde $n_1, n_2, \ldots, n_r$ son las observaciones correspondientes a cada uno de los r intervalos (con $n_1 + n_2 + \cdots + n_r = n$) y donde $X_1, X_2, \ldots, X_r$ son los puntos medios de dichos intervalos.

Usando logaritmos, tendremos

$$\log \bar{X}_g = \frac{n_1 \log X_1 + n_2 \log X_2 + \cdots + n_r \log X_r}{n} = \frac{\Sigma n_i \log X_i}{n}$$

b) *Media armónica*

Llamamos media armónica, $\bar{X}_a$, de n valores $X_1, X_2, \ldots, X_n$ al recíproco de la media aritmética de los recíprocos de esos n valores. Es decir,

$$\bar{X}_a = \frac{1}{\dfrac{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \cdots + \dfrac{1}{X_n}}{n}} = \frac{n}{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \cdots + \dfrac{1}{X_n}}$$

De la expresión anterior se deduce que

$$\frac{1}{\bar{X}_a} = \frac{\dfrac{1}{X_1} + \dfrac{1}{X_2} + \cdots + \dfrac{1}{X_n}}{n}$$

En otras palabras, la media armónica, $\bar{X}_a$, es un valor tal que su recíproco es igual a la media aritmética de los recíprocos de los datos. De aquí que digamos que la media armónica es una media aritmética generalizada.

EJEMPLO 5.13. Sea un impulso nervioso eferente que recorre 0,40 m para alcanzar el músculo y provocar una contracción. Debido a las características de la

vía nerviosa por la que ha de transmitirse, la mitad de la distancia la recorre a una velocidad de 2 m/seg y el resto a una velocidad de 10 m/seg. (Véase Osgood, 1953.) La velocidad media con que se traslada el impulso es la media armónica de las dos velocidades. En efecto, de acuerdo con la ecuación fundamental de la cinemática, $e = vt \left(\text{ó, } t = \dfrac{e}{v}\right)$, en los primeros 0,20 m se verificará $t_1 = \dfrac{0,20}{2}$ y en los últimos 0,20 m se verificará $t_2 = \dfrac{0,20}{10}$. Consiguientemente, la velocidad media buscada será igual al espacio total (0,20 + 0,20) recorrido, dividido por el tiempo $(t_1 + t_2)$ empleado en recorrerlo. Es decir,

$$v_m = \frac{(2)(0,20)}{\dfrac{0,20}{2} + \dfrac{0,20}{10}} = \frac{2}{\dfrac{1}{2} + \dfrac{1}{10}} = 3,33 \text{ m/seg}$$

que es la media armónica de las dos velocidades.

Si los datos estuvieran agrupados en intervalos,

$$\bar{X}_a = \frac{1}{\dfrac{\dfrac{n_1}{X_1} + \dfrac{n_2}{X_2} + \cdots + \dfrac{n_r}{X_r}}{n}} = \frac{n}{\dfrac{n_1}{X_1} + \dfrac{n_2}{X_2} + \cdots + \dfrac{n_r}{X_r}}$$

donde $n_1, n_2, \ldots, n_r$ son las observaciones correspondientes a cada uno de los r intervalos (con $n_1 + n_2 + \cdots + n_r = n$) y donde $X_1, X_2, \ldots, X_r$ son los puntos medios de dichos intervalos.

c) *Media cuadrática*

Llamamos media cuadrática, $\bar{X}_c$, de n valores $X_1, X_2, \ldots, X_n$ a la raíz cuadrada de la media aritmética de los cuadrados de esos n valores. Es decir,

$$\bar{X}_c = \sqrt{\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}}$$

De la expresión anterior se deduce que

$$\bar{X}_c^2 = \frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}$$

En otras palabras, la media cuadrática, $\bar{X}_c$, es un valor tal que su cuadrado es igual a la media aritmética de los cuadrados de los datos. De aquí que digamos que la media cuadrática es una media aritmética generalizada.

EJEMPLO 5.14. Calculemos la media cuadrática de 1, 0, 3, −1, 3.

$$\bar{X}_c = \sqrt{\frac{1 + 0 + 9 + 1 + 9}{5}} = \sqrt{\frac{20}{5}} = 2$$

Si los datos estuvieran agrupados en intervalos,

$$\bar{X}_c = \sqrt{\frac{n_1 X_1^2 + n_2 X_2^2 + \cdots + n_r X_r^2}{n}}$$

donde $n_1, n_2, \ldots, n_r$ son las observaciones correspondientes a cada uno de los r intervalos (con $n_1 + n_2 + \cdots + n_r = n$) y donde $X_1, X_2, \ldots, X_r$ son los puntos medios de dichos intervalos.

5.2.7. NOTA

La segunda propiedad de la media, «la suma de las diferencias cuadráticas de n puntuaciones respecto a su media es mínima», queda legitimada más sencillamente derivando $F(k) = \Sigma (X_i - k)^2$ respecto a k. En efecto, $F'(k) = -2 \Sigma (X_i - k)$, $F''(k) = 2n$.

Con estos supuestos,

a) Si $k = \bar{X}$, $F'(\bar{X}) = -2 \Sigma (X_i - \bar{X}) = 0$, $F''(\bar{X}) = 2n > 0$. Por tanto, para $k = \bar{X}$, F pasa por un mínimo.

b) Si F pasa por un mínimo, $F'(k) = -2 \Sigma (X_i - k) = 0$. Por tanto, $\Sigma X_i = nk$,

$$k = \frac{\Sigma X_i}{n} = \bar{X}$$

5.3. Mediana

5.3.1. Introducción previa

Suponemos que, antes de calcular la mediana, hemos ordenado las puntuaciones de menor a mayor o de mayor a menor. Ordinariamente, lo haremos de menor a mayor. Nos limitamos a variables cuantitativas continuas, porque, aunque es posible calcularla en el caso de variables cuantitativas discretas y aun en el de variables cuasi-cuantitativas (ordinales), sin embargo, de hecho, casi siempre suele ser calculada para variables cuantitativas continuas.

5.3.2. Definición

Punto o valor numérico que deja por encima y por debajo de sí el 50 por 100 de las observaciones.

5.3.3. Cálculo

a) *Datos agrupados en intervalos*

Comenzamos con datos agrupados en intervalos porque la aplicación del concepto propuesto de mediana es más obvia en este caso que en el de datos no agrupados.

EJEMPLO 5.15. Engelman (1970) investigó la influencia de la educación escolar sobre el cociente intelectual (CI) en un grupo de niños con bajo nivel social y cultural. Para ello les sometió a un programa educativo intensivo durante dos años y midió sus cocientes intelectuales al principio y al fin del programa. Comenzaron el programa 15 niños de cuatro años. Lo concluyeron 12. Los datos (agrupados por nosotros en intervalos) de los 15 niños al empezar son los siguientes:

TABLA 5.1

	CI	n_j
129,5		
	115-129	1
114,5		
	100-114	4
99,5		
	85-99	8
84,5		
	70-84	2
69,5		
		15

A partir de ellos, calculemos la mediana.

Comenzamos aceptando que dentro de cada intervalo, las observaciones en él contenidas se distribuyen homogéneamente. Es decir, que si en un intervalo tenemos cinco observaciones, suponemos que cada una de ellas ocupa la quinta parte del mismo.

La mediana tiene que ser un punto o valor numérico mayor que 69,5 y menor que 129,5, tal que deje por encima y por debajo de sí el 50 por 100 de las 15 observaciones, es decir, 7,5 observaciones. Evidentemente, 84,5 no puede ser la mediana, pues deja por debajo de sí 2 observaciones (y por encima 13). Tampoco 99,5 será

mediana, pues deja por debajo de sí 10 observaciones (y por encima 5). La mediana tiene que ser un valor entre 84,5 y 99,5. Llamemos «crítico» al intervalo (84,5 – – 99,5) y dibujémoslo ampliado (Fig. 5.1).

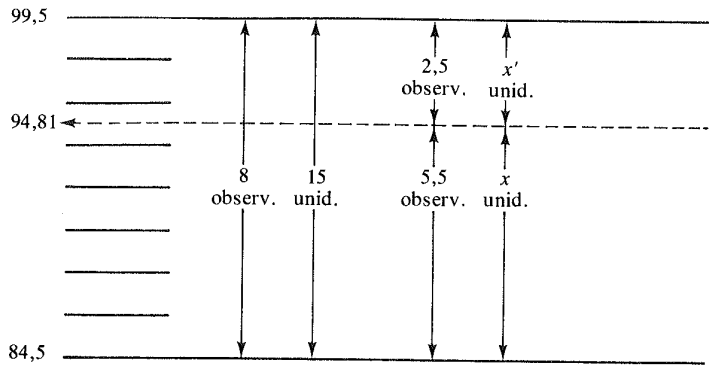


Fig. 5.1

Si las ocho observaciones del intervalo (84,5 – 99,5) ocupan homogéneamente una distancia igual a 15 unidades, las 5,5 observaciones (que con las 2 inferiores forman el 50 por 100) ocuparán una distancia igual a x unidades. Es decir,

$$\frac{8}{5,5} = \frac{15}{x}, \quad x = \frac{(5,5)(15)}{8} = 10,31$$

Por tanto, $Md = 84,5 + 10,31 = 94,81$

La vertical levantada sobre la mediana divide el área total en dos áreas de igual superficie (Fig. 5.2).

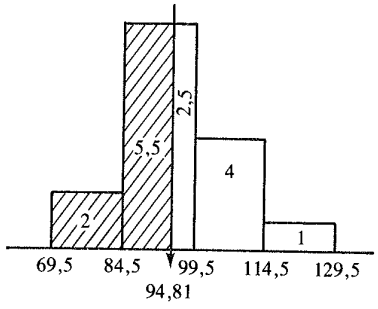


Fig. 5.2

Nótese que

$$Md = 84,5 + 10,31 = 84,5 + \frac{(5,5)(15)}{8} = 84,5 + \frac{7,5 - 2}{8}15 =$$

$$84,5 + \left(\frac{\frac{15}{2} - 2}{8}\right)15 = 94,81$$

Ahora bien,

$84,5 = l_i$: límite exacto inferior del intervalo «crítico».

$\frac{15}{2} = \frac{n}{2}$: mitad o 50 por 100 de las observaciones.

$2 = n_b$: número de observaciones *bajo* el intervalo «crítico».

$8 = n_d$: número de observaciones *dentro* del intervalo «crítico».

$15 = I$: amplitud del intervalo «crítico»*.

Como el razonamiento expuesto en este caso particular es válido para cualquier otro caso, podemos aceptar como fórmula de la mediana:

$$Md = l_i + \left(\frac{\frac{n}{2} - n_b}{n_d}\right)I \tag{5.5}$$

Podíamos, también, haber pensado así: Si las ocho observaciones del intervalo (84,5 – 99,5) ocupan homogéneamente una distancia igual a 15 unidades, las 2,5 observaciones (que con las cinco superiores forman el 50 por 100) ocuparán una distancia igual a x' unidades. Es decir,

$\frac{8}{2,5} = \frac{15}{x'}$, $x' = \frac{(2,5)(15)}{8} = 4,69$. Por tanto, $Md = 99,5 - 4,69 = 94,81$.

Nótese que

$$Md = 99,5 - 4,69 = 99,5 - \frac{(2,5)(15)}{8} = 99,5 - \frac{7,5 - 5}{8}15 =$$

$$= 99,5 - \left(\frac{\frac{15}{2} - 5}{8}\right)15 = 94,81$$

* Es puramente casual que coincida el valor del número de observaciones ($n = 15$) con el valor correspondiente a la amplitud del intervalo «crítico» ($I = 15$). Veremos cómo en el ejemplo 5.16 no se da tal coincidencia.

Ahora bien,

99,5 = l_s: límite exacto superior del intervalo «crítico».

5 = n_s: número de observaciones *sobre* el intervalo «crítico».

$\frac{15}{2}$, 8 y 15 significan lo mismo que en el caso anterior.

Por consiguiente, podemos proponer la siguiente como fórmula alternativa para la mediana

$$Md = l_s - \left(\frac{\frac{n}{2} - n_s}{n_d}\right) I \qquad (5.6)$$

EJEMPLO 5.16. Calculemos la mediana a partir de los datos (agrupados en intervalos por nosotros) correspondientes a los 12 niños que concluyeron los dos años del programa intensivo. La tabla adjunta nos muestra los *CI* de dichos niños al final del segundo año.

TABLA 5.2

	CI	n_j
142,5		
	133-142	2
132,5		
	123-132	3
122,5		
	113-122	3
112,5		
	103-112	4
102,5		
		12

La mediana tiene que ser un punto o valor numérico mayor que 102,5 y menor que 142,5, tal que deje por encima y por debajo de sí el 50 por 100 de las 12 observaciones, es decir, seis observaciones. Evidentemente, 112,5 no puede ser la mediana, pues deja por debajo de sí cuatro observaciones (y por encima ocho). Tampoco 122,5 será mediana, pues deja por debajo de sí siete observaciones (y por encima cinco). La mediana tiene que ser un valor entre 112,5 y 122,5. Llamemos «crítico» al intervalo (112,5 – 122,5) y dibujémoslo ampliado (Fig. 5.3).

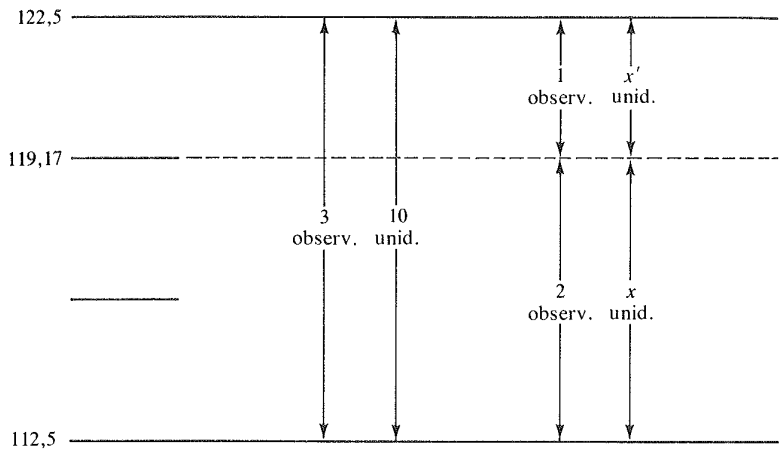


Fig. 5.3

Si las tres observaciones del intervalo (112,5 – 122,5) ocupan homogéneamente una distancia igual a diez unidades, las dos observaciones (que con las cuatro inferiores forman el 50 por 100) ocuparán una distancia igual a x unidades. Es decir,

$$\frac{3}{2} = \frac{10}{x}, \quad x = \frac{(2)(10)}{3} = 6{,}67$$

Por tanto, $Md = 112{,}5 + 6{,}67 = 119{,}17$.

Por supuesto, a este mismo resultado habríamos llegado aplicando (5.5). En efecto,

$$Md = 112{,}5 + \frac{\frac{12}{2} - 4}{3} 10 = 112{,}5 + 6{,}67 = 119{,}17$$

Podíamos, también, haber pensado así: Si las tres observaciones del intervalo (112,5 – 122,5) ocupan homogéneamente una distancia igual a 10 unidades, la única observación (que con las cinco superiores forma el 50 por 100) ocupará una distancia gual a x' unidades. Es decir,

$$\frac{3}{1} = \frac{10}{x'}, \quad x' = \frac{(1)(10)}{3} = 3{,}33$$

Por tanto, $Md = 122{,}5 - 3{,}33 = 119{,}17$

Por supuesto, a este mismo resultado habríamos llegado aplicando (5.6). En efecto,

$$Md = 122{,}5 - \frac{\frac{12}{2} - 5}{3} 10 = 122{,}5 - 3{,}33 = 119{,}17$$

La vertical levantada sobre la mediana divide el área total en dos áreas de igual superficie. (Fig. 5.4.)

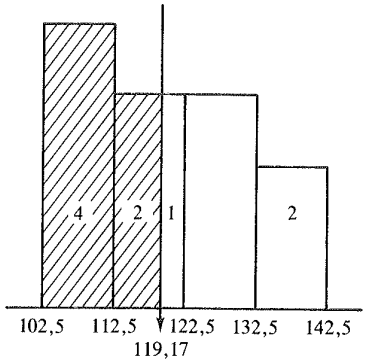


Fig. 5.4

La diferencia entre la mediana al principio del experimento y la mediana al fin del mismo (119,17 − 94,81 = 24,36 puntos del *CI*) manifiesta la influencia de la educación escolar sobre el *CI*.

b) *Datos no agrupados en intervalos*

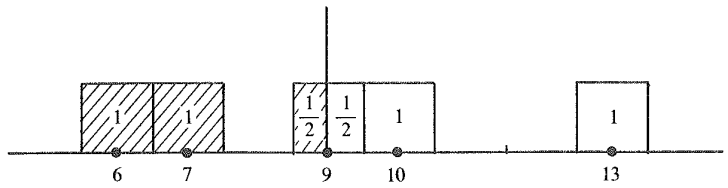
No es más que una aplicación de lo dicho en el apartado anterior para el caso particular en que $I = 1$ y $n_d = 1$ o $n_d = k$, según que exista una sola o existan k observaciones dentro del intervalo unitario «crítico». Consideremos por separado las dos alternativas: número impar de observaciones y número par de las mismas.

Número impar de observaciones. Tendremos un solo lugar «central».

1) La puntuación que ocupa el lugar «central» es distinta de todas las demás.

EJEMPLO 5.17. Calculemos la mediana de 6, 7, 9, 10, 13. La puntuación que ocupa el lugar tercero (el «central») es 9, que es distinta de todas las restantes. El intervalo «crítico» es el intervalo unitario (8,5 − 9,5). Por tanto,

$$Md = 8{,}5 + \left(\frac{\frac{5}{2} - 2}{1}\right) 1 = 8{,}5 + 0{,}5 = 9$$

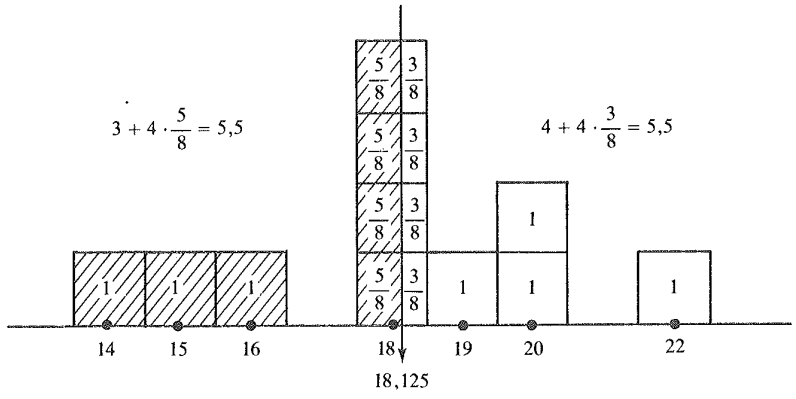


Es decir, la mediana es la puntuación que ocupa el lugar «central». La vertical levantada sobre la mediana (9) divide el área total en dos áreas de igual superficie (2 + 0,5 = 2,5; 2 + 0,5 = 2,5).

2) La puntuación que ocupa el lugar «central» es igual a una o varias de las restantes.

EJEMPLO 5.18. Calculemos la mediana de 14, 15, 16, 18, 18, 18, 18, 19, 20, 20, 22. La puntuación que ocupa el lugar sexto (el «central») es 18 que es igual a las que ocupan los lugares cuarto, quinto y séptimo. El intervalo «crítico» es el intervalo unitario (17,5 − 18,5). Por tanto,

$$Md = 17{,}5 + \left(\frac{\frac{11}{2} - 3}{4}\right) 1 = 17{,}5 + \frac{2{,}5}{4} = 18{,}125$$



La vertical levantada sobre la mediana (18,125) divide el área total en dos áreas de igual superficie (3 + 2,5 = 5,5; 4 + 1,5 = 5,5).

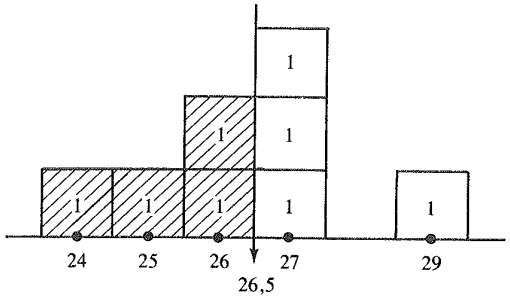
Número par de observaciones. Tendremos dos lugares «centrales».

1) Las puntuaciones que ocupan los lugares «centrales» son distintas entre sí y ambas son consecutivas o no lo son.

Bajo una y otra condición es posible seguir aplicando la fórmula general eligiendo como intervalo «crítico» unitario, bien el que contiene la puntuación central inferior, bien el que contiene la superior. Sin embargo, sólo si las puntuaciones centrales son consecutivas, el valor encontrado para la mediana es único y, además, coincide con la media aritmética de las dos puntuaciones centrales. Por el contrario, si no son consecutivas, el valor obtenido para la mediana es uno si elegimos como intervalo «crítico» unitario el que contiene la puntuación central inferior y es otro si elegimos el que contiene la puntuación central superior. Son, además, posibles otros infinitos valores distintos para la mediana, situados entre los dos anteriores. Todos ellos verifican la definición de mediana. Pero con el fin de fijar un valor único para la mediana, se suele tomar, arbitrariamente, como tal la media aritmética de las dos puntuaciones centrales.

En conclusión, en uno y otro caso la mediana será la media aritmética de las dos puntuaciones centrales.

EJEMPLO 5.19. Calculemos la mediana de 24, 25, 26, 26, 27, 27, 27, 29. Las puntuaciones que ocupan los lugares cuarto y quinto (los dos «centrales») son 26 y 27. Por tanto, la mediana valdrá 26,5, media aritmética de 26 y 27.



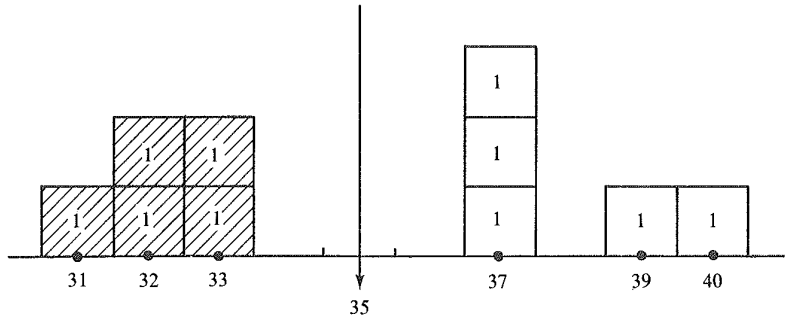
Si el lector aplica la fórmula general eligiendo como «crítico» tanto el intervalo (25,5 – 26,5) como el intervalo (26,5 – 27,5) llegará al mismo resultado, a saber, $Md = 26{,}5$.

La vertical levantada sobre la mediana (26,5) divide el área total en dos áreas de igual superficie (4, 4).

EJEMPLO 5.20. Calculemos la mediana de 31, 32, 32, 33, 33, 37, 37, 37, 39, 40. Las puntuaciones que ocupan los lugares quinto y sexto (los dos «centrales») son 33 y 37. Por tanto, la mediana valdrá 35, media aritmética de 33 y 37.

Si el lector aplica la fórmula general eligiendo como «crítico» el intervalo (32,5 – 33,5) llegará a $Md = 33{,}5$. Si, en cambio, elige como «crítico» el intervalo (36,5 – 37,5) llegará a $Md = 36{,}5$. Ciertamente, tanto 33,5 como 36,5 (y los infinitos situados entre ambos) verifican la definición de mediana. Sin embargo, solemos elegir como mediana el situado en el punto medio de todos ellos, es decir, 35.

La vertical levantada sobre la mediana (35) divide el área total en dos áreas de igual superficie (5, 5).



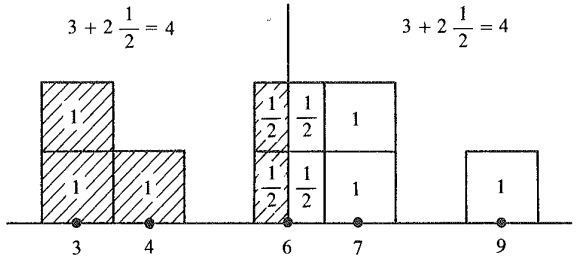
2) Las puntuaciones que ocupan los lugares «centrales» son iguales entre sí, pero distintas de todas las demás.

EJEMPLO 5.21. Calculemos la mediana de 3, 3, 4, 6, 6, 7, 7, 9. Las puntuaciones que ocupan los lugares cuarto y quinto (los «centrales») son las dos iguales a 6 y distintas de las restantes. El intervalo «crítico» es el intervalo unitario (5,5 – 6,5).

Por tanto,

$$Md = 5{,}5 + \left(\frac{\frac{8}{2} - 3}{2}\right) 1 = 5{,}5 + 0{,}5 = 6$$

Es decir, la mediana es el valor común a las dos puntuaciones «centrales». La vertical levantada sobre la mediana (6) divide el área total en dos áreas de igual superficie (3 + 1 = 4; 3 + 1 = 4).

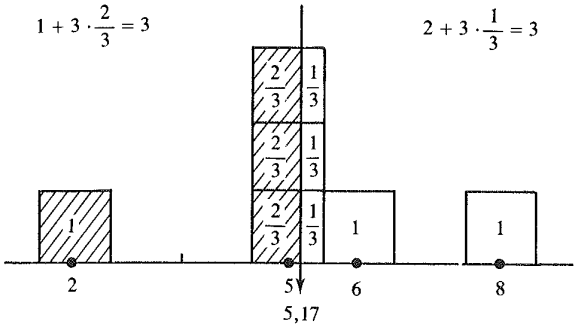


3) Las puntuaciones que ocupan los lugares «centrales» son iguales entre sí e iguales a una o varias de las restantes.

EJEMPLO 5.22. Calculemos la mediana de 2, 5, 5, 5, 6, 8. Las puntuaciones que ocupan los lugares tercero y cuarto (los «centrales») son iguales las dos a 5 e iguales a la que ocupa el lugar segundo. El intervalo «crítico» es el intervalo unitario (4,5 – 5,5). Por tanto,

$$Md = 4{,}5 + \left(\frac{\frac{6}{2} - 1}{3}\right) 1 = 4{,}5 + 0{,}67 = 5{,}17$$

La vertical levantada sobre la mediana (5,17) divide el área total en dos áreas de igual superficie (1 + 2 = 3; 2 + 1 = 3).



NOTA. Quizá, sean un poco superfluas las distinciones acabadas de exponer para el caso de datos no agrupados en intervalos y baste con el siguiente criterio:

a) Si el número de observaciones es impar, la mediana es la puntuación de la observación que ocupa el lugar «central».

b) Si el número de observaciones es par, la mediana es la media aritmética de las puntuaciones correspondientes a las dos observaciones que ocupan los dos valores «centrales».

5.3.4. Propiedades

a) La suma de las diferencias (en valor absoluto) de n puntuaciones respecto a su mediana es igual o menor que la suma de las diferencias (en valor absoluto) de esas puntuaciones respecto a cualquier otro valor. Para la legitimación de esta propiedad pueden ser consultados Freeman (1963), Horst (1966), Calot (1969). Aquí nos limitaremos a comprobarla con un ejemplo muy sencillo.

EJEMPLO 5.23. Dadas las puntuaciones 2, 4 y 9, calculemos la suma de sus diferencias (en valor absoluto) respecto de la media y de la mediana.

Su media vale 5 y su mediana vale 4. Esto supuesto,

$$|2 - 5| + |4 - 5| + |9 - 5| = 8 \quad ; \qquad |2 - 4| + |4 - 4| + |9 - 4| = 7$$

Vemos, en efecto, cómo la suma de las diferencias (en valor absoluto) respecto a la media es mayor que la suma respecto a la mediana (la primera suma vale 8 y la segunda vale 7).

b) Es menos sensible que la media a la variación de cada una de las puntuaciones. Aunque la variación de una sola puntuación (elegida a propósito) puede hacer variar la mediana, sin embargo, la variación, de modo bastante anárquico, de la mayoría y aun de todas las puntuaciones pueden dejar invariante la mediana. Así, por ejemplo, (2, 59, 61, 945) y (57, 58, 62, 63) tienen la misma mediana (60).

c) Es función de los intervalos elegidos (de su amplitud, de su número y de los límites de los mismos).

d) Es fundamento de diversas técnicas estadísticas. Sin embargo, el número de éstas es mucho menor que el de las técnicas basadas sobre la media.

e) Puede ser calculada aunque el intervalo máximo no tenga límite superior, ni el intervalo mínimo lo tenga inferior (o, al menos, uno de los dos carezca de su correspondiente límite extremo). Con todo, cuando uno de esos dos intervalos sin límite extremo contenga dentro de sí más del 50 por 100 de los casos, tampoco podrá ser calculada la mediana. Así, será imposible calcular la mediana, dada la siguiente distribución de frecuencias

X	n_j
46-48	2
43-45	1
40-42	2
37-39	14
36 ó menos	20

f) La mediana es un punto tal, que la vertical levantada sobre el mismo divide el área total del histograma en dos áreas con idéntica superficie.

g) Es más recomendable que la media cuando la distribución de frecuencias es muy asimétrica.

h) Dados r grupos con medianas Md_1, Md_2, . . . , Md_r, la mediana del grupo total es igual o mayor que la mediana mínima e igual o menor que la máxima. (Véase Calot, 1969.)

5.4. Moda

5.4.1. Definición

Nivel de intervalos o razón

a) Datos no agrupados en intervalos.

Puntuación a la que corresponde frecuencia máxima. Es decir, la puntuación

que más veces se repite. Así, por ejemplo, valdrá 2 la moda de las puntuaciones: 2, 5, 7, 2, 4, 2, 6, 6, 4, 2, 2, 3, 2.

b) Datos agrupados en intervalos.

Punto medio del intervalo al que corresponde frecuencia máxima.

EJEMPLO 5.24. Conde y Doménech (1976) estudiando una muestra de pacientes esquizofrénicos encontraron la siguiente distribución de frecuencias:

Edad	Total	Varones	Mujeres
80-89	9	3	6
70-79	11	2	9
60-69	15	9	6
50-59	16	5	11
40-49	7	3	4
30-39	11	9	2
20-29	31	25	6
10-19	7	6	1

La moda para el grupo total y para el de los varones vale 24,5 años (punto medio del intervalo 20 – 29) y vale 54,5 años (punto medio del intervalo 50 – 59) para el de las mujeres.

Nivel ordinal

Valor o categoría ordinal a los que corresponde frecuencia máxima.

EJEMPLO 5.25. Conde y Doménech (1976) en la misma investigación acabada de citar encontraron que el nivel cultural de 97 personas, de la muestra estudiada, era el siguiente:

Nivel cultural		Total	Varones	Mujeres
Superior	(4)	14	13	1
Medio	(3)	19	10	9
Inferior	(2)	63	30	33
Nulo	(1)	1	0	1

La moda es la categoría «inferior» (o valor ordinal 2) tanto para el grupo total, como para el grupo de los varones y de las mujeres por separado. Es la categoría «inferior» a la que corresponde frecuencia máxima. En otras palabras, el grupo con nivel cultural «inferior» es el más numeroso.

Nivel nominal

Modalidad o categoría nominal a las que corresponde frecuencia máxima.

EJEMPLO 5.26. Conde y Doménech (1976) proponen los siguientes resultados respecto a su muestra de esquizofrénicos.

Estado civil	Total	Varones	Mujeres
Solteros	80	52	28
Casados	20	10	10
Viudos	7	0	7

La moda es la categoría «solteros» tanto para el grupo total, como para el grupo de los varones y de las mujeres por separado. Es la categoría «solteros» a la que corresponde frecuencia máxima. En otras palabras, la mayoría de los esquizofrénicos estudiados son solteros.

5.4.2. Propiedades

a) Es muy sencilla de calcular.

b) Tiene el inconveniente de no ser necesariamente única. Dentro de una misma distribución de frecuencias pueden aparecer dos o más valores o dos o más categorías a los que corresponda frecuencia máxima.

c) Es función de los intervalos elegidos (de su amplitud, de su número y de los límites de los mismos).

d) Puede ser calculada aunque el intervalo máximo no tenga límite superior ni el mínimo lo tenga inferior (o, al menos, uno de los dos carezca de su correspondiente límite extremo). Con todo, cuando uno de esos dos intervalos sin límite extremo contenga dentro de sí la frecuencia máxima, tampoco será calculable la moda.

5.5. Percentiles

5.5.1. Definición

Hemos definido la mediana como un valor numérico que deja por debajo de sí el 50 por 100 de las observaciones. Pues bien, definiremos el percentil k como un valor numérico que deja por debajo de sí el k por 100 de las observaciones. Así, decir que 35 es el percentil 72 equivale a decir que dicho valor numérico deja por debajo de sí el 72 por 100 de las observaciones del grupo de que se trate.

Todo percentil sigue siendo índice de posición. Nos indica la posición dentro del grupo del objeto (persona, animal, cosa, familia, entidad, etc.) que tiene como puntuación dicho percentil. Si, por ejemplo, dicho objeto tiene como puntuación el percentil 80, sabemos que deja por debajo de sí el 80 por 100 de los objetos de su grupo. O, expuesto de otro modo, que su puntuación supera al 80 por 100 de las puntuaciones del grupo. Sin embargo, ya no podemos decir que sea índice de tendencia central. Un percentil, por su propia naturaleza, es una puntuación que puede estar situada tanto hacia el centro de las restantes puntuaciones de un grupo, como en uno cualquiera de los dos extremos.

5.5.2. Cálculo

Se siguen pasos análogos a los seguidos para el cálculo de la mediana. En realidad, la mediana es un percentil determinado, el percentil 50. Al igual que ésta, los percentiles son calculables a nivel ordinal. No obstante, sólo suelen ser aplicados a variables cuantitativas y, más concretamente, continuas.

Propondremos el cálculo de percentiles únicamente en el caso de datos agrupados en intervalos. El lector se encargará de hacer consideraciones análogas a las verificadas para el cálculo de la mediana en el caso de datos no agrupados en intervalos.

EJEMPLO 5.27. Cravioto y Robles (1965), mediante el test de Gesell, encontraron los siguientes cocientes de desarrollo motor en un grupo de niños de corta edad que sufrían una fuerte desnutrición de proteínas: 67, 25, 20, 20, 33, 33, 40, 69, 75, 42, 38, 46, 37, 52, 31, 57, 40, 39, 7, 26. Calculemos el percentil 60 de estos datos después de haberlos agrupado en los siguientes intervalos:

	X	n_j
80,5		
	66-80	3
65,5		
	51-65	2
50,5		
	36-50	7
35,5		
	21-35	5
20,5		
	6-20	3
5,5		
		20

Comenzamos aceptando, al igual que lo hicimos al tratar de la mediana, que dentro de cada intervalo las observaciones, en él contenidas, se distribuyen homogéneamente.

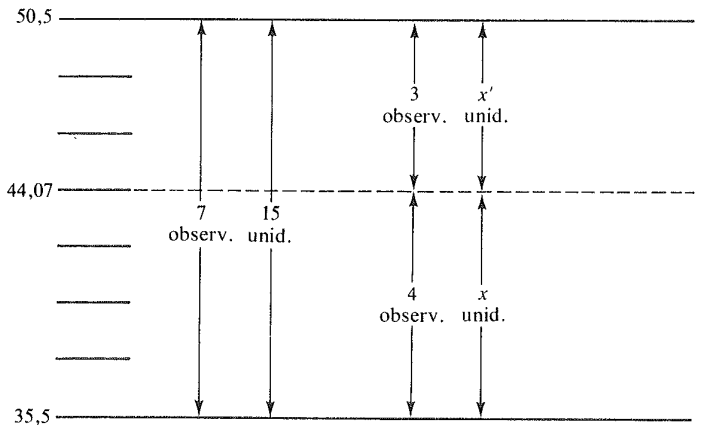


Fig. 5.5

El percentil 60, P_{60}, tiene que ser un punto o valor numérico mayor que 5,5 y menor que 80,5, tal que deje por debajo de sí el 60 por 100 de las 20 observaciones, es decir, $\frac{(60)(20)}{100} = 12$ observaciones. Evidentemente 35,5 no puede ser P_{60}, pues deja por debajo de sí ocho observaciones. Tampoco 50,5 será P_{60}, pues deja por debajo de sí 15 observaciones. P_{60} tiene que ser un valor entre 35,5 y 50,5. Llamemos «crítico» al intervalo (35,5 – 50,5) y dibujémoslo ampliado (Fig. 5.5).

Si las siete observaciones del intervalo (35,5 – 50,5) ocupan homogéneamente una distancia igual a 15 unidades, las cuatro observaciones (que con las ocho inferiores forman el 60 por 100) ocuparán una distancia igual a x unidades. Es decir,

$$\frac{7}{4} = \frac{15}{x} \quad ; \quad x = \frac{(4)(15)}{7} = 8{,}57$$

Por tanto, $P_{60} = 35{,}5 + 8{,}57 = 44{,}07$.

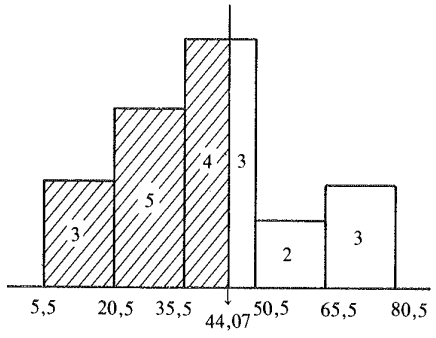


Fig. 5.6

La vertical levantada sobre P_{60} divide el área total en dos áreas iguales al 60 por 100 (la de la izquierda) y al 40 por 100 (la de la derecha) respectivamente (3 + 5 + 4 = 12; 3 + 2 + 3 = 8) (Fig. 5.6).

Nótese que

$$P_{60} = 35{,}5 + 8{,}57 = 35{,}5 + \frac{(4)(15)}{7} = 35{,}5 + \frac{12 - 8}{7}15 =$$

$$= 35{,}5 + \left(\frac{\frac{(60)(20)}{100} - 8}{7}\right)15 = 44{,}07$$

Ahora bien,

$35{,}5 = l_i$: límite exacto inferior del intervalo «crítico».

$\frac{(60)(20)}{100} = \frac{(k)(n)}{100}$: k por 100 de las observaciones.

$8 = n_b$: número de observaciones *bajo* el intervalo «crítico».

$7 = n_d$: número de observaciones *dentro* del intervalo «crítico».

$15 = I$: amplitud del intervalo «crítico».

Como el razonamiento expuesto en este caso particular es válido para cualquier otro caso, podemos aceptar como fórmula del percentil k, P_k:

$$P_k = l_i + \left(\frac{\frac{(k)(n)}{100} - n_b}{n_d}\right) I \qquad (5.7)$$

Podíamos, también, haber pensado así: Si las siete observaciones del intervalo (35,5 – 50,5) ocupan homogéneamente una distancia igual a 15 unidades, las tres observaciones (que con las cinco superiores forman el 40 por 100) ocuparán una distancia igual a x' unidades. Es decir,

$$\frac{7}{3} = \frac{15}{x'}, \quad x' = \frac{(3)(15)}{7} = 6{,}43$$

Por tanto, $P_{60} = 50{,}5 - 6{,}43 = 44{,}07$.

Nótese que

$$P_{60} = 50{,}5 - 6{,}43 = 50{,}5 - \frac{(3)(15)}{7} = 50{,}5 - \frac{8 - 5}{7}15 =$$

$$= 50{,}5 - \left(\frac{\frac{(40)(20)}{100} - 5}{7}\right)15 = 44{,}07$$

Ahora bien,

$50{,}5 = l_s$: límite superior del intervalo «crítico».

$\frac{(40)(20)}{100} = \frac{(k')(n)}{100}$: k' por 100 de las observaciones, donde $k' = 100 - k$.

$5 = n_s$: número de observaciones *sobre* el intervalo «crítico».

7 y 15 representan lo mismo que en el caso anterior.

Por consiguiente, podemos proponer la siguiente como fórmula alternativa para P_k.

$$P_k = l_s - \left(\frac{\frac{(k')(n)}{100} - n_s}{n_d}\right) I \qquad (5.8)^*$$

5.6. Resumen: Definiciones y fórmulas

Media aritmética: Llamamos media aritmética de n valores a la suma de ellos dividida por n.

$$\bar{X} = \frac{\Sigma X_i}{n} \qquad \text{(datos no agrupados en intervalos)}$$

$$\bar{X} = \frac{\Sigma n_j X_j}{n} \qquad \text{(datos agrupados en intervalos)}$$

Medias aritméticas generalizadas

a) *Media geométrica*: Llamamos media geométrica de n valores a la raíz enésima del producto de esos n valores.

$$\bar{X}_g = \sqrt[n]{(X_1)(X_2)\ldots(X_n)} \qquad \text{(datos no agrupados en intervalos)}$$

$$\bar{X}_g = \sqrt[n]{(X_1)^{n_1}(X_2)^{n_2}\ldots(X_r)^{n_r}} \qquad \text{(datos agrupados en intervalos)}$$

* El percentil 25 suele ser llamado primer cuartil, el percentil 50 segundo cuartil (o mediana) y el percentil 75 suele denominarse tercer cuartil.

b) *Media armónica*: Llamamos media armónica de n valores al recíproco de la media aritmética de los recíprocos de esos n valores.

$$\bar{X}_a = \frac{1}{\frac{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}}{n}} = \frac{n}{\frac{1}{X_1} + \frac{1}{X_2} + \cdots + \frac{1}{X_n}}$$

(datos no agrupados en intervalos)

$$\bar{X}_a = \frac{1}{\frac{\frac{n_1}{X_1} + \frac{n_2}{X_2} + \cdots + \frac{n_r}{X_r}}{n}} = \frac{n}{\frac{n_1}{X_1} + \frac{n_2}{X_2} + \cdots + \frac{n_r}{X_r}}$$

(datos agrupados en intervalos)

c) *Media cuadrática*: Llamamos media cuadrática de n valores a la raíz cuadrada de la media aritmética de los cuadrados de esos n valores.

$$\bar{X}_c = \sqrt{\frac{X_1^2 + X_2^2 + \cdots + X_n^2}{n}}$$

(datos no agrupados en intervalos)

$$\bar{X}_c = \sqrt{\frac{n_1 X_1^2 + n_2 X_2^2 + \cdots + n_r X_r^2}{n}}$$

(datos agrupados en intervalos)

Mediana: Punto o valor numérico que deja por encima y por debajo de sí el 50 por 100 de las observaciones.

$$Md = l_i + \left(\frac{\frac{n}{2} - n_b}{n_d}\right) I = l_s - \left(\frac{\frac{n}{2} - n_s}{n_d}\right) I$$

Moda: Puntuación a la que corresponde frecuencia máxima. O punto medio del intervalo al que corresponde frecuencia máxima. (Si nos encontramos a nivel de intervalos o de razón.)

Categoría ordinal a la que corresponde frecuencia máxima. (Si nos encontramos a nivel ordinal.)

Categoría nominal a la que corresponde frecuencia máxima. (Si nos encontramos a nivel nominal.)

Percentil: Llamamos percentil k al punto o valor numérico que deja por debajo de sí el k por 100 de las observaciones.

$$P_k = l_i + \left(\frac{\frac{(k)(n)}{100} - n_b}{n_d}\right) I = l_s - \left(\frac{\frac{(100 - k)(n)}{100} - n_s}{n_d}\right) I$$

EJERCICIOS

5.1. Calcular la media aritmética a partir de los siguientes datos *no agrupados* en intervalos.

a) 8, 7, 2, 8, 6, 5, 2, 4
b) 18, 20, 18, 17, 24, 20, 23
c) 5, −1, 4, 8, −2
d) 2, −3, 8, −9, 0, 2, 1, −4, 0, −2
e) 1/2, 1, 2/5
f) 1/2, −3, −1/4, 2/5
g) 0,050, −0,200, 0,005, −0,010, −0,100.

5.2. Calcular la media aritmética a partir de los siguientes datos *agrupados* en intervalos

a)

X	n_j
11-13	3
8-10	6
5-7	7
2-4	4

b)

X	n_j
100-104	1
95-99	5
90-94	10
85-89	7
80-84	2

c)

X	n_j
80-83	8
76-79	12
72-75	14
68-71	6

d)

X	n_j
41-47	7
34-40	14
27-33	12
20-26	5

e)

X	n_j
45-49	6
40-44	10
35-39	17
30-34	18
25-29	10
20-24	10
15-19	9

f)

X	n_j
0,20-0,22	9
0,17-0,19	15
0,14-0,16	18
0,11-0,13	20
0,08-0,10	14
0,05-0,07	6

g)

X	n_j
19-21	2
16-18	4
13-15	7
10-12	8
7-9	6
4-6	5
1-3	4

5.3. Sabiendo que la media aritmética vale 8,85, calcular las dos frecuencias que faltan en el cuadro siguiente:

X	n_j	$n_j X_j$
14-16	1	
11-13		
8-10	8	
5-7	4	
2-4		
		177

5.4. La media aritmética de dos números vale 8 y uno de ellos es tres veces mayor que el otro, ¿cuánto valen ambos números?

5.5. Demostrar que, siendo $X_i = A_i + B_i + C_i$, $\bar{X} = \bar{A} + \bar{B} + \bar{C}$.

5.6. Deseamos transformar las puntuaciones 8, 13, 9, 15, 10 en otras (sumándoles a todas ellas una misma constante) de modo que su media valga 26. ¿Cuál debe ser esa constante aditiva?

5.7. Ocho personas han comparado dos frases, *A* y *B*, acerca del divorcio, valorando cuántas veces más favorece *A* que *B* al divorcio. Los datos obtenidos han sido los siguientes: 2; 2,5; 2,7; 1,8; 3; 1,5; 2; 2,5. ¿En qué proporción, por término medio, ha resultado ser más favorable al divorcio la frase *A* que la *B*?

5.8. Calcular dos números tales que su media aritmética valga 7,5 y su media geométrica valga 6.

5.9. Sea *k* el valor de uno de dos números cuya media aritmética es igual que su media armónica. Esto supuesto, ¿cuánto vale el otro número?

5.10. Un ciclista va de *A* a *B* a 20 km por hora y vuelve de *B* a *A* a 60 km por hora. Esto supuesto, ¿cuál es la velocidad media a la que ha recorrido la distancia total? (Sugerencia: utilizar la media armónica.)

5.11. Calcular el peso medio de un grupo de personas, conocido el número de personas y el peso medio de cada uno de los subgrupos en los que se divide el grupo total primero, teniendo en cuenta el cuadro siguiente:

Subgrupo «i»	Personas (n_i)	Media ($\bar{X}_i$)
1	150	75
2	220	60
3	180	65

5.12. Calcule la media aritmética en lenguaje del primer curso de Bachillerato de un centro escolar, sabiendo que está dividido en cuatro secciones, *A*, *B*, *C* y *D*, según el cuadro siguiente:

Sección	Número de alumnos	Nota media
A	20	10
B	50	6
C	60	5
D	30	8

5.13. Calcular la proporción de médicos, abogados e ingenieros dentro de un grupo de 800 personas, conociendo el número de personas y la proporción de médicos, abogados e ingenieros dentro de cada uno de los subgrupos en los que se divide el grupo primero, de acuerdo con el cuadro siguiente:

	Gr. 1.º, $n_1 = 210$	Gr. 2.º, $n_2 = 180$	Gr. 3.º, $n_3 = 220$	Gr. 4.º, $n_4 = 190$
Médicos	0,25	0,30	0,35	0,33
Abogados	0,50	0,52	0,45	0,40
Ingenieros	0,25	0,18	0,20	0,27
	1,00	1,00	1,00	1,00

5.14. Calcular la proporción de solteros, casados y viudos dentro de un grupo de 150 adultos, conociendo el número de personas y la proporción de solteros, casados y viudos dentro de cada uno de los subgrupos en los que se divide el grupo primero, de acuerdo con el cuadro siguiente:

	Gr. 1.º, $n_1 = 60$	Gr. 2.º, $n_2 = 90$
Solteros	0,30	0,25
Casados	0,45	0,55
Viudos	0,25	0,20
	1,00	1,00

5.15. Calcular la mediana a partir de los siguientes datos *agrupados* en intervalos:

a) X	n_j	b) X	n_j	c) X	n_j	d) X	n_j
7-8	7	30-34	8	40-45	6	57-64	7
5-6	10	25-29	18	34-39	12	49-56	11
3-4	8	20-24	20	28-33	16	41-48	18
1-2	5	15-19	16	22-27	14	33-40	15
		10-14	10	16-21	8	25-32	10
						17-24	3

e)

X	n_j
38-41	13
34-37	26
30-33	30
26-29	34
22-25	25
18-21	18
14-17	12

f)

X	n_j
137-141	10
132-136	17
127-131	33
122-126	44
117-121	40
112-116	24

g)

X	n_j
90-91	10
88-89	0
86-87	2
84-85	3
82-83	5

h)

X	n_j
122-128	5
115-121	10
108-114	23
101-107	38
94-100	35
87-93	30
80-86	14

5.16. Calcular la mediana a partir de los siguientes datos *no agrupados* en intervalos.

a) 8, 1, 3, 5, 14, 2, 25
b) 87, 42, 21, 105, 118, 8, 102, 20, 38
c) 20, 4, 14, 52, 39, 6, 10, 35
d) 32, 5, 9, 18, 44, 60, 18, 11
e) 12, 3, 7, 12, 12, 13, 8, 12
f) 5, 1, 2, 5, 14, 5, 5, 12, 1, 5
g) 1, 8, 9, 8, 11, 7, 8, 6, 3, 1, 8
h) 22, 20, 37, 22, 25
i) 4, 4, 2, 2, 4, 2, 2, 4, 4

5.17. ¿Es posible calcular la mediana a partir de los dos cuadros siguientes?

a)

X	n_j
25 puntos o más	3
21-24	10
17-20	26
13-16	38
12 puntos o menos	64

b)

X	n_j
25 puntos o más	3
21-24	16
17-20	18
13-16	24
12 puntos o menos	63

5.18. ¿Cuál de las dos, media o mediana, representa mejor a los valores numéricos: 1, 3, 4, 6, 8, 200, 8, 6, 4, 3, 1?

5.19. ¿Cuál eligiría usted en primer lugar, la media o la mediana, como medida de tendencia central, a partir de las siguientes distribuciones de frecuencias?

a)

X	n_j
34-39	2
28-33	0
22-27	1
16-21	6
10-15	15
4-9	12

b)

X	n_j
66-69	8
62-65	10
58-61	25
54-57	28
50-53	12
49 puntos o menos	9

c)

X	n_j
91-100	6
81-90	10
71-80	25
61-70	26
51-60	11
41-50	5

5.20. Calcular la mediana de las puntuaciones 4, 4, 5, X, 5, 4, 4, sabiendo que su media vale 5 y donde X es un valor desconocido.

5.21. Calcular los percentiles 15, 25, 36, 75 y 82, a partir de la siguiente distribución de frecuencias

X	n_j
100-104	2
95-99	10
90-94	21
85-89	30
80-84	40
75-79	38
70-74	34
65-69	18
60-64	7

5.22. Calcular los percentiles 25, 36 y 75, a partir de la siguiente distribución de frecuencias

X	n_j
103-108	8
97-102	11
91-96	16
85-90	10
79-84	5

5.23. A partir de la siguiente distribución de frecuencias, calcular las dos puntuaciones que dejan entre sí las 133 observaciones centrales y decir qué percentiles son esas dos puntuaciones.

X	n_j
40-44	13
35-39	40
30-34	77
25-29	25
20-24	20

5.24. Llamemos decil primero, decil segundo, . . . , decil noveno a los puntos de la escala que dejan por debajo de sí el 10 por 100, el 20 por 100, . . . , el 90 por 100 de las observaciones, respectivamente. Esto supuesto, ¿tiene que ser igual la distancia sobre la escala entre dos deciles cualesquiera consecutivos que la distancia entre otro par cualquiera de deciles también consecutivos?

5.25. Los percentiles son valores esencialmente positivos. ¿Sí? ¿No?

6

Estadísticos de variabilidad o dispersión

6.1. Introducción

Los estadísticos de tendencia central o de posición indican dónde se sitúa un grupo de puntuaciones (en zona alta, media o baja). Los de variabilidad o dispersión nos indican si esas puntuaciones se encuentran muy próximas entre sí o muy dispersas. Por ejemplo (7, 9, 11) y (1, 10, 16) tienen la misma media (posición), pero la variabilidad o dispersión de las puntuaciones del primer grupo es menor que la de las puntuaciones del segundo.

6.2. Desviación media

6.2.1. Definición

Es la media de las diferencias (en valor absoluto) de n puntuaciones respecto a su media aritmética. En otras palabras, dadas n puntuaciones $X_1, X_2, \ldots, X_n$, su desviación media viene definida por

$$DM = \frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + \cdots + |X_n - \bar{X}|}{n} = \frac{\Sigma |X_i - \bar{X}|}{n} \qquad (6.1)$$

6.2.2. Cálculo

a) *Datos no agrupados en intervalos*

Aplicación directa de la fórmula (6.1) a los datos originales.

EJEMPLO 6.1. Diez ratas blancas podían obtener un poco de comida al mover una palanca situada dentro de una caja. Cada día, después de veintidós horas de ayuno, las ratas eran introducidas en la caja y se medía el tiempo transcurrido desde que se les abría el acceso a la palanca hasta que la movían. Los datos de la tabla

adjunta corresponden al quinto día de entrenamiento. (Véase Hull, Felsinger, Gladstone y Yamaguchi (1947).)

Rata n.º 1 2 3 4 5 6 7 8 9 10

Tiempo (en seg.): 25; 1,43; 20; 1; 8,56; 16,5; 16; 34; 4,43; 44 $\bar{X} = 17{,}09$ segundos

$$DM = \frac{|25 - 17{,}09| + |1{,}43 - 17{,}09| + \cdots + |44 - 17{,}09|}{10} = \frac{109{,}26}{10} =$$

$$= 10{,}926 \text{ segundos}$$

b) *Datos agrupados en intervalos*

$$DM = \frac{\Sigma n_j |X_j - \bar{X}|}{\Sigma n_j} = \frac{\Sigma n_j |X_j - \bar{X}|}{n} \qquad (6.2)$$

donde n_j y X_j son, respectivamente, el número de observaciones y el punto medio del intervalo j.

EJEMPLO 6.2. Agrupemos los datos del ejemplo 6.1 en tres intervalos y calculemos la desviación media.

Tiempo	n_j	X_j	$n_j X_j$	$\lvert X_j - \bar{X}\rvert$	$n_j \lvert X_j - \bar{X}\rvert$
31-45	2	38	76	18	36
16-30	4	23	92	3	12
1-15	4	8	32	12	48
	10		200		96

$$\bar{X} = \frac{200}{10} = 20 \text{ segundos}$$

$$DM = \frac{96}{10} = 9{,}6 \text{ segundos}$$

Junto a la fórmula (6.2) y mediante la transformación $x'_j = \frac{X_j - X_0}{I}$ de que hemos hablado en 5.2.4, tendremos la siguiente fórmula

$$DM = I \frac{\Sigma n_j |x'_j - \bar{x}'|}{\Sigma n_j} = I \frac{\Sigma n_j |x'_j - \bar{x}'|}{n} \qquad (6.3)$$

EJEMPLO 6.3. Elijamos como origen el intervalo 16 – 30 en el cuadro correspondiente al ejemplo 6.2 y calculemos la desviación media.

Tiempo	n_j	x'_j	$n_j x'_j$	$\lvert x'_j - \bar{x}'\rvert$	$n_j \lvert x'_j - \bar{x}'\rvert$
31-45	2	1	2	1,2	2,4
16-30	4	0	0	0,2	0,8
1-15	4	–1	–4	0,8	3,2
	10		–2		6,4

$$\bar{x}' = \frac{-2}{10} = -0{,}2$$

$$DM = 15 \frac{6{,}4}{10} = 9{,}6 \text{ segundos}$$

Queda como ejercicio para el lector la legitimación de (6.2) y (6.3).

6.2.3. Propiedades

a) Es fácilmente inteligible y fácilmente calculable.
b) Raramente usada debido a que los valores absolutos son muy poco manejables matemáticamente.

6.3. Varianza y desviación típica

6.3.1. Introducción

Si sumamos las diferencias de n puntuaciones respecto a su media, sabemos que dicha suma vale siempre cero. Para evitar este inconveniente, podemos tomar dichas puntuaciones en valor absoluto (caso de la desviación media) o podemos, también, elevar dichas diferencias al cuadrado y sumar estas diferencias cuadráticas. Esta última táctica es la que seguiremos en el caso de la varianza y de la desviación típica.

La varianza, referida a una muestra, será designada por s^2 y, referida a una población, por σ^2. Como en este tomo 1 nos limitamos al estudio de muestras, usaremos s^2 al tratar de la varianza. Excepcionalmente usaremos el símbolo σ^2, cuando de modo incidental nos refiramos a la varianza de la población. A su vez, la desviación típica, referida a una muestra, será designada por s y referida a una población por σ. Llamaremos s_x^2, s_y^2, ..., a la varianza de la variable X, Y, ...

6.3.2. Definición

a) *Varianza*

Es la media de las diferencias (al cuadrado) de n puntuaciones respecto a su media aritmética. En otras palabras, dadas n puntuaciones X_1, X_2, ..., X_n, su varianza viene definida por:

$$s_x^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n} = \frac{\Sigma (X_i - \bar{X})^2}{n} \qquad (6.4)$$

b) *Desviación típica*

Es la raíz cuadrada de la varianza. Será considerada siempre como positiva.

6.3.3. Cálculo

a) *Datos no agrupados en intervalos*

Aplicación directa de (6.4) a los datos originales.
Desarrollando (6.4) podemos llegar a otras dos fórmulas equivalentes.

$$\frac{\Sigma (X_i - \bar{X})^2}{n} = \frac{\Sigma (X_i^2 - 2\bar{X}X_i + \bar{X}^2)}{n} = \frac{\Sigma X_i^2}{n} - \frac{2\bar{X}\Sigma X_i}{n} + \frac{n\bar{X}^2}{n} =$$

$$= \frac{\Sigma X_i^2}{n} - 2\bar{X}^2 + \bar{X}^2 = \frac{\Sigma X_i^2}{n} - \bar{X}^2 = \frac{\Sigma X_i^2}{n} - \left(\frac{\Sigma X_i}{n}\right)^2 = \frac{n\Sigma X_i^2 - (\Sigma X_i)^2}{n^2}$$

Por consiguiente,

$$s_x^2 = \frac{\Sigma X_i^2}{n} - \bar{X}^2 \qquad (6.5)$$

$$s_x^2 = \frac{n\Sigma X_i^2 - (\Sigma X_i)^2}{n^2} \qquad (6.6)$$

EJEMPLO 6.4. Rachman (1968) en un experimento con cinco jóvenes pudo confirmar la hipótesis de que el fetichismo sexual es un comportamiento que puede ser establecido tras un proceso de condicionamiento clásico. El número de ensayos de condicionamiento que fueron necesarios para implantar una conducta fetichista en cada uno de los cinco sujetos fue: 35, 36, 21, 45, 38. Calculemos la varianza y la desviación típica de estos datos.

X_i	X_i^2	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
35	1.225	0	0
36	1.296	1	1
21	441	−14	196
45	2.025	10	100
38	1.444	3	9
175	6.431	0	306

$$\bar{X} = \frac{175}{5} = 35 \text{ ensayos}$$

Según (6.4): $s_x^2 = \frac{306}{5} = 61{,}20$, $s_x = 7{,}823$ ensayos

Según (6.5): $s_x^2 = \frac{6.431}{5} - 35^2 = 1.286{,}20 - 1.225 = 61{,}20$, $s_x = 7{,}823$ ensayos

Según (6.6): $s_x^2 = \frac{(5)(6.431) - 175^2}{5^2} = \frac{32.155 - 30.625}{25} = 61{,}20$, $s_x = 7{,}823$ ensayos

b) *Datos agrupados en intervalos*

Supongamos n observaciones agrupadas en r intervalos, todos ellos de igual amplitud. Sea X_1 el punto medio del intervalo primero y n_1 el número de observaciones dentro del mismo. Sea X_2 el punto medio del intervalo segundo y n_2 el número de observaciones dentro del mismo. Sea X_r el punto medio del intervalo r y n_r el número de observaciones dentro del mismo. Según sabemos, al agrupar las observaciones en intervalos, atribuimos a cada una de ellas (como puntuación) el punto medio del intervalo dentro del que se encuentra. Restemos ahora cada una de las puntuaciones de la media del grupo y elevemos al cuadrado estas diferencias. Dentro del intervalo primero tendremos n_1 puntuaciones iguales a X_1 y, consiguientemente, n_1 diferencias cuadráticas iguales a $(X_1 - \bar{X})^2$ cuya suma valdrá $n_1(X_1 - \bar{X})^2$. Dentro del intervalo segundo tendremos n_2 puntuaciones iguales a X_2 y, consiguientemente, n_2 diferencias cuadráticas iguales a $(X_2 - \bar{X})^2$ cuya suma valdrá $n_2(X_2 - \bar{X})^2$. Dentro del intervalo r tendremos n_r puntuaciones iguales a X_r y, consiguientemente, n_r diferencias cuadráticas iguales a $(X_r - \bar{X})^2$ cuya suma valdrá $n_r(X_r - \bar{X})^2$. En conclusión, la suma de las $n_1 + n_2 + \cdots + n_r = n$ diferencias cuadráticas valdrá: $n_1(X_1 - \bar{X})^2 + n_2(X_2 - \bar{X})^2 + \cdots + n_r(X_r - \bar{X})^2 = \Sigma n_j(X_j - \bar{X})^2$ y su media (es decir, la varianza de las n puntuaciones) valdrá:

$$s_x^2 = \frac{\Sigma n_j(X_j - \bar{X})^2}{\Sigma n_j} = \frac{\Sigma n_j(X_j - \bar{X})^2}{n} \qquad (6.7)$$

De (6.7) es fácil deducir las dos fórmulas siguientes, de acuerdo con lo dicho para el caso de datos no agrupados en intervalos,

$$s_x^2 = \frac{\Sigma n_j X_j^2}{n} - \bar{X}^2 \qquad (6.8)$$

$$s_x^2 = \frac{n\Sigma n_j X_j^2 - (\Sigma n_j X_j)^2}{n^2} \qquad (6.9)$$

EJEMPLO 6.5. Agrupando en cuatro intervalos los datos de Conde y Doménech (1976) propuestos en el ejemplo 5.24, nos queda

Edad	n_j	X_j	X_j^2	n_jX_j	$n_jX_j^2$	$X_j - \bar{X}$	$(X_j - \bar{X})^2$	$n_j(X_j - \bar{X})^2$
70-89	20	79,5	6.320,25	1.590,00	126.405,00	33,832	1.144,604	22.892,080
50-69	31	59,5	3.540,25	1.844,50	109.747,75	13,832	191,324	5.931,044
30-49	18	39,5	1.560,25	711,00	28.084,50	−6,168	38,044	684,792
10-29	38	19,5	380,25	741,00	14.449,50	−26,168	684,764	26.021,032
	107			4.886,50	278.686,75			55.528,948

$$\bar{X} = \frac{4.886,50}{107} = 45,668 \text{ años}$$

Según (6.7):

$$s_x^2 = \frac{55.528,948}{107} = 518,962\ , \quad s_x = 22,781 \text{ años}$$

Según (6.8):

$$s_x^2 = \frac{278.686,75}{107} - 45,668^2 = 2.604,549 - 2.085,566 = 518,983,\ s_x = 22,781 \text{ años}$$

Según (6.9):

$$s_x^2 = \frac{(107)(278.686,75) - 4.886,5^2}{107^2} = \frac{29.819.482,25 - 23.877.882,25}{11.449} = 518,962$$

$$s_x = 22,781 \text{ años}$$

La diferencia entre 518,962, en (6.7) y 518,983, en (6.8) y (6.9), es debida a efectos de redondeo.

6.3.4. Propiedades

a) La varianza de $Y_1 = AX_1 + B$, $Y_2 = AX_2 + B$, ..., $Y_n = AX_n + B$, siendo A y B dos constantes arbitrarias, es igual a la varianza de X multiplicada por A^2. Consiguientemente, la desviación típica de Y_1, Y_2, ..., Y_n, es igual a la desviación típica de X multiplicada por $|A|$.

En efecto, sabemos que $\bar{Y} = A\bar{X} + B$ (véase 5.2.3.c). Por tanto, la varianza de las nuevas puntuaciones, s_y^2, valdrá, por definición,

$$s_y^2 = \frac{\Sigma\,[(AX_i + B) - (A\bar{X} + B)]^2}{n} = \frac{\Sigma\,(AX_i - A\bar{X})^2}{n} = \frac{A^2\,\Sigma\,(X_i - \bar{X})^2}{n} = A^2 s_x^2$$

Consiguientemente, $s_y = |A| s_x$.

Si $A = 1$ y $B \neq 0$, $s_y^2 = (1)s_x^2$. Es decir, si sumamos a todas las puntuaciones una constante, B, la varianza (y la desviación típica) de las nuevas puntuaciones es igual que la varianza (y que la desviación típica) de las antiguas. Lo cual es obvio, ya que añadir una constante, B, a todas las puntuaciones, equivale a moverlas como un bloque rígido, sin alterar su dispersión o variabilidad, bien hacia la derecha (si B es positiva), bien hacia la izquierda (si B es negativa).

Si $A \neq 0$ y $B = 0$, $s_y^2 = A^2 s_x^2$. Es decir, si multiplicamos todas las puntuaciones por una constante, A, la varianza de las nuevas puntuaciones es igual que la varianza de las antiguas multiplicada por A^2. Consiguientemente, la desviación típica de las nuevas es igual que la desviación típica de las antiguas multiplicada por A, pero tomada ésta en valor absoluto. La razón de esta restricción es la siguiente. Veremos enseguida que una desviación típica, no nula, tiene que ser positiva. Por tanto, s_x será positiva. Ello lleva consigo que si A fuera negativa, $s_y = As_x$ debería ser negativa, lo cual no tiene sentido, según lo acabado de indicar. De aquí se sigue que A debe ser tomada en valor absoluto.

b) La varianza y la desviación típica son sensibles a la variación de cada una de las puntuaciones. Baste con que varíe una de éstas, para que varíen aquellas. Ello es debido a que varianza y desviación típica dependen de todas y cada una de las puntuaciones y, consiguientemente, de la media.

c) Son fundamento de muchas técnicas estadísticas que tienen gran importancia en Psicología.

d) Son función de los intervalos elegidos (de su amplitud, de su número y de los límites de los mismos).

e) Fuera del intervalo $(\bar{X} - 2s_x, \bar{X} + 2s_x)$ se encuentra, a lo más, el $[(1/2^2)100]$ por $100 = 25$ por 100 de las observaciones. Fuera del intervalo $(\bar{X} - 3s_x, \bar{X} + 3s_x)$ se encuentra, a lo más, el $[(1/3^2)100]$ por $100 = 11$ por 100 de las observaciones.

En general, fuera del intervalo $(\bar{X} - ks_x, \bar{X} + ks_x)$ se encuentra, a lo más, el $[(1/k^2)100]$ por 100 de las observaciones, sea cual sea la forma de la distribución de frecuencias.

Esta propiedad quedará legitimada en el tomo 2: Estadística Inferencial.

f) No serán calculables, o no serán recomendables, cuando no sea calculable, o no sea recomendable, la media como medida de posición o tendencia central.

g) La desviación típica viene expresada en las mismas unidades en las que vienen expresados los datos. No ocurre lo mismo con la varianza. Si los datos, por ejemplo vienen dados en metros, la desviación típica vendrá dada en metros, pero la varianza vendrá dada en metros cuadrados.

h) Dados r grupos, el primero con n_1 puntuaciones, media $\bar{X}_1$ y varianza s_1^2, el segundo con n_2 puntuaciones, media $\bar{X}_2$ y varianza s_2^2, ..., el r-ésimo con n_r puntuaciones, media $\bar{X}_r$ y varianza s_r^2, la varianza, s_x^2, de las $n_1 + n_2 + \cdots + n_r = n$ puntuaciones, vale

$$s_x^2 = \frac{\Sigma\, n_j s_j^2}{n} + \frac{\Sigma\, n_j(\bar{X}_j - \bar{X})^2}{n}$$

Es decir, la varianza de las n puntuaciones es igual a la media de las varianzas más la varianza de las medias.

En efecto, sean

$X_{11}, X_{21}, \ldots, X_{n_1 1}$ las n_1 puntuaciones del grupo 1.º con media $\bar{X}_1$ y varianza s_1^2.

$X_{12}, X_{22}, \ldots, X_{n_2 2}$ las n_2 puntuaciones del grupo 2.º con media $\bar{X}_2$ y varianza s_2^2.

. . . . . .

$X_{1r}, X_{2r}, \ldots, X_{n_r r}$ las n_r puntuaciones del grupo r-ésimo con media $\bar{X}_r$ y varianza s_r^2.

Sean $\bar{X}$ y s_x^2 la media y la varianza del grupo total, es decir, del compuesto por las $n_1 + n_2 + \cdots + n_r = n$ puntuaciones. Esto supuesto,

$$s_x^2 = \frac{\sum_j \sum_i (X_{ij} - \bar{X})^2}{n} = \frac{\sum_j \sum_i [(X_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X})]^2}{n} =$$

$$= \frac{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2}{n} + \frac{2 \sum_j (\bar{X}_j - \bar{X}) \sum_i (X_{ij} - \bar{X}_j)}{n} + \frac{\sum_j n_j (\bar{X}_j - \bar{X})^2}{n} =$$

$$= \frac{\sum_j \sum_i (X_{ij} - \bar{X}_j)^2}{n} + 0 + \frac{\sum_j n_j (\bar{X}_j - \bar{X})^2}{n} = \frac{\sum n_j s_j^2}{n} + \frac{\sum n_j (\bar{X}_j - \bar{X})^2}{n}$$

Es decir, la varianza del grupo total es igual a la media de las varianzas de cada uno de los r subgrupos más la varianza de las medias de cada uno de estos mismos r subgrupos.

Recuérdese que, según 5.2.3. a),

$$\sum_i (X_{i1} - \bar{X}_1) = \sum_i (X_{i2} - \bar{X}_2) = \cdots = \sum_i (X_{ir} - \bar{X}_r) = 0$$

Por tanto,

$$\sum_j (\bar{X}_j - \bar{X}) \sum_i (X_{ij} - \bar{X}_j) = (\bar{X}_1 - \bar{X}) \sum_i (X_{i1} - \bar{X}_1) + (\bar{X}_2 - \bar{X}) \sum_i (X_{i2} - \bar{X}_2) +$$

$$+ \cdots + (\bar{X}_r - \bar{X}) \sum_i (X_{ir} - \bar{X}_r) = 0 + 0 + \cdots + 0 = 0$$

EJEMPLO 6.6.

Grupo primero	Grupo segundo	Grupo tercero
1 −1 3 5 2	6 8	3 7 5 5
10 $n_1 = 5$ $\bar{X}_1 = 2$ $s_1^2 = 4$	14 $n_2 = 2$ $\bar{X}_2 = 7$ $s_2^2 = 1$	20 $n_3 = 4$ $\bar{X}_3 = 5$ $s_3^2 = 2$

$n = 11$

$\bar{X} = 4$

$s_x^2 = \frac{72}{11}$

$$\frac{\sum n_j s_j^2}{n} = \frac{(5)(4) + (2)(1) + (4)(2)}{11} = \frac{30}{11}$$

$$\frac{\sum n_j (\bar{X}_j - \bar{X})^2}{n} = \frac{(5)(2-4)^2 + (2)(7-4)^2 + (4)(5-4)^2}{11} = \frac{20 + 18 + 4}{11} = \frac{42}{11}$$

Vemos, en efecto, cómo $\frac{72}{11}$ es igual a $\frac{30}{11}$ más $\frac{42}{11}$.

6.3.5. Método abreviado para el cálculo de la varianza

Supongamos n puntuaciones agrupadas en intervalos, todos ellos de amplitud I. Sea X_0 el punto medio de uno de ellos, elegido arbitrariamente, al que llamaremos intervalo origen. Hagamos $A = \frac{1}{I}$, $B = -\frac{X_0}{I}$. Según 6.3.4.a), las puntuaciones $x'_j = \frac{1}{I} X_j + \left(-\frac{X_0}{I}\right) = \frac{X_j - X_0}{I}$ tendrán como varianza $s_{x'}^2 = \frac{1}{I^2} s_x^2$. De donde, $s_x^2 = I^2 s_{x'}^2$. Esta última fórmula nos permite obtener s_x^2 mediante $s_{x'}^2$, cuyo cálculo suele ser más sencillo y breve que el de s_x^2. Recordando las indicaciones propuestas en 5.2.4, vamos a aplicar el método abreviado a los mismos datos del ejemplo (6.2) y de los que hemos calculado, allí, la varianza mediante el método no abreviado.

EJEMPLO 6.7.

	n_j	x'_j	x'^2_j	$n_j x'_j$	$n_j x'^2_j$	$x'_j - \bar{x}'$	$(x'_j - \bar{x}')^2$	$n_j(x'_j - \bar{x}')^2$
70-89	20	2	4	40	80	1,6916	2,8615	57,2300
50-69	31	1	1	31	31	0,6916	0,4783	14,8273
30-49	18	0	0	0	0	−0,3084	0,0951	1,7118
10-29	38	−1	1	−38	38	−1,3084	1,7119	65,0522
	107			33	149			138,8213

$$\bar{x}' = \frac{33}{107} = 0{,}3084$$

Calcular $s^2_{x'}$ equivale a aplicar (6.7), (6.8) y (6.9) a las puntuaciones x'_j.

Según (6.7):

$$s^2_{x'} = \frac{138{,}8213}{107} = 1{,}2974$$

Según (6.8):

$$s^2_{x'} = \frac{149}{107} - 0{,}3084^2 = 1{,}3925 - 0{,}0951 = 1{,}2974$$

Según (6.9):

$$s^2_{x'} = \frac{(107)(149) - 33^2}{107^2} = \frac{14.854}{11.449} = 1{,}2974$$

Por tanto,

$$s^2_x = (20)^2(1{,}2974) = 518{,}96$$

6.3.6. NOTA: sobre la definición de s^2

Supongamos una población finita compuesta de N elementos equiprobables. Por definición, su varianza vale $\sigma^2_x = \frac{\Sigma (X_i - \mu)^2}{N}$, donde μ es la media de dicha población. Formemos ahora todas las muestras ordenadas posibles de tamaño n y calculemos sus varianzas mediante la fórmula $s^2_x = \frac{\Sigma (X_i - \bar{X})^2}{n}$. Pues bien, resulta que la media de todas estas varianzas, así definidas, vale $\frac{n-1}{n}\sigma^2_x$ y no σ^2_x.

Para alcanzar que esa media valga σ^2_x, basta con multiplicarla por $\frac{n}{n-1}$ o, lo que es equivalente, basta con definir $s^2_x = \frac{n}{n-1}\frac{\Sigma (X_i - \bar{X})^2}{n} = \frac{\Sigma (X_i - \bar{X})^2}{n-1}$.

Esta última definición de s^2_x es la que suele aparecer en ciertos libros de Estadística Descriptiva y es la que usan muchas máquinas de calcular. La discusión de este problema la dejamos para el tomo 2: Estadística Inferencial. Por ahora nos limitamos a comprobar lo dicho con el siguiente ejemplo.

EJEMPLO 6.8. Sea la población finita compuesta por los tres elementos 1, 2, 3. Su media vale $\mu = 2$ y su varianza vale $\sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = \frac{2}{3}$.

Formemos todas las muestras binarias posibles, pero de modo que extraído un elemento de la población, lo repongamos en la misma antes de extraer el segundo. Tendremos (3)(3) = 9 muestras binarias.

Muestras binarias	$\bar{X}$	$\frac{\Sigma (X_i - \bar{X})^2}{n}$	$\frac{\Sigma (X_i - \bar{X})^2}{n-1}$
1, 1	1,00	0,00	0,00
1, 2	1,50	0,25	0,50
1, 3	2,00	1,00	2,00
2, 1	1,50	0,25	0,50
2, 2	2,00	0,00	0,00
2, 3	2,50	0,25	0,50
3, 1	2,00	1,00	2,00
3, 2	2,50	0,25	0,50
3, 3	3,00	0,00	0,00
		3,00	6,00
		$\frac{3}{9} = \frac{1}{3}$	$\frac{6}{9} = \frac{2}{3}$

Comprobamos cómo la media de las varianzas definidas por $\frac{\Sigma (X_i - \bar{X})^2}{n-1}$ vale $\frac{2}{3}$, es decir, un valor igual que el de la varianza de la población y cómo la media de las varianzas definidas por $\frac{\Sigma (X_i - \bar{X})^2}{n}$ vale $\frac{1}{3}$, es decir, un valor distinto (menor) que el de la varianza de la población.

6.4. Amplitud total

6.4.1. Definición

Diferencia entre la puntuación máxima y la mínima. O, teniendo en cuenta los límites exactos de los intervalos elementales o compuestos, diferencia entre la puntuación máxima y la mínima más una unidad.

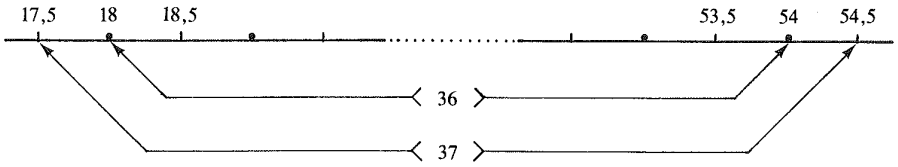
6.4.2. Cálculo

a) *Datos no agrupados en intervalos*

EJEMPLO 6.9. La fenilcetonuria es una enfermedad causante de un deterioro mental que se agudiza con la edad. Los datos siguientes, tomados de Berman, Waisman y Graham (1966) representan los cocientes intelectuales de cinco niños afectados por dicha enfermedad y cuya edad media aproximada era de cinco años.

18, 29, 39, 40, 54

$AT = 54 - 18 = 36$ o $AT = 54 - 18 + 1 = 54{,}5 - 17{,}5 = 37$



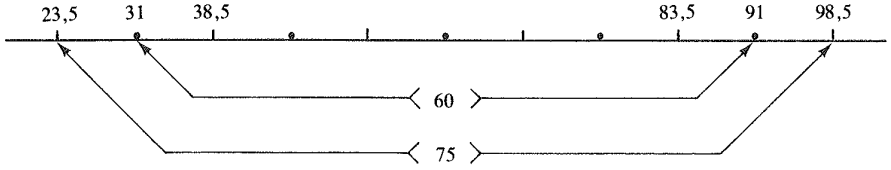
b) *Datos agrupados en intervalos*

EJEMPLO 6.10. Calculemos la amplitud total de los cocientes intelectuales de otro grupo de 22 niños fenilcetonúricos más jóvenes que los del ejemplo 6.9 y, como puede apreciarse, con menor deterioro mental. Los datos expuestos a continuación están tomados de Berman y otros (1966) y han sido agrupados en intervalos por nosotros del modo siguiente:

CI	X_j	n_j
84-98	91	6
69-83	76	5
54-68	61	2
39-53	46	5
24-38	31	4
		22

$AT = 91 - 31 = 60$ o $AT = 91 - 31 + 15 = 98{,}5 - 23{,}5 = 75$

Ahora, en realidad, las únicas puntuaciones existentes son los puntos medios de los intervalos. Por tanto, la amplitud total será la diferencia entre el punto medio del intervalo máximo (91) y el punto medio del intervalo mínimo (31). A su vez, la unidad que puede ser añadida es una unidad de intervalo, medio intervalo por debajo del punto medio mínimo y medio intervalo por encima del punto medio máximo. En otras palabras, añadir una unidad de intervalo, equivale a aceptar como amplitud total la diferencia entre el límite exacto superior del intervalo máximo y el límite exacto inferior del intervalo mínimo.



6.4.3. Propiedades

a) Muy fácilmente calculable.

b) Presenta el inconveniente de tener en cuenta únicamente dos puntuaciones: las dos extremas. Si éstas se mantienen constantes, la amplitud total se mantendrá constante aunque varíen de cualquier modo las restantes, siempre claro está, que éstas queden dentro del intervalo limitado por las dos primeras.

Así, dejando intactas las dos puntuaciones extremas (18 y 54) del ejemplo 6.9, hagamos variar las tres interiores de dos maneras distintas. Comparemos los resultados del caso primitivo con los de los dos nuevos y veamos cómo permanece invariante la amplitud total en los tres y cómo varían claramente la desviación media y la desviación típica.

Puntuaciones	*AT*	*DM*	s_x
18, 29, 39, 40, 54	36 ó 37	10	12,017
18, 36, 36, 36, 54	36 ó 37	7,2	11,384
18, 54, 54, 54, 54	36 ó 37	11,52	14,40

6.4.4. Nota

La amplitud total suele ser llamada, también, recorrido o rango.

6.5. Amplitud semiintercuartil

6.5.1. Definición

Semidistancia entre el tercer cuartil y el primer cuartil, es decir, entre el percentil 75 y el percentil 25.

6.5.2. Cálculo

Basta con calcular los percentiles 75 y 25 (según lo dicho en 5.5.2) y hallar la semidiferencia entre ambos.

EJEMPLO 6.11. Calculemos la amplitud semiintercuartil, a partir de los datos del ejemplo 6.10.

En dichos datos, $P_{75} = 84{,}75$, $P_{25} = 43$. Por tanto

$$ASI = \frac{84{,}75 - 43}{2} = 20{,}875$$

6.5.3. Propiedades

a) Es preferible a la desviación típica en el caso de distribuciones muy asimétricas. (Recuérdese que en este caso la mediana era preferible a la media.)

b) Cuando el intervalo máximo carece de límite superior y/o el mínimo carece de límite inferior, es imposible calcular la desviación típica. Bajo estas condiciones es posible calcular la amplitud semiintercuartil, siempre que el primero y el tercer cuartil no se encuentren en esos intervalos extremos. (Recuérdese que bajo estas condiciones era posible calcular la mediana y no era posible el cálculo de la media.)

c) Definida como distancia entre dos puntos, sólo es calculable a nivel de intervalos o de razón, pero no a nivel meramente ordinal.

d) Menos sensible que la desviación media y que la desviación típica a la variación de los datos. Veamos confirmada esta afirmación en los siguientes datos:

Puntuaciones	*ASI*	*DM*	s_x
30, 30, 30, 40, 40, 40, 40, 40, 40, 50, 50, 50	$\frac{(45 - 35)}{2} = 5$	5	7,07
1, 1, 98, 100, 100, 100, 108, 108, 108, 110, 207, 207	$\frac{(109 - 99)}{2} = 5$	37,33	59,58

Mientras que la amplitud semiintercuartil ha permanecido invariante frente al cambio de puntuaciones, la desviación media ha pasado de 5 a 37,33 y la desviación típica ha pasado de 7,07 a 59,58.

6.6. Coeficiente de variación

6.6.1. Nota previa

Consideremos dos variables distintas, por ejemplo, peso (X) y altura (Y). Es claro que s_x vendrá dada en unidades de peso (por ejemplo, gramos) y s_y en unidades de longitud (por ejemplo, metros). Ambas desviaciones típicas no son comparables. Si s_x valiera 5 gramos y s_y 3 metros, 5 gramos no son más ni menos que 3 metros, son cosas distintas. Para hacer comparables las variabilidades de ambos grupos, con variables de distinta naturaleza, es necesario que vengan expresadas en números abstractos (es decir, ni en metros, ni en gramos). Un modo de conseguirlo es tomar como medidas de variabilidad $s_x/\bar{X}$ y $s_y/\bar{Y}$. Estos cocientes son números abstractos. Lo único que nos indican es el número de veces que el numerador contiene al denominador, independientemente de lo que ambos signifiquen (gramos, metros, etc.).

Consideremos ahora una misma variable y dos grupos distintos, cuyas medias en esa variable son muy distintas entre sí. Por ejemplo, el peso de 100 elefantes y el de 100 hormigas. Es evidente que una misma desviación típica igual a 1 kg representa una variabilidad insignificante para el grupo de los elefantes e inconcebiblemente alta para el de las hormigas. Por tanto, la desviación típica será algo equívoca, en este caso, como medida de variabilidad. Un modo de evitar esta dificultad es dividir la desviación típica de cada grupo por su correspondiente media, $s_{x_1}/\bar{X}_1$, $s_{x_2}/\bar{X}_2$.

6.6.2. Definición

Resultado de dividir la desviación típica por la media. Ordinariamente, este cociente viene multiplicado por 100. Es decir,

$$CV = \frac{s_x}{\bar{X}} \quad \text{o, más frecuentemente,} \quad CV = \frac{s_x}{\bar{X}}\,100$$

6.6.3. Cálculo

Mera aplicación de la fórmula anterior.

EJEMPLO 6.12. El peso medio (media aritmética) de un grupo de elefantes es 6.000 kg y el de un grupo de hormigas es 2 dg. Suponiendo que la desviación típica

en ambos grupos fuera 1 dg, ¿cuánto valdrá el coeficiente de variación en uno y otro grupo?

$$CV = \frac{1}{60.000.000} = 0{,}0000000167 \quad \text{ó, multiplicado por 100,} \quad 0{,}00000167$$

$$CV = \frac{1}{2} = 0{,}5 \quad \text{o, multiplicado por 100,} \quad 50$$

El segundo CV es 30 millones (!) mayor que el primero. Esto quiere decir que la variabilidad relativa de ambos grupos es enormemente distinta, a pesar de que son iguales sus desviaciones típicas. Una desviación típica de 1 decigramo representa una variabilidad apreciable en relación con el grupo de las hormigas y no representa, prácticamente, variabilidad alguna en relación con el grupo de los elefantes.

6.6.4. Propiedades

a) Es un valor abstracto, como cociente de dos números concretos (es decir, dados en ciertas unidades concretas de medida) del mismo tipo. Recuérdese que la media y la desviación típica vienen dadas en las mismas unidades en las que vienen dadas las puntuaciones a partir de las cuales aquellas son calculadas.

b) Si a unas puntuaciones dadas les sumamos una cantidad positiva, el coeficiente de variación disminuirá, ya que s_x se mantendrá constante, pero $\bar{X}$ aumentará en esa cantidad. Por tanto, el cociente $s_x/\bar{X}$ disminuirá después de dicha suma.

Si, por el contrario, les restamos una cantidad positiva, el coeficiente de variación aumentará, por análoga razón.

c) Si multiplicamos unas puntuaciones dadas por cualquier constante positiva, el coeficiente de variación se mantendrá constante, pues el numerador, s_x, y el denominador, $\bar{X}$, quedarán multiplicados por la misma cantidad.

Es recomendable que, junto al coeficiente de variación, se ofrezcan las correspondientes s_x y $\bar{X}$ a partir de las cuales ha sido calculado.

6.7. Notas

a) Recordemos que, a nivel de intervalos o de razón, todo índice de tendencia central era un valor numérico que venía representado por un punto sobre el eje de abscisas. Pues bien, todo índice de variabilidad es una distancia que viene representada por un segmento rectilíneo.

b) Tengamos presente que todo índice de variabilidad es esencialmente no negativo. Las puntuaciones pueden ser positivas o negativas, pero su variabilidad o dispersión será siempre positiva (no son todas las puntuaciones iguales entre sí, hay alguna variabilidad) o nula (todas las puntuaciones son iguales entre sí, no hay variabilidad), pero es inconcebible una variabilidad negativa. De aquí que sean esencialmente no negativos los valores que pueden tomar los índices de variabilidad o dispersión. Por esta razón, supondremos $\bar{X} \neq 0$ y tomaremos $\bar{X}$ en valor absoluto.

6.8. Resumen: Definiciones y fórmulas

Desviación media: Media de las diferencias (en valor absoluto) de n puntuaciones respecto a su media aritmética.

$$DM = \frac{\Sigma |X_i - \bar{X}|}{n} \qquad \text{(datos no agrupados en intervalos)}$$

$$DM = \frac{\Sigma n_j |X_j - \bar{X}|}{n} \qquad \text{(datos agrupados en intervalos)}$$

Varianza: Media de las diferencias (al cuadrado) de n puntuaciones respecto a su media aritmética.

$$s_x^2 = \frac{\Sigma (X_i - \bar{X})^2}{n} \qquad \text{(datos no agrupados en intervalos)}$$

$$s_x^2 = \frac{\Sigma n_j (X_j - \bar{X})^2}{n} \qquad \text{(datos agrupados en intervalos)}$$

Desviación típica: Raíz cuadrada de la varianza.

Amplitud total: Diferencia entre la puntuación máxima y la mínima (o, entre la máxima y la mínima más una unidad).

Amplitud semiintercuartil: Semidiferencia entre el tercer cuartil, Q_3, y el primer cuartil, Q_1.

$$ASI = \frac{Q_3 - Q_1}{2}$$

Coeficiente de variación: Cociente entre la desviación típica y la media (multiplicado, ordinariamente, por 100).

$$CV = \frac{s_x}{\bar{X}} 100$$

EJERCICIOS

6.1. Calcular la varianza y la desviación típica a partir de los siguientes datos *no agrupados* en intervalos.

a) 3, 5, 1, 6, 10
b) 3, 4, 1, 4
c) 2, 5, 6, 1, 1
d) 1, 0, 3, 1, 3, 4
e) 5, 2, 1, 5, 3, 8
f) 1, 9, 3, 7, 8, 8

6.2. Calcular la desviación media a partir de los datos del ejercicio anterior.

6.3. Calcular la varianza y la desviación típica a partir de los siguientes datos *agrupados* en intervalos.

a)

X	n_j
9-11	1
6-8	2
3-5	5
0-2	4

b)

X	n_j
10-12	1
7-9	4
4-6	3
1-3	2

c)

X	n_j
9-10	1
7-8	2
5-6	1
3-4	4
1-2	2

d)

X	n_j
15-21	4
8-14	10
1-7	6

e)

X	n_j
11-13	2
8-10	3
5-7	6
2-4	4

f)

X	n_j
109-113	2
104-108	4
99-103	8
94-98	5
89-93	1

g)

X	n_j
110-112	2
107-109	4
104-106	3
101-103	1

h)

X	n_j
75-78	2
71-74	3
67-70	8
63-66	7
59-62	4
55-58	1

6.4. Calcular la desviación media a partir de los datos del ejercicio anterior.

6.5. Calcular la amplitud semiintercuartil a partir de los datos de los ejercicios 5.21 y 5.22.

6.6. Demostrar que $\Sigma (X_i - \bar{X})^2 = \Sigma (X_i - k)^2 - n(\bar{X} - k)^2$.

6.7. Comprobar la igualdad anterior para $X_1 = 1$, $X_2 = 2$, $X_3 = 6$ y $k = 5$.

6.8. Válgase de la igualdad expuesta en 6.6 para calcular $\Sigma (X_i - \bar{X})^2$ usando $k = 5$ y teniendo en cuenta los datos siguientes: 3, 4, 5, 4, 6, 7, 4, 5.

6.9. ¿Qué haría Vd. para simplificar el cálculo de la varianza de las siguientes puntuaciones: 2,75; −3,25; 6,75; 1,75; 5,75; −2,25; 4,75; −4,25? ¿Cuánto vale dicha varianza?

6.10. Sean 7 y 20, respectivamente, la media y la varianza de las puntuaciones $X_1, X_2, \ldots, X_n$. Calcule la media de las nuevas puntuaciones $X_1^2, X_2^2, \ldots, X_n^2$.

6.11. Sean 10 y 4, respectivamente, la media y la desviación típica de las puntuaciones $X_1, X_2, \ldots, X_n$. Calcule la media de las nuevas puntuaciones $(X_1 - 3)^2$, $(X_2 - 3)^2, \ldots, (X_n - 3)^2$.

6.12. Calcule el coeficiente de variación a partir de los datos del ejercicio 6.1.

6.13. Sean 4 y 3, respectivamente, la media y la desviación típica de las puntuaciones $X_1, X_2, \ldots, X_n$. Calcule la media y la desviación típica de las puntuaciones $X_1^2, X_2^2, \ldots, X_n^2$, sabiendo que el coeficiente de variación de estas últimas puntuaciones vale 96.

6.14. Sean 25 y 15, respectivamente, la media y la desviación típica de las puntuaciones $3X_1 - 5, 3X_2 - 5, \ldots, 3X_n - 5$. Calcular la desviación típica de las nuevas puntuaciones $X_1^2, X_2^2, \ldots, X_n^2$, sabiendo que el coeficiente de variación de estas últimas vale 80.

6.15. Sabiendo que las cinco puntuaciones $-4, 2, X, 4, 0$ tienen como varianza 16 y como coeficiente de variación 200, calcular el valor de la tercera puntuación desconocida, X, y la mediana de las cinco.

6.16. Sea un grupo compuesto de n_1 personas con media $\bar{X}_1$ y varianza s_1^2 y otro grupo compuesto de n_2 personas con media $\bar{X}_2$ y varianza s_2^2. Demostrar que la varianza del grupo total, s^2, compuesto de las $n_1 + n_2$ personas, vale:

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \dfrac{n_1 n_2}{n_1 + n_2} (\bar{X}_1 - \bar{X}_2)^2}{n_1 + n_2}$$

(Sugerencia: recordar la propiedad 6.3.4.*i* de la varianza.)

6.17. Deducir las fórmulas (6.2) y (6.3).

6.18. Deducir las fórmulas (6.8) y (6.9).

6.19. Supongamos una población compuesta de los cuatro elementos: 1, 2, 3 y 4. Formemos a partir de la misma las (4)(4)(4) = 64 muestras ternarias posibles, entendiendo que, extraído un elemento de la población, es devuelto a la misma antes de extraer un segundo elemento y, a su vez, éste es devuelto a la misma antes de extraer un tercero. Esto supuesto, calcular la media de las varianzas de las 64 muestras, habiendo definido, primero, la varianza como $\dfrac{\Sigma (X_i - \bar{X})^2}{n}$. Hacer lo mismo, habiendo definido la varianza como $\dfrac{\Sigma (X_i - \bar{X})^2}{n - 1}$ (donde $n = 3$). Comprobar, finalmente, cómo $\dfrac{\Sigma (X_1 - \bar{X})^2}{3} = \dfrac{480/9}{64} = \dfrac{5}{6}$ y cómo $\dfrac{\Sigma (X_i - \bar{X})^2}{2} = \dfrac{720/9}{64} = \dfrac{5}{4}$

que es, precisamente, el valor de la varianza de la población:

$$\frac{(1-2{,}5)^2+(2-2{,}5)^2+(3-2{,}5)^2+(4-2{,}5)^2}{4}=\frac{5}{4}$$

(Recuerde el lector lo dicho en 6.3.6.)

7

Estadísticos de asimetría y apuntamiento

7.1. Introducción

Una distribución de frecuencias queda bastante bien caracterizada mediante los estadísticos de posición o tendencia central y de variabilidad o dispersión, pero quedará aún mejor caracterizada si conocemos la simetría o asimetría de la misma y su apuntamiento.

7.2. Asimetría

7.2.1. Idea general

Sabemos que la mediana divide el histograma (representación gráfica de la distribución de frecuencias) en dos áreas iguales, es decir, de igual superficie. Pues bien, diremos que la distribución de frecuencias es simétrica si una de las áreas es imagen de la otra. Nótese que si un área es imagen de la otra, ambas tienen la misma superficie, pero pueden tener ambas la misma superficie y no ser una imagen de la otra. Así, por ejemplo, en la figura 7.1, A y A' tienen igual superficie y una es imagen de la otra. En cambio, en la figura 7.2, A y A', teniendo igual superficie, una no es imagen de la otra.

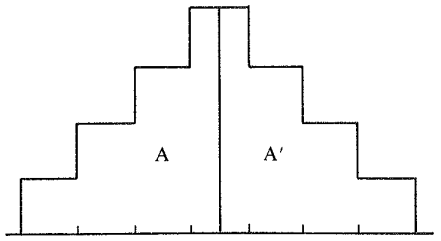


Fig. 7.1

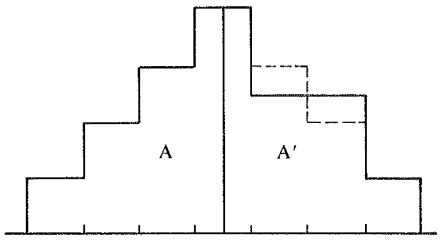
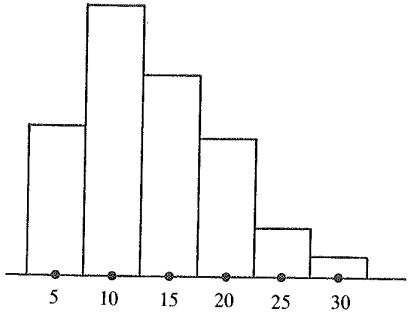


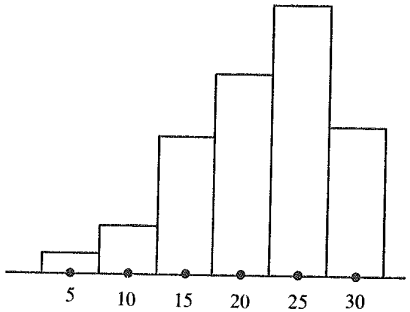
Fig. 7.2

Es claro que si la distribución es simétrica, la mediana es, también, centro de gravedad, es decir, la mediana coincide con la media. Por otra parte, si la distribución de frecuencias es unimodal, esa única moda coincide con la mediana y, consiguientemente, con la media. En conclusión, si la distribución es simétrica, media, mediana y moda coinciden.

Diremos que la asimetría es positiva si tenemos muchas puntuaciones bajas y pocas altas. Diremos que es negativa si sucede lo contrario. Un test difícil dará lugar a una distribución asimétrica positiva. Un test fácil dará lugar a una distribución asimétrica negativa:



Asimetría positiva



Asimetría negativa

Teniendo en cuenta estas consideraciones, presentamos a continuación diversos estadísticos o índices de asimetría.

7.2.2. Índice basado en los tres cuartiles

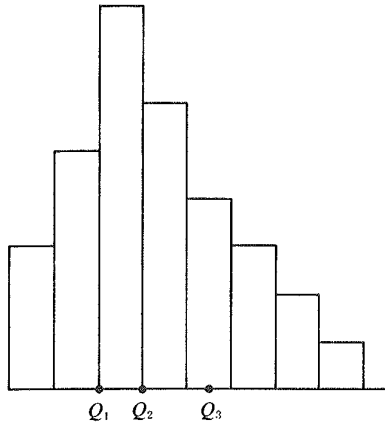
Llamaremos Q_1, Q_2 y Q_3 a los cuartiles primero (o percentil 25), segundo (o percentil 50, o mediana) y tercero (o percentil 75), respectivamente. Esto supuesto:

Si la distribución es simétrica, $Q_3 - Q_2 = Q_2 - Q_1$. Si es asimétrica positiva, $Q_3 - Q_2 > Q_2 - Q_1$. Si es asimétrica negativa, $Q_3 - Q_2 < Q_2 - Q_1$ (véase figuras adjuntas). De aquí tomar como índice de asimetría el siguiente criterio:

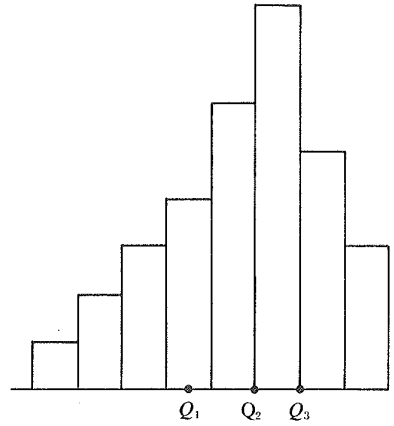
$$A_s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \qquad (7.1)$$

Si la distribución es asimétrica positiva, $A_s > 0$. Si es asimétrica negativa, $A_s < 0$. Si es simétrica, $A_s = 0$. Además, $-1 < A_s < 1$. En efecto, al tender $(Q_3 - Q_2)$ a cero, A_s tiende a -1; y al tender $(Q_2 - Q_1)$ a cero, A_s tiende a 1. El número así obtenido es invariante frente a cualquier transformación del origen y de la unidad de medida. En otras palabras, el número obtenido a partir de unas

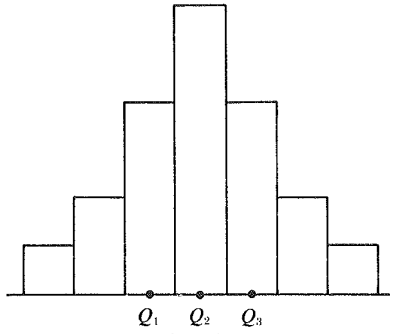
puntuaciones $X_1, X_2, \ldots, X_n$ sigue siendo el mismo que el obtenido a partir de las puntuaciones $Y_1 = AX_1 + B$, $Y_2 = AX_2 + B$, $\ldots$, $Y_n = AX_n + B$.



Asimetría positiva



Asimetría negativa



Simétrica

Nótese que al dividir por $(Q_3 - Q_1) = (Q_3 - Q_2) + (Q_2 - Q_1)$, hacemos que A_s sea un número abstracto, además de conseguir que quede dentro del intervalo $(-1, +1)$.

Conviene advertir lo siguiente acerca del criterio (7.1):

a) Si la distribución es simétrica, necesariamente $Q_3 - Q_2 = Q_2 - Q_1$ y, consiguientemente, $A_s = 0$. Pero si $Q_3 - Q_2 = Q_2 - Q_1$, la distribución puede ser simétrica o asimétrica.

b) Si $Q_3 - Q_2 \neq Q_2 - Q_1$, necesariamente la distribución es asimétrica. Pero, siendo la distribución asimétrica, tanto puede verificarse $Q_3 - Q_2 = Q_2 - Q_1$, como $Q_3 - Q_2 \neq Q_2 - Q_1$.

EJEMPLO 7.1. Los siguientes datos (una vez agrupados en intervalos) corresponden a la latencia media de una respuesta motora manifestada por 59 ratas blancas (véase Felsinger, Gladstone, Yamaguchi y Hull, 1947).

TABLA 7.1

Latencia (en segundos)	n_j	X_j	n_jX_j	$X_j - \bar{X}$	$(X_j - \bar{X})^2$	$n_j(X_j - \bar{X})^2$	$(X_j - \bar{X})^3$	$n_j(X_j - \bar{X})^3$
9,00-10,49	1	9,75	9,75	7,91	62,57	62,57	494,91	494,91
7,50-8,99	1	8,25	8,25	6,41	41,09	41,09	263,37	263,37
6,00-7,49	1	6,75	6,75	4,91	24,11	24,11	118,37	118,37
4,50-5,99	3	5,25	15,75	3,41	11,63	34,89	39,65	118,95
3,00-4,49	7	3,75	26,25	1,91	3,65	25,55	6,97	48,79
1,50-2,99	5	2,25	11,25	0,41	0,17	0,85	0,07	0,35
0,00-1,49	41	0,75	30,75	−1,09	1,19	48,79	−1,30	−53,30
	59		108,75			237,85		991,44

$\bar{X} = 1{,}84$ seg. $Md = Q_2 = 1{,}074$ seg. $Mod = 0{,}75$ seg. $Q_1 = 0{,}535$ seg. $Q_3 = 2{,}470$ seg. $S_x = 2{,}01$ seg.

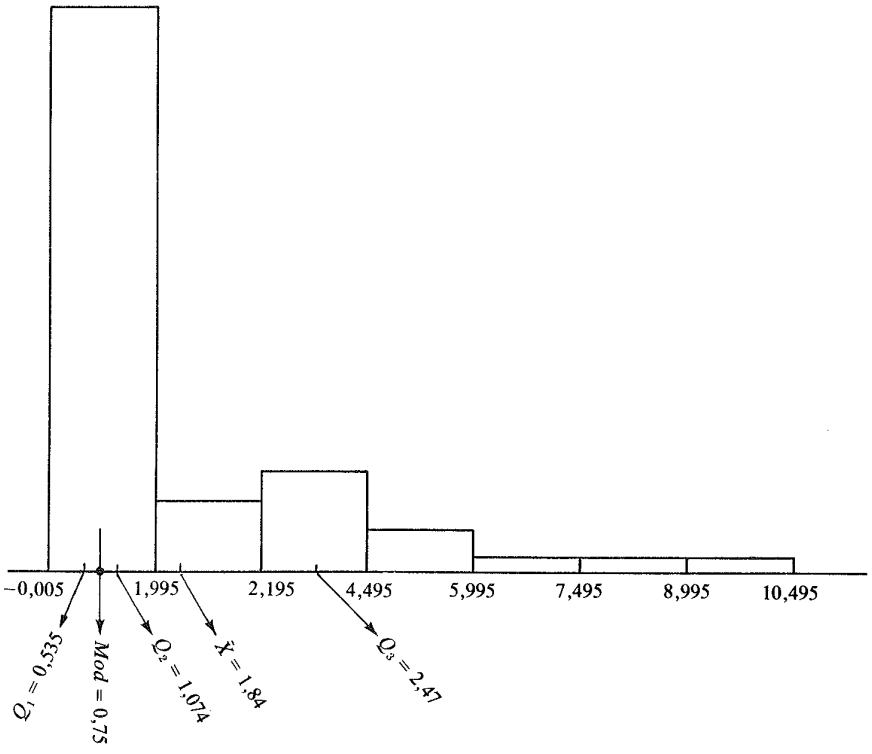


Fig. 7.3

Según (7.1):

$$A_s = \frac{(2{,}470 - 1{,}074) - (1{,}074 - 0{,}535)}{(2{,}470 - 1{,}074) + (1{,}074 - 0{,}535)} = \frac{1{,}396 - 0{,}539}{1{,}396 + 0{,}539} = \frac{0{,}857}{1{,}935} = 0{,}443$$

La asimetría es positiva. Consideremos el histograma correspondiente a la distribución de frecuencias anterior (Fig. 7.3).

EJEMPLO 7.2. En la tabla 7.2, dentro de cada intervalo tenemos los niños nacidos que, entre 10.000, probablemente presentarán el síndrome de Down (mongolismo) y cuyas madres tienen edades comprendidas dentro de dicho intervalo (véase Collman y Stoller, 1962). Como se ve, el riesgo de tener un hijo mongólico aumenta con la edad, por lo cual la distribución es asimétrica negativa.

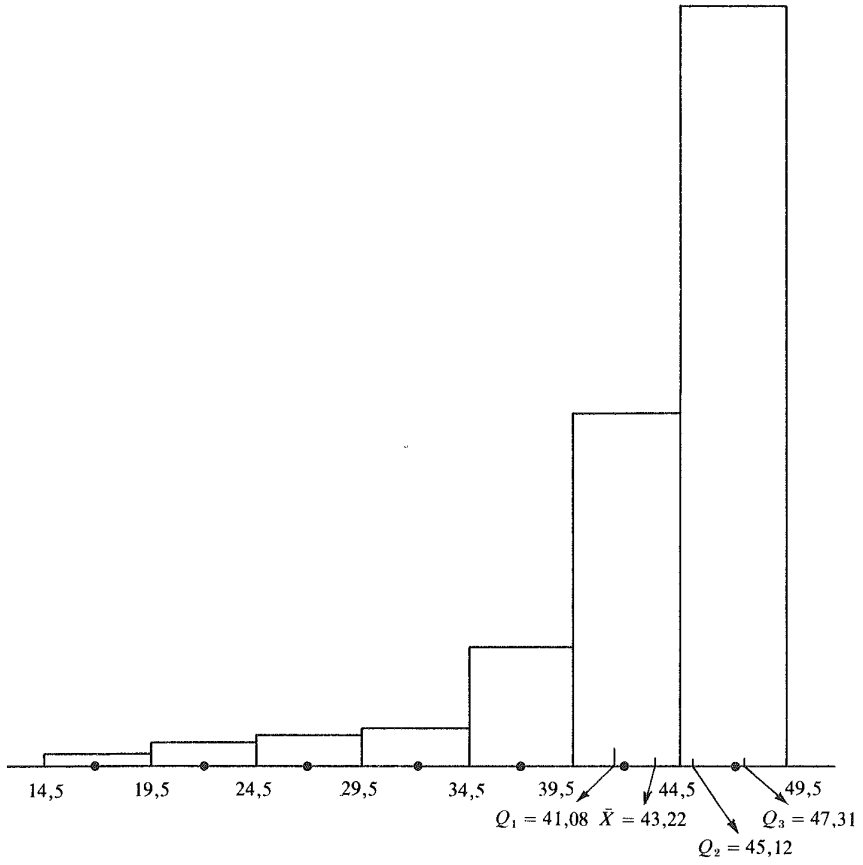


Fig. 7.4

TABLA 7.2

Edad	n_j	X_j	n_jX_j	$X_j - \bar{X}$	$(X_j - \bar{X})^2$	$n_j(X_j - \bar{X})^2$	$(X_j - \bar{X})^3$	$n_j(X_j - \bar{X})^3$
45-49	218	47	10.246	3,78	14,29	3.115,22	54,01	11.774,18
40-44	100	42	4.200	−1,22	1,49	149,00	−1,82	−182,00
35-39	35	37	1.295	−6,22	38,69	1.354,15	−240,64	−8.422,40
30-34	11	32	352	−11,22	125,89	1.384,79	−1.412,47	−15.537,17
25-29	8	27	216	−16,22	263,09	2.104,72	−4.267,29	−34.138,32
20-24	6	22	132	−21,22	450,29	2.701,74	−9,555,12	−57.330,72
15-19	4	17	68	−26,22	687,49	2.749,96	−18.025,95	−72.103,08
	382		16.509			13.559,58		−175.939,51

$\bar{X} = 43{,}22, \quad Md = Q_2 = 45{,}12, \quad Q_1 = 41{,}08, \quad Q_3 = 47{,}31, \quad s_x = 5{,}96$

Según 7.1:

$$A_s = \frac{(47{,}31 - 45{,}12) - (45{,}12 - 41{,}08)}{(47{,}31 - 45{,}12) + (45{,}12 - 41{,}08)} = \frac{2{,}19 - 4{,}04}{2{,}19 + 4{,}04} = \frac{-1{,}85}{6{,}23} = -0{,}297$$

La asimetría es negativa. Consideremos el histograma correspondiente a la distribución de frecuencias anterior (Fig. 7.4).

7.2.3. Índice basado en el momento de tercer orden

Llamaremos momento empírico de orden p respecto a la media, a la expresión:

$$m_p = \frac{\Sigma (X_i - \bar{X})^p}{n} \quad \text{ó} \quad m_p = \frac{\Sigma n_j(X_j - \bar{X})^p}{n}$$

En particular,

$$m_1 = \frac{\Sigma (X_i - \bar{X})}{n} \quad \text{ó} \quad m_1 = \frac{\Sigma n_j(X_j - \bar{X})}{n}$$

$$m_2 = \frac{\Sigma (X_i - \bar{X})^2}{n} \quad \text{ó} \quad m_2 = \frac{\Sigma n_j(X_j - \bar{X})^2}{n}$$

$$m_3 = \frac{\Sigma (X_i - \bar{X})^3}{n} \quad \text{ó} \quad m_3 = \frac{\Sigma n_j(X_j - \bar{X})^3}{n}$$

$$m_4 = \frac{\Sigma (X_i - \bar{X})^4}{n} \quad \text{ó} \quad m_4 = \frac{\Sigma n_j(X_j - \bar{X})^4}{n}$$

Pues bien, aceptaremos como índice de asimetría la expresión

$$a_3 = \frac{m_3}{m_2\sqrt{m_2}} = \frac{\Sigma (X_i - \bar{X})^3/n}{s_x^3} \quad \text{ó} \quad a_3 = \frac{\Sigma n_j (X_j - \bar{X})^3/n}{s_x^3} \qquad (7.2)$$

Según sabemos, $m_1 = 0$ siempre. El momento m_2 (o varianza) es siempre no negativo. Ninguno de los dos nos vale para medir la asimetría. El momento de tercer orden, m_3, puede ser positivo, negativo o nulo. Si la distribución es simétrica, $m_3 = 0$. En efecto, debido a la simetría, a cada diferencia $(X_i - \bar{X})$ positiva, le corresponde otra igual pero negativa. Consiguientemente, a cada cubo $(X_i - \bar{X})^3$ positivo le corresponde otro cubo igual pero negativo. Por tanto, la suma de todos los cubos será nula y $a_3 = 0$.

Por su parte, si la distribución es asimétrica positiva, las diferencias $(X_i - \bar{X})$ máximas son positivas y estas diferencias quedarán muy aumentadas al ser elevadas al cubo. Ello hace que la suma de las diferencias positivas, elevadas al cubo, sea mayor que la suma de las diferencias negativas, elevadas al cubo, aunque el número de estas últimas sea mayor que el de las diferencias positivas. Lo contrario ocurre cuando la asimetría es negativa.

En conclusión, a_3 es positivo cuando la asimetría es positiva y es negativo cuando la asimetría es negativa.

La razón de dividir por s_x^3 es conseguir que a_3 sea un número abstracto e independiente de la variabilidad del grupo.

EJEMPLO 7.3. Calculemos a_3 a partir de la tabla 7.1.

$$s_x^2 = \frac{237{,}85}{59} = 4{,}03, \quad s_x = 2{,}01, \quad s_x^3 = 8{,}10$$

$$a_3 = \frac{991{,}44/59}{8{,}10} = 2{,}08$$

Nótese que hay 41 diferencias $(X_i - \bar{X})$ negativas y sólo 18 positivas, pero la suma de éstas (elevadas al cubo) es mucho mayor que la suma de las primeras (elevadas, también, al cubo).

EJEMPLO 7.4. Calculemos a_3 a partir de la tabla 7.2.

$$s_x^2 = \frac{13.559{,}58}{382} = 35{,}50, \quad s_x = 5{,}96, \quad s_x^3 = 211{,}58$$

$$a_3 = \frac{-175.939{,}51/382}{211{,}58} = -2{,}18$$

Nótese que hay 218 diferencias $(X_i - \bar{X})$ positivas y sólo 164 negativas, pero la suma de éstas (elevadas al cubo) es mucho mayor que la suma de las primeras (elevadas, también, al cubo).

Conviene advertir lo siguiente acerca del criterio (7.2):

a) Si la distribución es simétrica, necesariamente $\Sigma (X_i - \bar{X})^3 = 0$ y, consiguientemente, $a_3 = 0$. Pero si $a_3 = 0$, la distribución puede ser simétrica o asimétrica.

b) Si $\Sigma (X_i - \bar{X})^3 \neq 0$, necesariamente la distribución es asimétrica, pero siendo la distribución asimétrica, puede verificarse tanto $\Sigma (X_i - X)^3 = 0$, como $\Sigma (X_i - \bar{X})^3 \neq 0$.

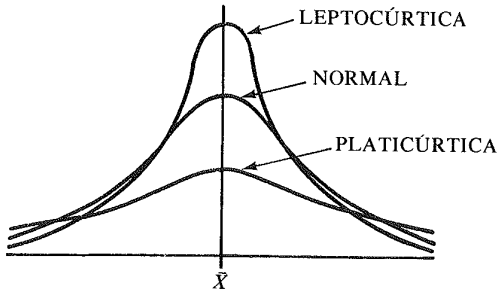
7.2.4. Nota

Hemos visto que si la distribución es simétrica y unimodal, media, mediana y moda coinciden y que si no coinciden, la distribución tiene que ser asimétrica. Fundados en ello, podemos introducir índices de asimetría. De hecho, han sido propuestos como índices de asimetría $A_s = \dfrac{\bar{X} - Mod}{s_x}$ y $A_s = \dfrac{3(\bar{X} - Md)}{s_x}$, esperando que $A_s > 0$ si la asimetría es positiva y que $A_s < 0$ si la asimetría es negativa. Lo que indudablemente es cierto es que $A = 0$ si la distribución es simétrica y que si $A_s \neq 0$, la distribución es asimétrica.

7.3. Apuntamiento

7.3.1. Idea previa

De un modo no muy riguroso diremos que una curva es muy apuntada si es muy alta y estrecha. Diremos que es poco apuntada si es baja y ancha. Eligiendo como patrón la curva normal (de la que hablaremos pronto), diremos que una curva es leptocúrtica si es más apuntada que la normal y que es platicúrtica si es menos apuntada que la normal. Diremos que la curva normal es mesocúrtica*.



* Del griego, λεπτός (leptós): delgado, estrecho; πλατύς (platís): ancho, extendido; μέσος (mesos): medio; χνρτός (kirtós): encorvado, convexo; χυρτότης (kirtotes): curvatura.

7.3.2. Índice basado en el momento de cuarto orden

Para dos curvas con la misma desviación típica, la más apuntada deberá contener más observaciones bajo los dos extremos alejados de la media que la menos apuntada. Para estas observaciones extremas, las diferencias $(X_i - \bar{X})$ serán grandes y serán mucho mayores al ser elevadas a la cuarta potencia. Por tanto, $\Sigma (X_i - \bar{X})^4$ será mayor para la curva leptocúrtica que para la platicúrtica. Por otra parte, se demuestra en Estadística Inferencial que $\dfrac{\Sigma (X_i - \bar{X})^4/n}{s_x^4} - 3 = 0$ en el caso de la curva normal. Por ello, suele ser elegido como criterio de apuntamiento o curtosis la expresión

$$a_4 = \frac{\Sigma (X_i - \bar{X})^4/n}{s_x^4} - 3$$

De lo expuesto se infiere que $a_4 > 0$ para las curvas leptocúrticas y $a_4 < 0$ para las platicúrticas.

La razón de dividir $\Sigma (X_i - \bar{X})^4/n$ por s_x^4 es análoga a la ofrecida en 7.2.3 respecto al índice de asimetría a_3.

EJEMPLO 7.5. Burt (1963) midió el cociente intelectual de un grupo numeroso de niños y obtuvo la distribución de frecuencias que, con ligeras modificaciones, exponemos a continuación.

CI	n_j	X_j	n_jX_j	$(X_j - \bar{X})$	$(X_j - \bar{X})^2$	$n_j(X_j - \bar{X})^2$	$n_j(X_j - \bar{X})^4/4.659$
150-159	5	154,5	772,5	55,74	3.106,95	15.534,75	10.359,652
140-149	18	144,5	2.601,0	45,74	2.092,15	37.658,70	16.910,811
130-139	84	134,5	11.298,0	35,74	1.277,35	107.297,40	29.417,433
120-129	253	124,5	31.498,5	25,74	662,55	167.625,15	23.837,570
110-119	747	114,5	85.531,5	15,74	247,75	185.069,25	9.841,172
100-109	1.217	104,5	127.176,5	5,74	32,95	40.100,15	283,560
90-99	1.148	94,5	108.486,0	−4,26	18,15	20.836,20	81,150
80-89	719	84,5	60.755,5	−14,26	203,35	146.208,65	6.381,375
70-79	294	74,5	21.903,0	−24,26	588,55	173.033,70	21.858,372
60-69	99	64,5	6.385,5	−34,26	1.173,75	116.201,25	29.274,664
50-59	47	54,5	2.561,5	−44,26	1.958,95	92.070,65	38.712,462
40-49	21	44,5	934,5	−54,26	2.944,15	61.827,15	39.070,210
30-39	7	34,5	241,5	−64,26	4.129,35	28.905,45	25.619,356
	4.659		460.145,5			1.192.368,45	251.647,787

$$\bar{X} = \frac{460.145,5}{4.659} = 98,76, \quad s_x^2 = \frac{1.192.368,45}{4.659} = 255,93, \quad s_x^4 = 65.500,16$$

$$a_4 = \frac{251.647,787}{65.500,16} - 3 = 3,84 - 3 = 0,84$$

La distribución es leptocúrtica, lo cual indica que existen más subnormales y más superdotados que los que habría habido si la distribución hubiera sido normal.

7.4. Resumen: Definiciones y fórmulas

Asimetría: Diremos que una distribución de frecuencias es simétrica si a cada puntuación X_{i+} distante de la media, por la derecha, una distancia k, le corresponde otra puntuación X_{i-} distante de la media, por la izquierda, la misma distancia k y tales que la frecuencia correspondiente a X_{i+} es la misma que la correspondiente a X_{i-}.

Diremos que es asimétrica toda distribución que no cumpla con lo acabado de indicar.

Índices de asimetría:

a) $$A_s = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$

b) $$a_3 = \frac{\Sigma (X_i - \bar{X})^3/n}{s_x^3}$$ (para datos no agrupados en intervalos)

$$a_3 = \frac{\Sigma n_j(X_j - \bar{X})^3/n}{s_x^3}$$ (para datos agrupados en intervalos)

Apuntamiento: Diremos que una distribución de frecuencias es muy apuntada o poco apuntada según que el correspondiente histograma sea alto y estrecho o bajo y ancho.

Índice de apuntamiento:

$$a_4 = \frac{\Sigma (X_i - \bar{X})^4/n}{s_x^4} - 3$$ (para datos no agrupados en intervalos)

$$a_4 = \frac{\Sigma n_j(X_j - \bar{X})^4/n}{s_x^4} - 3$$ (para datos agrupados en intervalos)

EJERCICIOS

7.1. Calcular el índice de asimetría basado en el momento de tercer orden a partir de los siguientes datos: 10, 12, 12, 14, 10, 10, 16, 12, 14, 10.

7.2. Calcular el índice de apuntamiento a partir de los datos del ejercicio anterior.

7.3. Sean n puntuaciones $X_1, X_2, \ldots, X_n$. Transformemos estas puntuaciones en las siguientes: $Y_1 = AX_1 + B$, $Y_2 = AX_2 + B$, $\ldots$, $Y_n = AX_n + B$. Demostrar que el índice de curtosis o apuntamiento, propuesto en el texto, de las nuevas puntuaciones Y_i es el mismo que el de las puntuaciones primitivas X_i.

7.4. Calcular los índices de asimetría, propuestos en el texto, a partir de las siguientes distribuciones de frecuencias:

a)

X	n_j
10-12	5
7-9	10
4-6	20
1-3	15

b)

X	n_j
9-11	2
6-8	10
3-5	6
0-2	2

7.5. Calcular el índice de curtosis, propuesto en el texto, a partir de las distribuciones del ejercicio anterior.

7.6. Supongamos que, siendo simétrica la distribución de frecuencias, el percentil 25 vale 6,5, $\Sigma n_j = 12$, $\Sigma n_j X_j = 114$, $\Sigma n_j X_j^2 = 1.259$. Esto supuesto, calcular el percentil 75 y el coeficiente de variación.

7.7. En una distribución simétrica de 40 elementos, $X_1, X_2, \ldots, X_{40}$, $\Sigma (X_i - 11)^2 = 1.040$, $\Sigma |X_i - k|$ es mínima para $k = 10$. Esto supuesto, calcular el coeficiente de variación de dichas puntuaciones.

8

Puntuaciones típicas

8.1. Puntuaciones directas, diferenciales y típicas

a) *Puntuación directa*

La atribuida directamente a cada objeto al ser sometido a cualquier tipo de prueba. Estas puntuaciones suelen ser designadas en Estadística Descriptiva, por letras mayúsculas latinas.

b) *Puntuación diferencial*

Puntuación directa menos la media. Estas puntuaciones suelen ser designadas, en Estadística Descriptiva, por letras minúsculas latinas.

c) *Puntuación típica*

Puntuación diferencial dividida por la desviación típica. Estas puntuaciones suelen ser designadas en Estadística Descriptiva, por la letra minúscula latina z.

EJEMPLO 8.1. Sean 6, 4, 2, 5, 8 los valores obtenidos directamente en una prueba por cinco personas. Su media vale 5 y su desviación típica vale 2. Esto supuesto, veamos cuánto valen las puntuaciones diferenciales y típicas.

Puntuaciones directas:	X_i	$= 6$	4	2	5	8
Puntuaciones diferenciales:	$x_i = X_i - \bar{X}$	$= 1$	-1	-3	0	3
Puntuaciones típicas:	$z_i = x_i/s_x = (X_i - \bar{X})/s_x$	$= 0{,}5$	$-0{,}5$	$-1{,}5$	0	1,5

Nótese que

$$x_i = X_i - \bar{X}$$

equivale a

$$x_i = AX_i + B$$

con $A = 1$ y $B = -\bar{X}$;

$$z_i = \frac{1}{s_x} X_i - \frac{\bar{X}}{s_x}$$

equivale a

$$z_i = AX_i + B$$

con $A = \dfrac{1}{s_x}$ y $B = -\dfrac{\bar{X}}{s_x}$.

8.2. Propiedades de las puntuaciones típicas

a) La media de las puntuaciones típicas vale cero.
En efecto,

$$\Sigma\, z_i = \Sigma\, (X_i - \bar{X})/s_x = (1/s_x)\, \Sigma\, (X_i - \bar{X}) = 0$$

pues según sabemos (ver 5.2.3.*a*)

$$\Sigma\, (X_i - \bar{X}) = 0$$

Consiguientemente,

$$\bar{z} = \Sigma\, z_i/n = (1/n)\, \Sigma\, z_i = 0$$

Podíamos, también, haber pensado así:

$$z_i = \frac{1}{s_x} X_i - \frac{\bar{X}}{s_x}$$

Por consiguiente, según 5.2.3.*c*,

$$\bar{z} = \frac{1}{s_x}\bar{X} - \frac{\bar{X}}{s_x} = 0$$

EJEMPLO 8.2. Comprobemos esta propiedad con los datos del ejemplo 8.1.

$$\bar{z} = \frac{(0,5) + (-0,5) + (-1,5) + (0) + (1,5)}{5} = \frac{0}{5} = 0$$

b) La varianza (y la desviación típica) de las puntuaciones típicas vale uno. En efecto,

$$s_z^2 = \frac{\Sigma\,(z_i - \bar{z})^2}{n} = \frac{\Sigma\, z_i^2}{n} = \frac{\Sigma\,(x_i/s_x)^2}{n} = \frac{\Sigma\, x_i^2}{(s_x^2)(n)} = \frac{1}{s_x^2}\frac{\Sigma\, x_i^2}{n} = \frac{1}{s_x^2}s_x^2 = 1$$

De lo acabado de ver se deduce que $\Sigma\, z_i^2 = n$.

Podíamos, también, haber pensado así:

$$z_i = \frac{1}{s_x}X_i - \frac{\bar{X}}{s_x}$$

Por consiguiente, según 6.3.4.*a*,

$$s_z^2 = \frac{1}{s_x^2}s_x^2 = 1$$

EJEMPLO 8.3. Comprobemos esta propiedad con los datos del ejemplo 8.1.

$$s_z^2 = \frac{(0,5)^2 + (-0,5)^2 + (-1,5)^2 + (0)^2 + (1,5)^2}{5} = \frac{5}{5} = 1$$

c) Si multiplicamos las puntuaciones típicas por una constante A y sumamos a esos productos otra constante B, las nuevas puntuaciones tienen como media B y como desviación típica $|A|$.

En efecto, acabamos de ver que

$$\bar{z} = 0 \quad \text{y} \quad s_z = 1$$

Ahora bien, las nuevas puntuaciones son

$$Y_i = Az_i + B$$

Por tanto,

$$\begin{aligned} \bar{Y} &= A\bar{z} + B = (A)(0) + B = B \text{ (recordando 5.2.3.}c\text{)} \\ s_y &= |A|s_z \quad = |A|(1) \quad = |A| \text{ (recordando 6.3.4.}a\text{)} \end{aligned}$$

EJEMPLO 8.4. Transformemos las puntuaciones típicas del ejemplo 8.1 mediante $Y_i = 10z_i + 20$ y comprobemos cómo la media y la desviación típica de las nuevas puntuaciones Y_i valen, respectivamente, $\bar{Y} = 20$, $s_y = 10$.

$$\begin{aligned} Y_1 &= (10)(0,5) + 20 = 25 \quad , \quad Y_2 = (10)(-0,5) + 20 = 15 \\ Y_3 &= (10)(-1,5) + 20 = 5 \quad , \quad Y_4 = (10)(0) + 20 = 20 \\ Y_5 &= (10)(1,5) + 20 = 35 \end{aligned}$$

$$\bar{Y} = \frac{25 + 15 + 5 + 20 + 35}{5} = \frac{100}{5} = 20$$

$$s_y^2 = \frac{(25 - 20)^2 + (15 - 20)^2 + (5 - 20)^2 + (20 - 20)^2 + (35 - 20)^2}{5} =$$

$$= \frac{500}{5} = 100 \quad , \quad s_y = 10$$

Basándonos en esta propiedad podemos convertir unas puntuaciones dadas $X_1, X_2, \ldots, X_n$, con media $\bar{X}$ y desviación típica s_x, en otras puntuaciones $Y_1, Y_2, \ldots, Y_n$ cuya media $\bar{Y}$ y cuya desviación típica s_y sean dos valores fijados de antemano por nosotros. En efecto, basta con transformar en típicas las puntuaciones primitivas, multiplicando, después, dichas puntuaciones típicas por s_y y sumando $\bar{Y}$ a cada uno de esos productos, donde s_y e $\bar{Y}$ son dos valores elegidos a nuestro arbitrio. Es decir, las nuevas puntuaciones serán:

$$Y_i = \frac{X_i - \bar{X}}{s_x}s_y + \bar{Y} = \frac{s_y}{s_x}X_i + \left(\bar{Y} - \frac{s_y}{s_x}\bar{X}\right)$$

EJEMPLO 8.5. Deseamos transformar las puntuaciones 21, 11, 1, 16, 31 obtenidas en un test en otras cuya media sea 100 y cuya desviación típica valga 20. Esto supuesto, ¿en qué puntuaciones se convertirán las puntuaciones dadas?

La media de estas puntuaciones vale 16 y la desviación típica vale 10. Por consiguiente,

$$Y_i = \frac{20}{10}X_i + (100 - \frac{20}{10}16) = 2X_i + 68$$

Es decir,

$$Y_1 = (2)(21) + 68 = 110 \quad , \quad Y_2 = (2)(11) + 68 = 90$$
$$Y_3 = (2)(1) + 68 = 70 \quad , \quad Y_4 = (2)(16) + 68 = 100$$
$$Y_5 = (2)(31) + 68 = 130$$

Las nuevas puntuaciones cumplen con las condiciones exigidas, $\bar{Y} = 100$, $s_y = 20$. En efecto:

$$\bar{Y} = \frac{110 + 90 + 70 + 100 + 130}{5} = \frac{500}{5} = 100$$

$$s_y^2 = \frac{(110 - 100)^2 + (90 - 100)^2 + (70 - 100)^2 + (100 - 100)^2 + (130 - 100)^2}{5} =$$

$$= \frac{2.000}{5} = 400$$

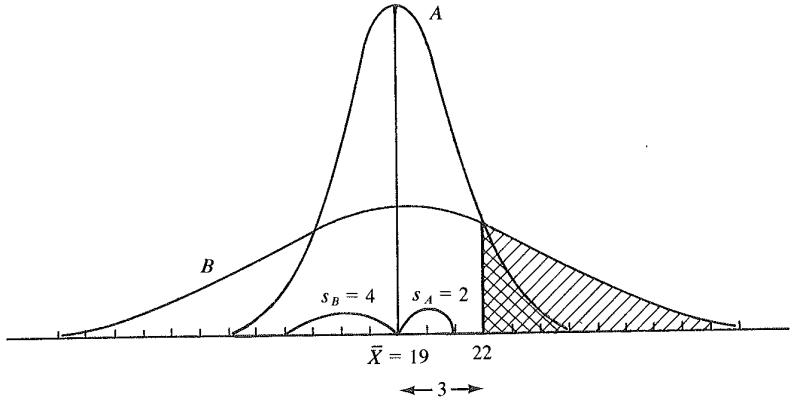
$$s_y = 20$$

8.3. Significado de las puntuaciones directas, diferenciales y típicas

Sabiendo que Pedro obtiene una puntuación directa igual a 22 en una prueba de memoria, nada podemos afirmar sobre su memoria. Es necesario conocer las puntuaciones obtenidas por las restantes personas del grupo al que pertenece Pedro. Este grupo puede ser reducido o amplio, puede constar de sus compañeros de clase, de los muchachos de su propia edad, de las personas de su nación, de todos los habitantes de la Tierra. Pero, pequeño o grande, es necesario un grupo de referencia para poder hacer la más mínima afirmación sobre la memoria de Pedro.

Supongamos que, definido dicho grupo, la media de éste vale 19. La puntuación diferencial de Pedro valdrá $22 - 19 = 3$. Por ser positiva comprobamos que Pedro está encima de la media de su grupo. Si hubiera sido negativa, Pedro se habría encontrado por debajo de la media. La puntuación diferencial nos permite afirmar algo sobre la memoria de Pedro, pero aún esta interpretación es bastante imprecisa. Superar la media en tres unidades, ¿es mucho o poco? Depende de los casos. Si nadie o casi nadie del grupo se aparta de la media tres o más unidades, tres significa mucho. Pero si bastantes la superan en más de tres unidades, tres significa mucho menos. Ahora bien, en general, en el primer caso la variabilidad del grupo (y, en concreto, la desviación típica) será pequeña. Por el contrario, en el segundo será grande. Por consiguiente, la interpretación de una misma puntuación diferencial será distinta según sea una u otra la variabilidad del grupo y, en concreto, la desviación típica.

Supongamos dos grupos A y B, tales que $s_A = 2$ y $s_B = 4$. La misma puntuación diferencial tres, significa más referida a A que a B. Ahora bien, esta diferente significación viene dada precisamente por las correspondientes puntuaciones típicas, $z_A = 3/2 = 1{,}5$, $z_B = 3/4 = 0{,}75$. Por consiguiente, la puntuación típica admite una interpretación más completa sobre la memoria de Pedro.



En conclusión, el significado de las puntuaciones directas, consideradas en sí mismas, es prácticamente nulo. Admiten un cierto significado, consideradas en relación con la tendencia central (media). Éste es aún más completo, consideradas en relación con la tendencia central (media) y con la variabilidad (desviación típica). Es decir, las puntuaciones típicas significan más que las diferenciales y éstas más que las directas.

Pronto veremos que, muy frecuentemente, en Psicología cada puntuación típica será, además, traducible en un porcentaje. Dada una puntuación típica, podremos calcular cuántas personas del grupo de referencia se encuentran por debajo de ella. Así, mediante las puntuaciones típicas podemos obtener una interpretación muy razonable sobre la memoria de Pedro.

8.4. Comparabilidad de las puntuaciones típicas

En principio, dos puntuaciones directas (o diferenciales) referidas a dos características distintas, no son comparables entre sí. Si, por ejemplo, las dos características son peso y altura, 70 kg ni son más ni son menos que 180 cm. Son dos cosas distintas y, por tanto, no comparables. Por el contrario, dos puntuaciones típicas son siempre comparables, al ser números abstractos; es decir, al no venir expresadas en ninguna unidad concreta de medida.

Más aún. En el caso de una sola característica, serían ya comparables dos puntuaciones directas y diferenciales, pues ambas vendrían expresadas en una misma unidad de medida. Sin embargo, aun en este caso, las puntuaciones típicas son más comparables entre sí que las directas y diferenciales. Supongamos que dos estudiantes, M y N, pertenecientes a dos grupos distintos realizan una misma prueba. Es claro que las puntuaciones típicas del primer grupo y las del segundo tienen una misma media (igual a cero) y una misma desviación típica (igual a uno). Por esta razón, las puntuaciones típicas de M y N, referidas a una misma media y a una misma desviación típica, son más comparables entre sí que sus puntuaciones directas, referidas probablemente a distintas medias y a distintas desviaciones típicas, y que sus puntuaciones diferenciales, referidas probablemente a distintas desviaciones típicas. Por consiguiente, en general, dos puntuaciones típicas admiten una comparabilidad mayor que las directas y diferenciales.

Con todo, esta mayor comparabilidad no implica que dos o más personas, pertenecientes a dos o más grupos distintos, sean iguales bajo cualquier aspecto por el mero hecho de haber obtenido las mismas puntuaciones típicas. Vamos a limitarnos a un par de casos elementales.

a) Dos grupos distintos realizan una misma prueba y obtienen la misma media y la misma desviación típica. Pues bien, si dos personas, una del primero y otra del segundo, obtienen la misma puntuación típica, obtendrán la misma puntuación directa (es decir, dentro de lo posible, manifestarán en el mismo grado la característica de que se trata), pero no dejarán necesariamente por debajo de sí el mismo porcentaje. Recíprocamente, si dos personas, una de cada grupo, dejan por debajo de sí el mismo porcentaje, no obtendrán necesariamente la misma puntuación directa (es decir, no manifestarán la característica en el mismo grado) ni, en consecuencia, obtendrán necesariamente la misma puntuación típica. Consideremos, por ejemplo, los dos grupos siguientes que, realizando una misma prueba, X, alcanzan la misma media y la misma desviación típica, pero tales que el primero muestra una clara asimetría positiva y el segundo una clara asimetría negativa.

Grupo 1.º

X_i	n_i	n_iX_i	$n_iX_i^2$
5	10	50	250
4	20	80	320
3	30	90	270
2	40	80	160
	100	300	1.000

$\bar{X}_1 = 3, \quad s_1 = 1$

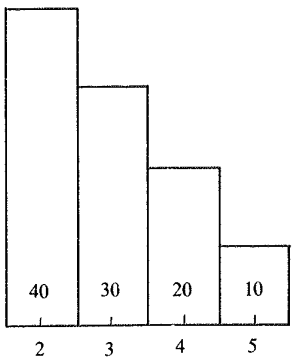
Grupo 2.º

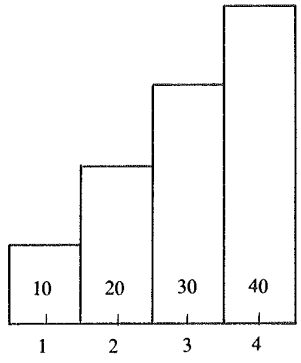
X_i	n_i	n_iX_i	$n_iX_i^2$
4	40	160	640
3	30	90	270
2	20	40	80
1	10	10	10
	100	300	1.000

$\bar{X}_2 = 3, \quad s_2 = 1$

Dos personas, A_1 (del grupo 1.º) y A_2 (del grupo 2.º), con la misma puntuación típica 0,5, obtendrán la misma puntuación directa: $X_1 = X_2 = 3 + (1)(0,5) = 3,5$.

Pero, A_1 dejará por debajo de sí el 70 por 100 de los casos y A_2 sólo el 60 por 100. Es decir, a iguales puntuaciones típicas (y directas) corresponden distintos porcentajes.





A su vez, A_1 con $z_1 = 0,5$, y A_2 con $z_2 = 0,75$, obtendrán puntuaciones directas distintas $X_1 = 3 + (1)(0,5) = 3,5$, $X_2 = 3 + (1)(0,75) = 3,75$, pero dejarán por debajo de sí el 70 por 100 en ambos casos. Es decir, a iguales porcentajes corresponden puntuaciones típicas (y directas) distintas.

b) Dos grupos distintos realizan una misma prueba y alcanzan diversa media y/o diversa desviación típica. Pues bien, si dos personas, una del grupo primero y otra del segundo, obtienen la misma puntuación típica, obtendrán, en general, distintas puntuaciones directas (es decir, manifestarán la característica en distinto grado) y dejarán por debajo de sí distintos porcentajes. Si dos personas, una del primer grupo y otra del segundo, obtienen la misma puntuación directa, (manifiestan la característica en el mismo grado), obtendrán, en general, distintas puntuaciones típicas y dejarán por debajo de sí distintos porcentajes. Consideremos, por ejemplo, los dos grupos siguientes que, realizando una misma prueba, X, alcanzan distinta media y distinta desviación típica, pero tales que el primero muestra una clara asimetría positiva y el segundo una clara asimetría negativa.

Grupo 1.º

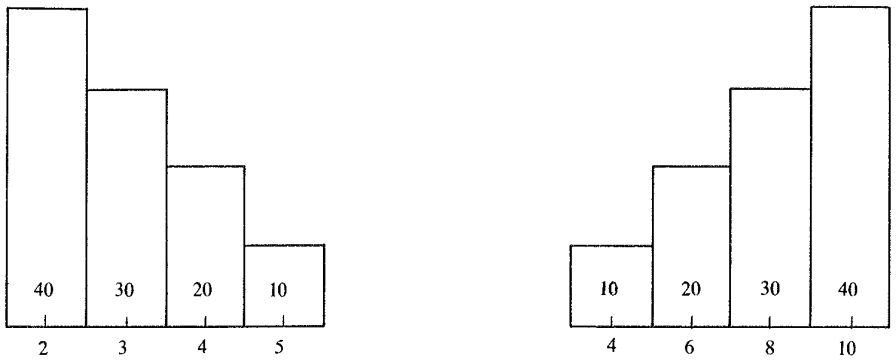
X_i	n_i	n_iX_i	$n_iX_i^2$
5	10	50	250
4	20	80	320
3	30	90	270
2	40	80	160
	100	300	1.000

$\bar{X}_1 = 3, \quad s_1 = 1$

Grupo 2.º

X_i	n_i	n_iX_i	$n_iX_i^2$
10	40	400	4.000
8	30	240	1.920
6	20	120	720
4	10	40	160
	100	800	6.800

$\bar{X}_2 = 8, \quad s_2 = 2$



Dos personas, A_1 (del grupo 1.º) y A_2 (del grupo 2.º), con la misma puntuación típica 0,5, obtendrán distintas puntuaciones directas: $X_1 = 3 + (1)(0,5) = 3,5$, $X_2 = 8 + (2)(0,5) = 9$ y dejarán por debajo de sí distintos porcentajes: 70 por 100 y 60 por 100 respectivamente. Es decir, a iguales puntuaciones típicas corresponden distintas puntuaciones directas y distintos porcentajes.

Dos personas, A_1 (del grupo 1.º) y A_2 (del grupo 2.º) con la misma puntuación directa 4,5, obtendrán distintas puntuaciones típicas: $z_1 = \frac{4,5 - 3}{1} = 1,5$, $z_2 = \frac{4,5 - 8}{2} = -1,75$ y dejarán por debajo de sí distintos porcentajes: 90 por 100 y 7,5 por 100 respectivamente. Es decir, a iguales puntuaciones directas corresponden distintas puntuaciones típicas y distintos porcentajes.

Dos personas, A_1 (del grupo 1.º) y A_2 (del grupo 2.º) que dejan por debajo de sí el mismo porcentaje (30 por 100), obtendrán distintas puntuaciones directas: $X_1 = 2,25$, $X_2 = 7$, y distintas puntuaciones típicas: $z_1 = \frac{2,25 - 3}{1} = -0,75$, $z_2 = \frac{7 - 8}{2} = -0,5$. Es decir, a iguales porcentajes, corresponden distintas puntuaciones directas y distintas puntuaciones típicas.

8.5. Nota

Acabamos de considerar dos criterios indicadores de la posición relativa de una persona respecto a un grupo de referencia: *a*) Posición relativa como distancia de esa persona a la media del grupo (medida en unidades típicas); *b*) Posición relativa como porcentaje de personas del grupo que deja por debajo de sí esa persona. Hemos visto que dos personas, pertenecientes a dos grupos distintos, cuya posición relativa es idéntica según uno de los dos criterios, no lo es necesariamente según el otro. Sin embargo, si las distribuciones de frecuencias de ambos grupos son iguales,

a posición relativa idéntica de dos personas según uno de los dos criterios, corresponde posición relativa idéntica de las mismas según el otro criterio. Y a posición distinta, según uno, corresponde posición relativa distinta según el otro.

Ahora bien, las distribuciones de frecuencias en Psicología suelen ser normales o aproximadamente normales con bastante frecuencia. En otras palabras, dos grupos distintos sometidos a la misma o distinta prueba suelen distribuirse de la misma manera, de acuerdo con la distribución normal, de la que hablaremos enseguida. Por consiguiente, con gran frecuencia, dos personas con la misma puntuación típica dejarán por debajo de sí, aproximadamente, el mismo porcentaje. Más aún, conociendo esa puntuación típica, podremos determinar cuál es el porcentaje que queda por debajo.

8.6. Combinación de puntuaciones

Para combinar en una sola puntuación total, las puntuaciones de varias pruebas, conviene operar con puntuaciones típicas. Es decir, reducir las unidades de las distintas pruebas a una unidad común. Esto es conveniente aun en el caso en que las pruebas sean muy semejantes (por ejemplo, varias notas de una misma asignatura combinadas para dar la nota final de curso) ya que superar la media en una misma distancia significa más cuando la desviación típica es pequeña que cuando es grande. De aquí la necesidad de operar con puntuaciones típicas, de tener en cuenta la desviación típica de cada prueba.

EJEMPLO 8.6. Sean tres pruebas I, II, III y dos alumnos A y B. Sean $\bar{X}_I = 45$, $s_I = 2$; $\bar{X}_{II} = 60$, $s_{II} = 4$; $\bar{X}_{III} = 60$, $s_{III} = 6$. Sean finalmente las puntuaciones siguientes las obtenidas por A y B en las tres pruebas:

	DIRECTAS				DIFERENCIALES				TIPICAS			
	I	II	III	TOTAL	I	II	III	TOTAL	I	II	III	TOTAL
A	40	64	66	170	−5	4	6	5	−2,5	1	1	−0,5
B	50	56	54	160	5	−4	−6	−5	2,5	−1	−1	0,5

El alumno A tiene una puntuación total (a partir de las directas y diferenciales) mayor que B. Sin embargo, el alumno B tiene una puntuación total (a partir de las típicas) mayor que A. El alumno B está por encima de la media en una sola prueba, pero donde la variabilidad es muy pequeña. En cambio, el alumno A está por encima de la media en dos pruebas, pero donde la variabilidad es muy grande. En conjunto, el alumno B debería ser preferido al A, pues su puntuación típica total es mayor que la de A.

8.7. Desviación típica y puntuaciones típicas

Desviación típica y puntuación típica son conceptos distintos. La primera es propia del grupo. La segunda es propia de cada persona. En un grupo de n personas, tenemos n puntuaciones típicas (algunas de las cuales pueden ser iguales entre sí) y una sola desviación típica. Sin embargo, es equivalente decir que una persona obtiene una puntuación típica igual a dos o que supera a la media en dos desviaciones típicas. Por ejemplo, suponiendo $\bar{X} = 60$, $s_x = 3$, una persona con puntuación directa $X_i = 66$, tendrá como puntuación típica:

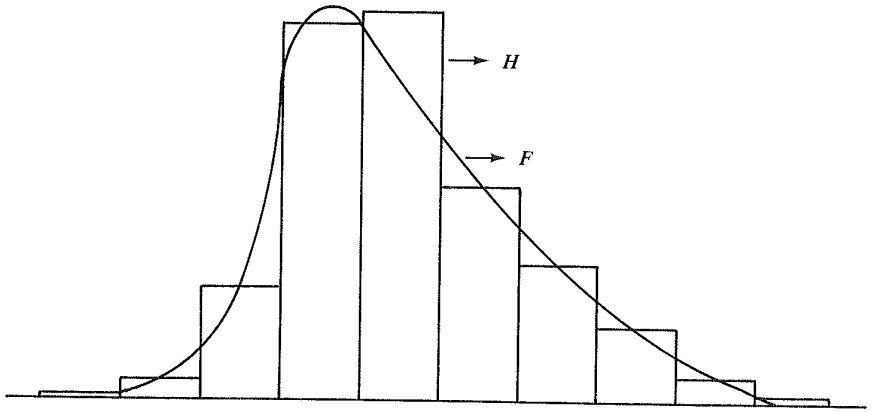
$$z_i = \frac{66 - 60}{3} = \frac{6}{3} = 2$$

Esto equivale a decir que su distancia de la media ($66 - 60 = 6$ unidades) contiene dos veces la desviación típica: $3, \left(\frac{6}{3} = 2\right)$, o, a decir que supera la media en dos desviaciones típicas.

8.8. Puntuaciones típicas y curva normal

8.8.1. Límite del histograma con intervalos infinitamente pequeños

Al aumentar indefinidamente el número de intervalos, disminuye su amplitud y los rectángulos del histograma se adelgazan más y más. En el límite, la línea quebrada H se identificará con la curva F. Esta curva es, con gran frecuencia, en Psicología la llamada curva normal o campana de Gauss.



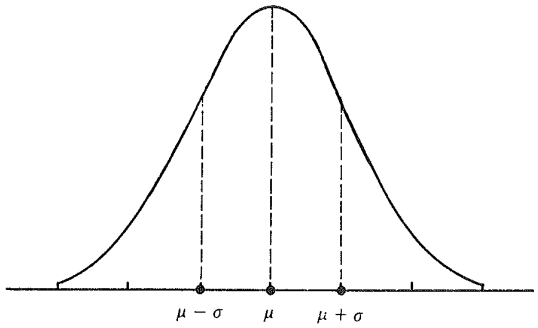
8.8.2. Curva normal

De ella trataremos ampliamente en el tomo 2 (Estadística Inferencial). Sin embargo, vamos a ofrecer ahora una idea sucinta sobre la misma. Suponemos que operamos con proporciones. Ello significa que vale la unidad el área limitada por el histograma o (en el límite) por la curva F y el eje de abscisas. Bajo esta condición, la curva normal es la representación gráfica de la siguiente función:

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

μ = media de la población, σ = desviación típica de la población, e = base de los logaritmos neperianos: 2,718281828..., π = relación de la circunferencia a su diámetro: 3,141592...

Para cada par de valores de μ y de σ tendremos una curva normal distinta. Es decir, tenemos una familia de curvas. Pero todas ellas coinciden en algunas características (véase la figura adjunta):



a) Tienen un único máximo para $X = \mu$.

b) Tienen dos puntos de inflexión, para $x = \mu - \sigma$ y para $X = \mu + \sigma$. Es decir, en el punto $X = \mu - \sigma$ la curva pasa de ser cóncava hacia arriba a ser cóncava hacia abajo y en el punto $X = \mu + \sigma$ la curva pasa de ser cóncava hacia abajo a ser cóncava hacia arriba.

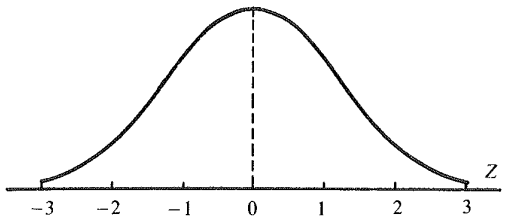
c) Se acercan asintóticamente al eje de abscisas. En otras palabras, se acercan más y más a ese eje, tanto por la derecha como por la izquierda, sin llegar a tocarlo en ningún punto finito.

d) Son simétricas respecto al eje vertical que pasa por la media.

Suele ser útil operar con puntuaciones típicas. En este caso la ecuación de la curva normal viene dada por

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

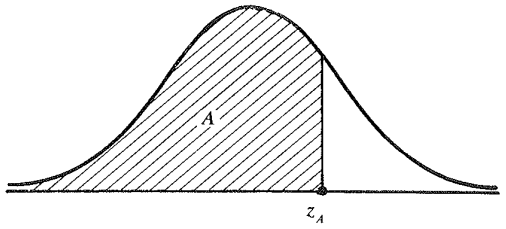
y su representación gráfica es la siguiente:



Entre tres desviaciones típicas por la izquierda y tres por la derecha (es decir, entre la puntuación típica -3 y la puntuación típica 3) se encuentra el 99,74 por 100 del área total contenida bajo la curva normal.

8.8.3. Relación entre las áreas bajo la curva normal y proporciones o probabilidades

En la figura adjunta el área A representa la probabilidad de obtener una puntuación igual o menor que z_A. Interpretada esta probabilidad como proporción, el área anterior representa la proporción de observaciones con puntuaciones iguales o menores que z_A, en el supuesto, claro está, de que dichas observaciones se distribuyan normalmente.



Por ahora, serán equivalentes para nosotros probabilidad y proporción. En el tomo 2 discutiremos más detalladamente esta equivalencia. Naturalmente, las áreas representadas serán porcentajes si multiplicamos las proporciones por cien.

8.8.4. Uso de la tabla de las áreas bajo la curva normal

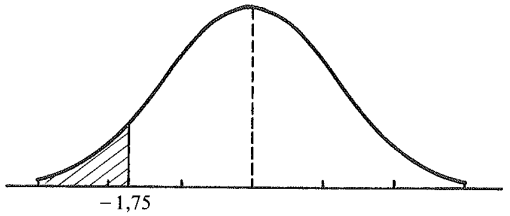
La tabla A (Apéndice III) ofrece con cada puntuación típica la proporción (porcentaje si la multiplicamos por 100) o área situadas bajo dicha puntuación, suponiendo, naturalmente, que la distribución de frecuencias es normal.

EJEMPLO 8.7. ¿Qué porcentaje de observaciones queda por debajo de la puntuación directa $X = 23$, valiendo 30 la media y 4 la desviación típica?

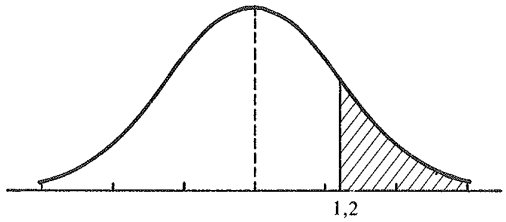
Transformemos esa puntuación directa en típica:

$$z_{23} = \frac{23 - 30}{4} = -1{,}75$$

En la tabla, a $z = -1{,}75$ le corresponde un porcentaje igual a 4,01. Por tanto, por debajo de la puntuación típica $z = -1{,}75$ (es decir de la directa $X = 23$) queda el 4,01 por 100 de las observaciones.



EJEMPLO 8.8. ¿Qué porcentaje de observaciones queda por encima de la puntuación directa $X = 54$, valiendo 48 la media y 5 la desviación típica?

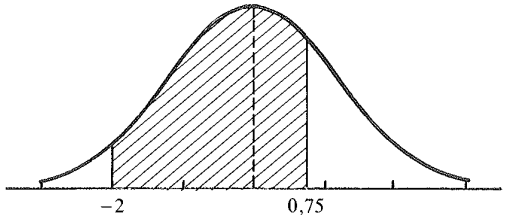


Transformemos esa puntuación directa en típica:

$$z_{54} = \frac{54 - 48}{5} = 1{,}2$$

En la tabla, a $z = 1{,}2$ le corresponde un porcentaje igual a 88,49. Por tanto, por encima de la puntuación típica $z = 1{,}2$ (es decir, de la directa $X = 54$) queda el (100 − 88,49) por 100 = 11,51 por 100 de las observaciones.

EJEMPLO 8.9. ¿Qué porcentaje de observaciones queda por encima de la puntuación directa $X = 9$ y, simultáneamente, por debajo de la puntuación directa, $X = 31$, valiendo 25 la media y 8 la desviación típica?



Transformemos esas puntuaciones directas en típicas.

$$z_9 = \frac{9 - 25}{8} = -2 \quad , \quad z_{31} = \frac{31 - 25}{8} = 0{,}75$$

En la tabla, a $z = -2$ le corresponde un porcentaje igual a 2,28 y a $z = 0{,}75$ un porcentaje igual a 77,34. Por tanto, por encima de la puntuación típica $z = -2$ (es decir, de la directa $X = 9$) y, simultáneamente, por debajo de la puntuación típica $z = 0{,}75$ (es decir, de la puntuación directa $X = 31$) queda el (77,34 − − 2,28) por 100 = 75,06 por 100 de las observaciones.

EJEMPLO 8.10. ¿Qué puntuación directa deja por debajo de sí el 64 por 100 de las observaciones, valiendo 30 la media y 5 la desviación típica?

En la tabla, el porcentaje más próximo al 64 por 100 es el 64,06 por 100 y a este porcentaje le corresponde la puntuación típica $z = 0{,}36$. Por tanto:

$$0{,}36 = \frac{X - 30}{5} \quad ; \quad X = (5)(0{,}36) + 30 = 1{,}8 + 30 = 31{,}8$$

EJEMPLO 8.11. ¿Qué puntuación directa deja por encima de sí el 61 por 100 de las observaciones, valiendo 40 la media y 6 la desviación típica?

Dejar por encima de sí el 61 por 100 de las observaciones equivale a dejar por debajo de sí el 39 por 100 de las mismas. En la tabla, el porcentaje más próximo al 39 por 100 es el 38,97 por 100 y a este porcentaje le corresponde la puntuación típica $z = -0{,}28$. Por tanto,

$$-0{,}28 = \frac{X - 40}{6} \quad ; \quad X = (6)(-0{,}28) + 40 = 40 - 1{,}68 = 38{,}32$$

EJEMPLO 8.12. Calculemos dos puntuaciones directas, X_1 y X_2, tales que la primera deje por debajo de sí un 10 por 100 de las observaciones y la segunda deje por encima de sí otro 10 por 100 de los casos, valiendo 40 la media y 7 la desviación típica.

En la tabla, el porcentaje más próximo al 10 por 100 es el 10,03 por 100 y a este porcentaje le corresponde la puntuación típica $z = -1{,}28$. Por otra parte, dejar por encima el 10 por 100 equivale a dejar por debajo el 90 por 100. En la tabla el porcentaje más próximo al 90 por 100 es el 89,97 por 100 y a este porcentaje le corresponde la puntuación típica $z = 1{,}28$. Por tanto,

$$-1{,}28 = \frac{X_1 - 40}{7} \quad ; \quad X_1 = (7)(-1{,}28) + 40 = 40 - 8{,}96 = 31{,}04$$

$$1{,}28 = \frac{X_2 - 40}{7} \quad ; \quad X_2 = (7)(1{,}28) + 40 = 40 + 8{,}96 = 48{,}96$$

8.9. Puntuaciones T

Las puntuaciones típicas ofrecen un doble inconveniente. En primer lugar, unas son positivas y otras negativas (circunstancia que puede ocasionar errores en los cálculos). En segundo lugar, casi todas las observaciones suelen quedar contenidas dentro de tres desviaciones típicas a la derecha de la media (igual a cero) y otras tres desviaciones típicas a la izquierda de la misma. Es decir, sólo tendremos 7 puntuaciones enteras posibles (−3, −2, −1, 0, 1, 2, 3); todas las demás serán decimales (con los consiguientes inconvenientes para el cálculo).

Para evitar los decimales (o, al menos, bastantes de ellos) multiplicamos las puntuaciones típicas por una constante apropiada. Para evitar los valores negativos, sumamos a los productos obtenidos otra constante adecuada. En particular, suelen ser usadas, respectivamente, 10 y 50. En otras ocasiones, 100 y 500 u otras constantes que nos sirvan para conseguir el fin pretendido.

Dentro de este contexto, llamaremos puntuaciones T a las obtenidas mediante la constante multiplicadora 10 y la constante aditiva 50, pero tras previa norma-

lización de la distribución de frecuencias. Ello equivale a calcular las puntuaciones típicas valiéndonos de la tabla de las áreas bajo la curva normal y no mediante la media y la desviación típica de las puntuaciones directas que nos son dadas. Las puntuaciones típicas así calculadas son las que se multiplican por 10 y se les añade la constante 50.

EJEMPLO 8.13. Calculemos las puntuaciones T a partir de la siguiente distribución de frecuencias:

X_j	n_j	Frec. ac. (pm)	Porc. ac. (pm)	z_j	$(10)(z_j)$	$(10)(z_j) + 50$
18-20	7	76,5	95,63	1,71	17,1	67,1
15-17	8	69	86,25	1,09	10,9	60,9
12-14	15	57,5	71,88	0,58	5,8	55,8
9-11	16	42	52,50	0,06	0,6	50,6
6-8	22	23	28,75	−0,56	−5,6	44,4
3-5	12	6	7,50	−1,44	−14,4	35,6
	80					

Frec. ac. (pm): frecuencia acumulada hasta el punto medio. Es decir, hasta la mitad del intervalo (3-5) habrá $\frac{12}{2} = 6$ observaciones. Hasta la mitad del intervalo (6-8) habrá $12 + \frac{22}{2} = 23$. Etc.

Porc. ac. (pm): porcentaje acumulado hasta el punto medio. Basta con dividir cada frecuencia acumulada hasta el punto medio por 80 y multiplicar este cociente por 100.

z_j: puntuación típica normalizada. Es decir, $-1,44$ es la puntuación típica que, supuesta una distribución normal, deja por debajo de sí el 7,5 por 100 de las observaciones. Etc.

8.10. Resumen: Definiciones y fórmulas

Puntuación directa: la atribuida directamente a cada elemento de un grupo.

Puntuación diferencial: puntuación directa menos la media del grupo.

Puntuación típica: puntuación diferencial dividida por la desviación típica del grupo.

Curva normal: representación gráfica de la siguiente función.

$$y = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \qquad \text{(en puntuaciones directas)}$$

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \qquad \text{(en puntuaciones típicas)}$$

EJERCICIOS

8.1. Calcular las correspondientes puntuaciones diferenciales y típicas a partir de las siguientes puntuaciones directas.

a) 2, 6, 8, 5, 4 *b*) 7, 4, 1, 5, 3 *c*) 1, 11, 11, 1

d) 2, 3, 5, 2 *e*) 1, 4, 7

8.2. Suponiendo $\bar{X} = 20$ y $s_x = 4$, ¿qué puntuaciones diferenciales y directas corresponderán a las siguientes puntuaciones típicas?

a) $z_x = 2$; *b*) $z_x = 1,5$; *c*) $z_x = -1$; *d*) $z_x = -0,25$; *e*) $z_x = 0,9$

8.3. Supongamos que $X_1, X_2, \ldots, X_n$ son ciertas puntuaciones directas y que $x_1, x_2, \ldots, x_n$ son sus correspondientes puntuaciones diferenciales. Esto supuesto, demostrar que $\Sigma X_i y_i = \Sigma x_i Y_i$

8.4. Calcular el coeficiente de variación de X, sabiendo que

$$z_{x_1} = z_{y_1} \quad ; \quad z_{x_2} = z_{y_2}, \ldots, z_{x_5} = z_{y_5} \quad ; \quad \text{que } s_y = 4,$$

que el coeficiente de variación de Y vale 20 y teniendo en cuenta el cuadro adjunto.

X_i	Y_i
·	18
6	22
·	26
2	·
·	20

8.5. ¿Son típicas las puntuaciones $-3, 1, 1, 0, 1, 0$, obtenidas por seis personas?

8.6. Transformar las puntuaciones 9, 5, 7, 1, 13 en otras cuya media valga 50 y cuya desviación típica valga 12.

8.7. Sean $X_1, X_2, \ldots, X_n$, n puntuaciones con media $\bar{X} = 10$. Consideremos las puntuaciones $3X_1, 3X_2, \ldots, 3X_n$ cuya desviación típica nos es dada y vale 6. Esto supuesto, calcular la varianza de las $2n$ puntuaciones $X_1, X_2, \ldots, X_n, 3X_1, 3X_2, \ldots, 3X_n$.

8.8. Transformar las puntuaciones $X_1 = 7$, $X_2 = 5$, $X_3 = 3$, $X_4 = 6$, $X_5 = 9$ en otras puntuaciones Y tales que $\bar{Y} = 2\bar{X} - 5$, $s_y = 2s_x$.

8.9. Ponga las puntuaciones que faltan en el cuadro siguiente, sabiendo que $z_{x_1} = z_{y_1}, z_{x_2} = z_{y_2}, \ldots, z_{x_5} = z_{y_5}$, que $s_x = 2$ y que los coeficientes de variación de X y de Y valen 20 y 50, respectivamente.

X_i	Y_i
13	·
·	·
7	4
10	·
11	·

8.10. Sean $\bar{X} = 50$ y $s_x = 4$ la media y la desviación típica de n puntuaciones $X_1, X_2, \ldots, X_n$. Esto supuesto, ¿cuánto valdrá el coeficiente de variación de las puntuaciones $Y_1 = 3X_1 - 30$, $Y_2 = 3X_2 - 30, \ldots, Y_n = 3X_n - 30$?

8.11. Sabiendo que los coeficientes de variación de X e Y valen, respectivamente, 40 y 50, que $s_y = (3/4)s_x$ y que la puntuación típica correspondiente a 24 en X es la misma que la correspondiente a 15 en Y, calcular $\bar{X}$, $\bar{Y}$, s_x y s_y.

8.12. Aplicado un test H a un grupo normativo de personas, las puntuaciones directas obtenidas por éstas han sido transformadas mediante la ecuación $3X + 45$ con el fin de que su media fuera 100 y su desviación típica, 20. El mismo test H es aplicado a tres nuevas personas que obtienen las puntuaciones directas 15, 22, 25. En este supuesto: *a*) ¿Qué haría usted para comparar estas tres puntuaciones con las puntuaciones transformadas del grupo normativo? *b*) Realizada la transformación apropiada ¿a cuántas desviaciones típicas se encuentran estas tres personas por encima o por debajo de la media?

8.13. Suponiendo que $\bar{X} = 30, s_x = 4, n = 150$ y que la distribución de frecuencias es normal, calcular el porcentaje y el correspondiente número de observaciones con puntuaciones: *a*) Menores que 24. *b*) Menores que 34. *c*) Mayores que 28. *d*) Mayores que 33. *e*) Mayores que 22 y menores que 28. *f*) Mayores que 32 y menores que 35. *g*) Mayores que 28 y menores que 34.

8.14. Suponiendo que $\bar{X} = 50$, $s_x = 8$ y $n = 500$ y que la distribución de frecuencias es normal, calcular la puntuación directa que deja por debajo o por encima de sí los siguientes porcentajes o número de observaciones. *a*) Que deja por debajo el 11 por 100. *b*) Que deja por debajo 220 observaciones. *c*) Que deja por encima el 48 por 100. *d*) Que deja por encima 320 observaciones. *e*) Dos puntuaciones (equidistantes de la media, a uno y otro lado de la misma) que dejan entre ambas el 54 por 100. *f*) Dos puntuaciones (equidistantes de la media) que dejan entre ambas 100 observaciones.

8.15. Calcular la media de un grupo de personas, suponiendo que la distribución de frecuencias es normal y sabiendo que la desviación típica vale 10 y que el 40 por 100 obtiene puntuaciones menores que 28.

8.16. Calcular la media de un grupo de 200 personas, distribuidas normalmente, sabiendo que la desviación típica vale 8 y que 15 personas obtienen puntuaciones mayores que 28.

8.17. Calcular la desviación típica de un grupo de personas, distribuidas normalmente, sabiendo que la media vale 47,7 y que el 98 por 100 obtiene puntuaciones menores que 60.

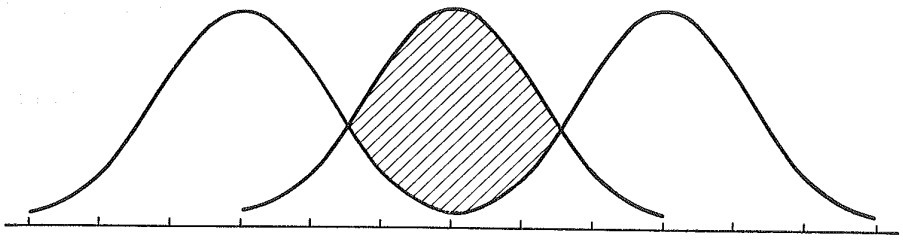
8.18. Calcular la desviación típica de un grupo de 120 personas, distribuidas normalmente, sabiendo que la media vale 44,24 y que 12 personas obtienen puntuaciones mayores que 50.

8.19. Calcular la media y la desviación típica de un grupo de 500 personas, distribuidas normalmente, sabiendo que 100 personas de dicho grupo han obtenido puntuaciones directas mayores que 64,20 y 130 han obtenido puntuaciones directas menores que 56,80.

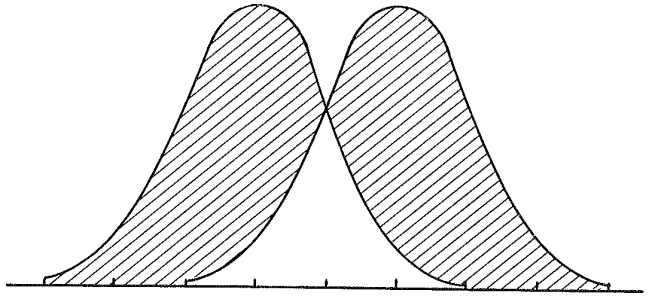
8.20. Calcular la media y la desviación típica de un grupo de 5.000 personas, distribuidas normalmente, sabiendo que el primer cuartil vale 30,65 y que 3.203 personas han obtenido puntuaciones directas mayores que 32,2.

8.21. Calcular el número de personas que obtendrán puntuaciones directas mayores que 27,9 en un grupo de 50.000 cuya distribución de frecuencias es normal, cuya amplitud semiintercuartil vale 2,01 y cuyo coeficiente de variación vale 10.

8.22. Sean 20, 26 y 32 las medias obtenidas por tres grupos distintos, siendo normales las tres distribuciones de frecuencias. Cada grupo consta de 150 personas y éstas se extienden desde tres desviaciones típicas a la izquierda de su media hasta tres desviaciones típicas a la derecha de la misma. Esto supuesto, ¿cuántas personas quedarán dentro del área rayada en el gráfico adjunto, siendo $s_1 = s_2 = s_3$?



8.23. Sean dos grupos A y B, cada uno de ellos con 300 personas y distribuido normalmente. En cada grupo las 300 personas se extienden desde tres desviaciones típicas a la izquierda de su media hasta tres desviaciones típicas a la derecha de la misma. Suponiendo, además, que $s_A = s_B$, ¿cuántas personas quedarán dentro del área rayada en el gráfico adjunto?



8.24. Hemos aplicado una prueba de memoria a 5.000 varones y a 2.000 mujeres separadamente. Son normales tanto la distribución de los varones como la de las mujeres. Del total de los varones, 1.148 superan la media de las mujeres. Además, 134 mujeres han obtenido puntuaciones superiores a 70. Suponiendo que la media y la desviación típica de los varones valen 54,3 y 5 respectivamente, calcular la media y la desviación típica de las mujeres.

8.25. Un grupo de personas se distribuye normalmente en la variable X. Para estas personas el percentil 61 vale 44 y la desviación típica vale 5. Calcular la media y la desviación típica de unas nuevas puntuaciones Y relacionadas con las primeras mediante $4X - 2Y - 20{,}4 = 0$.

8.26. De un grupo de 600 personas, distribuidas normalmente, 114 obtienen una puntuación directa menor que 20. Sabiendo que la varianza vale 16, calcular la moda de dicha distribución y el número de personas con puntuaciones mayores que 24 y menores que 28.

8.27. Las puntuaciones de un grupo de 500 niños en inteligencia espacial, X, y en aptitud mecánica, Y, se distribuyen normalmente. Calcular el porcentaje de niños cuyas puntuaciones en X superan la media de las puntuaciones en Y, sabiendo que $Md_y - Md_x = 15$, $Md_y + Md_x = 95$ y $\Sigma X_i^2 = 850.000$.

8.28. Calcular el coeficiente de variación de un grupo de puntuaciones, sabiendo que la distribución es normal, que $Q_3 - Q_1 = 5{,}36$ y que $P_{67} = 62{,}96$.

8.29. Un grupo de 400 personas responde a un cuestionario y sus respuestas se distribuyen normalmente. El coeficiente de variación vale 25. Por encima de la puntuación directa 23,85 se encuentran 88 personas. Calcular la media y la desviación típica.

8.30. ¿Es simétrica toda curva normal?

8.31. ¿Es normal toda curva simétrica?

8.32. Siempre que conozcamos unas puntuaciones típicas, ¿podemos determinar el porcentaje que deja cada una de ellas por debajo o por encima de sí, acudiendo a la tabla de las áreas bajo la curva normal?

8.33. Sean $X_1, X_2, \ldots, X_n$ las puntuaciones directas obtenidas por n personas en un test de extroversión. Supongamos que su distribución de frecuencias es asimétrica positiva. Transformamos en típicas las puntuaciones directas anteriores. Esto supuesto, ¿será normal (o, al menos, simétrica) la distribución de frecuencias de estas puntuaciones típicas?

8.34. Transforme en puntuaciones T (normalizadas) las siguientes puntuaciones directas:

a)

X_j	n_j
35	5
33	6
31	7
29	12
27	16
25	4

b)

X	n_j
26 – 28	4
23 – 25	8
20 – 22	10
17 – 19	14
14 – 16	12
11 – 13	2

III

Estudio conjunto de dos variables

9

Organización de datos e índices de tendencia central y variabilidad

9.1. Distribución conjunta de frecuencias

Hasta aquí hemos considerado una sola variable. Ahora vamos a estudiar conjuntamente dos variables. Por ejemplo, peso y altura de un grupo de estudiantes, aptitud para una asignatura y aprovechamiento en la misma, provincia de origen y carrera estudiada, etc. Con cada persona tenemos dos modalidades, una perteneciente a la primera variable y otra a la segunda.

Desde luego, podíamos ir estudiando por separado cada uno de los casos posibles: ambas variables nominales, ambas ordinales, . . . ; una nominal y otra ordinal, una nominal y otra de intervalos, etc. Sin embargo, no seguiremos este camino por una doble razón. En primer lugar, sería enormemente prolijo irnos deteniendo en cada uno de los casos posibles, dado su gran número. En segundo lugar, ello nos llevaría a repeticiones superfluas ya que lo dicho para uno de los casos, vale prácticamente para los restantes, salvo diferencias accidentales fácilmente comprensibles. Consiguientemente, nos limitaremos por ahora a exponer la distribución conjunta de frecuencias respecto al caso en que las variables sean estrictamente cuantitativas (es decir, a nivel, al menos, de intervalos), por ser el más común en Psicología.

EJEMPLO 9.1. Supongamos que 50 personas han obtenido, según podemos ver en la tabla de la página siguiente, los siguientes resultados en un test de inteligencia abstracta (X) y en una prueba de aritmética (Y).

Podemos considerar estas puntuaciones tal como vienen dadas, es decir, no agrupadas en intervalos ni en X ni en Y. Cada persona aparece con el par de puntuaciones que ha obtenido directamente en el test y en el examen. Pero, también, podemos considerarlas agrupadas en intervalos. Para ello, estudiando por separado cada una de las dos variables, agrupamos sus puntuaciones en intervalos, siguiendo los criterios ya expuestos para agrupar datos en el caso de una sola variable. El número de intervalos en cada una de las dos variables puede ser el mismo o distinto. La amplitud de los intervalos en una de las dos variables puede ser la misma o distinta que la amplitud de los intervalos en la otra. Elijamos para X los intervalos:

X	Y	X	Y	X	Y	X	Y	X	Y
10	51	8	51	13	51	12	56	10	52
12	51	10	53	19	55	18	55	14	53
10	54	9	53	12	53	17	56	16	56
20	56	15	54	11	52	16	57	14	55
13	54	14	56	17	54	18	58	21	57
21	58	18	57	16	54	13	53	17	53
13	52	11	54	10	50	9	51	17	55
20	57	9	50	12	55	16	55	15	56
18	56	19	54	20	58	15	53	17	57
19	57	14	52	14	54	15	55	8	52

7-10, 11-14, 15-18, 19-22 (con puntos medios 8,5; 12,5; 16,5 y 20,5, respectivamente) y elijamos para Y los intervalos: 50-52, 53-55, 56-58 (con puntos medios 51, 54 y 57, respectivamente). Tendremos:

TABLA 9.1

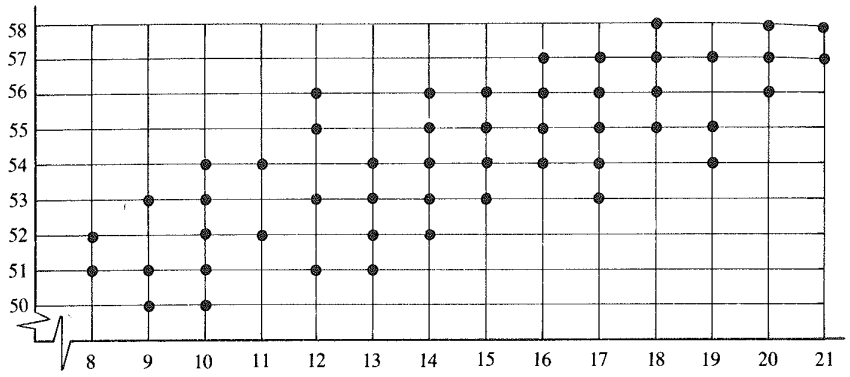
		X: 7-10	11-14	15-18	19-22
Y	56-58		//	/// ///	/// /
	53-55	///	/// ///	/// ////	//
	50-52	/// //	///		

		X: 7-10	11-14	15-18	19-22	
Y	56-58	0	2	8	6	16
	53-55	3	8	9	2	22
	50-52	7	5	0	0	12
		10	15	17	8	50

En general, tendremos r intervalos en X, y s intervalos en Y. Llamemos (I_i, I_j') al par de intervalos i de X y j de Y. A cada (I_i, I_j') le corresponderá una frecuencia n_{ij} (número de observaciones con puntuaciones dentro del intervalo i de X y dentro del intervalo j de Y). Pues bien, llamaremos distribución conjunta de frecuencias al conjunto de pares de intervalos (I_i, I_j') y de sus correspondientes frecuencias (proporciones o porcentajes).

9.2. Representación gráfica

Comencemos considerando las puntuaciones anteriores no agrupadas en intervalos. En este caso a cada puntuación de X (eje de abscisas) y a cada puntuación Y (eje de ordenadas), tomadas conjuntamente, les corresponde un punto en el plano, representante de la persona que ha obtenido esas dos puntuaciones. En nuestro ejemplo tendremos 50 puntos. Este conjunto o nube de puntos constituye el diagrama de dispersión.



Otra manera de representar gráficamente los datos no agrupados sería la siguiente. A cada valor X_i le corresponde un intervalo unitario y a cada valor Y_j otro intervalo unitario. Es decir, a cada par de puntuaciones (X_i, Y_j) les corresponderá un rectángulo unitario limitado por los intervalos unitarios. Pues bien, sobre cada rectángulo unitario podemos levantar un prisma con una altura proporcional al número de personas cuyo par de puntuaciones coincide con el par correspondiente a ese rectángulo unitario. Nótese que la unidad elegida en el eje de las X no tiene por qué ser igual que la elegida en el eje de las Y. Si fuera igual, el rectángulo unitario se convertiría en un cuadrado.

Consideremos ahora los datos agrupados en intervalos. La representación gráfica correspondiente al histograma (propuesto para el caso de una sola variable), es la siguiente. El plano quedará dividido en $(r) \times (s)$ rectángulos, donde r es el número de intervalos en X y s el de intervalos en Y. La base de esos rectángulos será la amplitud de los intervalos de X, y su altura la amplitud de los intervalos de Y. Pues bien, sobre cada uno de esos rectángulos (incidentalmente cuadrados) levantamos un prisma cuya altura sea proporcional a la frecuencia correspondiente a dicho rectángulo. Es decir, dados el intervalo con punto medio X_i y el intervalo con punto medio Y_j, la altura será proporcional a la frecuencia de personas que, a la vez, se encuentran dentro del intervalo primero y del segundo. En nuestro ejemplo (véase tabla 9.1) tendríamos nueve prismas con alturas proporcionales a 2, 8, 6, 3, 8, 9, 2, 7, 5. (En rigor, tendríamos doce prismas, aunque tres de ellos con altura nula.)

9.3. Distribuciones marginales de X e Y

Llamamos distribución marginal de X a la distribución en X de todas las observaciones, independientemente de sus puntuaciones en Y. Viene dada, en la tabla 9.1, por la fila situada en el margen inferior. Es decir, es la siguiente:

TABLA 9.2*a*

X	n_j	X_j
19-22	8	20,5
15-18	17	16,5
11-14	15	12,5
7-10	10	8,5
	50	

Llamamos distribución marginal de Y a la distribución en Y de todas las observaciones, independientemente de sus puntuaciones en X. Viene dada, en la tabla 9.1, por la columna situada en el margen derecho. Es decir, es la siguiente:

TABLA 9.2*b*

Y	n_j	Y_j
56-58	16	57
53-55	22	54
50-52	12	51
	50	

Con la distribución marginal de X tendremos una media, $\bar{X}$, y una varianza, s_x^2, que llamaremos media y varianza marginales de X.

Con la distribución marginal de Y tendremos una media, $\bar{Y}$, y una varianza, s_y^2, que llamaremos media y varianza marginales de Y.

De acuerdo con los datos de las tablas 9.2*a* y 9.2*b*, tendremos:

$$\bar{X} = \frac{(10)(8,5) + (15)(12,5) + (17)(16,5) + (8)(20,5)}{50} = \frac{717}{50} = 14,34$$

$$s_x^2 = \frac{(10)(8,5)^2 + (15)(12,5)^2 + (17)(16,5)^2 + (8)(20,5)^2}{50} - (14,34)^2 =$$

$$= \frac{11.056,5}{50} - 205,636 = 15,494$$

$$\bar{Y} = \frac{(12)(51) + (22)(54) + (16)(57)}{50} = \frac{2.712}{50} = 54,24$$

$$s_y^2 = \frac{(12)(51)^2 + (22)(54)^2 + (16)(57)^2}{50} - (54,24)^2 = \frac{147.348}{50} - 2.941,978 = 4,982$$

9.4. Distribuciones condicionales de *X* e *Y*

Llamamos distribución condicional de X, para $Y = Y_k$, a la distribución en X de todas, y solas, las observaciones con puntuación Y_k en Y (bajo la condición de tener puntuación Y_k en Y). En la tabla 9.1 tendremos tres distribuciones condicionales de X, correspondientes a $Y = 51$, $Y = 54$, $Y = 57$. Son las siguientes:

TABLA 9.3

Para $Y = 51$			Para $Y = 54$			Para $Y = 57$		
X	n_j	X_j	X	n_j	X_j	X	n_j	X_j
19-22	0	20,5	19-22	2	20,5	19-22	6	20,5
15-18	0	16,5	15-18	9	16,5	15-18	8	16,5
11-14	5	12,5	11-14	8	12,5	11-14	2	12,5
7-10	7	8,5	7-10	3	8,5	7-10	0	8,5
	12			22			16	

Llamamos distribución condicional de Y, para $X = X_k$, a la distribución en Y de todas, y solas, las observaciones con puntuación X_k en X (bajo la condición de tener puntuación X_k en X).

De acuerdo con la tabla 9.1 tendremos cuatro distribuciones condicionales de Y, correspondientes a $X = 8,5$, $X = 12,5$, $X = 16,5$, $X = 20,5$. Son las siguientes:

TABLA 9.4

Para $X = 8,5$			Para $X = 12,5$			Para $X = 16,5$			Para $X = 20,5$		
Y	n_j	Y_j	Y	n_j	Y_j	Y	n_j	Y_j	Y	n_j	Y_j
56-58	0	57	56-58	2	57	56-58	8	57	56-58	6	57
53-55	3	54	53-55	8	54	53-55	9	54	53-55	2	54
50-52	7	51	50-52	5	51	50-52	0	51	50-52	0	51
	10			15			17			8	

Con las tres distribuciones condicionales de X, tendremos tres medias y tres varianzas que llamaremos medias y varianzas condicionales de X.

Con las cuatro distribuciones condicionales de Y, tendremos cuatro medias y cuatro varianzas que llamaremos medias y varianzas condicionales de Y.

De acuerdo con la tabla 9.3, las tres medias y las tres varianzas condicionales son las siguientes:

$$\bar{X}_{Y=51} = \frac{(7)(8,5) + (5)(12,5) + (0)(16,5) + (0)(20,5)}{12} = \frac{122}{12} = 10,167$$

$$\bar{X}_{Y=54} = \frac{(3)(8,5) + (8)(12,5) + (9)(16,5) + (2)(20,5)}{22} = \frac{315}{22} = 14,318$$

$$\bar{X}_{Y=57} = \frac{(0)(8,5) + (2)(12,5) + (8)(16,5) + (6)(20,5)}{16} = \frac{280}{16} = 17,500$$

$$s^2_{x,\,Y=51} = \frac{(7)(8,5)^2 + (5)(12,5)^2 + (0)(16,5)^2 + (0)(20,5)^2}{12} - (10,167)^2 =$$

$$= \frac{1.287}{12} - 103,37 = 3,88$$

$$s^2_{x,\,Y=54} = \frac{(3)(8,5)^2 + (8)(12,5)^2 + (9)(16,5)^2 + (2)(20,5)^2}{22} - (14,318)^2 =$$

$$= \frac{4.757,5}{22} - 205,00 = 11,250$$

$$s^2_{x,\,Y=57} = \frac{(0)(8,5)^2 + (2)(12,5)^2 + (8)(16,5)^2 + (6)(20,5)^2}{16} - (17,500)^2 =$$

$$= \frac{5.012}{16} - 306,25 = 7$$

$$\bar{Y}_{X=8,5} = \frac{(7)(51) + (3)(54) + (0)(57)}{10} = \frac{519}{10} = 51,90$$

$$\bar{Y}_{X=12,5} = \frac{(5)(51) + (8)(54) + (2)(57)}{15} = \frac{801}{15} = 53,40$$

$$\bar{Y}_{X=16,5} = \frac{(0)(51) + (9)(54) + (8)(57)}{17} = \frac{942}{17} = 55,4118$$

$$\bar{Y}_{X=20,5} = \frac{(0)(51) + (2)(54) + (6)(57)}{8} = \frac{450}{8} = 56,25$$

$$s^2_{y,\,X=8,5} = \frac{(7)(51)^2 + (3)(54)^2 + (0)(57)^2}{10} - (51,90)^2 =$$

$$= \frac{26.955}{10} - 2.693,61 = 1,89$$

$$s^2_{y,\,X=12,5} = \frac{(5)(51)^2 + (8)(54)^2 + (2)(57)^2}{15} - (53,40)^2 =$$

$$= \frac{42.831}{15} - 2.851,56 = 3,84$$

$$s^2_{y,\,X=16,5} = \frac{(0)(51)^2 + (9)(54)^2 + (8)(57)^2}{17} - (55,4118)^2 =$$

$$= \frac{52.236}{17} - 3.070,468 = 2,24$$

$$s^2_{y,\,X=20,5} = \frac{(0)(51)^2 + (2)(54)^2 + (6)(57)^2}{8} - (56,25)^2 =$$

$$= \frac{25.326}{8} - 3.164,06 = 1,69$$

Comprobemos ahora cómo la varianza marginal de X es igual a la media de las varianzas condicionales de X más la varianza de las medias condicionales de X y cómo la varianza marginal de Y es igual a la media de las varianzas condicionales de Y más la varianza de las medias condicionales de Y. Nótese que estas dos relaciones no son más que una aplicación de la propiedad 6.3.4.*i*) de la varianza, donde los r grupos en que allí se descomponía el grupo total son ahora, bien las distribuciones condicionales de X, bien las distribuciones condicionales de Y.

Hemos visto que la varianza marginal de X, s^2_x, valía 15,494. Pues bien:

$$\frac{(12)(3,88) + (22)(11,25) + (16)(7)}{50} + \frac{(12)(10,167)^2 + (22)(14,318)^2 + (16)(17,5)^2}{50} -$$

$$- (14,34)^2 = 15,496$$

Hemos visto que la varianza marginal de Y, s^2_y, valía 4,982. Pues bien:

$$\frac{(10)(1,89) + (15)(3,84) + (17)(2,24) + (8)(1,69)}{50} +$$

$$+ \frac{(10)(51,9)^2 + (15)(53,4)^2 + (17)(55,4118)^2 + (8)(56,25)^2}{50} - (54,24)^2 = 4,983$$

EJEMPLO 9.2. Consideremos las puntuaciones de 40 estudiantes de Educación General Básica en dos pruebas de razonamiento (García Méndez, 1976, comunicación personal). La primera (X) mide la capacidad de razonamiento espacial. En ella se les proponían diversas sucesiones de imágenes, cada una de las cuales representaba varias posiciones de un cuerpo geométrico moviéndose en el espacio según una ley determinada que ellos debían descubrir. La otra prueba (Y) era de razona-

miento abstracto y en ella les eran presentadas a los estudiantes distintas sucesiones de figuras planas, pidiéndoles que averiguaran en cada sucesión cuál era la ley de transformación de unas figuras en otras.

X	Y	X	Y	X	Y	X	Y
7	22	14	31	9	27	11	29
5	8	6	21	15	35	8	10
9	9	17	32	13	31	3	9
9	31	4	11	16	34	10	27
13	30	8	12	16	31	14	36
16	35	11	37	16	35	18	22
12	29	12	36	19	37	18	33
12	31	14	34	6	17	16	34
20	25	13	20	16	29	14	29
10	30	3	23	5	25	10	26

Elijamos para X los intervalos (3-8), (9-14), (15-20) y para Y los intervalos (8-16), (17-25), (26-34), (35-43).

Tendremos la tabla siguiente:

Y \ X	3-8	9-14	15-20	
35-43	0	3	4	7
26-34	0	13	6	19
17-25	5	1	2	8
8-16	5	1	0	6
	10	18	12	40

Distribuciones marginales de X y de Y

a) de X:

X	n_j	X_j
15-20	12	17,5
9-14	18	11,5
3-8	10	5,5
	40	

b) de Y:

Y	n_j	Y_j
35-43	7	39
26-34	19	30
17-25	8	21
8-16	6	12
	40	

$\bar{X} = 11{,}80 \qquad s_x^2 = 19{,}71 \qquad \bar{Y} = 27{,}075 \qquad s_y^2 = 70{,}42$

Distribuciones condicionales de X.

Para $Y = 12$

X	n_j	X_j
15-20	0	17,5
9-14	1	11,5
3-8	5	5,5
	6	

$\bar{X}_{Y=12} = 6{,}50 \qquad s^2_{x,\,Y=12} = 5{,}00$

Para $Y = 21$

X	n_j	X_j
15-20	2	17,5
9-14	1	11,5
3-8	5	5,5
	8	

$\bar{X}_{Y=21} = 9{,}25 \qquad s^2_{x,\,Y=21} = 26{,}44$

Para $Y = 30$

X	n_j	X_j
15-20	6	17,5
9-14	13	11,5
3-8	0	5,5
	19	

$\bar{X}_{Y=30} = 13{,}3947 \qquad s^2_{x,\,Y=30} = 7{,}78$

Para $Y = 39$

X	n_j	X_j
15-20	4	17,5
9-14	3	11,5
3-8	0	5,5
	7	

$\bar{X}_{Y=39} = 14{,}9286 \qquad s^2_{x,\,Y=39} = 8{,}82$

Distribuciones condicionales de Y.

Para $X = 5{,}5$

Y	n_j	Y_j
35-43	0	39
26-34	0	30
17-25	5	21
8-16	5	12
	10	

$\bar{Y}_{X=5,5} = 16{,}50 \qquad s^2_{y,\,X=5,5} = 20{,}25$

Para $X = 11{,}5$

Y	n_j	Y_j
35-43	3	39
26-34	13	30
17-25	1	21
8-16	1	12
	18	

$\bar{Y}_{X=11,5} = 30 \qquad s^2_{y,\,X=11,5} = 36$

Para $X = 17{,}5$

Y	n_j	Y_j
35-43	4	39
26-34	6	30
17-25	2	21
8-16	0	12
	12	

$\bar{Y}_{X=17,5} = 31{,}5 \qquad s^2_{y,\,X=17,5} = 38{,}25$

Comprobemos cómo la varianza marginal de X es igual a la media de las varianzas condicionales de X más la varianza de las medias condicionales de X y cómo la varianza marginal de Y es igual a la media de las varianzas condicionales de Y más la varianza de las medias condicionales de Y.

Hemos visto que la varianza marginal de X, s_x^2, valía 19,71. Pues bien:

$$\frac{(6)(5) + (8)(26{,}44) + (19)(7{,}78) + (7)(8{,}82)}{40} +$$

$$+ \frac{(6)(6{,}5)^2 + (8)(9{,}25)^2 + (19)(13{,}3947)^2 + (7)(14{,}9286)^2}{40} - (11{,}80)^2 = 19{,}71$$

Hemos visto que la varianza marginal de Y, s_y^2, valía 70,42. Pues bien:

$$\frac{(10)(20,25)+(18)(36)+(12)(38,25)}{40}+\frac{(10)(16,5)^2+(18)(30)^2+(12)(31,5)^2}{40}-$$
$$-(27,075)^2=70,42$$

9.5. Covarianza de X e Y

9.5.1. Definición

Media aritmética de los productos entre la diferencia $(X_i - \bar{X})$ y la diferencia $(Y_i - \bar{Y})$ correspondientes a cada uno de los n elementos que componen un grupo. La designaremos por $\text{cov}(X, Y)$ o por s_{xy}. Por tanto,

$$\text{cov}(X, Y) \equiv s_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n} =$$
$$= \frac{\sum_{i=1}^{n} X_i Y_i}{n} - \bar{X}\bar{Y} \qquad (9.1)$$

$$\text{cov}(X, Y) \equiv s_{xy} = \frac{\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij}(X_i - \bar{X})(Y_j - \bar{Y})}{n} =$$
$$= \frac{\sum_{j=1}^{s}\sum_{i=1}^{r} n_{ij} X_i Y_j}{n} - \bar{X}\bar{Y} \qquad (9.2)$$

La fórmula (9.1) es la apropiada para datos no agrupados en intervalos y la (9.2) es la apropiada para datos agrupados en intervalos. Se entiende que r es el número de intervalos en que ha sido clasificada la variable X, s el número de intervalos en que ha sido clasificada la variable Y y n_{ij} es el número de observaciones dentro del intervalo i en la variable X y del intervalo j en la variable Y.

9.5.2. Cálculo

Aplicación de las fórmulas (9.1) y (9.2).

EJEMPLO 9.3. Calculemos la covarianza entre el rendimiento en lectura y el rendimiento en aritmética a partir de los siguientes datos:

X_i	Y_i	X_iY_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$
11	7	77	0,53	−5,24	−2,7772
26	23	598	15,53	10,76	167,1028
9	12	108	−1,47	−0,24	0,3528
3	14	42	−7,47	1,76	−13,1472
13	10	130	2,53	−2,24	−5,6672
5	10	50	−5,47	−2,24	12,2528
18	16	288	7,53	3,76	28,3128
15	12	180	4,53	−0,24	−1,0872
13	14	182	2,53	1,76	4,4528
12	15	180	1,53	2,76	4,2228
12	15	180	1,53	2,76	4,2228
0	3	0	−10,47	−9,24	96,7428
9	12	108	−1,47	−0,24	0,3528
13	14	182	2,53	1,76	4,4528
9	14	126	−1,47	1,76	−2,5872
7	11	77	−3,47	−1,24	4,3028
3	6	18	−7,47	−6,24	46,6128
178	208	2.526	0,01	0,02	348,1176

$$\bar{X} = \frac{178}{17} = 10,47$$

$$\bar{Y} = \frac{208}{17} = 12,24$$

$$s_{xy} = \frac{348,1176}{17} = 20,4775$$
$$= \frac{2.526}{17} - \frac{178}{17}\,\frac{208}{17}$$
$$= 20,4775$$

EJEMPLO 9.4. Calculemos s_{xy} a partir de los siguientes datos:

X	Y	XY	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
2	10	20	−3	1	−3
4	9	36	−1	0	0
4	10	40	−1	1	−1
2	7	14	−3	−2	6
8	13	104	3	4	12
4	5	20	−1	−4	4
9	12	108	4	3	12
4	11	44	−1	2	−2
3	8	24	−2	−1	2
8	10	80	3	1	3
7	11	77	2	2	4
5	8	40	0	−1	0
6	13	78	1	4	4
1	5	5	−4	−4	16
2	6	12	−3	−3	9
8	9	72	3	0	0
5	6	30	0	−3	0
5	11	55	0	2	0
6	6	36	1	−3	−3
7	10	70	2	1	2
100	180	965	0	0	65
			$\bar{X} = 5$	$\bar{Y} = 9$	

$$s_{xy} = \frac{65}{20} = 3,25$$

$$= \frac{965}{20} - (5)(9) =$$

$$= 48,25 - 45 = 3,25$$

Agrupemos en intervalos los datos anteriores del modo siguiente:

Y \ X	1-3	4-6	7-9	
9-13	1	5	6	12
4-8	4	4	0	8
	5	9	6	20

$$\bar{X} = \frac{(5)(2) + (9)(5) + (6)(8)}{20} = \frac{103}{20} = 5{,}15$$

$$\bar{Y} = \frac{(8)(6) + (12)(11)}{20} = \frac{180}{20} = 9$$

$$s_{xy} = \frac{(4)(2)(6) + (4)(5)(6) + (0)(8)(6) + (1)(2)(11) + (5)(5)(11) + (6)(8)(11)}{20} - (5{,}15)(9) =$$

$$= 49{,}65 - 46{,}35 = 3{,}30$$

EJEMPLO 9.5. Calculemos s_{xy} a partir del cuadro siguiente:

X	Y	XY	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
1	0	0	−1,1	−1,6	1,76
2	1	2	−0,1	−0,6	0,06
2	1	2	−0,1	−0,6	0,06
2	2	4	−0,1	0,4	−0,04
3	3	9	0,9	1,4	1,26
1	1	1	−1,1	−0,6	0,66
2	1	2	−0,1	−0,6	0,06
3	2	6	0,9	0,4	0,36
3	3	9	0,9	1,4	1,26
2	2	4	−0,1	0,4	−0,04
2	3	6	−0,1	1,4	−0,14
2	1	2	−0,1	−0,6	0,06
1	1	1	−1,1	−0,6	0,66
2	1	2	−0,1	−0,6	0,06
3	3	9	0,9	1,4	1,26
1	0	0	−1,1	−1,6	1,76
3	2	6	0,9	0,4	0,36
2	1	2	−0,1	−0,6	0,06
3	2	6	0,9	0,4	0,36
2	2	4	−0,1	0,4	−0,04
42	32	77	0,0	0,0	9,80
			$\bar{X} = 2{,}1$	$\bar{Y} = 1{,}6$	

$$s_{xy} = \frac{9{,}80}{20} = 0{,}49$$

$$= \frac{77}{20} - (2{,}1)(1{,}6) =$$

$$= 3{,}85 - 3{,}36 = 0{,}49$$

En este caso podemos agrupar los datos en intervalos de amplitud unidad del modo siguiente:

Y \ X	1	2	3	
3	0	1	3	4
2	0	3	3	6
1	2	6	0	8
0	2	0	0	2
	4	10	6	20

$$\bar{X} = \frac{(4)(1) + (10)(2) + (6)(3)}{20} = \frac{42}{20} = 2{,}1$$

$$\bar{Y} = \frac{(2)(0) + (8)(1) + (6)(2) + (4)(3)}{20} = \frac{32}{20} = 1{,}6$$

$$s_{xy} = \frac{(0)(1)(3) + (1)(2)(3) + (3)(3)(3) + (0)(1)(2) + (3)(2)(2) + (3)(3)(2)}{20} +$$

$$+ \frac{(2)(1)(1) + (6)(2)(1) + (0)(3)(1) + (2)(1)(0) + (0)(2)(0) + (0)(3)(0)}{20} - 2{,}1(1{,}6) =$$

$$= \frac{77}{20} - 3{,}36 = 3{,}85 - 3{,}36 = 0{,}49$$

Ahora el resultado es el mismo agrupando los datos que sin agruparlos porque al tener los intervalos amplitud unidad, cada puntuación coincidirá con el punto medio del intervalo (unitario) dentro del cual se encuentra.

9.5.3. Propiedades

a) Sean $V_i = aX_i + b$ y $W_i = cY_i + d$, siendo a, b, c y d cuatro constantes arbitrarias. Pues bien, $s_{vw} = (ac)\, s_{xy}$.

En efecto:

$$s_{vw} = \frac{\Sigma\,(V_i - \bar{V})\,(W_i - \bar{W})}{n} = \frac{\Sigma\,[(aX_i + b) - (a\bar{X} + b)]\,[(cY_i + d) - (c\bar{Y} + d)]}{n} =$$

$$= \frac{\Sigma\,(aX_i - a\bar{X})\,(cY_i - c\bar{Y})}{n} = \frac{\Sigma\, a(X_i - \bar{X})c(Y_i - \bar{Y})}{n} =$$

$$= ac\,\frac{\Sigma\,(X_i - \bar{X})\,(Y_i - \bar{Y})}{n} = (ac)\, s_{xy}$$

EJEMPLO 9.6. Transformemos los datos de la tabla 9.5 mediante la transformación $V_i = 2X_i + 1$, $W_i = 3Y_i - 8$ y comprobemos cómo, en efecto, $s_{vw} = (2)(3)s_{xy}$.

TABLA 9.5

X_i	Y_i	X_iY_i
3	5	15
1	3	3
5	5	25
7	11	77
16	24	120

$\bar{X} = \frac{16}{4} = 4$

$\bar{Y} = \frac{24}{4} = 6$

V_i	W_i	V_iW_i
7	7	49
3	1	3
11	7	77
15	25	375
36	40	504

$\bar{V} = \frac{36}{4} = 9$

$\bar{W} = \frac{40}{4} = 10$

$$s_{xy} = \frac{120}{4} - (4)(6) = 6 \qquad s_{vw} = \frac{504}{4} - (9)(10) = 36$$

Es decir:

$$s_{vw} = 36 = (2)(3)(6) = (2)(3)\, s_{xy}$$

b) Sean:

$X_{11}, X_{21}, \ldots, X_{n_1 1}$ e $Y_{11}, Y_{21}, \ldots, Y_{n_1 1}$ las puntuaciones de n_1 personas en dos variables X e Y.

$X_{12}, X_{22}, \ldots, X_{n_2 2}$ e $Y_{12}, Y_{22}, \ldots, Y_{n_2 2}$ las puntuaciones de n_2 personas en dos variables X e Y.

.....

$X_{1r}, X_{2r}, \ldots, X_{n_r r}$ e $Y_{1r}, Y_{2r}, \ldots, Y_{n_r r}$ las puntuaciones de n_r personas en dos variables X e Y.

Sean $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_r$ e $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_r$ las medias de cada uno de los r grupos en X y en Y.

Sean $\text{cov}_1(X, Y)$, $\text{cov}_2(X, Y), \ldots, \text{cov}_r(X, Y)$ las covarianzas entre X e Y respecto al grupo primero, al segundo, ..., al r.

Sean $\bar{X}$ e $\bar{Y}$ las medias en X y en Y del grupo total. Sea $\text{cov}(X, Y)$ la covarianza entre X e Y respecto al grupo total.

En este supuesto,

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n_j} (X_{ij} - \bar{X})(Y_{ij} - \bar{Y}) = \frac{1}{n} \sum_{j=1}^{r} \sum_{i=1}^{n_j} [(X_{ij} - \bar{X}_j) +$$

$$+ (\bar{X}_j - \bar{X})][(Y_{ij} - \bar{Y}_j) + (\bar{Y}_j - \bar{Y})] =$$

$$= \frac{1}{n}\left[\sum_{j=1}^{r} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j) + \sum_{j=1}^{r} (\bar{X}_j - \bar{X}) \sum_{i=1}^{n_j} (Y_{ij} - \right.$$

$$\left. - \bar{Y}_j) + \sum_{j=1}^{r} (\bar{Y}_j - \bar{Y}) \sum_{i=1}^{n_j} (X_{ij} - X_j) + \sum_{j=1}^{r} n_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})\right] =$$

$$= \frac{\sum_{j=1}^{r} \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)}{n} + \frac{\sum_{j=1}^{r} n_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})}{n}$$

pues:

$$\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j) = \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j) = 0$$

para

$$j = 1, 2, \ldots, r$$

Pero:

$$\frac{\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)(Y_{ij} - \bar{Y}_j)}{n_j} = \text{cov}_j(X, Y) \quad \text{y} \quad \frac{\sum_{j=1}^{r} n_j(\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})}{n} =$$

$$= \text{cov}(\bar{X}_j, \bar{Y}_j)$$

Por tanto,

$$\text{cov}(X, Y) = \frac{\sum_{j=1}^{r} n_j \text{cov}_j(X, Y)}{n} + \frac{\sum_{j=1}^{r} n_j (\bar{X}_j - \bar{X})(\bar{Y}_j - \bar{Y})}{n}$$

En conclusión, la covarianza del grupo total es igual a la media de las covarianzas, más la covarianza de las medias.

EJEMPLO 9.6. Consideremos el siguiente grupo total compuesto por los tres subgrupos siguientes:

Grupo 1.º $n_1 = 4$

X_1	Y_1	X_1Y_1
2	4	8
0	2	0
4	4	16
6	10	60
12	20	84
$\bar{X}_1 = 3$	$\bar{Y}_1 = 5$	

Grupo 2.º $n_2 = 3$

X_2	Y_2	X_2Y_2
0	2	0
1	2	2
5	5	25
6	9	27
$\bar{X}_2 = 2$	$\bar{Y}_2 = 3$	

Grupo 3.º $n_3 = 5$			Grupo total $n = 12$		
X_3	Y_3	X_3Y_3	X	Y	XY
0	2	0	2	4	8
2	0	0	0	2	0
3	4	12	4	4	16
5	6	30	6	10	60
5	8	40	0	2	0
			1	2	2
15	20	82	5	5	25
$\bar{X}_3 = 3$	$\bar{Y}_3 = 4$		0	2	0
			2	0	0
			3	4	12
			5	6	30
			5	8	40
			33	49	193
			$\bar{X} = \frac{33}{12}$	$\bar{Y} = \frac{49}{12}$	

$$\text{cov}(X, Y) = \frac{193}{12} - \frac{33}{12}\frac{49}{12} = (2.316 - 1.617)/144 = 699/144 = \frac{233}{48}$$

$$\text{cov}_1(X, Y) = \frac{84}{4} - (3)(5) = 21 - 15 = 6$$

$$\text{cov}_2(X, Y) = \frac{27}{3} - (2)(3) = 9 - 6 = 3$$

$$\text{cov}_3(X, Y) = \frac{82}{5} - (3)(4) = 16{,}4 - 12 = 4{,}4$$

$$\frac{\sum_{j=1}^{r} n_j \text{cov}_j(X, Y)}{n} = \frac{(4)(6) + (3)(3) + (5)(4{,}4)}{12} = \frac{55}{12} = \frac{220}{48}$$

$$\frac{\sum_{j=1}^{r} n_j(\bar{X}_j - \bar{X})\ (\bar{Y}_j - \bar{Y})}{n} = \frac{1}{12}\left[(4)\left(3 - \frac{33}{12}\right)\left(5 - \frac{49}{12}\right) + \right.$$

$$\left. + (3)\left(2 - \frac{33}{12}\right)\left(3 - \frac{49}{12}\right) + (5)\left(3 - \frac{33}{12}\right)\left(4 - \frac{49}{12}\right)\right] =$$

$$= \frac{1}{12}\left[\frac{132 + 351 - 15}{144}\right] = \frac{468}{1.728} = \frac{13}{48}$$

Comprobamos cómo, en efecto:

$$\frac{\sum_{j=1}^{r} n_j \text{cov}_j(X, Y)}{n} + \frac{\sum_{j=1}^{r} n_j(\bar{X}_j - \bar{X})\ (\bar{Y}_j - \bar{Y})}{n} = \frac{220}{48} + \frac{13}{48} = \frac{233}{48} = \text{cov}(X, Y)$$

NOTA

Si $X_1 = Y_1, X_2 = Y_2, \ldots, X_n = Y_n$, las covarianzas se convertirían en varianzas y la propiedad acabada de demostrar para las covarianzas quedaría traducida así: la varianza del grupo total es igual a la media de las varianzas más la varianza de las medias, propiedad de las varianzas que ya conocíamos (véase 6.3.4.*i*).

9.6. Resumen: Definiciones y fórmulas

Distribución conjunta de frecuencias (*dos variables*): Supuestas clasificadas n observaciones en r intervalos (respecto a una variable X) y en s intervalos (respecto a una variable Y), tendremos $(r) \times (s)$ pares de intervalos (I_i, I'_j). Pues bien, llamaremos distribución conjunta de frecuencias al conjunto de esos $(r) \times (s)$ pares de intervalos (I_i, I'_j) y de las frecuencias (proporciones o porcentajes) correspondientes a cada uno de ellos.

Distribución marginal de X: Distribución en X de todas las observaciones, independientemente de sus puntuaciones en Y.

A la media, $\bar{X}$, y a la varianza, s_x^2, de esta distribución las llamaremos media y varianza marginales de X.

Distribución marginal de Y: Distribución en Y de todas las observaciones, independientemente de sus puntuaciones en X.

A la media, $\bar{Y}$, y a la varianza, s_y^2, de esta distribución las llamaremos media y varianza marginales de Y.

Distribución condicional de X (*para* $Y = Y_k$): Distribución en X de todas, y solas, las observaciones con puntuación $Y = Y_k$.

A la media, $\bar{X}_{Y=Y_k}$, y a la varianza, $s^2_{X,\,Y=Y_k}$, las llamaremos media y varianza condicionales de X (para $Y = Y_k$).

Distribución condicional de Y (para $X = X_k$): Distribución en Y de todas, y solas, las observaciones con puntuación $X = X_k$.

A la media, $\bar{Y}_{X=X_k}$, y a la varianza, $s^2_{Y,\,X=X_k}$, las llamaremos media y varianza condicionales de Y (para $X = X_k$).

Covarianza de X e Y: Media aritmética de los productos entre la diferencia $(X_i - \bar{X})$ y la diferencia $(Y_i - \bar{Y})$ correspondientes a cada uno de los n elementos que componen un grupo. La designaremos por $\text{cov}(X, Y)$ o por s_{xy}. Por tanto:

$$\text{cov}(X, Y) \equiv s_{xy} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{n} = \frac{\Sigma X_i Y_j}{n} - \bar{X}\bar{Y}$$

(para datos no agrupados en intervalos)

$$\text{cov}(X, Y) \equiv s_{xy} = \frac{\Sigma \Sigma n_{ij} (X_i - \bar{X})(Y_j - \bar{Y})}{n} = \frac{\Sigma \Sigma n_{ij} X_i Y_j}{n} - \bar{X}\bar{Y}$$

(para datos agrupados en intervalos)

EJERCICIOS

9.1. A partir de los datos siguientes, construir una tabla de frecuencias tal que en la variable X existan tres intervalos de amplitud 4 y en la variable Y existan dos intervalos de amplitud 5. Dibujar el correspondiente diagrama de dispersión.

X	Y	X	Y	X	Y	X	Y
8	15	12	14	8	10	7	7
12	11	6	12	4	10	2	8
13	13	4	8	3	7	10	14
8	13	3	6	12	15	5	9
9	14	11	12	3	7	5	8

9.2. A partir de los datos anteriores (sin agrupar), calcular la covarianza de X e Y.

9.3. A partir de los datos anteriores (agrupados según 9.1), calcular:

a) Las medias marginales de X y de Y.
b) Las medias condicionales de X y de Y.
c) Las varianzas marginales de X y de Y.
d) Las varianzas condicionales de X y de Y.
e) La covarianza de X e Y.

9.4. A partir del cuadro siguiente, calcular:

a) Las medias marginales de X y de Y.
b) Las medias condicionales de X y de Y.
c) Las varianzas marginales de X y de Y.
d) Las varianzas condicionales de X y de Y.
e) La covarianza de X e Y.

		X	
	2	4	
Y 8	36	44	80
5	4	16	20
	40	60	100

9.5. A partir del cuadro siguiente, calcular:

a) Las medias marginales de X y de Y.
b) La media condicional de X para $Y = 4$.
c) La media condicional de Y para $X = 8$.
d) Las varianzas marginales de X y de Y.
e) La varianza condicional de X para $Y = 4$.
f) La varianza condicional de Y para $X = 8$
g) La covarianza de X e Y.

		X		
	1-3	4-6	7-9	
9-11	0	5	15	20
Y 6-8	5	13	6	24
3-5	5	10	5	20
0-2	10	6	0	16
	20	34	26	80

9.6. Demostrar que $\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$.

9.7. Demostrar que $\sum_{j=1}^{r} \sum_{i=1}^{n_j} n_{ij} (X_i - \bar{X})(Y_j - \bar{Y}) = \sum_{j=1}^{r} \sum_{i=1}^{n_j} n_{ij} X_i Y_j - n\bar{X}\bar{Y}$.

9.8. Sea un grupo compuesto de n_1 personas con medias $\bar{X}_1$ e $\bar{Y}_1$ en dos variables X e Y.

Sea otro grupo compuesto de n_2 personas con medias $\bar{X}_2$ e $\bar{Y}_2$ en dos variables X e Y.

Sean $\text{cov}_1(X, Y)$ y $\text{cov}_2(X, Y)$ las covarianzas de X e Y para el grupo primero y para el grupo segundo.

Sea $\text{cov}(X, Y)$ la covarianza de X e Y para el grupo total.

Esto supuesto, demostrar que

$$\text{cov}(X, Y) = \frac{1}{n_1 + n_2} \left[n_1 \text{cov}_1(X, Y) + n_2 \text{cov}_2(X, Y) + \frac{n_1 n_2 (\bar{X}_1 - \bar{X}_2)(\bar{Y}_1 - \bar{Y}_2)}{n_1 + n_2} \right]$$

9.9. Comprobar la propiedad anterior con el ejemplo siguiente:

Grupo 1.º		Grupo 2.º		Grupo total	
X	Y	X	Y	X	Y
2	1	8	6	2	1
6	3	14	10	6	3
10	5			10	5
				8	6
				14	10

9.10. Sabiendo que $\bar{X} = 6$, $\bar{Y} = 8$, $\text{cov}(X, Y) = 13$, poner los valores que faltan en el cuadro siguiente:

X	Y	XY
2	.	8
4	.	.
.	.	80
10	.	.

10

Relación (lineal) entre dos variables

10.1. Idea general

Intentamos medir la posible relación entre dos variables. Estudiaremos bajo el título «correlación» los problemas referentes a la variación conjunta de dos variables, su intensidad y su sentido (positivo o negativo). Estudiaremos bajo el título «regresión» los problemas referentes a la predicción o pronóstico de los resultados en una de las dos variables, conocidos los resultados en la otra.

Diremos que existe correlación entre dos variables, si cierta o ciertas modalidades de una de las dos variables están ligadas a cierta o ciertas modalidades de la otra. Así, a nivel nominal, diremos que existe correlación entre el lugar de origen y la carrera universitaria elegida si, sistemáticamente, las personas de cierta o ciertas provincias (diversas modalidades de la variable lugar de origen) tienden a estudiar cierta o ciertas carreras universitarias (diversas modalidades de la variable carrera universitaria). Así, por ejemplo, los de la región A estudian preferentemente Economía, los de la B Medicina, los de la C Filosofía, etc. La idea de correlación aparece más clara a nivel de intervalos. Diremos que existe correlación positiva entre el peso y la altura, si los de mucho peso tienden a ser altos (y recíprocamente); si los de peso medio, tienden a ser medianamente altos (y recíprocamente); si los de poco peso, tienden a ser bajos (y recíprocamente). Diremos que existe, también, correlación positiva entre un test de aptitud, X, y el rendimiento en cierta asignatura, Y, si los alumnos con puntuaciones altas en X, tienden a a ser altos en Y; si los alumnos con puntuaciones medias en X, tienden a ser medios en Y; si los alumnos con puntuaciones bajas en X, tienden a ser bajos en Y. Existe una correlación negativa (perfecta) entre la velocidad media a la que hemos recorrido cierta distancia y el tiempo empleado en recorrerla. A velocidad alta, corto tiempo; a velocidad baja, largo tiempo.

Comenzaremos con variables cuantitativas. Más adelante consideraremos variables ordinales y variables nominales.

En este capítulo vamos a introducir un índice que nos mida el grado de correlación entre dos variables, X e Y, pero limitándonos a variables cuantitativas y que, además, estén relacionadas linealmente. Es decir, tales que los puntos del diagrama

de dispersión correspondiente a X e Y están situados aproximadamente a lo largo de una línea recta. Este índice va a ser el coeficiente de correlación de Pearson y que designaremos por r_{xy}. Históricamente, la regresión aparece antes que la correlación y, desde luego, tiene sus ventajas definir r_{xy} a partir de la regresión. Nosotros, sin embargo, por razones didácticas, propondremos r_{xy} antes de introducir las rectas de regresión. Con todo, una vez introducidas dichas rectas, volveremos a redefinir r_{xy}, dentro del contexto de la regresión, para alcanzar una idea más clara y más cabal del mismo.

10.2. Coeficiente de correlación de Pearson

10.2.1. Definición

Según lo indicado, la correlación entre X e Y es positiva cuando X e Y covarían en el mismo sentido, es decir, cuando a estar por encima de la media en X corresponde estar, también, por encima de la media en Y, y cuando a estar por debajo de la media en X corresponde estar, también, por debajo de la media en Y. La correlación entre X e Y es negativa cuando covarían en sentido opuesto, es decir, cuando a estar por encima de la media en X corresponde estar por debajo de la media en Y, y cuando a estar por debajo de la media en X corresponde estar por encima de la media en Y. La correlación es nula cuando no covarían en ninguno de los dos sentidos, es decir, cuando a estar por encima de la media en X corresponde tanto estar por encima como por debajo de la media en Y, e, igualmente, cuando a estar por debajo de la media en X corresponde tanto estar por encima como por debajo de la media en Y. Por todo ello, la covarianza parece ser un índice apropiado para medir la correlación. En efecto, en el caso de correlación positiva, $(X - \bar{X})$ e $(Y - \bar{Y})$ serán para todas o casi todas las observaciones del mismo signo y, por tanto, su producto será positivo; consiguientemente, $s_{xy} = \Sigma (X - \bar{X})(Y - \bar{Y})/n$ será positiva. Por el contrario, en el caso de correlación negativa, $(X - \bar{X})$ e $(Y - \bar{Y})$ serán, para todas o casi todas las observaciones de distinto signo y, por tanto, su producto $(X - \bar{X})(Y - \bar{Y})$ será negativo; consiguientemente, $s_{xy} = \Sigma (X - \bar{X})(Y - \bar{Y})/n$ será negativa. Finalmente, en el caso de correlación nula, aproximadamente, para la mitad de observaciones, $(X - \bar{X})$ e $(Y - \bar{Y})$ serán del mismo signo y para la otra mitad, lo serán de distinto signo; por tanto, más o menos, la mitad de los productos $(X - \bar{X})(Y - \bar{Y})$ serán positivos y la otra mitad serán negativos. Consiguientemente, s_{xy} será nula o muy próxima a cero. En conclusión, a correlación positiva, negativa o nula, corresponde covarianza positiva, negativa o nula, respectivamente.

Sin embargo, s_{xy} como índice de correlación presenta algunos inconvenientes.

En primer lugar, el valor de s_{xy} es función de las unidades de medida elegidas para X e Y. Supongamos, por ejemplo, que X es «aprovechamiento en gramática» e Y es «aprovechamiento en aritmética». Calculemos s_{xy} bajo dos condiciones distintas: *a*) las puntuaciones posibles van de 0 a 10, *b*) las puntuaciones posibles van de 0 a 100. Supongamos ahora los dos cuadros siguientes, el primero bajo la condición *a*) y el segundo bajo la condición *b*). Ambos son equivalentes. Sólo difieren en la unidad de medida elegida.

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
8	9	1,5	2	3
6	5	−0,5	−2	1
7	8	0,5	1	0,5
5	6	−1,5	−1	1,5
26	28			6

$s_{xy} = 6/4 = 1{,}5$

X	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
80	90	15	20	300
60	50	−5	−20	100
70	80	5	10	50
50	60	−15	−10	150
260	280			600

$s_{xy} = 600/4 = 150$

Teniendo en cuenta el cuadro de la derecha, s_{xy} vale 100 veces más que considerando el de la izquierda, por el mero hecho de haber elegido unidades de medida distintas.

En segundo lugar, es algo discutible la interpretación de los números 6 y 600, como suma que son de productos cuyo par de factores representan cosas de naturaleza distinta.

Para evitar estas dificultades, una táctica obvia es dividir $(X - \bar{X})$ por s_x e $(Y - \bar{Y})$ por s_y. Según sabemos, $(X - \bar{X})$ y s_x vienen dadas en una misma unidad de medida. Por tanto, el cociente $(X - \bar{X})/s_x$ es el mismo, sea cual sea la unidad de medida elegida para medir la variable X. Así, es fácil comprobar cómo en el cuadro de la izquierda $s_x = 1{,}118$ y cómo en el cuadro de la derecha $s_x = 11{,}18$. De donde se deduce que $1{,}5/1{,}118 = 15/11{,}18$, $-0{,}5/1{,}118 = -5/11{,}18$, etc. De modo análogo es fácil comprobar cómo en el cuadro de la izquierda $s_y = 1{,}581$ y cómo en el cuadro de la derecha $s_y = 15{,}81$. De donde se deduce que $2/1{,}581 = 20/15{,}81$, $-2/1{,}581 = -20/15{,}81$, etc. En conclusión, $\Sigma (X - \bar{X})(Y - \bar{Y})/s_x s_y$ es invariante frente a cualquier cambio de unidad de medida. Así, en nuestro ejemplo,

$$6/(1{,}118)(1{,}581) = 600/(11{,}18)(15{,}81)$$

Por otra parte, tanto $(X - \bar{X})$ como s_x son números concretos que representan alguna característica (ambos la misma) como «peso», «aprovechamiento en Estadística», «actitud frente a la guerra», etc., y ambos vienen dados en una misma unidad de medida elegida arbitrariamente. Mientras que $(X - \bar{X})/s_x$ es un mero número abstracto que se limita a indicar cuántas veces el numerador contiene al denominador, independientemente de lo que éstos signifiquen. Lo mismo sucede con $(Y - \bar{Y})/s_y$.

Finalmente, $\dfrac{(X - \bar{X})}{s_x}\dfrac{(Y - \bar{Y})}{s_y}$ es un número abstracto, como producto de dos números abstractos.

Por todas estas razones, en vez de usar

$$s_{xy} = \frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n}$$

como índice de correlación, usaremos

$$\frac{\Sigma\left(\frac{X - \bar{X}}{s_x}\right)\left(\frac{Y - \bar{Y}}{s_y}\right)}{n} = \frac{\frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n}}{s_x s_y} = \frac{s_{xy}}{s_x s_y} \qquad (10.1)$$

Le llamaremos coeficiente de correlación de Pearson y le designaremos por r_{xy}. La fórmula (10.1) adopta diversas versiones, todas ellas equivalentes:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n}}{s_x s_y} = \frac{\Sigma\,\frac{(X - \bar{X})}{s_x}\frac{(Y - \bar{Y})}{s_y}}{n} = \frac{\Sigma\, z_x z_y}{n} \qquad (10.2)$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n}}{s_x s_y} = \frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n\, s_x s_y} = \frac{\Sigma\, xy}{n\, s_x s_y} = \qquad (10.3)$$

$$= \frac{\Sigma\, xy}{\sqrt{\Sigma\, x^2}\sqrt{\Sigma\, y^2}} \qquad (10.3')$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{\Sigma\,(X - \bar{X})\,(Y - \bar{Y})}{n}}{s_x s_y} = \frac{\Sigma\, XY - n\,\bar{X}\bar{Y}}{n\sqrt{\frac{n\,\Sigma\, X^2 - (\Sigma\, X)^2}{n^2}}\sqrt{\frac{n\,\Sigma\, Y^2 - (\Sigma\, Y)^2}{n^2}}} =$$

$$= \frac{n\,\Sigma\, XY - \Sigma\, X\,\Sigma\, Y}{\sqrt{n\,\Sigma\, X^2 - (\Sigma\, X)^2}\sqrt{n\,\Sigma\, Y^2 - (\Sigma\, Y)^2}} \qquad (10.4)$$

10.2.2. Cálculo

a) *Datos no agrupados*

Mera aplicación de las fórmulas anteriores, a los datos originales, es decir, considerando una a una las puntuaciones dadas.

EJEMPLO 10.1. Calculemos r_{xy} a partir de los cinco pares de puntuaciones siguientes:

X	Y	X^2	Y^2	XY	x	y	x^2	y^2	xy	z_x	z_y	$z_x z_y$
3	9	9	81	27	−1	0	1	0	0	−0,5	0	0
5	12	25	144	60	1	3	1	9	3	0,5	0,5	0,25
4	0	16	0	0	0	−9	0	81	0	0	−1,5	0
7	18	49	324	126	3	9	9	81	27	1,5	1,5	2,25
1	6	1	36	6	−3	−3	9	9	9	−1,5	−0,5	0,75
20	45	100	585	219	0	0	20	180	39	0	0	3,25

$s_{xy} = 39/5 = 7{,}8 \qquad s_x = \sqrt{20/5} = 2 \qquad s_y = \sqrt{180/5} = 6$

Según (10.1):

$$r_{xy} = \frac{7{,}8}{(2)(6)} = 0{,}65$$

Según (10.2):

$$r_{xy} = \frac{3{,}25}{5} = 0{,}65$$

Según (10.3):

$$r_{xy} = \frac{39}{(5)(2)(6)} = \frac{39}{60} = 0{,}65$$

Según (10.3′):

$$r_{xy} = \frac{39}{\sqrt{20}\sqrt{180}} = \frac{39}{60} = 0{,}65$$

Según (10.4):

$$r_{xy} = \frac{(5)(219) - (20)(45)}{\sqrt{(5)(100) - 20^2}\sqrt{(5)(585) - 45^2}} = \frac{1.095 - 900}{\sqrt{100}\sqrt{900}} = \frac{195}{(10)(30)} = 0{,}65$$

a) *Datos agrupados en intervalos*

$$r_{xy} = \frac{n \Sigma \Sigma n_{xy} XY - (\Sigma n_x X)(\Sigma n_y Y)}{\sqrt{n \Sigma n_x X^2 - (\Sigma n_x X)^2}\sqrt{n \Sigma n_y Y^2 - (\Sigma n_y Y)^2}} \qquad (10.5)$$

Donde:

n_x: representa el número de observaciones o frecuencias marginales de X.

n_y: representa el número de observaciones o frecuencias marginales de Y.

n_{xy}: representa el número de observaciones dentro de cada una de las casillas interiores de la tabla de frecuencias. Es decir, el número de observaciones que simultáneamente pertenecen a un cierto intervalo de la variable X y a otro de la variable Y.

EJEMPLO 10.2. Calculemos r_{xy} a partir de la siguiente distribución conjunta de frecuencias.

Y \ X	0-2	3-5	6-8	9-11	12-14	n_y
10-14	0	0	2	2	4	8
5-9	1	2	6	3	0	12
0-4	4	3	2	1	0	10
n_x	5	5	10	6	4	30

Para calcular r_{xy} organicemos los datos anteriores del modo siguiente.

Y \ X	0-2	3-5	6-8	9-11	12-14	n_y	Y	n_yY	n_yY^2	$n_{xy}XY$
10-14	12 0 0	48 0 0	84 2 168	120 2 240	156 4 624	8	12	96	1.152	1.032
5-9	7 1 7	28 2 56	49 6 294	70 3 210	91 0 0	12	7	84	588	567
0-4	2 4 8	8 3 24	14 2 28	20 1 20	26 0 0	10	2	20	40	80
n_x	5	5	10	6	4	30		200	1.780	1.679
X	1	4	7	10	13					
n_xX	5	20	70	60	52	207				
n_xX^2	5	80	490	600	676	1.851				

La fila n_x está formada por las frecuencias marginales de X. (5, 5, 10, 6, 4). La fila X está formada por los puntos medios de los intervalos en los que ha sido distribuida la variable X (1, 4, 7, 10, 13). La columna n_y está formada por las frecuencias marginales de Y (10, 12, 8). La columna Y está formada por los puntos medios de los intervalos en los que ha sido distribuida la variable Y (2, 7, 12).

n_xX consta de los productos de la fila n_x por la fila X: (5)(1), (5)(4), (10)(7), (6)(10), (4)(13).

n_xX^2 consta de los productos de la fila n_xX por la fila X: (5)(1), (20)(4), (70)(7), (60)(10), (52)(13).

n_yY consta de los productos de la columna n_y por la columna Y: (8)(12), (12)(7), (10)(2).

n_yY^2 consta de los productos de la columna n_yY por la columna Y: (96)(12), (84)(7), (20)(2).

En cada una de las 15 casillas interiores de la tabla anterior tenemos tres números. El central indica el número de observaciones o frecuencia de cada casilla. El situado arriba a la derecha es el producto de los puntos medios de los dos intervalos correspondientes a cada casilla: (1)(2), (1)(7), (1)(12), (4)(2), (4)(7), (4)(12), . . . , (13)(2), (13)(7), (13)(12). El situado abajo a la izquierda es el producto de la frecuencia, n_{xy}, de cada casilla por el XY correspondiente a la misma casilla: (2)(4), (7)(1), (12)(0), . . . , (26)(0), (91)(0), (156)(4).

La columna $n_{xy}XY$ consta de la suma de los números $n_{xy}XY$, los situados abajo a la izquierda, pertenecientes a las casillas de la primera fila, (1.032), de la suma de los números $n_{xy}XY$ pertenecientes a las casillas de la segunda fila (567), de la suma de los números $n_{xy}XY$ pertenecientes a las casillas de la tercera fila (80).

Por consiguiente, aplicando la fórmula (10.5):

$$r_{xy} = \frac{(30)(1.679) - (207)(200)}{\sqrt{(30)(1.851) - (207)^2}\sqrt{(30)(1.780) - (200)^2}} = \frac{8.970}{13.035{,}5} = 0{,}688$$

10.2.3. Propiedades

a) El coeficiente de correlación de Pearson no puede valer menos que -1 ni más que 1. Es decir, $-1 \leq r_{xy} \leq 1$.

En efecto, sean $u = z_x + z_y$, $v = z_x - z_y$, donde z_x y z_y son puntuaciones típicas.

Evidentemente, $\bar{u} = \bar{v} = 0$. Por tanto,

$$s_u^2 = \frac{\Sigma (z_x + z_y)^2}{n} = \frac{\Sigma z_x^2}{n} + 2\frac{\Sigma z_x z_y}{n} + \frac{\Sigma z_y^2}{n} = 1 + 2r_{xy} + 1 = 2(1 + r_{xy})$$

$$s_v^2 = \frac{\Sigma (z_x - z_y)^2}{n} = \frac{\Sigma z_x^2}{n} - 2\frac{\Sigma z_x z_y}{n} + \frac{\Sigma z_y^2}{n} = 1 - 2r_{xy} + 1 = 2(1 - r_{xy})$$

Ahora bien, tanto s_u^2 como s_v^2 son esencialmente no negativos. Por tanto, $(1 + r_{xy}) \geq 0$. Es decir, $r_{xy} \geq -1$. Igualmente, $(1 - r_{xy}) \geq 0$. Es decir, $r_{xy} \leq 1$. En conclusión, $-1 \leq r_{xy} \leq 1$.

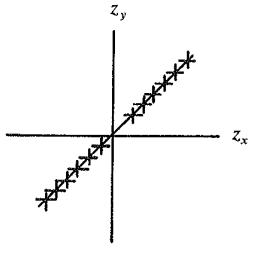
El coeficiente r_{xy} alcanza su valor extremo 1, cuando cada una de las personas obtiene la misma puntuación típica en X y en Y, o sea, cuando $z_x = z_y$ para toda persona del grupo. Evidentemente, en este caso

$$r_{xy} = \frac{\Sigma z_x z_y}{n} = \frac{\Sigma z_x z_x}{n} = \frac{\Sigma z_x^2}{n} = 1$$

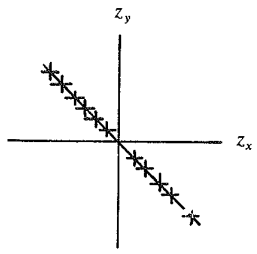
r_{xy} alcanza su valor extremo -1, cuando cada una de las personas obtiene la misma puntuación típica (pero con distinto signo) en X y en Y, es decir, cuando $z_y = -z_x$ para toda persona del grupo. Evidentemente, en este caso

$$r_{xy} = \frac{\Sigma z_x z_y}{n} = \frac{\Sigma z_x(-z_x)}{n} = -\frac{\Sigma z_x^2}{n} = -1$$

En el primer caso, los puntos que representan a las n personas están sobre la diagonal de los cuadrantes 1.º y 3.º (figura *A*). En el segundo, sobre la diagonal de los cuadrantes 2.º y 4.º (figura *B*).



(A)

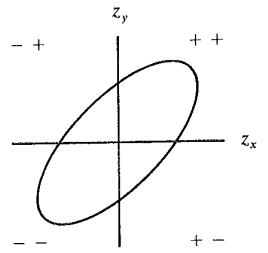


(B)

Siempre que $r_{xy} = 1$ o $r_{xy} = -1$, los puntos siguen estando sobre una línea recta (no necesariamente las dos diagonales anteriores) si nos valemos de puntuaciones diferenciales o directas. Y siempre que los puntos estén sobre una línea recta, $r_{xy} = 1$ o $r_{xy} = -1$. Estas afirmaciones quedarán legitimadas en el capítulo 12.

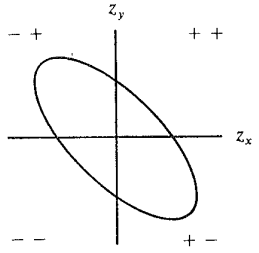
Veamos ahora lo que sucede cuando los puntos no están sobre una línea recta. Consideremos tres casos suponiendo puntuaciones típicas.

a.1) La mayor parte de las personas están situadas en el 1.º y 3.º cuadrantes, es decir, las dos puntuaciones típicas de cada una de estas personas son ambas positivas o negativas. Además, bastantes de ellas tienen puntuaciones típicas mayores (en valor absoluto) que las personas del 2.º y 4.º cuadrantes. Esto quiere decir que los productos del par de puntuaciones típicas de la mayoría de las personas



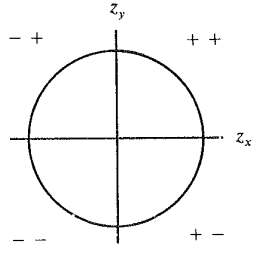
serán positivos y grandes. En cambio, los productos del par de puntuaciones típicas de la minoría serán negativos y pequeños. En conclusión, la suma de los n productos, $\Sigma z_x z_y$, será positiva y, por tanto, r_{xy} será positiva.

a.2) La mayor parte de las personas están situadas en el 2.º y 4.º cuadrantes, es decir, las dos puntuaciones típicas de cada una de estas personas son una positiva y la otra negativa. Además, bastantes de ellas tienen puntuaciones típicas mayores (en valor absoluto) que las personas del 1.º y 3.º cuadrantes. Esto quiere



decir que los productos del par de puntuaciones típicas de la mayoría de las personas serán negativos y grandes. En cambio, los productos del par de puntuaciones típicas de la minoría serán positivos y pequeños. En conclusión, la suma de los n productos, $\Sigma z_x z_y$ será negativa y, por tanto, r_{xy} será negativa.

a.3) Aproximadamente la cuarta parte de las n personas están situadas en cada uno de los cuatro cuadrantes. Por tanto, la mitad de ellas tendrán puntuaciones típicas del mismo signo y la otra mitad de signo distinto. Además, en los cuatro



cuadrantes tenemos puntuaciones típicas (en valor absoluto) altas, medias y bajas. Esto quiere decir que los productos del par de puntuaciones típicas de la mitad de las personas serán positivos y los de la otra mitad serán negativos. Por otra parte, la suma de los productos de la primera mitad será aproximadamente igual (y de signo contrario) que la suma de los de la segunda. En conclusión, la suma de los n productos, $\Sigma z_x z_y$, será nula o, aproximadamente, nula.

b) El coeficiente r_{xy}, en valor absoluto, entre dos variables es invariante frente a cualquier transformación lineal de ambas.

Esto quiere decir lo siguiente:

Sean $X_1, X_2, \ldots, X_n$; $Y_1, Y_2, \ldots, Y_n$ las puntuaciones obtenidas por n personas en las dos variables X e Y. Sean $\bar{X}$, $\bar{Y}$, s_x, s_y las correspondientes medias y desviaciones típicas. Formemos las nuevas puntuaciones:

$$V_1 = AX_1 + B,\ V_2 = AX_2 + B, \ldots,\ V_n = AX_n + B$$
$$W_1 = CY_1 + D,\ W_2 = CY_2 + D, \ldots,\ W_n = CY_n + D$$

donde A, B, C y D son cuatro constantes arbitrarias. Supuesto esto, $|r_{vw}| = |r_{xy}|$.

En efecto, según ya sabemos, $\bar{V} = A\bar{X} + B$, $s_v = |A|s_x$, $\bar{W} = C\bar{Y} + D$, $s_w = |C|s_y$. Por consiguiente:

$$r_{vw} = \frac{\Sigma\, vw}{n s_v s_w} = \frac{\Sigma\,(V - \bar{V})(W - \bar{W})}{n s_v s_w} =$$

$$= \frac{\Sigma\,[(AX + B) - (A\bar{X} + B)][(CY + D) - (C\bar{Y} + D)]}{n|A|s_x|C|s_y} =$$

$$= \frac{\Sigma\, A(X - \bar{X})\, C(Y - \bar{Y})}{|AC| n s_x s_y} = \frac{AC}{|AC|}\,\frac{\Sigma\,(X - \bar{X})(Y - \bar{Y})}{n s_x s_y} = \frac{AC}{|AC|}\,\frac{\Sigma\, xy}{n s_x s_y} =$$

$$= \frac{AC}{|AC|}\, r_{xy} \begin{cases} = r_{xy} \text{ si el signo de } A \text{ es igual que el de } C \\ = -r_{xy} \text{ si el signo de } A \text{ es distinto que el de } C \end{cases}$$

Por tanto, $|r_{xy}| = |r_{vw}|$.

De esta propiedad se infiere fácilmente que en (10.4) las puntuaciones directas en X y en Y pueden ser sustituidas por puntuaciones diferenciales o típicas. El resultado final es el mismo usando directas-directas, que diferenciales-diferenciales, directas-diferenciales, diferenciales-típicas, etc. En efecto, $x = (1)(X) + (-\bar{X})$,

$$z_x = \left(\frac{1}{s_x}\right)X + \left(-\frac{\bar{X}}{s_x}\right),\ y = (1)(Y) + (-\bar{Y}),\ z_y = \left(\frac{1}{s_y}\right)Y + \left(-\frac{\bar{Y}}{s_y}\right) \text{ con } s_x \text{ y } s_y$$

positivas.

10.2.4. Método abreviado para el cálculo de r_{xy}

Basados en la propiedad 10.2.3.*b*), vamos a obtener un método que simplifica o abrevia el cálculo de r_{xy}.

Supongamos n puntuaciones agrupadas en r intervalos (en una variable X), todos ellos con la misma amplitud I_x y, a su vez, agrupadas en s intervalos (en otra variable Y), todos ellos con la misma amplitud I_y. Sean X_0 e Y_0 los puntos medios de dos intervalos (uno en X y otro en Y) elegidos arbitrariamente y a los que llamaremos intervalos origen. Hagamos $A = \frac{1}{I_x}$, $B = -\frac{X_0}{I_x}$, $C = \frac{1}{I_y}$, $D = -\frac{Y_0}{I_y}$, introduciendo $x' = \frac{1}{I_x}X - \frac{X_0}{I_x}$, $y' = \frac{1}{I_y}Y - \frac{Y_0}{I_y}$. Dado que I_x e I_y son siempre positivas, $r_{x'y'} = r_{xy}$, según 10.2.3.*b*).

Según sabemos (véase 5.2.4) las transformaciones anteriores hacen corresponder a X_0 el valor $x' = 0$, a las puntuaciones superiores a X_0 los valores $x' = 1$, $x' = 2, \ldots$, y a las puntuaciones inferiores a X_0 los valores $x' = -1$, $x' = -2, \ldots$ A su vez, hacen corresponder a Y_0 el valor $y' = 0$, a las puntuaciones superiores a Y_0 los valores $y' = 1$, $y' = 2, \ldots$ y a las puntuaciones inferiores a Y_0 los valores $y' = -1$, $y' = -2, \ldots$

En conclusión, las transformaciones anteriores nos permiten calcular el coeficiente de correlación de Pearson valiéndonos de las puntuaciones x' e y' que son, ordinariamente, mucho más manejables que las puntuaciones originales X e Y.

La fórmula (10.5) tomará ahora la forma siguiente

$$r_{xy} = \frac{n\,\Sigma\,\Sigma\, n_{xy}x'y' - (\Sigma\, n_x x')(\Sigma\, n_y y')}{\sqrt{n\,\Sigma\, n_x x'^2 - (\Sigma\, n_x x')^2}\sqrt{n\,\Sigma\, n_y y'^2 - (\Sigma\, n_y y')^2}} \qquad (10.6)$$

EJEMPLO 10.3. Apliquemos la fórmula (10.6) a los datos del ejemplo 10.2.

	0-2	3-5	6-8	9-11	12-14	n_y	y'	$n_y y'$	$n_y y'^2$	$n_{xy}x'y'$
10-14	0	0	2	2	4	8	1	8	8	10
5-9	1	2	6	3	0	12	0	0	0	0
0-4	4	3	2	1	0	10	−1	−10	10	10
n_x	5	5	10	6	4	30		−2	18	20
x'	−2	−1	0	1	2					
$n_x x'$	−10	−5	0	6	8	−1				
$n_x x'^2$	20	5	0	6	16	47				

El procedimiento para construir la tabla es análogo al utilizado en el ejemplo 10.2. Aplicando la fórmula (10.6) nos queda:

$$r_{xy} = \frac{(30)(20) - (-1)(-2)}{\sqrt{(30)(47) - (-1)^2}\sqrt{(30)(18) - (-2)^2}} = \frac{598}{869} = 0{,}688$$

El resultado es el mismo que el obtenido mediante el método no abreviado.

10.3. Factores de los que depende r_{xy}

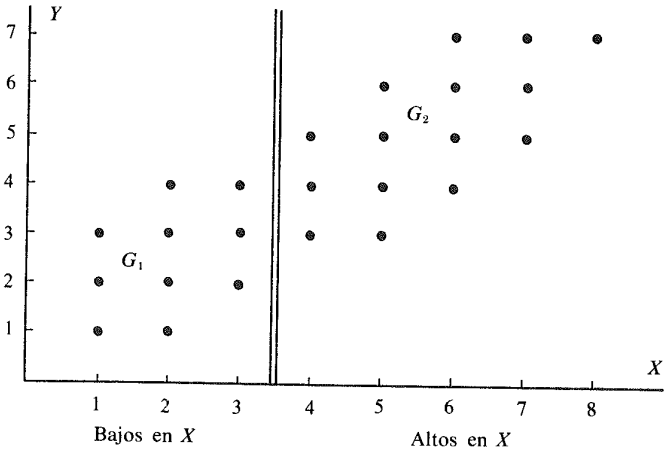
a) *Variabilidad del grupo*

Supongamos dos variables, X (capacidad intelectual) e Y (éxito escolar), relacionadas linealmente. Calculemos el coeficiente de correlación de Pearson entre ambas del modo siguiente. Consideremos, en primer lugar, el grupo total, G_T, y, luego, dos subgrupos del mismo: G_1, compuesto de los bajos en X (que, dada la relación lineal entre X e Y, tenderán a ser bajos, también, en Y) y G_2, compuesto de los altos en X (que, por la razón anterior, tenderán a ser altos, también, en Y). Es claro que la variabilidad en X y en Y de las personas del grupo G_T será mayor que la de las personas tanto de G_1 como de G_2. En efecto, G_T consta de personas altas y bajas en capacidad intelectual (y en éxito escolar). En cambio, G_1 sólo consta de personas bajas y G_2 sólo consta de personas altas. Pues bien, r_{xy} es mayor en G_T que en G_1 y en G_2. Es decir, el coeficiente de correlación de Pearson queda reducido al restringir la variabilidad del grupo en una variable (o en las dos).

EJEMPLO 10.4. Consideremos un grupo total, G_T, con 25 personas. Vamos a descomponerlo en dos subgrupos: uno, G_1, con las diez personas más bajas en X, y otro, G_2, con las 15 más altas en X.

La representación gráfica es la siguiente:

	X	Y
G_2 (Altos en X)	8	7
	7	7
	7	6
	7	5
	6	7
	6	6
	6	5
	6	4
	5	6
	5	5
	5	4
	5	3
	4	5
	4	4
	4	3
G_1 (Bajos en X)	3	4
	3	3
	3	2
	2	4
	2	3
	2	2
	2	1
	1	3
	1	2
	1	1



Intuitivamente se puede apreciar que la nube de puntos correspondiente al grupo total es más estrecha y alargada que las dos nubes correspondientes al grupo G_1 y al G_2.

Pues bien, para G_T, $r_{xy} = 0{,}84$
para G_1, $r_{xy} = 0{,}38$
para G_2, $r_{xy} = 0{,}67$

Nótese que este valor de r_{xy} menor para G_1 y G_2 que para G_T, no es debido a que los dos subgrupos consten de menos personas que el grupo total. En efecto, del grupo total podemos elegir sólo siete personas, pero que formen un grupo de gran variabilidad. Elijamos, por ejemplo, las tres más bajas en X y las tres más altas en X.

X	Y	X^2	Y^2	XY
8	7	64	49	56
7	7	49	49	49
7	6	49	36	42
7	5	49	25	35
1	3	1	9	3
1	2	1	4	2
1	1	1	1	1
32	31	214	173	188

$$r_{xy} = \frac{(7)(188) - (32)(31)}{\sqrt{(7)(214) - 32^2}\sqrt{(7)(173) - 31^2}} = \frac{324}{344{,}238} = 0{,}94$$

Para este grupo reducido, pero muy heterogéneo, $r_{xy} = 0{,}94$, mayor que el coeficiente de correlación de Pearson correspondiente al grupo total compuesto de 25 personas.

Nótese, también, que por razón del aumento de variabilidad, un solo dato alejado mucho de los restantes puede hacer que aumente espectacularmente el coeficiente de correlación de Pearson. Así, por ejemplo, el alumno puede comprobar cómo $r_{xy} = 0$ en la tabla A, y cómo $r_{xy} = 0{,}95$ en la tabla B que no es más que la tabla A con un nuevo punto, el (12, 12) muy alejado de los cuatro restantes.

TABLA A

X	Y
3	3
3	1
1	3
1	1

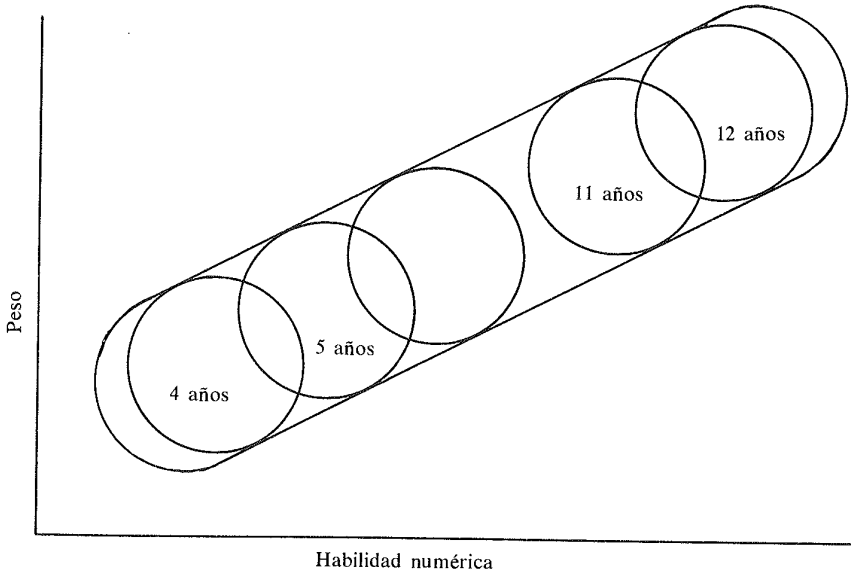
TABLA B

X	Y
12	12
3	3
3	1
1	3
1	1

Este influjo de la variabilidad del grupo sobre r_{xy} debe ser tenido en cuenta al valorar ésta. Contemplemos un caso muy ordinario en Psicología. De modo elemental diremos que un test es fiable cuando, aplicado en dos ocasiones distintas a un mismo grupo, da lugar a dos sucesiones de puntuaciones muy parecidas. Más concretamente, llamaremos coeficiente de fiabilidad o, simplemente, fiabilidad de un test, al coeficiente de correlación de Pearson, r_{12}, entre las puntuaciones obtenidas en la primera ocasión y las obtenidas en la segunda. Pues bien, según lo dicho anteriormente, r_{12} será función de la variabilidad del grupo en aquella variable medida por el test. Si se trata de un test de inteligencia y aplicamos el test a un grupo de niños muy homogéneos (o todos muy inteligentes, o todos medios, o todos poco inteligentes) obtendremos un valor para r_{12} mucho menor que si lo hubiéramos aplicado a un grupo muy heterogéneo (compuesto de niños, unos muy inteligentes, otros de inteligencia media y otros poco inteligentes). En otras palabras, el mismo test aplicado a un grupo muy homogéneo puede dar lugar a $r_{12} = 0{,}75$, por ejemplo, y aplicado a un grupo heterogéneo puede dar lugar a $r_{12} = 0{,}93$, por ejemplo.

b) *Influjo de una tercera variable*

Si calculamos el coeficiente de correlación de Pearson entre el peso y la habilidad en realizar operaciones aritméticas para un grupo de niños cuyas edades oscilen aproximadamente entre los cuatro y los doce años, veremos que r_{xy} suele ser positivo y alto. ¿Quiere decir esto que, en general, los niños de más peso tienen mayor facilidad para el cálculo numérico que los niños de menor peso? A primera vista y ateniéndonos al valor de r_{xy}, parece que sí. Sin embargo, la respuesta es



negativa. Este alto coeficiente de correlación entre peso y habilidad numérica es debido a la presencia de una tercera variable: la edad. Dentro del margen de edades propuesto, al ir aumentando la edad aumentan simultáneamente el peso y la habilidad en operar con números. Un modo de eliminar el influjo de la edad es dividir el grupo total en subgrupos, tales que los niños de cada uno de ellos tengan la misma edad y calcular r_{xy} dentro de cada uno de dichos subgrupos. Pues bien, siguiendo esta táctica veremos que el coeficiente de correlación de Pearson entre peso y habilidad numérica es muy bajo, por muchos niños que elijamos de cada subgrupo de la misma edad.

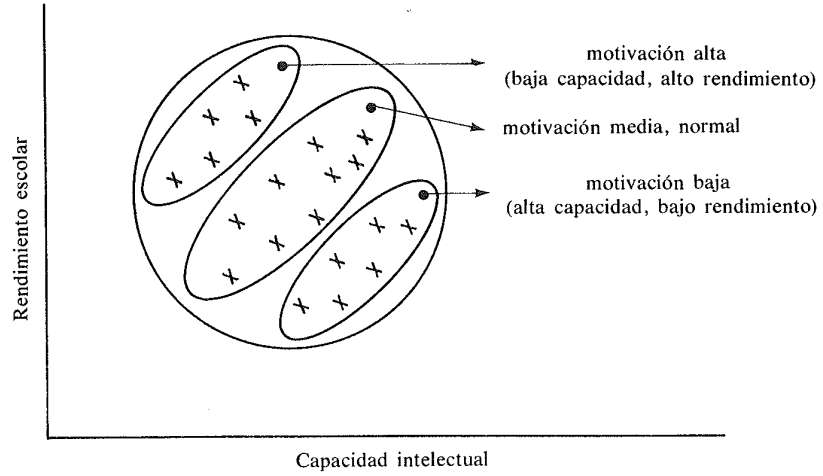
Más adelante propondremos una manera de controlar estadísticamente el influjo de la tercera variable de que se trate, sin tener que dividir el grupo total en varios subgrupos, de acuerdo con lo indicado.

La situación presente quedaría representada gráficamente según se ve en la figura de la página anterior.

Dentro de cada edad, r_{xy} es prácticamente nula (una circunferencia limita los puntos representativos de las personas de cada edad). Pero consideradas conjuntamente todas las edades, es decir, sin controlar el influjo de la edad, r_{xy} crece sensiblemente (una elipse estrecha y alargada limita los puntos representativos de las personas de todas las edades).

Vemos ahora una situación en la que el influjo de la tercera variable tiende a reducir la correlación entre dos variables. Consideremos la capacidad intelectual, X, y el rendimiento escolar en cierta disciplina, Y. Consideremos, además, la motivación como tercera variable, Z. Si controlamos Z, r_{xy} será, en general, alta. Pero sin este control es posible que alumnos con gran motivación logren alto rendimiento escolar, a pesar de tener capacidad intelectual media y aun baja. Por el contrario, otros alumnos, poco motivados, pueden lograr un bajo rendimiento escolar, a pesar de estar dotados de gran capacidad intelectual, y es probable que descienda r_{xy} si dejamos actuar a Z.

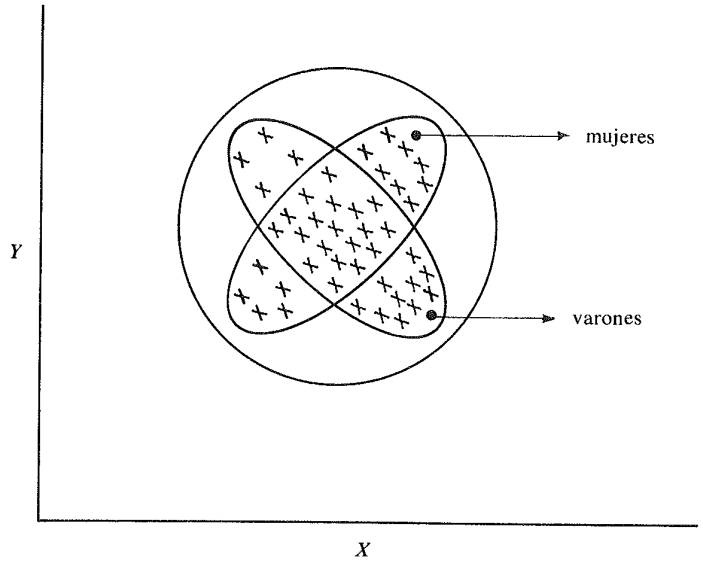
La situación presente quedaría gráficamente representada así:



Si eliminamos los alumnos con motivación alta y escasa capacidad intelectual y los alumnos con motivación baja y gran capacidad intelectual, r_{xy} será alta. (Una elipse estrecha y alargada limita los puntos representativos de las personas no eliminadas.) Pero si no los eliminamos, r_{xy} puede descender sensiblemente (una circunferencia limita los puntos de todas las personas, sin eliminación de ninguna).

Consideremos un último caso hipótetico. Supongamos que la relación entre dos variables, X e Y, es alta y positiva para las mujeres y alta y negativa para los varones y es nula para el grupo total. En otras palabras, si prescindimos del influjo del sexo, $r_{xy} = 0$. Pero si tenemos en cuenta dicho influjo, r_{xy} es alta y positiva o alta y negativa.

La situación presente quedaría representada gráficamente así:



Considerado el grupo formado simultáneamente por varones y mujeres, r_{xy} sería prácticamente nula (una circunferencia limita los puntos representativos de dichas personas). Pero considerado exclusivamente el grupo de las mujeres, r_{xy} sería alta y positiva (una elipse estrecha y alargada que va del 1.º al 3.er cuadrante limita los puntos representativos de las mujeres), y considerado el grupo de los varones, r_{xy} sería alta y negativa (una elipse estrecha y alargada que va del 2.º al 4.º cuadrante limita los puntos representativos de los varones).

De estas y otras consideraciones que pasamos por alto se infiere la gran utilidad de dibujar el diagrama de dispersión antes de calcular el coeficiente de correlación de Pearson, para luego poder interpretar éste de modo apropiado.

10.4. Condición esencial para poder calcular r_{xy}

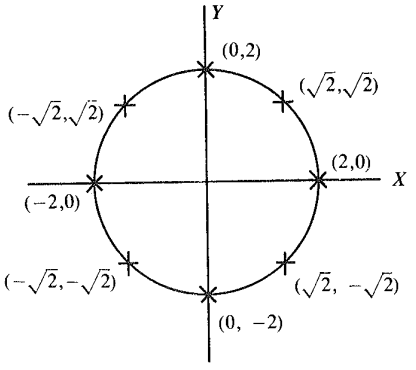
A nivel *meramente descriptivo* (en el cual nos movemos por ahora) la condición esencial para poder calcular r_{xy} es la linealidad de los puntos que representan a las n personas. Es decir, estos puntos deben estar situados próximos a una línea recta. No es necesario que las dos variables se distribuyan normalmente. Cualquier tipo de distribución es válido, a condición de que los puntos se sitúen cercanos a una recta. Y esto es posible aun cuando las dos distribuciones sean asimétricas. Por ejemplo,

X: 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5.
Y: 2, 2, 2, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4, 4, 6, 6, 6, 8, 8, 10.

son francamente asimétricas. Sin embargo, los puntos (1,2), (1,2), . . . , (4,8), (5,10) se encuentran sobre la recta $Y = 2X$ y r_{xy} vale 1.

Adviértase que r_{xy} puede ser calculado materialmente, siempre que tengamos n pares de números. Nuestro problema presente no es si puede ser calculado o no materialmente, sino si el resultado obtenido admite una interpretación razonable como coeficiente de correlación, es decir, como indicador de la relación existente entre las dos variables. Pues bien, la interpretación es razonable solamente cuando se verifica la condición de linealidad. Si los puntos representantes de las personas (es decir, de los pares de puntuaciones de cada persona) están sobre una línea curva, r_{xy} no sería recomendable, pues detectaría muy pobremente la relación «curvilínea» entre las dos variables. Más aún, a una perfecta relación curvilínea entre dos variables, puede corresponder $r_{xy} = 0$. Véase el ejemplo siguiente:

X	Y
2	0
$\sqrt{2}$	$\sqrt{2}$
0	2
$-\sqrt{2}$	$\sqrt{2}$
-2	0
$-\sqrt{2}$	$-\sqrt{2}$
0	-2
$\sqrt{2}$	$-\sqrt{2}$



Los ocho puntos anteriores están situados sobre la circunferencia $X^2 + Y^2 = 4$. Existe una correlación (curvilínea) perfecta entre X e Y y, sin embargo, $r_{xy} = 0$.

Por tanto, si no existe relación alguna entre dos variables, $r_{xy} = 0$. Pero si $r_{xy} = 0$, no se puede concluir que no exista relación alguna. Ciertamente no existe relación *lineal*, pero puede existir una relación no lineal. En resumen, r_{xy} mide la relación *lineal* entre dos variables. Pronto veremos que $r_{xy} = \pm 1$ si los puntos están sobre una recta y, recíprocamente, que si $r_{xy} = \pm 1$, los puntos están sobre una recta. De aquí que volvamos a inculcar la importancia de dibujar el diagrama de dispersión antes de calcular r_{xy} para comprobar si los puntos siguen una trayectoria más o menos rectilínea. Sólo bajo esta condición será recomendable r_{xy} como índice de correlación.

10.5. Interpretación de r_{xy}

Sabemos que $r_{xy} = \pm 1$ indica correlación lineal perfecta, y que $r_{xy} = 0$ indica correlación lineal nula. Esto supuesto, ¿qué significa 0,65?, ¿correlación alta, media o baja? Esta pregunta no tiene sentido considerada absolutamente. Depende de las circunstancias. Es baja si se trata de la fiabilidad de un test y es alta si se trata de la validez del mismo. Será baja si se trata de la correlación entre dos tests de inteligencia espacial parecidos, pero será alta si se trata de la correlación entre dos variables sociales como, por ejemplo, patriotismo y prejuicio religioso. En general, la única valoración razonable de un coeficiente de correlación, es compararlo con los coeficientes de correlación encontrados por otros investigadores entre las mismas variables y en circunstancias semejantes. El coeficiente de correlación encontrado por nosotros será bajo, si es inferior al encontrado por otros investigadores; será alto, si supera a los coeficientes obtenidos por estos. Así, por ejemplo, si se trata de las variables «prejuicio antiprotestante» y «religiosidad utilitaria», $r_{xy} = 0{,}65$ sería alto, pues no suelen alcanzar dicho valor los coeficientes encontrados por otros investigadores entre estas variables u otras muy semejantes como «prejuicio racial» y «religiosidad utilitaria».

Por estas razones, son muy equívocas las tablas en las que se valoran los coeficientes de correlación como bajos (por ejemplo, entre 0 y 0,30), medios (por ejemplo, entre 0,30 y 0,70), altos (por ejemplo, entre 0,70 y 1), o según otras categorías semejantes. Es evidente que el número 0,40 es menor que 0,50. No obstante, la relación expresada por 0,40 puede significar más que la expresada por 0,50. Depende de las variables en cuestión.

10.6. Correlación y causalidad

Por el hecho de que exista una alta correlación entre dos variables, no podemos decir que una de ellas sea causa de la otra. Es claro que hay correlación positiva entre el número de accidentes de tráfico y el número de teléfonos en las viviendas. En las regiones con mayor número de teléfonos suelen darse más accidentes de tráfico.

Lo cual no quiere decir que la abundancia de teléfonos sea la causa de los accidentes. Obviamente, a mayor nivel de vida corresponde simultáneamente mayor número de teléfonos y mayor número de coches y, como consecuencia, más accidentes. Pero a nadie se le ocurriría reducir el número de teléfonos pensando que así iba a evitar muchos de los accidentes.

La correlación indica una mera covariación entre dos variables y nada más.

10.7. Resumen: Definiciones y fórmulas

Coeficiente de correlación de Pearson entre X e Y, r_{xy}: Cociente entre la covarianza de X e Y y el producto de la desviación típica de X por la desviación típica de Y.

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Otras fórmulas equivalentes son:

$$r_{xy} = \frac{\Sigma z_x z_y}{n} \quad \text{(donde } z_x \text{ y } z_y \text{ son puntuaciones típicas)}$$

$$= \frac{\Sigma xy}{n s_x s_y} = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \quad \text{(donde } x \text{ e } y \text{ son puntuaciones diferenciales)}$$

$$= \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}} \quad \text{(donde } X \text{ e } Y \text{ son puntuaciones directas)}$$

$$= \frac{n\Sigma\Sigma n_{xy}XY - (\Sigma n_x X)(\Sigma n_y Y)}{\sqrt{n\Sigma n_x X^2 - (\Sigma n_x X)^2}\sqrt{n\Sigma n_y Y^2 - (\Sigma n_y Y)^2}} \quad \text{(donde } X \text{ e } Y \text{ son puntuaciones directas y los datos se encuentran agrupados en intervalos)}$$

EJERCICIOS

10.1. Calcular el coeficiente de correlación de Pearson entre las variables X e Y, a partir de los siguientes datos *no agrupados* en intervalos.

a)

X	Y
1	1
2	2
4	5
5	4

b)

X	Y
2	5
4	7
5	4
6	1
8	3

c)

X	Y
1	1
2	3
6	2
3	2

d)

X	Y
4	1
6	1
2	2
3	2
0	4

e)

X	Y
1	4
2	2
3	6
5	18
9	10

f)

X	Y
1	5
3	2
4	4
6	1

10.2. Calcular el coeficiente de correlación de Pearson entre las variables X e Y, a partir de los siguientes datos *agrupados* en intervalos.

a)

	2-4	5-7
4-6	4	10
1-3	4	2

b)

	6-9	10-13
13-17	3	0
8-12	1	2
3-7	0	4

c)

	55-60	61-66	67-72
94-100	1	2	6
87-93	2	5	3
80-86	6	4	1

d)

	62-66	67-71	72-76
35-39	4	2	1
30-34	1	3	1
25-29	1	4	3
20-24	0	1	4

e)

	18-20	21-23	24-26	27-29
40-44	6	4	2	0
35-39	2	4	4	2
30-34	0	3	6	6
25-29	0	1	3	7

10.3. Demostrar que la varianza de $V = X + Y$ es igual a la varianza de X más la varianza de Y, si $r_{xy} = 0$.

10.4. Demostrar que la varianza de $V = X - Y$ es igual a la varianza de X más la varianza de Y, si $r_{xy} = 0$.

10.5. Sean $\bar{X}$ e $\bar{Y}$ las medias de $X_1, X_2, \ldots, X_n$ y de $Y_1, Y_2, \ldots, Y_n$, respectivamente. Formemos las nuevas puntuaciones $X_1Y_1, X_2Y_2, \ldots, X_nY_n$. Supuesto esto, demostrar que la media de estas nuevas puntuaciones es igual a $\bar{X}\bar{Y}$ si $r_{xy} = 0$.

10.6. ¿Vale necesariamente 1 el coeficiente de correlación de Pearson entre X e Y si $\bar{X} = \bar{Y}$ y $s_x = s_y$?

10.7. ¿Se verifica necesariamente $\bar{X} = \bar{Y}$ y $s_x = s_y$ siempre que valga 1 el coeficiente de correlación de Pearson entre X e Y?

10.8. Demostrar que si sólo tenemos dos personas, con puntuaciones en X y en Y, $r_{xy} = 1$ ó $r_{xy} = -1$ ó $r_{xy} = 0$. (¿En qué únicos casos se verificará $r_{xy} = 0$?)

10.9. Deseamos calcular r_{xy} entre los resultados en una prueba de aritmética (X) y otra de gramática (Y). Nos valemos de un ordenador en cuya memoria se encuentra la fórmula de r_{xy} en puntuaciones directas. Por distracción introducimos en el ordenador puntuaciones diferenciales en gramática y típicas en aritmética. En este supuesto, *a*) ¿Tendrá algún sentido el resultado que nos ofrezca la máquina? *b*) ¿Será este resultado el mismo que el que hubiéramos obtenido introduciendo puntuaciones directas en aritmética y en gramática?

10.10. Sean $X_1, X_2, \ldots, X_n$ las puntuaciones directas obtenidas por n estudiantes en la primera mitad de un examen de Estadística, valiendo 4 la correspondiente desviación típica. Sean $Y_1, Y_2, \ldots, Y_n$ las puntuaciones directas obtenidas por esos mismos estudiantes en la segunda mitad de dicho examen, valiendo 5 la correspondiente desviación típica. Como notas finales en Estadística aceptamos las siguientes: $3X_1 + 2Y_1, 3X_2 + 2Y_2, \ldots, 3X_n + 2Y_n$. En este supuesto, ¿cuánto valdrá la varianza de estas notas finales, *a*) si $r_{xy} = 0$, *b*) si $r_{xy} = 1$, *c*) si $r_{xy} = -1$?

10.11. Sean $z_{11}, z_{12}, \ldots, z_{1n}$ las puntuaciones típicas obtenidas por n estudiantes en «Psicodiagnóstico de niños» y sean $z_{21}, z_{22}, \ldots, z_{2n}$ las puntuaciones típicas obtenidas por esos mismos estudiantes en «Psicodiagnóstico de adultos». Como nota final de «Psicodiagnóstico» atribuimos a cada persona «i» la puntuación $(z_{1i} + z_{2i})/2$. Demostrar que vale 0 la media de esas puntuaciones finales y que vale $(1 + r_{12})/2$ la varianza de las mismas.

10.12. Sean X e Y dos variables tales que $\bar{X} = 40$, $\bar{Y} = 15$, $s_y^2 = 25$, $r_{xy} = 0{,}125$, $CV_x = 10$. Calcular la desviación típica de las puntuaciones $W_i = X_i - Y_i$.

10.13. Consideremos las puntuaciones $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$. Introduzcamos la nueva variable $V_i = X_i - Y_i$. Calcular el coeficiente de correlación de Pearson entre X y V, y entre Y y V, en función de s_x, s_y y r_{xy}.

10.14. Calcular el coeficiente de correlación de Pearson introduciendo en la ecuación de r_{xy} en puntuaciones directas, las puntuaciones diferenciales (1, −1, 0, −3, 3) en X, y las puntuaciones típicas (0,5, 0, −1,5, −0,5, 1,5) en Y. Comprobar, además, cómo ese resultado coincide con el obtenido mediante $\Sigma z_x z_y/n$, una vez transformadas en típicas las puntuaciones diferenciales en X.

10.15. Siendo $z_{x_1}, z_{x_2}, \ldots, z_{x_n}$ puntuaciones típicas en X, y siendo $Y_1, Y_2, \ldots, Y_n$ puntuaciones directas en Y, calcular el valor de $\dfrac{\Sigma (Y_i - z_{x_i})^2}{n}$ para $r_{xy} = 1$, $r_{xy} = 0$, $r_{xy} = -1$, suponiendo que $s_y = 2$, $\bar{Y} = 6$.

10.16. Consideremos las puntuaciones $X_1, X_2, \ldots, X_n$ e $Y_1, Y_2, \ldots, Y_n$. Sean $s_x = 4$, $s_y = 2$, $r_{xy} = 0{,}25$. Calcular el coeficiente de correlación de Pearson entre X y $V = X - Y$.

10.17. Sean $W = X - V$, $\Sigma X^2 = 318$, $\Sigma X = 42$, $s_v = 3$, $n = 6$. Calcular el valor de s_w^2 cuando $r_{xv} = 0$ y cuando $r_{xv} = 1$.

10.18. Sabemos que el potencial excitatorio (E) se relaciona con la fuerza de hábito (H) y el drive (D) mediante la siguiente ecuación: $E = (H)(D)$. En un experimento con cinco personas hemos obtenido $\bar{E} = 30$, $\bar{D} = 6$, $\bar{H} = 5$. Calcular el coeficiente de correlación de Pearson entre D y H.

11

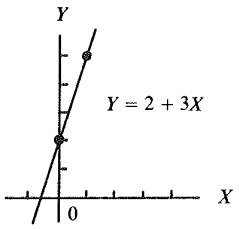
Ecuaciones de regresión

11.1. Regresión y predicción

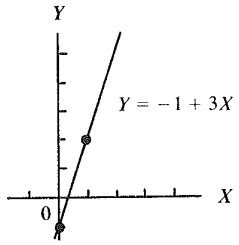
Regresión equivaldrá a predicción. La palabra regresión se debe a Galton. Estudiando la relación entre las características de los padres y las de sus hijos observó que a padres altos correspondían hijos altos, pero que, en general, se acercaban más a su media que los padres a la suya. Igualmente, a padres bajos correspondían hijos bajos, pero que, en general, se acercaban a su media más que los padres a la suya. Es decir, parecía darse cierta regresión hacia la media. Según este modo de pensar, los hijos serían más iguales, más homogéneos entre sí que lo eran sus padres. Pasamos por alto la discusión de esta interpretación y nos limitamos a constatar que el término regresión fue introducido con ocasión de estos estudios de Galton sobre la herencia y que hoy para nosotros equivale a predicción, pronóstico, estimación. Es decir, ecuación de regresión equivaldrá a ecuación de predicción, de pronóstico, de estimación.

11.2. Ecuación de una recta en el plano

Es de la forma $Y = a + bX$, o sea, de primer grado en X y en Y. Las constantes a y b son propias de cada recta. Al variar a y/o b, varía la recta; y al variar la recta, varían a y/o b. La constante a es llamada ordenada en el origen, pues representa el valor de Y (ordenada) cuando $X = 0$ (es decir, cuando en el eje de abscisas nos encontramos en el origen). La constante b es llamada pendiente de la recta, y representa la inclinación mayor o menor de la misma. Veamos algunas ecuaciones de rectas y sus correspondientes representaciones gráficas.

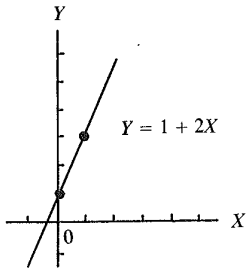


Para $X = 0$, $Y = 2$; para $X = 1$, $Y = 5$. Por tanto, la recta pasará por los puntos cuyas coordenadas son: (0,2) y (1,5).

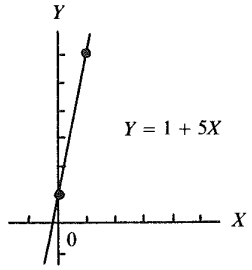


Para $X = 0$, $Y = -1$; para $X = 1$, $Y = 2$. Por tanto, la recta pasará por los puntos cuyas coordenadas son: (0,−1) y (1,2).

Las dos rectas anteriores son paralelas. Coinciden en tener la misma pendiente ($b = 3$). Tienen distinta ordenada en el origen (2 y -1), es decir, cortan al eje OY en distintos puntos.

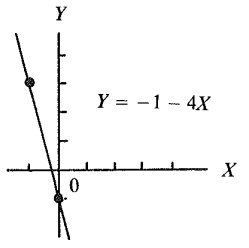


Para $X = 0$, $Y = 1$; para $X = 1$, $Y = 3$. Por tanto, la recta pasará por los puntos cuyas coordenadas son: (0,1) y (1,3).

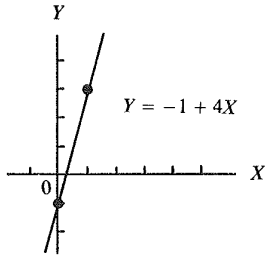


Para $X = 0$, $Y = 1$; para $X = 1$, $Y = 6$. Por tanto, la recta pasará por los puntos cuyas coordenadas son (0,1) y (1,6).

Las dos rectas anteriores tienen la misma ordenada en el origen ($a = 1$), es decir, cortan al eje OY en el mismo punto (0, 1). Tienen distintas pendientes (2 y 5). No son paralelas.

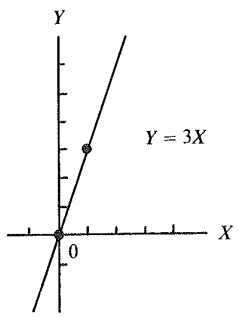


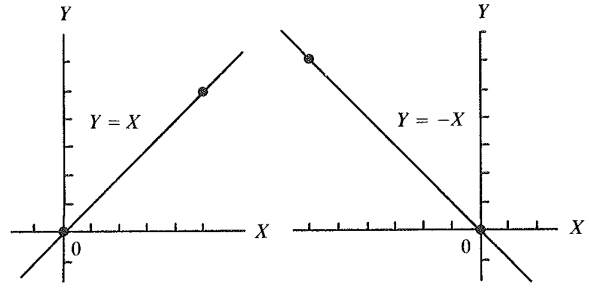
Para $X = 0$, $Y = -1$; para $X = -1$, $Y = 3$. Por tanto, la recta pasará por los puntos cuyas coordenadas son: (0,−1) y (−1,3).



Para $X = 0$, $Y = -1$; para $X = 1$, $Y = 3$. Por tanto, la recta pasará por los puntos cuyas coordenadas son: (0,−1) y (1,3).

Las dos rectas anteriores tienen la misma ordenada en el origen ($a = -1$), es decir, cortan al eje OY en el mismo punto (0, −1). Tienen la misma pendiente, pero con distinto signo (−4 y 4).



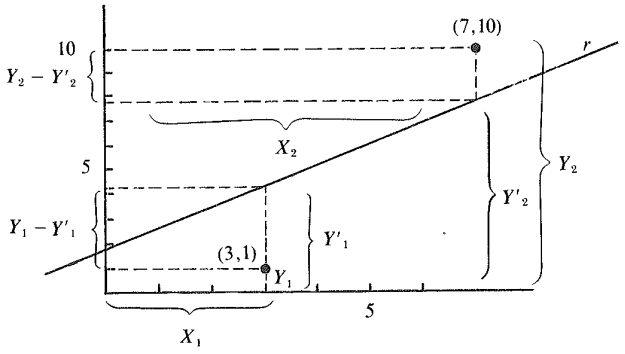


Las tres rectas anteriores tienen la misma ordenada en el origen ($a = 0$), es decir, cortan al eje OY en el mismo punto (0, 0). En otras palabras, las tres carecen de término independiente y pasan por el origen.

11.3. Ecuaciones de las rectas de regresión de *Y* sobre *X* según el criterio de mínimos cuadrados

Distinguiremos entre construcción y aplicación. Comenzamos con la construcción. Intentamos determinar una función matemática (una ecuación) que nos permita pronosticar la puntuación de cada persona en una variable Y (criterio), conocida su puntuación en otra variable X (variable predictora). Entre todas las funciones elegimos una muy sencilla: la función lineal, la ecuación de la línea recta.

X	Y
3	1
7	10
10	8
1	2
4	9



Pues bien, deseamos construir una línea recta tal que haga mínimo el error medio cometido en los pronósticos. Esta minimización del error la entendemos de la siguiente manera.

Supongamos un grupo de personas cuyas puntuaciones en X y en Y nos son dadas. Por ejemplo, consideremos las cinco personas de la tabla anterior.

En el gráfico anterior llamemos Y_i' a la ordenada del punto (sobre la recta r) cuya abscisa es X_i. En otras palabras, Y_i' es la puntuación pronosticada en Y mediante la recta r a la persona cuya puntuación en X es X_i. Es decir, Y_i es la puntuación obtenida en Y e Y_i' es la pronosticada.

Es claro que con cada recta tendremos en nuestro caso cinco diferencias: $Y_1 - Y_1'$, $Y_2 - Y_2'$, $Y_3 - Y_3'$, $Y_4 - Y_4'$, $Y_5 - Y_5'$ o errores entre la puntuación obtenida y la pronosticada. (En la figura adjunta tenemos las diferencias correspondientes a las dos primeras personas.) Si elevamos al cuadrado estos errores y sumamos estos cuadrados, tendremos una suma de errores cuadráticos. Pues bien, de todas las rectas posibles del plano pretendemos elegir aquella respecto a la cual sea mínima dicha suma. En esto consiste la construcción de las rectas de regresión de Y sobre X, según el criterio de mínimos cuadrados.

En conclusión, la recta de regresión de Y sobre X es una recta tal que, en nuestro ejemplo, haga mínima la suma:

$$(Y_1 - Y_1')^2 + (Y_2 - Y_2')^2 + (Y_3 - Y_3')^2 + (Y_4 - Y_4')^2 + (Y_5 - Y_5')^2$$

En general si tenemos n personas, intentamos construir una recta tal, que haga mínima la expresión

$$\Sigma (Y_i - Y_i')^2 \quad \text{donde} \quad i = 1, 2, \ldots, n.$$

Nótese que para construir la recta de regresión necesitamos un grupo de personas cuyas puntuaciones en X y en Y conozcamos. En cambio, la aplicaremos a otras personas, *semejantes* a las anteriores, de las que sólo conoceremos sus puntuaciones en X. Supongamos que X es un test de aptitud para la Estadística e Y el aprovechamiento en la misma manifestado mediante un examen. Queremos construir una ecuación que nos permita pronosticar del mejor modo posible el aprovechamiento, conocido el resultado en el test. Pues bien, para construir esa recta, necesitamos unas personas cuyas puntuaciones en el test y en el examen nos sean conocidas. Una vez construida, la aplicaremos a otras personas, *semejantes* a las anteriores, de las que sólo conoceremos sus puntuaciones en el test de aptitud.

Dada la semejanza entre los dos grupos de personas, es de esperar que la recta de regresión que fue óptima en reducir los errores cuadráticos respecto al primer grupo, será, también, razonablemente buena en reducir los errores cuadráticos respecto al segundo.

Expuestas estas consideraciones previas, veamos cuál es la recta de regresión de Y sobre X.

a) *Expresada en puntuaciones directas*

Comenzamos con la ecuación

$$Y' = A + BX \tag{11.1}$$

Nuestro propósito es determinar A y B, de modo que $\phi \equiv \Sigma (Y - Y')^2 = \Sigma (Y - A - BX)^2$ sea mínima. Según se demuestra en Cálculo, ello equivale a resolver las dos ecuaciones

$$\frac{\partial \phi}{\partial A} = 0, \quad \frac{\partial \phi}{\partial B} = 0 \qquad \text{donde} \qquad \frac{\partial \phi}{\partial A} \quad \text{y} \quad \frac{\partial \phi}{\partial B}$$

son las derivadas parciales de ϕ respecto a A y a B. Es decir,

$$\frac{\partial \Sigma (Y - A - BX)^2}{\partial A} = -2 \Sigma (Y - A - BX) = 0$$

$$\frac{\partial \Sigma (Y - A - BX)^2}{\partial B} = -2 \Sigma (Y - A - BX)X = 0$$

O, lo que es equivalente,

$$\Sigma (Y - A - BX) = 0, \quad \text{de donde,} \quad \Sigma Y = nA + B \Sigma X \tag{11.2}$$

$$\Sigma (Y - A - BX)X = 0, \quad \text{de donde,} \quad \Sigma XY = A \Sigma X + B \Sigma X^2 \tag{11.3}$$

Las ecuaciones (11.2) y (11.3) suelen ser llamadas normales. Ellas nos permiten despejar A y B. En efecto, dividiendo (11.2) por n nos queda

$$\bar{Y} = A + B\bar{X} \tag{11.4}$$

de donde

$$A = \bar{Y} - B\bar{X} \tag{11.5}$$

Multiplicando (11.4) por $n\bar{X}$ y restando de (11.3) nos queda,

$$\begin{aligned} \Sigma XY - n\bar{X}\bar{Y} &= (A \Sigma X + B \Sigma X^2) - (An\bar{X} + Bn\bar{X}^2) = \\ &= A(\Sigma X - n\bar{X}) + B(\Sigma X^2 - n\bar{X}^2) = \\ &= 0 + B(\Sigma X^2 - n\bar{X}^2) \end{aligned}$$

Por tanto,

$$B = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} \tag{11.6}$$

$$B = \frac{\Sigma XY - n\dfrac{\Sigma X}{n}\dfrac{\Sigma Y}{n}}{\Sigma X^2 - n\left(\dfrac{\Sigma X}{n}\right)^2} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad (11.7)$$

En resumen,

$$A = \bar{Y} - B\bar{X} \qquad (11.5)$$

$$B = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad (11.7)$$

Teniendo en cuenta (11.1) y (11.5),

$$Y' = A + BX = (\bar{Y} - B\bar{X}) + BX \qquad (11.8)$$

Ésta es la recta de regresión de Y sobre X expresada en puntuaciones directas, con B dada por (11.7)

De (11.8) se infieren inmediatamente las siguientes consecuencias:

1)
$$\bar{Y}' = \frac{\Sigma Y'}{n} = (\bar{Y} - B\bar{X}) + B\frac{\Sigma X}{n} = \bar{Y} - B\bar{X} + B\bar{X} = \bar{Y} \qquad (11.9)$$

Es decir, son iguales la media de las puntuaciones directas pronosticadas, $\bar{Y}'$, y la media de las puntuaciones directas obtenidas, $\bar{Y}$.

2) $s_{y'}^2 = B^2 s_x^2 =$ (teniendo en cuenta (11.8) y 6.3.4.*a*)

$$= \left[\frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2}\right]^2 s_x^2 = \quad \text{(teniendo en cuenta 11.6)}$$

$$= \left[\frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\Sigma (X-\bar{X})^2}\right]^2 s_x^2 = \left[\frac{\Sigma xy}{n s_x^2}\right]^2 s_x^2 = \left[\frac{n r_{xy} s_x s_y}{n s_x^2}\right]^2 s_x^2 = r_{xy}^2 s_y^2 \qquad (11.10)$$

Es decir, la varianza de las puntuaciones directas pronosticadas, $s_{y'}^2$, es igual o menor que la varianza de las puntuaciones directas obtenidas, s_y^2, pues $r_{xy}^2 \leq 1$.

3) Sustituyendo X por $\bar{X}$ en (11.8),

$$Y' = (\bar{Y} - B\bar{X}) + B\bar{X} = \bar{Y} \qquad (11.11)$$

Es decir, la recta de regresión de Y sobre X, en puntuaciones directas, pasa por el punto $(\bar{X}, \bar{Y})$. (Lo mismo sucede con la recta de regresión de X sobre Y.)

b) *Expresada en puntuaciones diferenciales*

Comenzamos con la ecuación

$$y' = a + bx \qquad (11.12)$$

Nuestro propósito es determinar a y b de modo que $\Sigma (y - y')^2 = \Sigma (y - a - bx)^2$ sea mínima. Según un razonamiento análogo al seguido en el caso de puntuaciones directas, llegamos a las dos ecuaciones normales

$$\Sigma (y - a - bx) = 0, \quad \text{de donde,} \quad \Sigma y = na + b\Sigma x \qquad (11.13)$$

$$\Sigma (y - a - bx)x = 0, \quad \text{de donde,} \quad \Sigma xy = a\Sigma x + b\Sigma x^2 \qquad (11.14)$$

Ahora bien, $\Sigma x = \Sigma y = 0$. Por consiguiente, las dos ecuaciones normales quedan reducidas a

$$0 = na + 0, \quad \text{de donde,} \quad na = 0, \quad a = 0$$

$$\Sigma xy = b\Sigma x^2, \quad \text{de donde,} \quad b = \frac{\Sigma xy}{\Sigma x^2}$$

En conclusión,

$$a = 0 \qquad (11.15)$$

$$b = \frac{\Sigma xy}{\Sigma x^2} \qquad (11.16)$$

$$= \frac{n r_{xy} s_x s_y}{n s_x^2} = r_{xy}\frac{s_y}{s_x} \qquad (11.17)$$

Nótese que

$$b = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma (X-\bar{X})(Y-\bar{Y})}{\Sigma (X-\bar{X})^2} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{\Sigma X^2 - n\bar{X}^2} = B \qquad (11.18)$$

En otras palabras, la recta de regresión en puntuaciones directas y la recta de regresión en puntuaciones diferenciales tienen la misma pendiente, son paralelas.

Teniendo en cuenta (11.12) y (11.15),

$$y' = bx \qquad (11.19)$$

Ésta es la recta de regresión de Y sobre X expresada en puntuaciones diferenciales, con b dada por (11.16) u (11.17).

De (11.19) se infieren inmediatamente las siguientes consecuencias:

1)

$$\bar{y}' = \frac{\Sigma y'}{n} = b\frac{\Sigma x}{n} = 0 \qquad (11.20)$$

Es decir, son iguales (valiendo ambas cero) la media de las puntuaciones diferenciales pronosticadas, $\bar{y}'$, y la media de las puntuaciones diferenciales obtenidas, $\bar{y}$.

2) $s_{y'}^2 = b^2 s_x^2$ (teniendo en cuenta (11.19) y 6.3.4.*a*)

$= r_{xy}^2 \frac{s_y^2}{s_x^2} s_x^2$ (teniendo en cuenta (11.17))

$$= r_{xy}^2 s_y^2 \qquad (11.21)$$

Es decir, llegamos al mismo resultado conseguido en (11.10).

3) Sustituyendo x por $\bar{x} = 0$ en (11.19),

$$y' = 0 \qquad (11.22)$$

Es decir, la recta de regresión de Y sobre X, en puntuaciones diferenciales, pasa por el origen ($\bar{x} = 0$, $\bar{y} = 0$). (Lo mismo sucede con la recta de regresión de X sobre Y.)

c) *Expresada en puntuaciones típicas*

Comenzamos con la ecuación

$$z_y' = a^* + b^* z_x \qquad (11.23)$$

Nuestro propósito es determinar a^* y b^* de modo que $\Sigma (z_y - z_y')^2 = \Sigma (z_y - a^* - b^* z_x)^2$ sea mínima. Según un razonamiento análogo al seguido en el caso de puntuaciones directas llegamos a las dos ecuaciones normales

$$\Sigma (z_y - a^* - b^* z_x) = 0, \quad \text{de donde,} \quad \Sigma z_y = na^* + b^* \Sigma z_x \qquad (11.24)$$

$$\Sigma (z_y - a^* - b^* z_x) z_x = 0, \quad \text{de donde,} \quad \Sigma z_x z_y = a^* \Sigma z_x + b^* \Sigma z_x^2 \qquad (11.25)$$

Ahora bien, $\Sigma z_x = \Sigma z_y = 0$, $\Sigma z_x^2 = n$. Por consiguiente, las dos ecuaciones normales quedan reducidas a

$$0 = na^* + 0, \quad \text{de donde,} \quad na^* = 0, \quad a^* = 0$$

$$\Sigma z_x z_y = b^* n, \quad \text{de donde,} \quad b^* = \Sigma z_x z_y / n = r_{xy}$$

En conclusión,

$$a^* = 0 \qquad (11.26)$$

$$b^* = r_{xy} \qquad (11.27)$$

Teniendo en cuenta (11.23), (11.26) y (11.27),

$$z_y' = r_{xy} z_x \qquad (11.28)$$

Esta es la recta de regresión de Y sobre X expresada en puntuaciones típicas.

De (11.28) se infieren inmediatamente las siguientes consecuencias:

1)

$$\bar{z}_y' = \frac{\Sigma z_y'}{n} = r_{xy}\frac{\Sigma z_x}{n} = 0 \qquad (11.29)$$

Es decir, son iguales (valiendo ambas cero) la media de las puntuaciones típicas pronosticadas, $\bar{z}_y'$, y la media de las puntuaciones típicas obtenidas, $\bar{z}_y$.

2)

$s_{z_y'}^2 = r_{xy}^2 s_{z_x}^2$ (teniendo en cuenta (11.28) y 6.3.4*a*)

$s_{z_y'}^2 = r_{xy}^2$ (teniendo en cuenta 8.2.*b*) (11.30)

Es decir, la varianza de las puntuaciones típicas pronosticadas es igual o menor que 1. Esto significa que las puntuaciones z_y' no cumplen con una de las propiedades esenciales de las puntuaciones típicas, a saber, que su varianza vale necesariamente 1. Por esta razón, estas puntuaciones z_y' deberían ser llamadas «pseudotípicas» en vez de típicas.

3) Sustituyendo z_x por $\bar{z}_x = 0$ en (11.28),

$$z_y' = 0$$

O sea, la recta de regresión de Y sobre X, en puntuaciones pseudotípicas pasa por el origen ($\bar{z}_x = 0$, $\bar{z}_y = 0$). (Lo mismo sucede con la recta de regresión de X sobre Y.)

EJEMPLO 11.1. Comenzaremos introduciendo un miniejemplo que ayude al lector a comprender mejor la aplicación de las fórmulas anteriores. Después, ofreceremos otro ejemplo algo más largo con unos datos obtenidos en la vida real.

Supongamos cinco personas con puntuaciones en una variable predictora (X) y en un criterio (Y), según la tabla 11.1. (De acuerdo con la costumbre seguida anteriormente, $(X - \bar{X}) = x$, e $(Y - \bar{Y}) = y$.)

TABLA 11.1

X	Y	X^2	Y^2	XY	x	y	x^2	y^2	xy	z_x	z_y	$z_x z_y$	Y'	y'	z'_y
3	9	9	81	27	−1	0	1	0	0	−0,5	0,0	0,00	7,05	−1,95	−0,325
5	12	25	144	60	1	3	1	9	3	0,5	0,5	0,25	10,95	1,95	0,325
4	0	16	0	0	0	−9	0	81	0	0,0	−1,5	0,00	9,00	0,00	0,000
7	18	49	324	126	3	9	9	81	27	1,5	1,5	2,25	14,85	5,85	0,975
1	6	1	36	6	−3	−3	9	9	9	−1,5	−0,5	0,75	3,15	−5,85	−0,975
20	45	100	585	219	0	0	20	180	39	0,0	0,0	3,25	45,00	0,00	0,000

$\bar{X} = \frac{20}{5} = 4, \quad \bar{Y} = \frac{45}{5} = 9, \quad s_x = \sqrt{20/5} = 2, \quad s_y = \sqrt{180/5} = 6, \quad r_{xy} = \frac{3,25}{5} = 0,65$

a) *Recta de regresión a partir de puntuaciones directas*

Según (11.7):

$$B = \frac{(5)(219) - (20)(45)}{(5)(100) - (20)^2} = \frac{195}{100} = 1,95$$

Según (11.5):

$$A = 9 - (1,95)(4) = 9 - 7,80 = 1,20$$

Por consiguiente,

$$Y'_i = 1,20 + 1,95\, X_i \qquad \text{según (11.8).}$$

Aplicando esta ecuación a las cinco puntuaciones directas en X (3, 5, 4, 7, 1), obtenemos la columna encabezada por Y' en la tabla 11.1.

Nótese cómo $\Sigma Y = \Sigma Y' = 45$, o, $\bar{Y} = \bar{Y}' = 9$.

b) *Recta de regresión a partir de puntuaciones diferenciales*

Según (11.16):

$$b = \frac{39}{20} = 1,95$$

Según (11.17):

$$b = 0,65\frac{6}{2} = 1,95$$

Según (11.15):

$$a = 0$$

Por consiguiente,

$$y'_i = 1,95\, x_i, \qquad \text{según (11.19).}$$

Aplicando esta ecuación a los cinco puntos diferenciales en X (−1, 1, 0, 3, −3), obtenemos la columna encabezada por y' en la tabla 12.1.

Nótese cómo $\Sigma y = \Sigma y' = 0$, o $\bar{y} = \bar{y}' = 0$.

c) *Recta de regresión a partir de puntuaciones típicas*

Según (11.27):

$$b^* = 0,65$$

Según (11.26):

$$a^* = 0$$

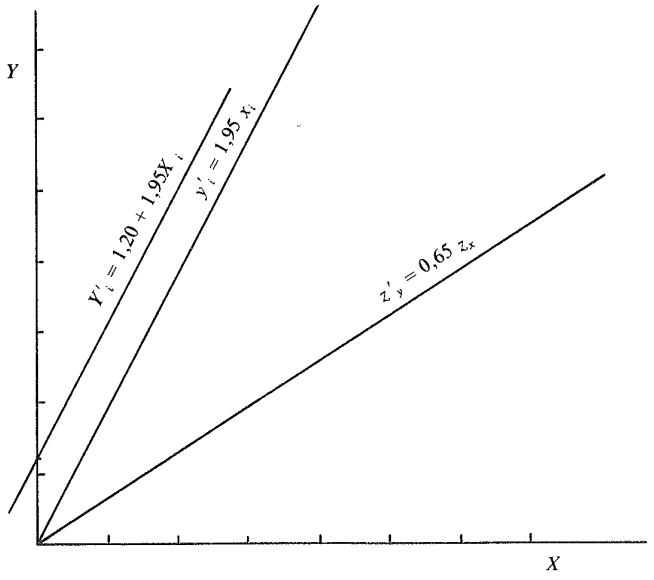
Por consiguiente,

$$z'_y = 0,65\, z_x, \qquad \text{según (11.28).}$$

Aplicando esta ecuación a las cinco puntuaciones típicas en X(−0,5; 0,5; 0,0; 1,5; −1,5), obtenemos la columna encabezada por z'_y en la tabla 11.1.

Nótese cómo $\Sigma z_y = \Sigma z'_y = 0$, o, $\bar{z}_y = \bar{z}'_y = 0$.

La representación gráfica de las tres rectas de regresión anteriores es la siguiente:



EJEMPLO 11.2. La tabla 11.2 nos presenta las puntuaciones obtenidas por 22 alumnos de Enseñanza General Básica de la ciudad de Oviedo en un test de razonamiento abstracto, X, y las notas alcanzadas por los mismos en rendimiento escolar, Y, a fin de curso (promedio de sus calificaciones en todas las materias cursadas). Los datos han sido ofrecidos por M. C. García Méndez.

TABLA 11.2

X	Y	Y'	y'	z'_y
2	4,3	2,8917	−3,1810	−1,9214
20	7,5	7,0083	0,9356	0,5651
12	5,0	5,1787	−0,8940	−0,5400
21	6,9	7,2370	1,1643	0,7032
19	5,2	6,7796	0,7069	0,4270
17	4,4	6,3222	0,2495	0,1507
14	6,3	5,6361	−0,4366	−0,2637
13	5,6	5,4074	−0,6653	−0,4018
21	8,7	7,2370	1,1643	0,7032
19	7,6	6,7796	0,7069	0,4270
14	7,0	5,6361	−0,4366	−0,2637
20	9,2	7,0083	0,9356	0,5651
16	6,5	6,0935	0,0208	0,0126
17	6,3	6,3222	0,2495	0,1507
18	8,8	6,5508	0,4782	0,2888
21	7,9	7,2370	1,1643	0,7032
11	5,0	4,9500	−1,1227	−0,6781
17	4,2	6,3222	0,2495	0,1507
16	4,1	6,0935	0,0208	0,0126
13	5,2	5,4074	−0,6653	−0,4018
13	3,6	5,4074	−0,6653	−0,4018
16	4,3	0,0208	0,0208	0,0126
350	133,6	133,5996	0,0002	0,0002

$$\Sigma X = 2 + 20 + \cdots + 16 = 350$$

$$\bar{X} = \frac{350}{22} = 15{,}9091$$

$$\Sigma Y = 4{,}3 + 7{,}5 + \cdots + 4{,}3 = 133{,}6$$

$$\bar{Y} = \frac{133{,}6}{22} = 6{,}0727$$

$$\Sigma X^2 = (2)^2 + (20)^2 + \cdots + (16)^2 = 5.972$$

$$\Sigma Y^2 = (4{,}3)^2 + (7{,}5)^2 + \cdots + (4{,}3)^2 = 871{,}62$$

$$\Sigma XY = (2)(4{,}3) + (20)(7{,}5) + \cdots + (16)(4{,}3) = 2.217{,}8$$

$$s_x = \sqrt{\frac{(22)(5.972) - (350)^2}{(22)^2}} = 4{,}2843$$

$$s_y = \sqrt{\frac{(22)(871{,}62) - (133{,}6)^2}{(22)^2}} = 1{,}6556$$

$$r_{xy} = \frac{(22)(2.217{,}8) - (350)(133{,}6)}{\sqrt{(22)(5.972) - (350)^2}\sqrt{(22)(871{,}62) - (133{,}6)^2}} = 0{,}5918$$

a) *Recta de regresión a partir de puntuaciones directas*

Según (11.7):

$$B = \frac{(22)(2.217{,}8) - (350)(133{,}6)}{(22)(5.972) - (350)^2} = 0{,}2287$$

Según (11.5):

$$A = 6{,}0727 - (0{,}2287)(15{,}9091) = 2{,}4343$$

Por consiguiente,

$$Y'_i = 2{,}4343 + 0{,}2287\, X_i$$

Nótese cómo en la tabla 11.2 $\Sigma Y_i = \Sigma Y'_i = 133{,}6$

b) *Recta de regresión a partir de puntuaciones diferenciales*

Según (11.18):

$$b = B = 0{,}2287$$

Según (11.15):

$$a = 0$$

Por consiguiente,

$$y'_i = 0{,}2287\, x_i$$

Nótese cómo en la tabla 11.2 $\Sigma y'_i = 0{,}0002 \simeq 0{,}000$.

c) *Recta de regresión a partir de puntuaciones típicas*

Según (11.27):

$$b^* = 0{,}5918$$

Según (11.26):

$$a^* = 0$$

Por consiguiente,

$$z'_y = 0{,}5918\, z_x$$

Nótese cómo en la tabla 11.2 $\Sigma\, z'_i = 0{,}0002 \simeq 0{,}000$.

11.4. Ecuaciones de las rectas de regresión de *X* sobre *Y*, según el criterio de mínimos cuadrados

Junto a las ecuaciones de regresión de Y sobre X, tenemos las ecuaciones de regresión de X sobre Y. Mediante ellas intentamos pronosticar X a partir de Y. Para ello nos valemos de una recta que haga mínima la expresión $\Sigma\,(X_i - X'_i)^2$.

De acuerdo con un razonamiento análogo al seguido en el caso de las rectas de regresión de Y sobre X, es fácil demostrar que ahora es:

a) *Expresada en puntuaciones directas*

$$A = \bar{X} - B\bar{Y} \tag{11.31}$$

$$B = \frac{n\,\Sigma\, YX - \Sigma\, Y\,\Sigma\, X}{n\,\Sigma\, Y^2 - (\Sigma\, Y)^2} \tag{11.32}$$

b) *Expresada en puntuaciones diferenciales*

$$a = 0 \tag{11.33}$$

$$b = \frac{\Sigma\, yx}{\Sigma\, y^2} \tag{11.34}$$

$$b = r_{xy}\frac{s_x}{s_y} \tag{11.35}$$

c) *Expresada en puntuaciones típicas*

$$a^* = 0 \tag{11.36}$$

$$b^* = r_{xy} \tag{11.37}$$

Ejemplo 11.3. Apliquemos estas ecuaciones a los datos del ejemplo 11.1.

a) *Expresada en puntuaciones directas*

Según (11.32):

$$B = \frac{(5)(219) - (20)(45)}{(5)(585) - (45)^2} = \frac{195}{900} = 0{,}2167$$

Según (11.31):

$$A = 4 - (0{,}2167)(9) = 2{,}0497$$

Por consiguiente,

$$X'_i = 2{,}0497 + 0{,}2167\, Y_i$$

b) *Expresada en puntuaciones diferenciales*

Según (11.34):

$$b = \frac{39}{180} = 0{,}2167$$

Según (11.33):

$$a = 0$$

Por consiguiente,

$$x'_i = 0{,}2167\, y_i$$

c) *Expresada en puntuaciones típicas*

Según (11.37):

$$b^* = 0{,}65$$

Según (11.36):

$$a^* = 0$$

Por consiguiente,

$$z'_x = 0{,}65\, z_y$$

Las ecuaciones anteriores dan lugar a los siguientes valores pronosticados:

Puntuaciones directas pronosticadas:	4,	4,65,	2,05,	5,95,	3,35
Puntuaciones diferenciales pronosticadas:	0,	0,65,	−1,95,	1,95,	−0,65
Puntuaciones pseudotípicas pronosticadas:	0,	0,325,	−0,975,	0,975,	−0,325

En adelante, hablaremos sólo de ecuaciones de regresión de Y sobre X. Con ellas podemos resolver los problemas que se nos presenten. Basta con llamar X a la variable que hará oficio de predictora y llamar Y a la que hará oficio de criterio, es decir, a la que será pronosticada a partir de la predictora. Por ejemplo, si intentamos pronosticar el peso a partir de la altura, llamaremos X a la altura e Y al peso. Pero si intentamos pronosticar la altura a partir del peso, llamaremos X al peso e Y a la altura.

11.5. Aplicación de las rectas de regresión

Una vez construidas las rectas de regresión, las podemos aplicar a otras personas con tal que sean semejantes a aquellas con las que las hemos construido. En realidad, suponemos que tanto el grupo con el que hemos construido las rectas, como el grupo al que se las aplicamos no son más que dos muestras de la misma población.

Sea X un test de aptitud y sea Y el aprovechamiento escolar o notas en el examen de fin de curso. Supongamos que para las personas del grupo primero (mediante las cuales construimos las rectas) $\bar{X} = 25$, $\bar{Y} = 35$, $s_x = 4$, $s_y = 6$, $r_{xy} = 0{,}80$. No olvidemos que de este primer grupo de personas conocemos sus puntuaciones en X y en Y. A partir de estos datos tendremos las siguientes ecuaciones de regresión:

$$Y' = 5 + (1{,}2)\,X \tag{11.38}$$

$$y' = (1{,}2)\,x \tag{11.39}$$

$$z_{y'} = (0{,}80)\,z_x \tag{11.40}$$

Un nuevo alumno (semejante a los primeros) hace el test de aptitud y obtiene una puntuación directa igual a 30. ¿Qué puntuación directa, diferencial y típica le pronosticaremos como nota de fin de curso?

Aplicando (11.38), $5 + (1{,}2)(30) = 5 + 36 = 41$. Es decir, le pronosticaremos la puntuación directa 41.

Para el pronóstico de su puntuación diferencial podemos seguir dos caminos:

1) Transformar su puntuación directa pronosticada en diferencial. Es decir, $Y' - \bar{Y}' = Y' - \bar{Y} = 41 - 35 = 6$.

2) Obtener su puntuación diferencial en X y aplicar (11.39) a esta puntuación diferencial. Es decir, $x = X - \bar{X} = 30 - 25 = 5$; $y' = (1{,}2)(5) = 6$. Como se ve llegamos al mismo resultado que antes.

Para el pronóstico de su puntuación «pseudotípica» podemos seguir dos caminos:

1) Transformar su puntuación directa o diferencial pronosticada en «pseudotípica». Es decir, $(Y' - \bar{Y})/s_y = y'/s_y = (41 - 35)/6 = 6/6 = 1$.

2) Obtener su puntuación típica en X y aplicar (11.40) a esta puntuación típica. Es decir, $z_x = (X - \bar{X})/s_x = x/s_x = (30 - 25)/4 = 5/4 = 1{,}25$; $z_{y'} = (0{,}80)(1{,}25) = 1$. Como se ve, llegamos al mismo resultado que antes.

Evidentemente, este nuevo alumno no pertenece al grupo primero mediante el cual hemos calculado $\bar{X}$, $\bar{Y}$, s_x, s_y. Sin embargo, podemos referir sus puntuaciones directas a esas medias y desviaciones típicas (para calcular sus puntuaciones diferenciales y típicas) porque suponemos que pertenece a la misma población a la que pertenecía el grupo primero.

NOTA 1.

Las puntuaciones diferenciales y típicas no son más que las directas sometidas a ciertas condiciones restrictivas. Por tanto, las relaciones válidas entre las directas, serán, también, válidas entre las diferenciales y típicas, con tal que impongamos a las directas las restricciones requeridas.

Ahora bien, introducir puntuaciones diferenciales equivale a imponer $\Sigma x = \Sigma y = \bar{x} = \bar{y} = 0$. Consiguientemente, (11.5) y (11.7) quedarán convertidas en

$$A = 0 + (B)(0) = 0 \qquad \text{y} \qquad B = \frac{n\,\Sigma xy - (0)(0)}{n\,\Sigma x^2 - (0)^2} = \frac{\Sigma xy}{\Sigma x^2}$$

que no son más que (11.15) y (11.16), respectivamente.

A su vez, introducir puntuaciones típicas equivale a imponer $\Sigma z_x = \Sigma z_y = \bar{z}_x = \bar{z}_y = 0$, $\Sigma z_x^2 = n$, $\Sigma z_x z_y = n r_{xy}$. Consiguientemente, (11.5) y (11.7) quedarán convertidas en

$$A = 0 + (B)(0) = 0 \quad \text{y} \quad B = \frac{n\,\Sigma z_x z_y - (0)(0)}{(n)(n) - (0)^2} = \frac{\Sigma z_x z_y}{n} = r_{xy}$$

que no son más que (11.26) y (11.27), respectivamente.

NOTA 2.

Según (11.12), (11.15) y (11.17)

$$y' = r_{xy}\frac{s_y}{s_x}\,x$$

Dividiendo ambos miembros por s_y, nos queda

$$\frac{y'}{s_y} = r_{xy}\frac{x}{s_x} = r_{xy} z_x$$

Ahora bien, $z'_y = r_{xy} z_x$, según (11.28). Por tanto, $z'_y = \dfrac{y'}{s_y}$. Esta igualdad nos vuelve a recordar algo que ya conocíamos, a saber, que z'_y no es auténtica puntuación típica, por ser cociente entre una puntuación diferencial *pronosticada* (y') y la desviación típica de las puntuaciones *obtenidas* (s_y). Es decir, las puntuaciones z'_y son puntuaciones pseudotípicas y no puntuaciones auténticamente típicas.

11.6. Resumen: Definiciones y fórmulas

Recta de regresión de Y sobre X: recta que nos permite hacer pronósticos en la variable Y a partir de las puntuaciones obtenidas en la variable X.

Ecuación de la recta de regresión de Y sobre X:

a) *Expresada en puntuaciones directas*

$$Y' = A + BX$$

donde

$$A = \bar{Y} - B\bar{X}$$

$$B = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}$$

b) *Expresada en puntuaciones diferenciales*

$$y' = a + bx$$

donde

$$a = 0$$

$$b = \frac{\Sigma xy}{\Sigma x^2}$$

$$b = r_{xy}\frac{s_y}{s_x}$$

c) *Expresada en puntuaciones típicas*

$$z'_y = a^* + b^* z_x$$

donde

$$a^* = 0$$

$$b^* = r_{xy}$$

APÉNDICE

A.1. Introducción

Hasta ahora hemos presentado líneas de regresión de la forma $Y' = a + bX$, es decir, en las que era lineal la relación entre X e Y'. Existen, sin embargo, en Psicología situaciones en las que no es el más adecuado este tipo de función lineal. Vamos ahora a presentar otras funciones no lineales de aplicación en las investigaciones psicológicas. En todos los casos siguientes nos valdremos del principio de mínimos cuadrados.

A.2. Función cuadrática, $Y' = a + bX + cX^2$

De acuerdo con el principio de mínimos cuadrados, derivaremos $\phi \equiv \Sigma (Y - Y')^2 = \Sigma (Y - a - bX - cX^2)^2$ respecto a a, b y c, y anulando estas derivadas, obtendremos las tres ecuaciones normales siguientes:

$$\frac{\partial\phi}{\partial a} = -2\Sigma(Y - a - bX - cX^2) = 0, \quad \text{de donde,} \quad \Sigma Y = na + b\Sigma X + c\Sigma X^2$$

$$\frac{\partial\phi}{\partial b} = -2\Sigma(Y - a - bX - cX^2)X = 0, \quad \text{de donde,} \quad \Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$$

$$\frac{\partial\phi}{\partial c} = -2\Sigma(Y - a - bX - cX^2)X^2 = 0, \quad \text{de donde,} \quad \Sigma X^2 Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$$

Despejando a, b y c en las tres ecuaciones anteriores, podemos determinar la ecuación $Y' = a + bX + cX^2$.

EJEMPLO 11.4. Comenzamos con un miniejemplo que aclare la táctica a seguir. Luego propondremos otro extraído de la literatura psicológica.

Construyamos la ecuación cuadrática $Y' = a + bX + cX^2$ a partir de los datos siguientes:

TABLA 11.3

X	Y	X^2	X^3	X^4	XY	X^2Y	Función cuadrática Y'	$(Y - Y')$	$(Y - Y')^2$	Función lineal Y'	$(Y - Y')$	$(Y - Y')^2$
0	0	0	0	0	0	0	0,5	−0,5	0,25	$\frac{8}{11}$	$-\frac{8}{11}$	$\frac{64}{121}$
0	1	0	0	0	0	0	0,5	0,5	0,25	$\frac{8}{11}$	$\frac{3}{11}$	$\frac{9}{121}$
1	2	1	1	1	2	2	2,0	0,0	0,0	$\frac{12}{11}$	$\frac{10}{11}$	$\frac{100}{121}$
2	1	4	8	16	2	4	1,0	0,0	0,0	$\frac{16}{11}$	$-\frac{5}{11}$	$\frac{25}{121}$
3	4	5	9	17	4	6	4,0	0,0	0,50	4	0	$\frac{198}{121} = 1{,}6364$

Las tres ecuaciones normales serán:

(I) $4 = 4a + 3b + 5c$
(II) $4 = 3a + 5b + 9c$
(III) $6 = 5a + 9b + 17c$

Multiplicando (I) por 3 y restando (III) de (3)(I), nos queda: $6 = 7a - 2c$.

Multiplicando (I) por 5, (II) por 3 y restando (3)(II) de (5)(I), nos queda: $8 = 11a - 2c$.

De aquí se deduce inmediatamente, $a = 0{,}5$, $c = -1{,}25$. Sustituyendo estos valores de a y c en (I) deducimos $b = 2{,}75$.

Por consiguiente, la ecuación buscada es la siguiente:

$$Y' = 0{,}5 + 2{,}75\,X - 1{,}25\,X^2$$

Sustituyendo en esta ecuación los valores (0, 0, 1, 2) de X, obtenemos los pronósticos (0,5, 0,5, 2, 1).

La correspondiente suma de errores cuadráticos vale 0,50. (Véase tabla 11.3.)

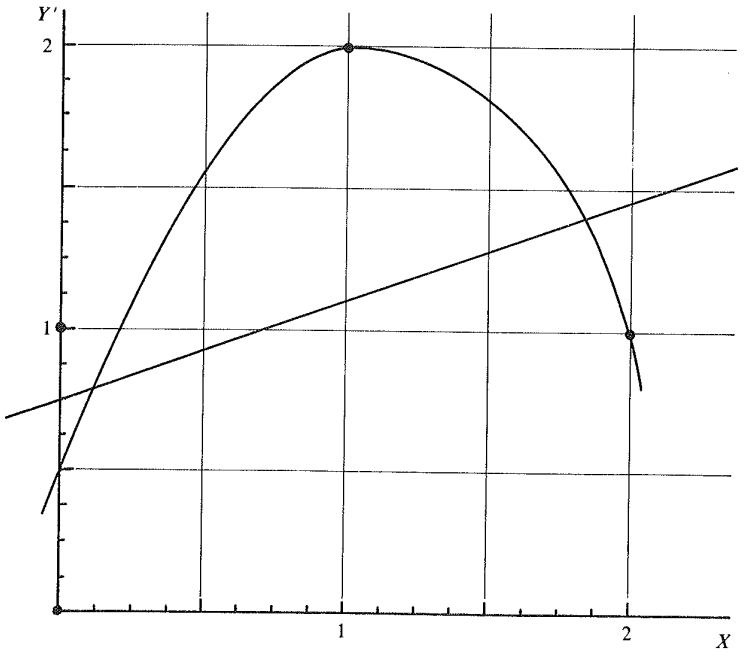
Calculemos ahora, para estos mismos datos, la ecuación $Y' = a + bX$:

$$b = \frac{(4)(4) - (3)(4)}{(4)(5) - (3)^2} = \frac{4}{11} \quad , \quad a = 1 - \frac{3}{11} = \frac{8}{11}$$

Por tanto,
$$Y' = \frac{8}{11} + \frac{4}{11}X$$

Sustituyendo en esta ecuación los valores (0, 0, 1, 2) de X, obtenemos los pronósticos (8/11, 8/11, 12/11, 16/11).

La correspondiente suma de errores cuadráticos vale 1,6364. (Véase tabla 11.3.)



Como se ve, es mucho mayor la suma de errores cuadráticos obtenida mediante la función lineal, que la obtenida mediante la función cuadrática. Ésta se aproxima mejor que la función lineal a los cuatro puntos dados. Esta aproximación puede ser observada en la figura de la página anterior.

EJEMPLO 11.5. Kitterle y Helson (1972) llevaron a cabo un experimento para probar que el tiempo de reacción ante un estímulo luminoso es función cuadrática del intervalo de tiempo transcurrido entre la aparición de ese estímulo y la aparición de otro estímulo perturbador, también luminoso. En la columna X de la tabla 11.4 tenemos, en centisegundos, el tiempo transcurrido entre la aparición del primer estímulo y la aparición del estímulo perturbador; en la columna Y se encuentran los tiempos de reacción correspondientes.

TABLA 11.4

X	Y	*Cuadrát.* Y'	*Lineal* Y'
2	17,3	17,4000	17,9182
4	17,7	17,6836	17,8564
6	17,9	17,8808	17,7945
8	18,2	17,9916	17,7327
10	18,2	18,0160	17,6709
12	17,6	17,9540	17,6091
14	17,8	17,8056	17,5473
16	17,5	17,5708	17,4855
18	17,3	17,2496	17,4236
20	16,9	16,8420	17,3618
110	176,4	176,3940	176,4000

El lector puede comprobar cómo:

$$\Sigma X^2 = 1.540, \quad \Sigma X^3 = 24.200, \quad \Sigma X^4 = 405.328, \quad \Sigma XY = 1.930{,}20, \quad \Sigma X^2 Y = 26.850$$

Las tres ecuaciones normales serán:

$$176{,}4 = 10\,a + 110\,b + 1.540\,c$$
$$1.930{,}20 = 110\,a + 1.540\,b + 24.200\,c$$
$$26.850 = 1.540\,a + 24.200\,b + 405.328\,c$$

De aquí se deduce,

$$a = 17{,}03, \quad b = 0{,}2066, \quad c = -0{,}0108$$

Por consiguiente, la ecuación cuadrática buscada es la siguiente:

$$Y' = 17{,}03 + 0{,}2066\,X - 0{,}0108\,X^2$$

Sustituyendo en esta ecuación los valores (2, 4, . . . , 20) de X, obtenemos los pronósticos (17,4000, 17,6836, . . . , 16,8420). (Véase tabla 11.4.)

La correspondiente suma de errores cuadráticos, $\Sigma\,(Y - Y')^2$, vale 0,2242, como puede comprobar el lector a partir de la tabla 11.4.

Calculemos ahora, para estos mismos datos, la ecuación $Y' = a + bX$.

$$b = \frac{(10)(1.930{,}20) - (110)(176{,}4)}{(10)(1.540) - (110)^2} = -0{,}0309$$

$$a = 17{,}64 - (-0{,}0309)(11) = 17{,}98$$

Por tanto,

$$Y' = 17{,}98 - 0{,}0309\,X$$

Sustituyendo en esta ecuación los valores (2, 4, . . . , 20) de X, obtenemos los pronósticos (17,9182, 17,8564, . . . , 17,3618). (Véase tabla 11.4.)

La correspondiente suma de errores cuadráticos, $\Sigma\,(Y - Y')^2$, vale 1,209, como puede comprobar el lector a partir de la tabla 11.4.

Como se ve, es bastante mayor la suma de errores cuadráticos obtenida mediante la función lineal que la obtenida mediante la función cuadrática. Ésta se aproxima mejor que la función lineal a los diez puntos dados.

A.3. Función potencial: $Y' = aX^b$

Es claro que, tomando logaritmos, la ecuación $Y = aX^b$ se transforma en $\log Y = \log a + b \log X$, ecuación lineal en $\log X$ y $\log Y$. Mediante esta última ecuación, podemos determinar $\log a$, y b, siguiendo la táctica adoptada al tratar de las rectas de regresión. Una vez calculado $\log a$, basta con encontrar su antilogaritmo, para determinar a.

EJEMPLO 11.6. Calculemos a y b, en la ecuación $Y' = aX^b$ a partir de la tabla siguiente:

TABLA 11.5

X	Y	$\log X$	$\log Y$	$(\log X)^2$	$(\log X)(\log Y)$	F. potencial Y'	F. lineal Y'
1	0,2	0,0000	−0,6990	0,0000	0,0000	0,1935	−3,28
2	1,5	0,3010	0,1761	0,0906	0,0530	1,5568	2,83
3	5,0	0,4771	0,6990	0,2276	0,3335	5,2718	8,94
4	13,0	0,6021	1,1139	0,3625	0,6707	12,5256	15,05
5	25,0	0,6990	1,3979	0,4886	0,9771	24,5088	21,16
15	44,7	2,0792	2,6879	1,1693	2,0343	44,0565	44,70

$$b = \frac{(5)(2{,}0343) - (2{,}0792)(2{,}6879)}{(5)(1{,}1693) - (2{,}0792)^2} = 3{,}0082$$

$$\log a = \frac{2{,}6879}{5} - (3{,}0082)\frac{2{,}0792}{5} = -0{,}7133$$

$$a = \text{antilog}\,(-0{,}7133) = 0{,}1935$$

Por consiguiente, la ecuación potencial buscada es:

$$Y' = (0{,}1935)\,X^{3{,}0082}.$$

Sustituyendo en esta ecuación los valores (1, 2, 3, 4, 5) de X, obtenemos los pronósticos expuestos en la tabla 11.5 (0,1935, 1,5568, . . .).

La correspondiente suma de errores cuadráticos, $\Sigma\,(Y - Y')^2$, vale 0,5435, como puede comprobar el lector a partir de la tabla 11.5.

Calculemos ahora, para los datos directos en X y en Y, la ecuación $Y' = a + bX$.

De la tabla 11.5 se deduce:

$$\Sigma X = 15, \quad \Sigma Y = 44{,}7, \quad \Sigma X^2 = 55, \quad \Sigma XY = 195{,}2$$

Por tanto,

$$b = \frac{(5)(195{,}2) - (15)(44{,}7)}{(5)(55) - (15)^2} = 6{,}11, \quad a = \frac{44{,}7}{5} - (6{,}11)\frac{15}{5} = -9{,}39$$

Consiguientemente,

$$Y' = -9{,}39 + 6{,}11\,X$$

Sustituyendo en esta ecuación los valores (1, 2, 3, 4, 5) de X, obtenemos los pronósticos expuestos en la tabla 11.5 (−3,28, 2,83, . . .).

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 48,3510, como puede comprobar el lector a partir de la tabla 11.5.

Como se ve, es enormemente superior la suma de errores cuadráticos obtenida mediante la función lineal que la obtenida mediante la función potencial. Ésta se aproxima mucho mejor que la función lineal a los cinco puntos dados.

EJEMPLO 11.7. Amón (1972) estudió los valores atribuidos a 16 frases acerca del divorcio, teniendo en cuenta el método seguido en la evaluación de las mismas. A continuación se exponen los valores atribuidos a dichas frases según el método de las comparaciones apareadas (decir cuál de las dos frases de cada par favorece más al divorcio) y según el método de la estimación de razones (decir cuántas veces favorece más al divorcio una frase del par que la otra). La variable X representa la estimación de razones y la variable Y las comparaciones apareadas. Los resultados obtenidos fueron los siguientes:

TABLA 11.6

X	Y	F. potencial Y'	F. lineal Y'
1,00	1,00	1,4747	2,1985
1,44	1,71	1,8078	2,3336
1,68	2,09	1,9703	2,4073
2,01	2,34	2,1779	2,5087
2,19	2,75	2,2848	2,5639
3,26	3,11	2,8532	2,8925
3,89	3,66	3,1491	3,0860
4,20	3,65	3,2869	3,1812
4,96	3,76	3,6069	3,4146
6,28	4,35	4,1149	3,8200
6,82	4,45	4,3090	3,9858
8,21	4,83	4,7793	4,4127
8,78	4,91	4,9619	4,5877
12,54	5,42	6,0548	5,7424
12,90	5,63	6,1513	5,8530
14,72	5,74	6,6219	6,4119
94,88	59,40	59,6047	59,3998

El lector puede comprobar cómo:

$$\Sigma \log X = 10{,}3155, \qquad \Sigma \log Y = 8{,}4607$$

$$\Sigma (\log X)^2 = 8{,}6290, \qquad \Sigma (\log X)(\log Y) = 6{,}5597$$

$$b = \frac{(16)(6{,}5597) - (10{,}3155)(8{,}4607)}{(16)(8{,}6290) - (10{,}3155)^2} = 0{,}5585$$

$$\log a = \frac{8{,}4607}{16} - (0{,}5585)\frac{10{,}3155}{16} = 0{,}1687$$

$$a = \text{antilog } (0{,}1687) = 1{,}4747$$

Por consiguiente, la ecuación potencial buscada es:

$$Y' = 1{,}4747\ X^{0{,}5585}$$

Sustituyendo en esta ecuación los valores (1, 1,44, . . . , 14,72) de X, obtenemos los pronósticos (1,4747, 1,8078, . . . , 6,6219) expuestos en la tabla 11.6.

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 2,5071, como puede comprobar el lector a partir de la tabla 11.6.

Calculemos ahora, para los datos directos en X y en Y, la ecuación

$$Y' = a + bX$$

De la tabla 11.6 se deduce:

$$\Sigma X = 94{,}88, \quad \Sigma Y = 59{,}4, \quad \Sigma X^2 = 853{,}5168, \quad \Sigma XY = 441{,}5728$$

Por tanto,

$$b = \frac{(16)(441{,}5728) - (94{,}88)(59{,}4)}{(16)(853{,}5168) - (94{,}88)^2} = 0{,}3071$$

$$a = \frac{59{,}44}{16} - (0{,}3071)\frac{94{,}88}{16} = 1{,}8914$$

Consiguientemente,

$$Y' = 1{,}8914 + 0{,}3071\ X$$

Sustituyendo en esta ecuación los valores (1, 1,44, . . . , 14,72) de X, obtenemos los pronósticos (2,1985, 2,3336, . . . , 6,4119).

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 4,0844, como puede comprobar el lector a partir de la tabla 11.6.

Como se ve, es superior la suma de errores cuadráticos obtenida mediante la función lineal, que la obtenida mediante la función potencial. Ésta se aproxima mejor que la función lineal a los 16 puntos dados.

A.4. Función exponencial; $Y' = ab^X$

Es claro que, tomando logaritmos, la ecuación $Y = ab^X$ se transforma en $\log Y = \log a + (\log b)\, X$, ecuación lineal en X y $\log Y$. Mediante esta última ecuación podemos determinar $\log a$ y $\log b$, siguiendo la táctica adoptada al tratar de las rectas de regresión. Una vez calculados $\log a$ y $\log b$, basta con encontrar sus antilogaritmos, para determinar a y b.

EJEMPLO 11.8. Los datos siguientes, propuestos por Ebbinghaus, están tomados de Lewis (1960). La variable X representa el número de sílabas sin sentido que deben ser memorizadas. La variable Y representa el número de repeticiones necesarias para memorizarlas.

TABLA 11.7

X	Y	X^2	$\log Y$	$X \log Y$	F. exponenc. Y'	F. lineal Y'
1,0	7	1,00	0,8451	0,8451	7,1267	4,5755
16,6	12	275,56	1,0792	17,9147	11,2505	12,5212
30,0	16	900,00	1,2041	36,1230	16,6532	19,3464
44,0	24	1.936,00	1,3802	60,7288	25,0869	26,4772
55,0	36	3.025,00	1,5563	85,5965	34,6150	32,0799
146,6	95	6.137,56	6,0649	201,2081	94,7323	95,0002

$$\log b = \frac{(5)(201{,}2081) - (146{,}6)(6{,}0649)}{(5)(6.137{,}56) - (146{,}6)^2} = 0{,}012715, \qquad b = \text{antilog } (0{,}012715) = 1{,}0297$$

$$\log a = \frac{6{,}0649}{5} - 0{,}012715\frac{146{,}6}{5} = 0{,}840176, \qquad a = \text{antilog } (0{,}840176) = 6{,}9211$$

Por tanto, la ecuación exponencial buscada es:

$$Y' = (6{,}9211)(1{,}0297)^X$$

Sustituyendo en esta ecuación los valores (1, 16,6, 30, 44, 55) de X, obtenemos los pronósticos expuestos en la tabla 11.7 (7,1267, . . .).

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 4,1042, como puede comprobar el lector a partir de la tabla 11.7.

Calculemos ahora, para los datos directos en X y en Y, la ecuación

$$Y' = a + bX$$

De la tabla 11.7 se deduce:

$$\Sigma X = 146{,}6, \quad \Sigma Y = 95, \quad \Sigma X^2 = 6.137{,}56, \quad \Sigma XY = 3.722{,}2.$$

Por tanto,

$$b = \frac{(5)(3.722{,}2) - (146{,}6)(95)}{(5)(6.137{,}56) - (146{,}6)^2} = 0{,}50934, \qquad a = \frac{95}{5} - 0{,}50934\frac{146{,}6}{5} = 4{,}0662$$

Consiguientemente,

$$Y' = 4{,}0662 + 0{,}50934\, X$$

Sustituyendo en esta ecuación los valores (1, 16,6, 30, 44, 55) de X, obtenemos los pronósticos expuestos en la tabla 11.7 (4,5755, . . .).

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 38,8519, como puede comprobar el lector a partir de la tabla 11.7.

Como se ve, es mucho mayor la suma de errores cuadráticos obtenida mediante la función lineal que la obtenida mediante la función exponencial. Ésta se aproxima mucho mejor que la función lineal a los cinco puntos dados.

A.5. Función logarítmica: $Y' = a + b \log X$

Como esta ecuación es lineal en Y y en $\log X$, seguiremos la táctica adoptada al tratar de las rectas de regresión.

EJEMPLO 11.9. Doce alumnos de la Universidad Complutense de Madrid, aceptando como criterio un estímulo luminoso, L, evaluaron la intensidad luminosa de otros respecto a L. La variable X representa la intensidad luminosa, medida en unidades arbitrarias. La variable Y representa las estimaciones de dicha intensidad ofrecidas por los sujetos experimentales *.

TABLA 11.8

X	Y	$\log X$	$(\log X)^2$	$(\log X)(Y)$	F. logarit. Y'	F. lineal Y'
0,1	0,3	−1,0000	1,0000	−0,3000	0,1517	11,2236
0,4	4,5	−0,3979	0,1583	−1,7906	5,4380	11,2611
9,1	18,0	0,9590	0,9197	17,2620	17,3527	12,3495
18,2	19,7	1,2601	1,5879	24,8240	19,9958	13,4879
36,4	25,0	1,5611	2,4370	39,0275	22,6389	15,7647
145,5	26,0	2,1629	4,6781	56,2354	27,9226	29,4132
209,7	93,5	4,5452	10,7810	135,2583	93,4997	93,5000

$$b = \frac{(6)(135{,}2583) - (4{,}5452)(93{,}5)}{(6)(10{,}7810) - (4{,}5452)^2} = 8{,}7803, \qquad a = \frac{93{,}5}{6} - 8{,}7803\frac{4{,}5452}{6} = 8{,}9320$$

* Los datos han sido ofrecidos por L. Jáñez.

Por tanto, la ecuación logarítmica buscada es:

$$Y' = 8{,}9320 + 8{,}7803 \log X$$

Sustituyendo en esta ecuación los valores (0,1, 0,4, ...) de X, obtenemos los pronósticos (0,1517, ...) expuestos en la tabla 11.8.

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 10,6795, como puede comprobar el lector a partir de la tabla 11.8.

Calculemos ahora, para los datos directos en X y en Y, la ecuación

$$Y' = a + bX$$

De la tabla 11.8 se deduce:

$$\Sigma X = 209{,}7, \quad \Sigma Y = 93{,}5, \quad \Sigma X^2 = 22.909{,}43, \quad \Sigma XY = 5.217{,}17$$

Por consiguiente,

$$b = \frac{(6)(5.217{,}17) - (209{,}7)(93{,}5)}{(6)(22.909{,}43) - (209{,}7)^2} = 0{,}1251, \quad a = \frac{93{,}5}{6} - 0{,}1251\,\frac{209{,}7}{6} = 11{,}2111$$

Consiguientemente,

$$Y' = 11{,}2111 + 0{,}1251\,X$$

Sustituyendo en esta ecuación los valores (0,1, 0,4, ...) de X, obtenemos los pronósticos (11,2236, ...) expuestos en la tabla 11.8.

La correspondiente suma de errores cuadráticos, $\Sigma (Y - Y')^2$, vale 332,4966, como puede comprobar el lector a partir de la tabla 11.8.

Como se ve, es mucho mayor la suma de errores cuadráticos obtenida mediante la función lineal que la obtenida mediante la función logarítmica. Ésta se aproxima mucho mejor que la función lineal a los seis puntos dados.

EJERCICIOS

11.1. Calcular las ecuaciones de regresión de Y sobre X y de X sobre Y, en puntuaciones directas, diferenciales y (pseudo)típicas, a partir de los datos contenidos en el ejercicio 10.1 del capítulo anterior.

11.2. A partir de los datos: X (2, 8, 14, 6, 10) e Y (6, 10, 22, -2, 14), calcular las puntuaciones directas (Y'), diferenciales (y') y pseudotípicas (z'_y) pronosticadas mediante las correspondientes rectas de regresión de Y sobre X.

11.3. Sean $\bar{X} = 10$, $\bar{Y} = 3\bar{X}$, $s_y = 2s_x$. Esto supuesto, ¿qué puntuación directa se podría pronosticar, mediante la recta de regresión de Y sobre X, a una persona con puntuación directa igual a 5,

a) si $r_{xy} = 1$, *b*) si $r_{xy} = 0$, *c*) si $r_{xy} = -1$?

11.4. Calcular la ecuación de regresión de Y sobre X, en puntuaciones directas y diferenciales, sabiendo que dicha recta debe hacer corresponder a $X = 10$, $Y' = 25$ y que $\bar{X} = 20$ e $\bar{Y} = 45$.

11.5. Supongamos que la recta de regresión de Y sobre X hace corresponder a las puntuaciones directas $X_1 = 2$ y $X_2 = -1$, las puntuaciones directas $Y'_1 = 2{,}2$ e $Y'_2 = -2{,}6$, respectivamente. Suponiendo, además, que $\bar{X} + \bar{Y} = 12$, calcular cuánto valen $\bar{X}$ e $\bar{Y}'$.

11.6. Sabiendo que $s_x = 6$, $s_y = 5$ y $s_{y'} = 3$, calcular la recta de regresión de Y sobre X, en puntuaciones diferenciales, suponiendo que es positiva la correlación entre X e Y.

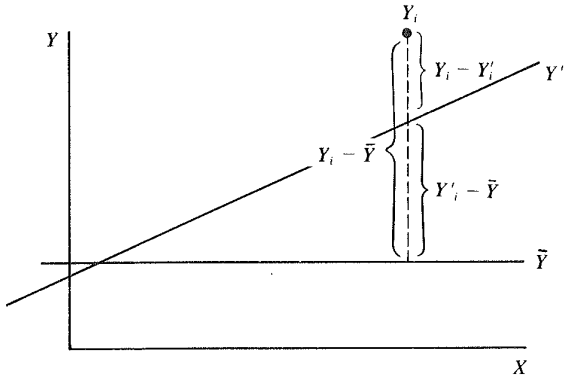
11.7. Suponiendo que $\bar{X} = 5$, que $\bar{Y} = 10$ y que $y' = 0{,}8x$, calcular la ecuación de regresión de Y sobre X, en puntuaciones directas.

12

El coeficiente de correlación de Pearson, r_{xy}, y las rectas de regresión

12.1. r_{xy}^2 como índice de reducción de error en los pronósticos

Vamos a considerar, en concreto, la recta de regresión de Y sobre X. Algo análogo valdrá para la recta de regresión de X sobre Y.



Sea Y_i la puntuación directa obtenida por la persona i e Y_i' la puntuación directa pronosticada a esa misma persona, mediante la recta de regresión de Y sobre X. El error cuadrático cometido en ese pronóstico individual valdrá $(Y_i - Y_i')^2$ y $\Sigma (Y_i - Y_i')^2$ será la suma de errores cuadráticos respecto a todas las personas de la muestra. Por otra parte, la suma $\Sigma (Y - \bar{Y})^2$ queda descompuesta en dos sumas de términos cuadráticos. En efecto,

$$\Sigma (Y - \bar{Y})^2 = \Sigma [(Y - Y') + (Y' - \bar{Y})]^2 = \Sigma (Y - Y')^2 + \Sigma (Y' - \bar{Y})^2 \quad (12.1)$$

pues $\Sigma (Y - Y')(Y' - \bar{Y}) = 0$, según se encargará de comprobar el lector en el ejercicio 12.6.

Ahora bien, $\dfrac{\Sigma (Y - \bar{Y})^2}{n}$, que hemos designado por s_y^2, es el error cuadrático medio, E_m^2, cometido al atribuir a cada persona, como puntuación, la media $\bar{Y}$.

A su vez, $\dfrac{\Sigma (Y - Y')^2}{n}$, que designaremos por $s_{y.x}^2$, es el E_m^2 cometido al atribuir a cada persona la puntuación Y' obtenida mediante la recta de regresión.

Finalmente, según (12.1), $\dfrac{\Sigma (Y' - \bar{Y})^2}{n} = \dfrac{\Sigma (Y - \bar{Y})^2}{n} - \dfrac{\Sigma (Y - Y')^2}{n}$, que designaremos por $s_{y'}^2$, no es más que la diferencia entre el E_m^2 cometido valiéndonos de $\bar{Y}$ y el E_m^2 cometido valiéndonos de Y'. En otras palabras, es aquella parte del E_m^2 que dejamos de cometer por el hecho de atribuir Y' a cada persona en vez de atribuirle $\bar{Y}$.

Por consiguiente,

$$\frac{s_{y'}^2}{s_y^2} = \frac{s_{y'}^2}{s_{y'}^2 + s_{y.x}^2} = \frac{\text{parte de } E_m^2 \text{ eliminada}}{\text{parte de } E_m^2 \text{ eliminada} + \text{parte de } E_m^2 \text{ no eliminada}} =$$

$$= \text{proporción de } E_m^2 \text{ eliminado}$$

o, de otro modo, proporción en que reducimos el E_m^2 primitivo, el que cometíamos valiéndonos de $\bar{Y}$. Si, por ejemplo, ese cociente fuera igual a 0,65, ello significaría que habíamos reducido el E_m^2 en un 65 por 100, es decir, que, valiéndonos de Y', sólo cometeríamos un 35 por 100 del que habríamos cometido valiéndonos de $\bar{Y}$.

Pero,

$$\frac{s_{y'}^2}{s_y^2} = \frac{\Sigma y'^2/n}{s_y^2} = \frac{b^2 \Sigma x^2/n}{s_y^2} = \frac{b^2 s_x^2}{s_y^2} = \frac{\left(\dfrac{\Sigma xy}{\Sigma x^2}\right)^2 s_x^2}{s_y^2} = \frac{\left(\dfrac{s_{xy}}{s_x^2}\right)^2 s_x^2}{s_y^2} =$$

$$= \frac{s_{xy}^2}{s_x^2 s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2 = r_{xy}^2 \quad (12.2)$$

Por tanto, r_{xy}^2 no es más que la proporción en que ha sido reducido el error cuadrático primitivo, el que habríamos cometido si nos hubiéramos valido de $\bar{Y}$ como puntuación pronosticada.

En conclusión,

$$r_{xy}^2 = \frac{s_{y'}^2}{s_y^2} = \frac{s_y^2 - s_{y \cdot x}^2}{s_y^2} = 1 - \frac{s_{y \cdot x}^2}{s_y^2} \quad (12.3)$$

Consideremos los dos casos extremos posibles: $s_{y \cdot x}^2 = 0$ (equivalente a $s_{y'}^2 = s_y^2$) y $s_{y'}^2 = 0$ (equivalente a $s_{y \cdot x}^2 = s_y^2$).

El primero nos indica que, valiéndonos de la recta de regresión, hemos eliminado o reducido por completo el E_m^2 primitivo, el que habríamos cometido atribuyendo $\bar{Y}$, como puntuación en Y, a cada una de las personas de la muestra. Ahora bien, si $s_{y.x}^2 = 0$, entonces, según (12.3), $r_{yx}^2 = 1$; luego a reducción total del E_m^2, corresponde $r_{xy}^2 = 1$. Recíprocamente, si $r_{xy}^2 = 1$, entonces, según (12.3), $s_{y.x}^2 = 0$; luego a $r_{xy}^2 = 1$, corresponde reducción total del E_m^2 primitivo.

El segundo caso extremo nos indica que, valiéndonos de la recta de regresión, no hemos reducido en nada el E_m^2 que cometíamos atribuyendo $\bar{Y}$, como puntuación en Y, a cada una de las personas de la muestra. Ahora bien, si $s_{y'}^2 = 0$, entonces, según (12.3), $r_{xy}^2 = 0$; luego a reducción nula del E_m^2 primitivo, corresponde $r_{xy}^2 = 0$. Recíprocamente, si $r_{xy}^2 = 0$, entonces, según (12.3), $s_{y'}^2 = 0$; luego a $r_{xy}^2 = 0$, corresponde reducción nula del E_m^2 primitivo.

En conclusión,

a $r_{xy}^2 = 1$, corresponde reducción total del E_m^2 primitivo y recíprocamente.
a $r_{xy}^2 = 0$, corresponde reducción nula del E_m^2 primitivo y recíprocamente.

El valor $s_{y.x}$ suele ser llamado error típico de estimación.

Teniendo en cuenta (12.3), nos queda

$$s_{y.x}^2 = s_y^2(1 - r_{xy}^2) \tag{12.4}$$

Nótese que para $r_{xy} = 0{,}20$, hemos reducido el E_m^2 primitivo en un 4 por 100 ($0{,}20^2 = 0{,}04$); y para $r_{xy} = 0{,}50$, lo hemos reducido en un 25 por 100 ($0{,}50^2 = 0{,}25$). En cambio, para $r_{xy} = 0{,}65$, lo hemos reducido en un 42 por 100 ($0{,}65^2 = 0{,}42$); y para $r_{xy} = 0{,}95$, lo hemos reducido en un 90 por 100 ($0{,}95^2 = 0{,}90$). Es decir, la diferencia en la proporción de E_m^2 reducido es mayor ($0{,}90 - 0{,}42 = 0{,}48$) en el segundo caso que en el primero ($0{,}25 - 0{,}04 = 0{,}21$), a pesar de que en ambos la diferencia entre las correspondientes correlaciones es la misma ($0{,}95 - 0{,}65 = 0{,}50 - 0{,}20$). En otras palabras, el paso de 0,20 a 0,50 no significa lo mismo que el paso de 0,65 a 0,95, en cuanto a reducción de E_m^2.

De acuerdo con (12.2) vemos que, definiendo r_{xy} como $\frac{s_{xy}}{s_x s_y}$, llegamos a $r_{xy}^2 = \frac{s_{y'}^2}{s_y^2}$.

Recíprocamente, vemos que, definiendo r_{xy}^2 como $\frac{s_{y'}^2}{s_y}$, llegamos a $r_{xy} = \frac{s_{xy}}{s_x s_y}$.

La definición de r_{xy}^2 como

$$\frac{s_{y'}^2}{s_y^2} = \frac{\Sigma (Y' - \bar{Y})^2}{\Sigma (Y - \bar{Y})^2} \tag{12.5}$$

$$= 1 - \frac{\Sigma (Y - Y')^2}{\Sigma (Y - \bar{Y})^2} \tag{12.6}$$

tiene la ventaja de su paralelismo con la definición de otros coeficientes de correlación como índices de reducción de error. Unos diferirán de otros en diversos aspectos accidentales, pero todos ellos coincidirán en ser índices de reducción de error.

Como ejercicios (también se encuentran propuestos al final del capítulo), el lector se encargará de probar que

$$\Sigma y^2 = \Sigma (y - y')^2 + \Sigma y'^2 \tag{12.7}$$

$$\Sigma z_y^2 = \Sigma (z_y - z'_y)^2 + \Sigma z_y'^2 \tag{12.8}$$

EJEMPLO 12.1. Comprobemos las relaciones acabadas de exponer, recordando la tabla 11.1 del ejemplo 11.1. Valiéndonos de esta tabla formemos la siguiente:

$(Y - \bar{Y})^2 = y^2$	z_y^2	Y'	y'	z'_y	$(Y - Y') = (y - y')$	$z_y - z'_y$	$(Y - Y')^2 = (y - y')^2$	$(z_y - z'_y)^2$	$(Y' - \bar{Y})^2 = y'^2$	$z_y'^2$
0	0,00	7,05	−1,95	−0,325	1,95	0,325	3,8025	0,105625	3,8025	0,105625
9	0,25	10,95	1,95	0,325	1,05	0,175	1,1025	0,030625	3,8025	0,105625
81	2,25	9,00	0,00	0,000	−9,00	−1,500	81,0000	2,250000	0,0000	0,000000
81	2,25	14,85	5,85	0,975	3,15	0,525	9,9225	0,275625	34,2225	0,950625
9	0,25	3,15	−5,85	−0,975	2,85	0,475	8,1225	0,225625	34,2225	0,950625
180	5,00	45,00	0,00	0,000	0,00	0,000	103,9500	2,887500	76,0500	2,112500

Comprobación de (12.1):

$$\Sigma (Y - Y')^2 + \Sigma (Y' - \bar{Y})^2 = 103{,}95 + 76{,}05 = 180 = \Sigma (Y - \bar{Y})^2$$

Comprobación de (12.7):

$$\Sigma (y - y')^2 + \Sigma y'^2 = 103{,}95 + 76{,}05 = 180 = \Sigma y^2$$

Según vimos en el ejemplo 11.1, $r_{xy}^2 = (0{,}65)^2 = 0{,}4225$. *

Comprobación de (12.2):

$$s_{y'}^2/s_y^2 = \frac{(76{,}05)/(5)}{(180)/(5)} = 0{,}4225 = r_{xy}^2$$

Comprobación de (12.3):

$$1 - s_{y.x}^2/s_y^2 = 1 - \frac{(103{,}95)/(5)}{(180)/(5)} = 1 - 0{,}5775 = 0{,}4225 = r_{xy}^2$$

---

* En el capítulo anterior, en (11.30) hemos visto que $r_{xy}^2 = s_{z'_y}^2$. Ahora vemos comprobada esta igualdad. En efecto,

$$r_{xy}^2 = 0{,}4225 = \frac{2{,}1125}{5} = \Sigma z_y'^2/n = s_{z'_y}^2.$$

Comprobación de (12.4):

$$s_y^2(1 - r_{xy}^2) = \frac{180}{5}(1 - 0{,}4225) = (36)(0{,}5775) = 20{,}79 = \frac{103{,}95}{5} = s_{y.x}^2$$

Comprobación de (12.8):

$$\Sigma (z_y - z'_y)^2 + \Sigma z_y'^2 = 2{,}8875 + 2{,}1125 = 5 = \Sigma z_y^2$$

NOTA. A pesar de lo expuesto, es posible que $r_{xy}^2 = 0$, valiendo $s_{y.x}^2 = 0$. Esto sucedería si todos los puntos estuvieran situados sobre la recta de regresión y ésta fuera paralela al eje OX. Ciertamente, $s_{y.x}^2 = 0$. Además, $s_y^2 = 0$. Por tanto, $s_{y'}^2 = 0 - 0 = 0$. Consiguientemente, $r_{xy}^2 = s_{y'}^2/s_y^2 = 0/0$, quedaría indeterminado. Sin embargo, tiene sentido admitir $r_{xy} = 0$, ya que en esta situación, tanto a puntuaciones altas como medias y bajas en X, corresponde una misma puntuación, $\bar{Y}$, en Y y, por tanto, la covariación entre X e Y es nula.

12.2. r_{xy}^2 como índice de aproximación de los puntos a la recta de regresión

Hemos visto en el párrafo anterior que $r_{xy}^2 = 1 - \Sigma (Y - Y')^2/\Sigma (Y - \bar{Y})^2$. De aquí se deduce inmediatamente lo siguiente:

a) Si todos los puntos (representantes de las puntuaciones obtenidas) se encuentran sobre la recta de regresión de Y sobre X, entonces, $Y = Y'$ para toda persona de la muestra. Por tanto, $Y - Y' = 0$. Consiguientemente, $\Sigma (Y - Y')^2 = 0$. Es decir, $r_{xy}^2 = 1$.

b) Si $r_{xy}^2 = 1$, entonces, $\Sigma (Y - Y')^2 = 0$. Ahora bien, como los n sumandos son no negativos (por ser cuadráticos), si su suma es cero, quiere decir que cada uno de ellos tiene que valer cero. Por consiguiente, $Y - Y' = 0$, es decir, $Y = Y'$, para toda persona de la muestra. En otras palabras, todos los puntos se encuentran sobre la recta de regresión de Y sobre X.

En conclusión:

Si todos los puntos se encuentran sobre la recta de regresión, $r_{xy}^2 = 1$.
Si $r_{xy}^2 = 1$, todos los puntos se encuentran sobre la recta de regresión.

Por consiguiente, r_{xy}^2 nos mide la aproximación de los puntos a la recta de regresión. Nótese, por tanto, que lo que mide r_{xy} es la aproximación de los puntos a una línea *recta*. Es posible que todos los puntos se encuentren exactamente sobre una línea *curva*, que exista una relación funcional (no lineal) perfecta entre X e Y que nos permita pronosticar un valor exacto de Y para cada valor dado de X, y que r_{xy} valga cero. Recuérdese a este respecto el ejemplo propuesto en el apartado 10.4.

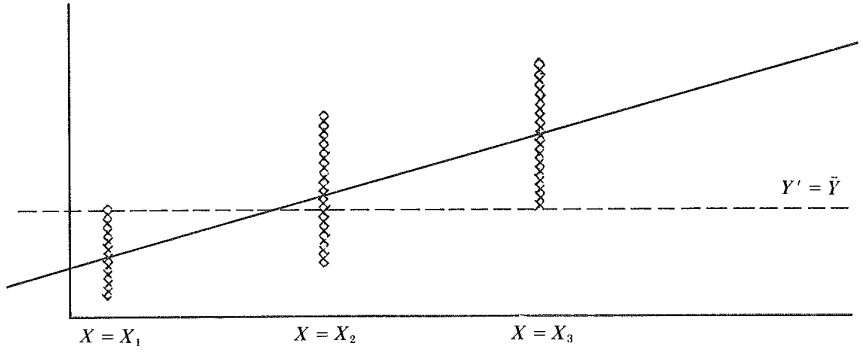
A pesar de lo acabado de exponer, si todos los puntos se encuentran sobre una recta paralela al eje OX, $r_{xy} = 0$. Vea el lector lo dicho sobre este caso particular en la NOTA del párrafo anterior.

12.3. r_{xy}^2 como proporción de la varianza de Y asociada a la variación de X

La relación $Y = Y' + (Y - Y')$ es una pura identidad que puede ser interpretada aceptando que la puntuación directa obtenida es igual al pronóstico (Y') más el error en dicho pronóstico ($Y - Y'$).

Sabemos que $Y' = a + bX$. Esto nos indica que Y' depende, es función de, está asociada a, X. Más concretamente, dado que a y b son dos constantes, valdrá 1 el coeficiente de correlación de Pearson entre X e Y' (es decir, entre X y $a + bX$), según lo dicho en 10.2.3. *b*. Puesto que $Y' = a + bX$, si suponemos que r_{xy} es positiva, a puntuación baja en X, corresponderá pronóstico bajo en Y; a puntuación media en X, corresponderá pronóstico medio en Y; a puntuación alta en X, corresponderá pronóstico alto en Y. Además, siempre y sólo cuando hay variación en la variable predictora X, la habrá en el pronóstico Y'. Todas las personas que no difieran en X, tampoco diferirán en Y'. A su vez, todas las personas que difieran en X, diferirán, también, en Y'.

Por el contrario, $Y - Y'$ no depende, no es función de, no está asociado a, X. De hecho, $r_{x(y-y')} = 0$. A puntuación baja en X, puede corresponder error bajo, medio o alto en Y. A puntuación media en X, puede corresponder error bajo, medio o alto en Y. A puntuación alta en X, puede corresponder error bajo, medio o alto en Y. Como se ve en la figura adjunta, tanto entre las personas con $X = X_1$ (puntuación baja en X), como entre las personas con $X = X_2$ (puntuación media en X), como entre las personas con $X = X_3$ (puntuación alta en X), a unas les corresponden errores grandes (a las que se alejen bastante de la recta de regresión), a otras medios (a las que se alejan moderadamente de la misma), y a otras pequeños (a las que se alejan muy poco de dicha recta).



Por otra parte, hay variación en los errores, cuando no la hay en la variable predictora, X. Es decir, a las personas con una misma puntuación en X, les corresponden, en general, distintos errores en los pronósticos (a unas, pequeños; a otras, medios; y a otras, altos).

Sabemos, además, que

$$\frac{\Sigma (Y - \bar{Y})^2}{n} = \frac{\Sigma (Y' - \bar{Y})^2}{n} + \frac{\Sigma (Y - Y')^2}{n}$$

$$s_y^2 = s_{y'}^2 + s_{y.x}^2$$

En otras palabras, la varianza total de Y (s_y^2) se descompone en dos partes aditivas. Una, $s_{y'}^2$, asociada a, dependiente de, explicada por, la variación de X, ya que Y' estaba asociada a, dependía de, era explicada por, la variación de X. Otra, $s_{y.x}^2$, no asociada a, no dependiente de, no explicada por, la variación de X, ya que $(Y - Y')$ no estaba asociada a, no dependía de, no era explicada por, la variación de X.

Por estas razones, llamaremos a $s_{y'}^2$ varianza asociada, y a $s_{y.x}^2$ la llamaremos varianza no asociada.

Ahora bien, sabemos que $r_{xy}^2 = \frac{s_{y'}^2}{s_y^2}$. Pero, $\frac{s_{y'}^2}{s_y^2} = \frac{s_{y'}^2}{s_{y'}^2 + s_{y.x}^2}$, dentro de este contexto, no es más que la proporción de varianza asociada (cociente entre la varianza asociada y la varianza total de Y). Por tanto, r_{xy}^2 representará esa proporción de varianza asociada. Si, por ejemplo, $r_{xy} = 0{,}70$, diremos que 0,49 es la proporción de varianza asociada y que 0,51 es la proporción de varianza no asociada. De otro modo, diremos que el 49 por 100 de las diferencias individuales en Y está asociado, depende de, es explicado por, es atribuible a, las diferencias individuales en X y que el 51 por 100 restante no está asociado a, no depende de, no es explicado por, no es atribuible a, las diferencias individuales en la variable predictora, X.

La idea de r_{xy}^2 como proporción de varianza asociada puede quedar aclarada así. Sea X un test de aptitud y sea Y el rendimiento a fin de curso. Este rendimiento depende de varios factores. Quien los posee en alto grado, obtiene puntuaciones altas a fin de curso, quien los posee en bajo grado, obtiene puntuaciones bajas a fin de curso. De estos factores, unos son tales que quien los posee en alto grado, obtiene, también, puntuaciones altas en el test de aptitud y quien los posee en bajo grado, obtiene, también, puntuaciones bajas en dicho test. En cambio, existen otros factores tales que quien los posee en alto grado (y obtiene altas puntuaciones a fin de curso), obtiene bien puntuaciones altas, bien medias, bien bajas en el test de aptitud; sucediendo algo parecido con los que los poseen en bajo grado. Por consiguiente, la varianza total del rendimiento escolar (es decir, la diferenciación de los alumnos en buenos y malos a fin de curso) queda descompuesta en dos partes: una debida a los primeros factores (varianza asociada) y otra debida a los segundos (varianza no asociada). La primera es debida a unos factores que hacen variar, diferenciarse sistemáticamente, a los estudiantes en el test de aptitud y en el rendimiento escolar. La segunda es debida a otros factores (profesor particular, salud, motivación, tiempo dedicable al estudio) que hacen diferenciarse sistemáticamente a los estudiantes en el rendimiento escolar, pero no en el test de aptitud. No por poseerlos en alto grado se obtendrán sistemáticamente puntuaciones altas en el test de aptitud, ni por poseerlos en bajo grado se obtendrán necesariamente puntuaciones bajas. Naturalmente, a nosotros nos interesa encontrar tests cuya correlación con el éxito escolar sea muy grande, es decir, que sea muy grande la proporción de varianza del éxito escolar que se encuentre asociada a la variación en los tests predictores.

NOTA 1. Dado que $s_y^2 = s_{y'}^2 + s_{y.x}^2$, necesariamente $s_{y'}^2 \leq s_y^2$. Por otra parte, $r_{xy}^2 = \frac{s_{y'}^2}{s_y^2}$. Consiguientemente, $r_{xy}^2 \leq 1$. De donde, $-1 \leq r_{xy} \leq 1$. Así vemos comprobada de otro modo la propiedad $-1 \leq r_{xy} \leq 1$.

NOTA 2. Nótese que $s_y = \sqrt{s_{y'}^2 + s_{y.x}^2} \neq s_{y'} + s_{y.x}$. Es decir, la desviación típica total de Y no es la suma de dos desviaciones típicas, una asociada y otra no asociada. De aquí que no podamos decir que r_{xy} es la proporción de desviación típica asociada a la variación de X.

NOTA 3. Es claro que la correlación entre puntuaciones obtenidas (y) y puntuaciones pronosticadas (y') indicará la precisión del pronóstico. Éste será tanto más preciso, cuanto mayor sea dicha correlación. Ahora bien, $r_{yy'} = \Sigma yy'/ns_y s_{y'} = \Sigma y\, bx/ns_y|b|s_x = |\Sigma xy|/ns_x s_y = |r_{xy}|$. Por tanto, $|r_{xy}|$ aparece claramente como índice de precisión en el pronóstico.

NOTA 4. La definición $r_{xy}^2 = \frac{s_{y'}^2}{s_y^2} = 1 - \frac{s_{y.x}^2}{s_y^2}$ nos puede ayudar a comprender cómo la variabilidad de las puntuaciones en dos variables correlacionadas influye en el valor del coeficiente de correlación de Pearson entre dichas variables.

Consideremos el grupo siguiente, compuesto de diez observaciones y dividido en dos subgrupos, cada uno de ellos compuesto de cinco observaciones.

Grupo total

	X	Y
Subg. 1.º	4	1
	2	3
	8	5
	5	4
	6	7
Subg. 2.º	18	11
	12	9
	14	7
	16	13
	15	10

subgrupo 1.º

$$s_{y.x}^2 = \frac{12{,}80}{5} = 2{,}56, \quad s_y^2 = \frac{20}{5} = 4$$

$$r_{xy}^2 = 1 - \frac{2{,}56}{4} = 1 - 0{,}64 = 0{,}36$$

subgrupo 2.º

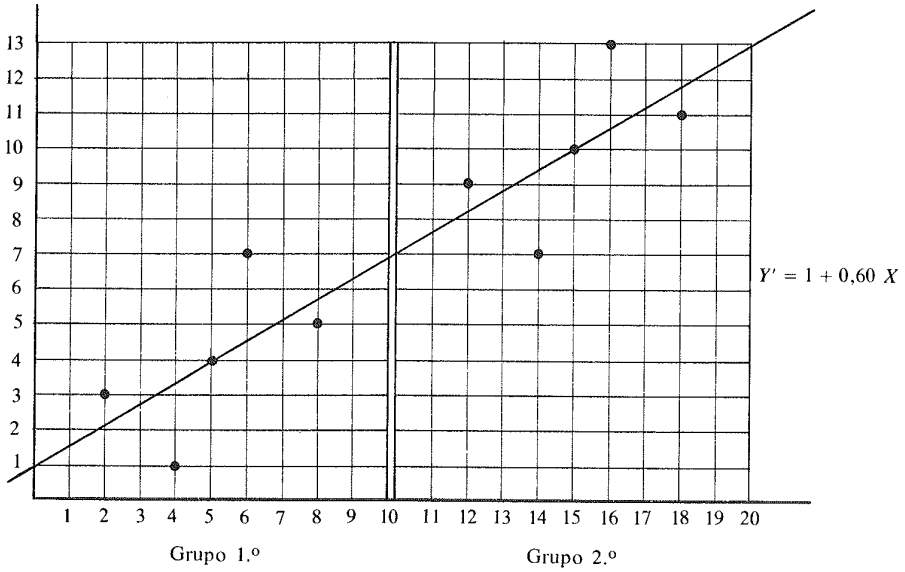
$$s_{y.x}^2 = \frac{12{,}80}{5} = 2{,}56, \quad s_y^2 = \frac{20}{5} = 4$$

$$r_{xy}^2 = 1 - \frac{2{,}56}{4} = 1 - 0{,}64 = 0{,}36$$

$$\underline{\text{grupo total}}$$

$$s^2_{y.x} = \frac{25{,}6}{10} = 2{,}56, \quad s^2_y = \frac{130}{10} = 13$$

$$r^2_{xy} = 1 - \frac{2{,}56}{13} = 1 - 0{,}197 = 0{,}803$$



La recta dibujada es de regresión (de Y sobre X) tanto para cada uno de los dos subgrupos, como para el grupo total. La situación acabada de exponer es artificial, pero bastante de acuerdo con lo que suele suceder en la realidad. El error cuadrático medio cometido en los pronósticos mediante dicha recta, es el mismo (2,56) en cada uno de los dos subgrupos y en el grupo total. En lo que difieren ambos subgrupos es en la variabilidad o dispersión. Mientras que $s^2_y = 4$ en cada uno de los dos subgrupos, $s^2_y = 13$ en el grupo total. Ello hace que $s^2_{y.x}/s^2_y = 2{,}56/4 = 0{,}64$ (para cada subgrupo) sea mayor que $s^2_{y.x}/s^2_y = 2{,}56/13 = 0{,}197$ (para el grupo total). Consiguientemente, $r^2_{xy} = 1 - 0{,}64 = 0{,}36$ (para cada subgrupo) será menor que $r^2_{xy} = 1 - 0{,}197 = 0{,}803$ (para el grupo total).

12.4. Resumen: Definiciones y fórmulas

$s^2_y = \dfrac{\Sigma (Y - \bar{Y})^2}{n}$: error cuadrático medio cometido al atribuir a cada persona, como puntuación, la media $\bar{Y}$. (Bajo otro punto de vista, varianza total de Y.)

$s^2_{y.x} = \dfrac{\Sigma (Y - Y')^2}{n}$: error cuadrático medio que aún cometemos al atribuir a cada persona la puntuación Y'. (Bajo otro punto de vista, varianza de Y no asociada a la variación de X.)

$s^2_{y'} = \dfrac{\Sigma (Y' - \bar{Y})^2}{n}$: error cuadrático medio que dejamos de cometer al atribuir a cada persona la puntuación Y'. (Bajo otro punto de vista, varianza de Y asociada a la variación de X.)

$$s^2_y = s^2_{y.x} + s^2_{y'}$$

$$r^2_{xy} = \frac{s^2_{y'}}{s^2_y} = 1 - \frac{s^2_{y.x}}{s^2_y}$$

EJERCICIOS

12.1. Sean b_{yx} y b_{xy} las pendientes de la recta de regresión de Y sobre X y de X sobre Y, respectivamente. Demostrar que $(b_{yx})(b_{xy}) = r^2_{xy}$.

12.2. Demostrar que el coeficiente de correlación de Pearson (al cuadrado) vale 1 entre las puntuaciones obtenidas en la variable predictora y las puntuaciones pronosticadas en el criterio. En otras palabras, demuestre que $r^2_{xy'} = 1$.

12.3. Demostrar que el coeficiente de correlación de Pearson vale 0 entre las puntuaciones obtenidas en la variable predictora y los errores cometidos en los pronósticos. En otras palabras, demostrar que $r_{x(y-y')} = 0$.

12.4. Demostrar que el coeficiente de correlación de Pearson vale 0 entre las puntuaciones pronosticadas en el criterio y los errores cometidos en los pronósticos. En otras palabras, demostrar que $r_{y'(y-y')} = 0$.

12.5. Demostrar que $Y - Y' = y - y'$.

12.6. Demostrar que $\Sigma (Y - Y')(Y' - \bar{Y}) = 0$.

12.7. Si $y' = (-0{,}25)x$, r_{xy} tiene que ser negativa. Sí ( ). No ( ).

12.8. Si en un grupo es $y' = 0{,}30x$, y en otro grupo es $y' = 0{,}20x$, r_{xy} valdrá necesariamente más en el primer grupo que en el segundo. Sí ( ). No ( ).

12.9. Sabiendo que vale 5,76 el error cuadrático medio cometido al pronosticar las puntuaciones típicas en un criterio, Y, a partir de las puntuaciones típicas en un test predictor, X, calcular la desviación típica de Y, teniendo en cuenta el cuadro adjunto.

x	y'
−1	·
·	4,8
−3	·
1	·
0	·

12.10. Demostrar que $\Sigma Xy' = \Sigma Yx' = \Sigma X'y = \Sigma Y'x = nr_{xy}s_x s_y$.

12.11. Demostrar que $\Sigma Xx' = \Sigma X'x' = \Sigma X'x = nr_{xy}^2 s_x^2$.

12.12. Demostrar que $\Sigma Yy' = \Sigma Y'y' = \Sigma Y'y = nr_{xy}^2 s_y^2$.

12.13. Supongamos:

a) X es un test de inteligencia numérica e Y un examen de aritmética.
b) $s_x = 4$, $s_y = 8$.
c) Valiéndonos de Y' (puntuaciones pronosticadas mediante la recta de regresión de Y sobre X) reducimos en un 56,25 por 100 el error cuadrático medio que habríamos cometido valiéndonos de $\bar{Y}$.

Bajo estas condiciones, poner los valores que faltan en el cuadro siguiente. (Suponemos que las cinco puntuaciones diferenciales en X son valores enteros.)

x	y'	y'^2
−6	·	·
·	−3	·
·	·	·
0	·	·
·	·	9

12.14. Supongamos que la pendiente de la recta de regresión (en puntuaciones directas) de Y sobre X vale 1/3, que $z_x = z_y$ (para toda persona de la muestra),

que $\bar{X} = 4$ y que $s_y = 2$. En este supuesto, poner las puntuaciones directas que faltan en el cuadro adjunto.

X	Y
·	6
·	10
·	·
−5	4
·	8

12.15. Sea X un test de habilidad manual e Y la habilidad manual manifestada en un oficio determinado. Sea 12,96 la parte de la varianza de los operarios en Y que no está asociada a su variación en X. Calcular r_{xy} teniendo en cuenta el cuadro adjunto y sabiendo que $\Sigma xy = 96$. (Suponemos que las cinco puntuaciones diferenciales en X son valores enteros.)

x^2	y'
·	·
36	·
4	−2,4
·	·
·	7,2

12.16. Demostrar que el cuadrado del coeficiente de correlación de Pearson entre las puntuaciones obtenidas en Y y los errores cometidos en los pronósticos, mediante la recta de regresión de Y sobre X, es igual a $1 - r_{xy}^2$. En otras palabras, demostrar que $r_{y(y-y')}^2 = 1 - r_{xy}^2$.

12.17. Comprobar cómo se verifica la anterior propiedad en el ejemplo siguiente:

X	Y
4	7
3	1
1	5
5	9
7	13

12.18. Calcular la pendiente de la recta de regresión de Y sobre X sabiendo que $r_{yy'} = 0{,}6$, $s_x^2 = 4$ y $s_{y.x}^2 = 10{,}24$.

12.19. Demostrar que $\Sigma y^2 = \Sigma (y - y')^2 + \Sigma y'^2$.

12.20. Demostrar que $\Sigma z_y^2 = \Sigma (z_y - z_y')^2 + \Sigma z_y'^2$.

12.21. Sabiendo que para cinco personas la pendiente de la recta de regresión (en puntuaciones directas) de Y sobre X vale 0,2, s_x vale 8 y $s_{y.x}^2/s_{y'}^2$ vale 9/16, calcular el error cuadrático medio que aún seguimos cometiendo al valernos de dicha recta de regresión, en vez de atribuir a cada persona, como puntuación, la media $\bar{Y}$.

12.22. Demostrar que $\Sigma y(y - y') = \Sigma (y - y')^2$.

12.23. Comprobar cómo se verifica la propiedad anterior con los datos del ejercicio 12.17.

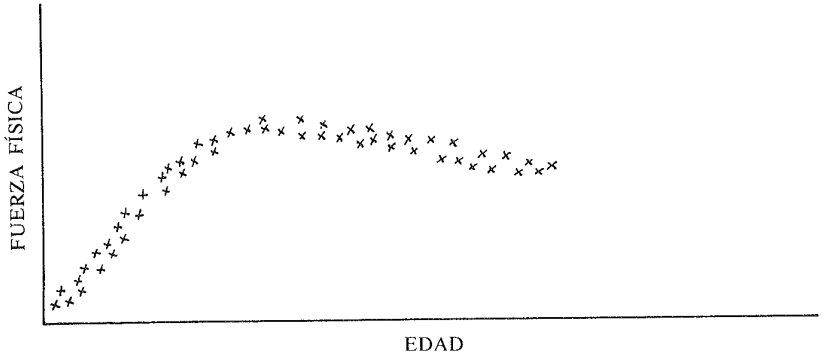
13

Relación (curvilínea) entre dos variables: Razón de correlación

13.1. Introducción

Cuando la relación entre dos variables, X e Y, es curvilínea, no es razonable calcular r_{xy}. Este coeficiente, según ya sabemos, mide la relación *lineal* entre dos variables y es poco apto para detectar relaciones curvilíneas. Es posible obtener $r_{xy} = 0$, o muy bajo, existiendo una alta relación entre X e Y, pero de tipo curvilíneo.

Suele ser curvilínea la relación existente entre bastantes características físicas y la edad, con tal que el período temporal considerado sea suficientemente amplio (en períodos cortos la relación puede ser aproximadamente lineal). Consideremos, por ejemplo, la relación entre la fuerza física y la edad. El niño al nacer carece prácticamente de fuerza física. Al ir creciendo, su fuerza física se va desarrollando progresivamente hasta alcanzar un máximo en cierta etapa de su vida. A partir de este momento, la fuerza física va decreciendo al ir aumentando la edad. En estos casos es bastante razonable calcular la razón de correlación.



13.2. Fundamento y definición (razón de correlación de *Y* sobre *X*).

Comencemos con un caso particular. Sea X la edad e Y una prueba de habilidad mecánica. Supongamos 50 personas distribuidas según la tabla siguiente:

TABLA 13.1

	X EDAD		
	5-14	15-24	25-34
Y HABILIDAD MECÁNICA	10 8 7 6 9 10 7 9 11 7 5 9 $n_1 = 15$ 8 9 5 120	10 8 8 14 12 12 6 8 10 8 10 8 $n_2 = 25$ 10 14 10 12 14 8 8 10 10 12 12 8 8 250	7 7 6 4 6 7 6 5 6 6 $n_3 = 10$ 60
	$\bar{Y}_1 = \frac{120}{15} = 8$	$\bar{Y}_2 = \frac{250}{25} = 10$	$\bar{Y}_3 = \frac{60}{10} = 6$

$$\bar{Y} = \frac{120 + 250 + 60}{15 + 25 + 10} = \frac{430}{50} = 8{,}6$$

Si sólo conocemos la media del grupo total, $\bar{Y} = 8{,}6$, pronosticaremos a cada una de esas 50 personas 8,6 como puntuación en Y. Si conocemos, además, las

medias particulares, $\bar{Y}_1 = 8$, $\bar{Y}_2 = 10$, $\bar{Y}_3 = 6$ de cada uno de los tres grupos, pronosticaremos a cada persona, como puntuación en Y, la media del grupo a que pertenece de acuerdo con su edad. Así, por ejemplo, si sabemos que una persona tiene dieciocho años, le atribuiremos 10 como puntuación en Y, por ser 10 la media de las personas cuyas edades oscilan entre quince y veinticuatro años. Esperamos que, valiéndonos de la media del propio grupo, en vez de valernos de la media del grupo total, mejoraremos nuestros pronósticos en Y.

Pasemos ahora al caso general. Sea una variable X descompuesta en s categorías (en el ejemplo anterior, $s = 3$). Sea n_c el número de personas pertenecientes a la categoría c; es decir, sea n_1 el número de personas pertenecientes a la categoría 1, sea n_2 el número de personas pertenecientes a la categoría 2, . . . , sea n_s el número de personas pertenecientes a la categoría s (en nuestro ejemplo, $n_1 = 15$, $n_2 = 25$, $n_3 = 10$).

TABLA 13.2

	Categoría 1	Categoría 2	. . .	Categoría c	. . .	Categoría s
Persona 1.ª	Y_{11}	Y_{12}	. . .	Y_{1c}	. . .	Y_{1s}
Persona 2.ª	Y_{21}	Y_{22}	. . .	Y_{2c}	. . .	Y_{2s}
Persona 3.ª	Y_{31}	Y_{32}	. . .	Y_{3c}	. . .	Y_{3s}
⋮	⋮	⋮		⋮		⋮
persona i	Y_{i1}	Y_{i2}	. . .	Y_{ic}	. . .	Y_{is}
⋮	⋮	⋮		⋮		⋮
	$Y_{n_1 1}$	$Y_{n_2 2}$	. . .	$Y_{n_c c}$	. . .	$Y_{n_s s}$
	$\bar{Y}_1 = \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1}$	$\bar{Y}_2 = \frac{\sum_{i=1}^{n_2} Y_{i2}}{n_2}$	. . .	$\bar{Y}_c = \frac{\sum_{i=1}^{n_c} Y_{ic}}{n_c}$	. . .	$\bar{Y}_s = \frac{\sum_{i=1}^{n_s} Y_{is}}{n_s}$
	$\bar{Y} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + \cdots + n_c\bar{Y}_c + \cdots + n_s\bar{Y}_s}{n_1 + n_2 + \cdots + n_c + \cdots + n_s}$					

En la tabla 13.2,

Y_{ic}: puntuación directa de la persona i perteneciente a la categoría c.
$\bar{Y}_c$: media en Y de las n_c personas pertenecientes a la categoría c.
$\bar{Y}$: media en Y de las n personas pertenecientes a todas las categorías, es decir, media total.

Bajo estos supuestos, $\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2$ será la suma de errores cuadráticos (que designaremos por E^2) cometida al atribuir a cada persona con puntuación Y_{ic} la media del grupo total, $\bar{Y}$.

$\Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)^2$ será la suma de errores cuadráticos (E^2) cometida al atribuir a cada persona con puntuación Y_{ic} la media, $\bar{Y}_c$, de la categoría c a la que pertenece.

Ahora bien,

$$\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2 = \Sigma\Sigma\,\{(Y_{ic} - \bar{Y}_c) + (\bar{Y}_c - \bar{Y})\}^2 = \Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)^2 + \\ + \Sigma\, n_c(\bar{Y}_c - \bar{Y})^2 \qquad (13.1)$$

pues

$$2\,\Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)(\bar{Y}_c - \bar{Y}) = 2\sum_c\,(\bar{Y}_c - \bar{Y})\sum_i\,(Y_{ic} - \bar{Y}_c) = 0$$

ya que para categoría c, es decir, para cada valor fijo de c,

$$\sum_i\,(Y_{ic} - \bar{Y}_c) = 0$$

de acuerdo con 5.2.3.*a*.

Por consiguiente,

$$\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2 = \Sigma\Sigma\,(Y_{ic} - \bar{Y})^2 - \Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)^2$$

no es más que:

(suma de E^2 cometida atribuyendo a cada persona la media total, $\bar{Y}$) − (suma de E^2 cometida atribuyéndole la media de su propia categoría, $\bar{Y}_c$).

Es decir, no es más que la parte en que ha sido reducida la suma de E^2 primitiva. Consiguientemente,

$$\frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2}$$

representa la proporción en que ha sido reducida la suma de E^2 primitiva (la cometida mediante $\bar{Y}$) al valernos de $\bar{Y}_c$. Si, por ejemplo, este cociente valiera 0,60, quiere decir que la suma de E^2 primitiva ha sido reducida en un 60 por 100. Es decir, que atribuyendo a cada persona la media, $\bar{Y}_c$, de su propia categoría, sólo cometemos un 40 por 100 de la suma de E^2 que cometíamos atribuyéndole la media total, $\bar{Y}$. Pues bien, por definición, llamaremos razón de correlación (de Y sobre X), o η_{yx}, a

$$\eta_{yx} = \sqrt{\frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2}} \qquad (13.2)$$

cuando los datos no están agrupados en intervalos en la variable Y.

$$\eta_{yx} = \sqrt{\frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\, n_j(Y_j - \bar{Y})^2}} \qquad (13.3)$$

cuando los datos están agrupados en intervalos en la variable Y.

En la segunda fórmula, Y_j es el punto medio del intervalo j de la variable Y, n_j es el número de personas dentro del mismo intervalo j. El índice j va de 1 a r, siendo r el número de intervalos.

Observemos el gran paralelismo entre η^2_{yx} y r^2_{yx}. Según la ecuación (12.1),

$$\Sigma\,(Y_i - \bar{Y})^2 = \Sigma\,(Y_i - Y_i')^2 + \Sigma\,(Y_i' - \bar{Y})^2$$

Llamemos Y_{ic} a la puntuación en Y de cada persona i cuya puntuación en X es X_c. Sean n_c las personas con $X = X_c$. A todas ellas las deberemos atribuir el mismo pronóstico, que llamaremos Y_c'. Supuesto esto, la anterior igualdad quedará así:

$$\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2 = \Sigma\Sigma\,(Y_{ic} - Y_c')^2 + \Sigma\, n_c(Y_c' - \bar{Y})^2 \qquad (13.4)$$

Pues bien, en la página anterior veíamos:

$$\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2 = \Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)^2 + \Sigma\, n_c(\bar{Y}_c - \bar{Y})^2$$

La única diferencia está en que en el apartado 12.1 atribuíamos a cada persona, en vez de $\bar{Y}$, la puntuación Y_c' pronosticada mediante la recta de regresión. Aquí atribuimos a cada persona, en vez de $\bar{Y}$, la media $\bar{Y}_c$ de su propia categoría.

Considerando la ecuación (13.4) y la ecuación (12.5) con los retoques acabados de indicar, tendremos:

$$r^2_{yx} = \frac{\Sigma\, n_c(Y_c' - \bar{Y})^2}{\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2} \qquad (13.5)$$

$$\eta^2_{yx} = \frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2} \qquad (13.6)$$

Las expresiones (13.5) y (13.6) representan lo mismo: proporción en que reducimos la suma de errores cuadráticos primitiva (la cometida valiéndonos de $\bar{Y}$) al valernos bien de la puntuación dada por la recta de regresión (13.5), bien de la media de cada una de las categorías (13.6).

13.3. Cálculo

a) *Datos no agrupados en intervalos*

Es una mera aplicación de (13.2).

EJEMPLO 13.1. Calculemos la razón de correlación, de Y sobre X, (de la habilidad mecánica sobre la edad) a partir de la tabla 13.1.

$$\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2 = (10 - 8{,}6)^2 + \cdots + (5 - 8{,}6)^2 + (10 - 8{,}6)^2 + \cdots + (8 - 8{,}6)^2 + (7 - 8{,}6)^2 + \cdots + (6 - 8{,}6)^2 = 51{,}4 + 169 + 75{,}6 = 296$$

$$\Sigma\Sigma\,(Y_{ic} - \bar{Y}_c)^2 = (10 - 8)^2 + \cdots + (5 - 8)^2 + (10 - 10)^2 + \cdots + (8 - 10)^2 + (7 - 6)^2 + \cdots + (6 - 6)^2 = 46 + 120 + 8 = 174$$

$$\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2 = (15)(8 - 8{,}6)^2 + (25)(10 - 8{,}6)^2 + (10)(6 - 8{,}6)^2 = 5{,}4 + 49 + 67{,}6 = 122$$

Por consiguiente,

$$\eta^2_{yx} = \frac{122}{296} = 0{,}412 \quad , \quad \eta_{yx} = 0{,}642$$

Nótese que $296 = 174 + 122$.

b) *Datos agrupados en intervalos*

Es una mera aplicación de (13.3).

EJEMPLO 13.2. Calculemos la razón de correlación, de Y sobre X (de la habilidad mecánica sobre la edad) a partir de la tabla 13.1, agrupando los datos según la tabla 13.3.

TABLA 13.3

Y \ X	5-14	15-24	25-34	n_j	Y_j	n_jY_j	$(Y_j - \bar{Y})$	$(Y_j - \bar{Y})^2$	$n_j(Y_j - \bar{Y})^2$
13-15	0	3	0	3	14	42	5,4	29,16	87,48
10-12	3	12	0	15	11	165	2,4	5,76	86,40
7-9	9	9	3	21	8	168	−0,6	0,36	7,56
4-6	3	1	7	11	5	55	−3,6	12,96	142,56
n_c	15	25	10			430			324,00
$\bar{Y}_c$	8	10,04	5,9						
$(\bar{Y}_c - \bar{Y})$	−0,6	1,44	−2,7						
$(\bar{Y}_c - \bar{Y})^2$	0,36	2,0736	7,29						
$n_c(\bar{Y}_c - \bar{Y})^2$	5,4 + 51,84 + 72,9 = 130,14								

$$\eta^2_{yx} = \frac{130{,}14}{324} = 0{,}402; \quad \eta_{yx} = 0{,}634$$

Nótese cómo $(3)(11 - 8)^2 + (9)(8 - 8)^2 + (3)(5 - 8)^2 + (3)(14 - 10{,}04)^2 + (12)(11 - 10{,}04)^2 + (9)(8 - 10{,}04)^2 + (1)(5 - 10{,}04)^2 + (3)(8 - 5{,}9)^2 + (7)(5 - 5{,}9)^2 = 193{,}86$.

Nótese, consiguientemente, cómo $324 = 193{,}86 + 130{,}14$.
La interpretación de la tabla anterior es la siguiente:

n_j: personas dentro del intervalo j en la variable Y. En los cuatro intervalos de nuestro ejemplo tenemos $3 + 1 + 7 = 11$; $9 + 9 + 3 = 21$; $3 + 12 + 0 = 15$; $0 + 3 + 0 = 3$.

Y_j: punto medio del intervalo j en la variable Y; los cuatro puntos medios son 5, 8, 11, 14.

$\bar{Y}$: media de todas las personas en Y; es decir, $430/50 = 8{,}6$.

$Y_j - \bar{Y}$: puntuaciones diferenciales, $5 - 8{,}6 = -3{,}6$; $8 - 8{,}6 = -0{,}6$; $11 - 8{,}6 = 2{,}4$; $14 - 8{,}6 = 5{,}4$.

$(Y_j - \bar{Y})^2$: $(-3{,}6)^2 = 12{,}96$; $(-0{,}6)^2 = 0{,}36$; $(2{,}4)^2 = 5{,}76$; $(5{,}4)^2 = 29{,}16$.

$n_j(Y_j - \bar{Y})^2$: $(11)(12{,}96) = 142{,}56$; $(21)(0{,}36) = 7{,}56$; $(15)(5{,}76) = 86{,}40$; $(3)(29{,}16) = 87{,}48$.

n_c: personas dentro de la categoría c en la variable X (edad). Dentro de cada una de las tres categorías de nuestro ejemplo tenemos $3 + 9 + 3 = 15$; $3 + 12 + 9 + 1 = 25$; $3 + 7 = 10$.

$\bar{Y}_c$: media en Y de las personas dentro de la categoría c.

$$\frac{(3)(5) + (9)(8) + (3)(11)}{15} = 8$$

$$\frac{(1)(5) + (9)(8) + (12)(11) + (3)(14)}{25} = 10{,}04; \quad \frac{(7)(5) + (3)(8)}{10} = 5{,}9$$

$(\bar{Y}_c - \bar{Y})$: $8 - 8{,}6 = -0{,}6$; $10{,}04 - 8{,}6 = 1{,}44$; $5{,}9 - 8{,}6 = -2{,}7$.

$(\bar{Y}_c - \bar{Y})^2$: $(-0{,}6)^2 = 0{,}36$; $(1{,}44)^2 = 2{,}0736$; $(-2{,}7)^2 = 7{,}29$.

$n_c(\bar{Y}_c - \bar{Y})^2$: $(15)(0{,}36) = 5{,}4$; $(25)(1{,}44) = 51{,}84$; $(10)(7{,}29) = 72{,}9$.

Nótese que la variable X puede ser cuantitativa o no serlo. Si es cuantitativa, sus valores numéricos no aparecen para nada en el cálculo de η_{yx}.

En lugar de utilizar las puntuaciones Y_j, podíamos haber introducido las puntuaciones y'_j, de acuerdo con lo visto en el caso de la media, de la desviación típica y del coeficiente de correlación de Pearson, cuando usábamos el método abreviado. El resultado obtenido con estas nuevas puntuaciones es el mismo que el obtenido con las Y_j.

13.4. Propiedades

a) η^2_{yx} es igual o mayor que cero e igual o menor que 1. Es decir, $0 \leq \eta^2_{yx} \leq 1$

En efecto, según (13.1), $\Sigma n_c(\bar{Y}_c - \bar{Y})^2 \leq \Sigma\Sigma (Y_{ic} - \bar{Y})^2$. Por tanto, teniendo en cuenta (13.2), $\eta^2_{yx} \leq 1$.

Por otra parte, η^2_{yx} es el cociente de dos expresiones no negativas y, por tanto, no negativo. Es decir, $\eta^2_{yx} \geq 0$.

Consideremos ahora los dos casos extremos posibles: $\eta^2_{yx} = 1$, $\eta^2_{yx} = 0$.

En primer lugar, teniendo en cuenta (13.2) y (13.1), podemos escribir:

$$\eta^2_{yx} = \frac{\Sigma\Sigma (Y_{ic} - \bar{Y})^2 - \Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2} = 1 - \frac{\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2} \qquad (13.7)$$

Por tanto, si $\eta^2_{yx} = 1$, teniendo en cuenta (13.7), $\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2 = 0$. Ahora bien, al tratarse de una suma nula compuesta de términos cuadráticos (es decir, no negativos), todos ellos tienen que ser nulos. En otras palabras, $Y_{ic} = \bar{Y}_c$, para toda persona de la muestra. Esto significa que, valiéndonos de $\bar{Y}_c$, no cometemos error alguno en nuestros pronósticos, que reducimos a cero la suma de errores cuadráticos primitiva, la que habríamos cometido si nos hubiéramos valido de $\bar{Y}$. Recíprocamente, si reducimos a cero la suma de errores cuadráticos primitiva, si no cometemos errores en nuestros pronósticos, quiere decir que $Y_{ic} = \bar{Y}_c$ para toda persona de la muestra y, consiguientemente, $\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2 = 0$. Pero si esto sucede, $\eta^2_{yx} = 1$, de acuerdo con (13.7).

En conclusión, a $\eta^2_{yx} = 1$, corresponde reducción total del error primitivo y a reducción total del error primitivo corresponde $\eta^2_{yx} = 1$.

Si $\eta^2_{yx} = 0$, teniendo en cuenta (13.7), $\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2 = \Sigma\Sigma (Y_{ic} - \bar{Y})^2$. Esto significa que nos da lo mismo usar $\bar{Y}_c$ que $\bar{Y}$ en nuestros pronósticos, ya que valiéndonos de $\bar{Y}_c$, seguimos cometiendo la misma suma de errores cuadráticos que cometíamos valiéndonos de $\bar{Y}$. Recíprocamente, si la suma de errores cuadráticos sigue siendo la misma valiéndonos de $\bar{Y}_c$ que valiéndonos de $\bar{Y}$, tendremos $\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2 = \Sigma\Sigma (Y_{ic} - \bar{Y})^2$ y, teniendo en cuenta (13.7), $\eta^2_{yx} = 0$.

En conclusión, a $\eta^2_{yx} = 0$, corresponde reducción nula del error primitivo y a reducción nula del error primitivo, corresponde $\eta^2_{yx} = 0$.

b) Para unos mismos datos, $r^2_{yx} \leq \eta^2_{yx}$.

En efecto, recordando (13.1), (13.4), (13.5) y (13.6),

$$r^2_{yx} = 1 - \frac{\Sigma\Sigma (Y_{ic} - Y'_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2} \qquad \eta^2_{yx} = 1 - \frac{\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2}$$

Ahora bien, $\Sigma\Sigma (Y_{ic} - Y'_c)^2 \geq \Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2$, dado que para cada valor de c $(1, 2, \ldots, s)$ la suma de errores cuadráticos respecto a la media $(\bar{Y}_c)$ es menor que respecto a cualquier otro valor, en particular que Y'_c, salvo el caso $\bar{Y}_c = Y'_c$ en el cual ambas sumas serían iguales. Por consiguiente,

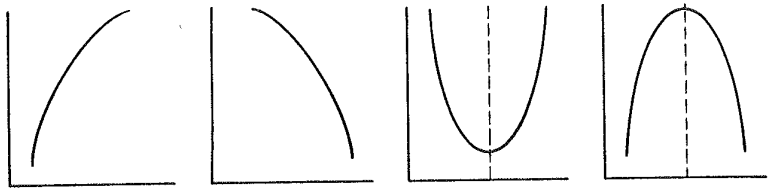
$$\frac{\Sigma\Sigma (Y_{ic} - Y'_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2} \geq \frac{\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2}$$

Es decir,

$$\left(1 - \frac{\Sigma\Sigma (Y_{ic} - Y'_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2}\right) \leq \left(1 - \frac{\Sigma\Sigma (Y_{ic} - \bar{Y}_c)^2}{\Sigma\Sigma (Y_{ic} - \bar{Y})^2}\right) \quad \text{o} \quad r^2_{yx} \leq \eta^2_{yx}$$

La diferencia $\eta^2_{yx} - r^2_{yx}$ nos puede medir el alejamiento mayor o menor de unos datos de la linealidad.

c) La razón de correlación η_{yx} es tomada como positiva; nos mide la aproximación de los puntos a la línea que une las medias $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_s$, es decir, la intensidad de la relación, sin indicarnos el sentido de esta última. El signo de la relación entre X e Y lo tenemos que inferir del diagrama de dispersión. Así, en (A)



la relación será positiva, en (B) negativa, en (C) y en (D) no tiene sentido hablar de signo, a no ser que consideremos por separado la parte izquierda y la parte derecha de cada una de las dos curvas. Para cada una de estas dos mitades podemos atribuir signo tanto en (C) como en (D).

d) η^2_{yx} es función del número de categorías. En general, η^2_{yx} tenderá a ser mayor, al aumentar el número de categorías en la variable X. Sin embargo, si éstas son muchas, la media en Y dentro de cada una de las categorías será poco fiable por estar calculada con pocas personas. En general, tenderá a ser menor, al disminuir el número de categorías en X, pero en este caso la media en Y de cada categoría será más fiable. Esta fiabilidad tiene importancia a nivel inferencial. Algunos autores sugieren que para 100 o más datos sean elegidas entre 6 y 12 categorías.

Si hiciéramos una sola categoría, $\eta^2_{yx} = 0$, tendríamos $\bar{Y}_c = \bar{Y}$. Si todas las personas tuvieran distintas puntuaciones e hiciéramos tantas categorías como personas, $\eta^2_{yx} = 1$, tendríamos en ese caso $Y_{ic} = \bar{Y}_c$.

13.5. Razón de correlación de *X* sobre *Y*

En el coeficiente de correlación de Pearson, $r_{xy} = r_{yx}$. Aquí, en general, $\eta_{xy} \neq \eta_{yx}$. Anteriormente teníamos para calcular η_{yx}:

$\bar{Y}_c$: media en habilidad mecánica de las personas cuya edad estaba dentro de la categoría c.
$\bar{Y}$: media en habilidad mecánica de todas las personas.
η^2_{yx}: proporción en que quedaba reducida la suma de errores cuadráticos por el hecho de pronosticar $\bar{Y}_c$ a cada persona de la categoría c en vez de pronosticarle $\bar{Y}$.

Ahora tendremos, para calcular η_{xy}:

$\bar{X}_c$: media en edad de las personas cuya habilidad mecánica está dentro de la categoría c.
$\bar{X}$: media en edad de todas las personas.
η^2_{xy}: proporción en que queda reducida la suma de errores cuadráticos por el hecho de pronosticar $\bar{X}_c$ a cada persona de la categoría c en vez de pronosticarle $\bar{X}$.

El cálculo de η_{xy} es análogo al expuesto para η_{yx}. La diferencia está en poner como puntuaciones $X(Y)$ las que antes poníamos como puntuaciones $Y(X)$. Es decir, poniendo como intervalos en Y las que antes eran categorías en X, y como categorías en X los que antes eran intervalos en Y. En particular, en nuestro ejemplo, la habilidad mecánica sería la variable X con 4 categorías y la edad sería Y con tres intervalos que antes no habíamos explicitado (porque no era necesario para calcular η_{yx}) y que ahora tendríamos que expresar de modo explícito de acuerdo con los datos en cuestión. Supongamos que éstos fueran 5-14 años, 15-24, 25-34. Pues bien, calcularíamos en primer lugar la edad media de todas las personas. Después la edad media de las personas con habilidad mecánica (4-6), la de las personas con habilidad mecánica (7-9), . . . , la de las personas con habilidad mecánica (13-15). Con este esquema previo es ya fácil calcular η_{xy}.

13.6. Interpretación de η^2_{yx} o de η^2_{xy}

Aquí vale lo dicho respecto a la interpretación de r_{xy}. La valoración de una η^2_{yx} (o η^2_{xy}) determinada como alta, media o baja, sólo puede ser hecha en función de las variables en cuestión. Para un par de variables, η^2_{yx} (o η^2_{xy}), igual a 0,35, puede significar una gran relación y para otro par de variables distintas puede significar una relación baja. El único juicio valorativo razonable es compararla con las η^2_{yx} (o η^2_{xy}) obtenidas por otros investigadores entre las mismas o semejantes variables. Pueden ser engañosas las tablas en las que se clasifican las η^2_{yx} (o η^2_{xy}) como bajas, medias y altas en sentido absoluto.

13.7. Resumen: Definición y fórmulas

Razón de correlación (al cuadrado): proporción en que reducimos la suma de errores cuadráticos cometida atribuyendo a cada persona la puntuación del grupo total, al atribuirle la media de su propio grupo.

$$\eta^2_{yx} = \frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\Sigma\,(Y_{ic} - \bar{Y})^2} \qquad \text{(para datos no agrupados en intervalos)}$$

$$\eta^2_{yx} = \frac{\Sigma\, n_c(\bar{Y}_c - \bar{Y})^2}{\Sigma\, n_j(Y_j - \bar{Y})^2} \qquad \text{(para datos agrupados en intervalos)}$$

EJERCICIOS

13.1. Calcular la razón de correlación η_{yx}, entre las variables X e Y, a partir de los siguientes datos *no agrupados* en intervalos (en la variable Y).

Los valores numéricos dentro de las tablas siguientes representan puntuaciones obtenidas en la variable Y.

a)

Y \ X			
	2	6	3
	1	8	7
	3	7	5
	2		

b)

Y \ X			
	1	3	2
	2	5	4
	3	4	1
	2	5	5
	1	3	3
	2		3
	3		
	2		
	2		

c)

Y \ X				
	1	5	7	1
	2	6	6	2
	4	4	8	4
	1	6	7	1
	2	4	7	
	1		8	
	3		6	
			9	
			5	

13.2. Calcular la razón de correlación η_{yx}, entre las variables X e Y, a partir de los siguientes datos *agrupados* en intervalos (en la variable Y).

Los valores numéricos dentro de las tablas representan, ahora, frecuencias.

a)

Y \ X				
10-12	0	1	0	
7-9	0	5	0	
4-6	4	2	3	
1-3	4	0	1	

b)

Y \ X			
8-10	1	0	2
5-7	1	1	2
2-4	0	3	0

c)

Y \ X	8	1	6
8-12			
3-7	2	4	4

d)

Y \ X				
9-11	4	0	0	2
6-8	7	1	0	8
3-5	4	8	1	0
0-2	0	11	4	0

13.3. Demostrar que η^2_{xy} (razón de correlación, al cuadrado, de Y sobre X) es la misma que η^2_{wv} (razón de correlación, al cuadrado de $V = AX + B$ sobre $W = CY + D$), donde A, B, C, D son cuatro constantes arbitrarias, con A y C distintas de cero.

Nótese que con este ejercicio queda legitimado el método abreviado.

13.4. Sabiendo que $\eta^2_{yx} = 0$, completar la tabla adjunta, donde X y x representan puntuaciones directas y diferenciales, respectivamente, en una variable predictora X e Y' representa puntuaciones directas pronosticadas mediante la recta de regresión de Y sobre X. Decir cuál es la pendiente de dicha recta.

X	x	Y'
.	.	.
.	2	.
4	0	5
5	.	.

14

Relación entre variables ordinales

14.1. Idea previa

Según sabemos, una variable es llamada ordinal cuando a lo largo de ella únicamente podemos ordenar un conjunto de objetos, es decir, solamente podemos decir cuál es el primero, cuál es el segundo, . . . , cuál es el último, pero no podemos atribuirles auténticos números que nos permitan establecer qué distancias existen entre el primero y el segundo, entre el segundo y el tercero, etc. Pues bien, en este capítulo vamos a proponer algunos índices de correlación entre dos variables de tipo ordinal.

14.2. Coeficiente de correlación de Spearman, r_s

14.2.1. Fundamento y fórmula

Comencemos considerando algunos ejemplos a los que es aplicable dicho coeficiente.

a) Dos directores de orquesta ordenan 15 canciones según su valor artístico.

b) Un capataz ordena a unos operarios según su puntualidad y su responsabilidad.

c) Conociendo las puntuaciones de un grupo de niños en aritmética y en geometría, consideramos únicamente sus posiciones o valores ordinales en ambas asignaturas.

d) Conociendo las puntuaciones de los alumnos en aritmética y sus posiciones o valores ordinales en laboriosidad, convertimos las primeras puntuaciones en valores ordinales, teniendo en cuenta únicamente los valores ordinales en aritmética y en laboriosidad.

En todos estos ejemplos nos encontramos, en definitiva, con dos sucesiones de valores ordinales. Pues bien, el coeficiente de correlación de Spearman no es

más que el coeficiente de correlación de Pearson entre estas dos sucesiones de valores ordinales. Para que tenga sentido la aplicación de Pearson, consideraremos el valor ordinal 1.º como puntuación 1, el valor ordinal 2.º como puntuación 2, etc.

Suponiendo que no hay empates (o sea, que dos o más objetos no ocupan una misma posición, no obtienen un mismo valor ordinal), tendremos tantos valores ordinales, en cada una de las dos variables, como objetos. En otras palabras, tendremos dos columnas de números, cada una de las cuales constará de los n primeros números naturales (1, 2, . . . , n). En conclusión, r_s no es más que r_{xy} entre los n primeros números naturales (dados en un cierto orden) y esos mismos números naturales (dados en el mismo o en distinto orden).

EJEMPLO 14.1. Cinco personas han quedado ordenadas según su laboriosidad (X) y su responsabilidad (Y) de acuerdo con la tabla siguiente:

	X	Y
Antonio	2.º	3.º
Carlos	5.º	4.º
Pedro	1.º	2.º
Juan	3.º	1.º
Luis	4.º	5.º

El coeficiente de correlación de Spearman, r_s, no es más que r_{xy} entre los cinco primeros números naturales, dados en el orden 2, 5, 1, 3, 4 y esos mismos números naturales, dados en el orden 3, 4, 2, 1, 5.

Se demuestra (véase 14.6. Apéndice) que la fórmula de r_{xy} aplicada a estas dos series de los n primeros números naturales (es decir, r_s) viene expresada del modo siguiente:

$$r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2 - 1)} \qquad (14.1)$$

donde d_i es la diferencia entre el valor ordinal en X y el valor ordinal en Y del objeto i.

Para el caso de empates hay otra fórmula distinta de (14.1). Sin embargo, con frecuencia (sobre todo, si el número de empates es reducido) suele ser aplicada (14.1), atribuyendo como valor ordinal a cada objeto empatado con otros la media aritmética de los valores ordinales que hubieran ocupado esos objetos en caso de no haber estado empatados. Por ejemplo, si un objeto ocupa el lugar primero, cuatro quedan empatados detrás de él y otros tres quedan empatados detrás de los anteriores, los valores ordinales de estos ocho objetos serían: 1 para el primero, $(2 + 3 + 4 + 5)/4 = 3{,}5$ para los cuatro siguientes, $(6 + 7 + 8)/3 = 7$ para los tres últimos.

Nótese que en el caso de empates obtendremos, en general, distinto resultado aplicando (14.1) a los valores ordinales, conseguidos según lo acabado de indicar, que aplicando r_{xy} a esos mismos valores. Si el número de empates es grande, conviene usar otra fórmula distinta de (14.1). (Véase Kendall, 1970.)

Téngase en cuenta que r_s más que la relación entre X e Y lo que nos mide es la relación entre el orden de los objetos en una variable y su orden en la otra.

14.2.2. Cálculo

Es una mera aplicación de (14.1).

EJEMPLO 14.2. La tabla adjunta nos muestra el orden de preferencia, según el cual, por término medio, un grupo de niños y un grupo de niñas entre diez y once años colocaron diversos objetos ofrecidos como premio. (Bisset y Rieber, 1966.)

Objetos	Niños X	Niñas Y	d_i	d_i^2
Joyas	3	1	2	4
Monedas de poco valor	2	2	0	0
Coches	1	4	−3	9
Cartas pequeñas	5	3	2	4
Dijes	4	5	−1	1
Canicas	7	6	1	1
Arandelas	6	8	−2	4
Clips	8	7	1	1
			0	24

$$r_s = 1 - \frac{(6)(24)}{(8)(8^2 - 1)} = 1 - \frac{144}{504} = 1 - 0{,}2857 = 0{,}7143$$

Es claro que $\Sigma d_i = 0$, pues no es más que la diferencia entre la suma de los n primeros números naturales y la suma de esos mismos números.

EJEMPLO 14.3. En el primer cuadro de la página siguiente tenemos las puntuaciones en nivel de lectura (X) y la conducta en clase (Y) de diez niños de clase media, con problemas en ambas variables, sometidos a un programa de recuperación cuyo fin era manifestar la eficacia de las técnicas de modificación de conducta en la superación de dichos problemas (Wadsworth, 1971).

Es indiferente atribuir en ambas variables el valor 1 a la persona con máxima puntuación, el 2 a la inmediatamente inferior, etc., que atribuir en ambas variables el valor 1 a la persona con mínima puntuación, el 2 a la inmediatamente superior, etc. El valor de r_s será el mismo en uno y otro caso.

Nivel lectura X	Conducta en clase Y	Órdenes X	Órdenes Y	d_i	d_i^2
2,7	40	4	1	3	9
2,2	14	8	9	−1	1
2,3	18	7	7	0	0
2,6	20	5	5	0	0
3,1	22	3	4	−1	1
3,4	36	2	3	−1	1
1,9	17	10	8	2	4
2,1	13	9	10	−1	1
2,4	39	6	2	4	16
3,9	19	1	6	−5	25
				0	58

$$r_s = 1 - \frac{(6)(58)}{(10)(10^2 - 1)} = 1 - 0{,}35 = 0{,}65$$

Distíngase claramente entre r_{xy} aplicado a las auténticas puntuaciones (cuando éstas nos sean dadas), y r_{xy} aplicado a los órdenes originados a partir de dichas puntuaciones (es decir, r_s). Ambos valores, en general, serán distintos.

EJEMPLO 14.4. Pavlov (1929) hacía sonar el tic-tac de un metrónomo en la habitación donde se hallaba un perro y, a continuación, le presentaba comida. Tras varios ensayos consiguió que la puesta en marcha del metrónomo provocase rápidamente la secreción de abundante saliva, como respuesta condicionada. Posteriormente procedió a la extinción de dicha respuesta haciendo sonar el metrónomo, sin presentar comida, con lo cual el perro, al oírlo, segregaba aún saliva,

Ensayo	X (segundos)	Y (gotas)	Órdenes X	Órdenes Y	d_i	d_i^2
1.º	3	10	7	1	6	36
2.º	7	7	3	3,5	−0,5	0,25
3.º	5	8	4,5	2	2,5	6,25
4.º	4	5	6	5	1	1
5.º	5	7	4,5	3,5	1	1
6.º	9	4	2	6	−4	16
7.º	13	3	1	7	−6	36
						96,5

$$r_s = 1 - \frac{(6)(96{,}5)}{(7)(7^2 - 1)} = 1 - \frac{579}{336} = 1 - 1{,}72 = -0{,}72$$

pero ésta era cada vez más escasa y tardaba más en aparecer (mayor latencia). Los datos sobre la latencia (X) y las gotas de saliva segregadas (Y) correspondientes a siete ensayos de extinción, se hallan en la tabla anterior. Calculemos r_s entre X e Y. Tendremos que transformar los valores cuantitativos en valores ordinales.

La relación es negativa, indicando que a mayor latencia en la segregación de saliva, menor número de gotas segregadas.

14.2.3. Propiedades

a) El coeficiente de correlación de Spearman no puede valer menos que −1 ni más que 1. Es decir,

$$-1 \leq r_s \leq 1$$

En efecto, basta con advertir que r_s no es más que un caso particular de r_{xy} (a saber, r_{xy} aplicado a unos valores ordinales considerados como puntuaciones).

Esta propiedad puede ser comprobada, también, a partir de la fórmula (14.1). Es claro que los dos casos extremos de correlación son los siguientes:

Cada objeto ocupa el mismo lugar ordinal en ambas variables (el que es primero en X, es primero en Y; el que es segundo en X, lo es, también, en Y; . . . ; el que es último en X, lo es, también, en Y).

Cada objeto ocupa un lugar opuesto en ambas variables (el que es primero en X, es último en Y; el que es segundo en X, es penúltimo en Y; . . . ; el que es último en X, es primero en Y).

Como ejercicio, verifique el lector cómo, efectivamente, en el primer caso $r_s = 1$ y cómo en el segundo $r_s = -1$.

b) Su cálculo es más sencillo que el de r_{xy} para unos mismos datos, si éstos no son muy numerosos y son escasos los empates.

14.3. Coeficiente de correlación de Kendall, τ

14.3.1. Fundamento y definición

Se diferencia de r_s en que no se funda en r_{xy}. Kendall (1970) considera el orden de n objetos en una variable y su orden en otra e intenta medir «el grado de correspondencia entre estos dos órdenes» (pág. 3).

Supongamos n personas y dos variables X e Y. Elijamos dos personas, A y B. Si A es superior a B en X e inferior a B en Y o inferior a B en X y superior a B en Y, diremos que se da una inversión. Si, por el contrario, A es superior a B en X y superior a B en Y o inferior a B en X e inferior a B en Y, diremos que no se da inversión alguna o que se da una no-inversión. Hagamos la misma comparación

con todos los pares posibles, es decir, con los $n(n-1)/2$, ya que éste es el número de pares que se pueden formar con n elementos de manera que cada par difiera de los restantes en uno, al menos, de sus elementos. Llamemos P al número de no-inversiones y Q al de inversiones. Bajo estas premisas, por definición:

$$\tau = \frac{P-Q}{P+Q} = \frac{P-Q}{\frac{n(n-1)}{2}} \tag{14.2}$$

14.3.2. Cálculo

Es una mera aplicación de (14.2).

EJEMPLO 14.5. Bingham, Moore y Gustad (1959) hicieron que 22 personas que solicitaban un empleo fueran entrevistadas por distintos jueces, cada uno de los cuales debía ordenarlas según la presumible aptitud de cada una para el empleo en cuestión. Por sencillez en los cálculos hemos elegido 5 de las 22. En la tabla adjunta aparece el orden relativo de estas 5 personas propuesto por dos de los jueces.

Personas	Juez 1	Juez 2
	Órdenes	
	X	Y
A	4	4
B	1	3
C	2	2
D	3	1
E	5	5

Hagamos las $(5)(5-1)/2 = 10$ comparaciones posibles: AB, AC, AD, AE, BC, BD, BE, CD, CE, DE. Los pares BC, BD y CD nos ofrecen inversión. Los pares restantes nos ofrecen no-inversión. Consiguientemente, $P = 7$, $Q = 3$. Por tanto,

$$\frac{7-3}{7+3} = \frac{4}{10} = 0{,}40$$

En la práctica, la manera más sencilla de calcular el número de inversiones es la siguiente. Las personas son intercambiadas entre sí de manera que en una de las dos variables queden colocadas de superior a inferior (o de inferior a superior). Supongamos que están colocadas en X de superior a inferior. Es decir, en la tabla está situada en primer lugar la persona que es superior a todas en X,

en segundo lugar la que supera a todas menos a la anterior, etc. Siempre que en la tabla una persona tenga por debajo de sí alguna persona superior a ella en Y, tendremos inversión, pues todas las situadas debajo de ella en la tabla son inferiores en X. En el ejemplo acabado de exponer intercambiemos las cinco personas de modo que en primer lugar se encuentre la que es superior en X a las cuatro restantes, en segundo lugar la que es superior en X a las tres restantes, etc. Según este criterio, la tabla anterior quedará reorganizada así:

	Órdenes	
	X	Y
B	1	3
C	2	2
D	3	1
A	4	4
E	5	5

Comparemos B con las cuatro restantes, fijándonos en la variable Y:

$$BC\,(3-2) : I, \quad BD\,(3-1) : I, \quad BA\,(3-4) : NI, \quad BE\,(3-5) : NI$$

Comparemos C con las tres restantes, fijándonos en la variable Y:

$$CD\,(2-1) : I, \quad CA\,(2-4) : NI, \quad CE\,(2-5) : NI$$

Comparemos D con las dos restantes, fijándonos en la variable Y:

$$DA\,(1-4) : NI, \quad DE\,(1-5) : NI$$

Comparemos A con la única restante, fijándonos en la variable Y:

$$AE\,(4-5) : NI$$

En conclusión, tenemos $2+2+2+1 = 7$ no-inversiones (NI) y $2+1+0+0 = 3$ inversiones (I).

14.3.3. Propiedades

a) El coeficiente de correlación de Kendall no puede valer menos que -1 ni más que 1. Es decir, $-1 \leq \tau \leq 1$.

En efecto, los dos casos extremos posibles son o que no haya inversión alguna, o que todos los pares sean inversos. En el primer caso $Q = 0$ y, por tanto, $\tau = P/P = 1$. En el segundo caso, $P = 0$ y, por tanto, $\tau = -Q/Q = -1$.

b) A partir de τ pueden ser calculados coeficientes de correlación parcial, semejantes a los que calcularemos más adelante a partir de r_{xy}.

c) Según Kendall (1970, pág. 12), «en la práctica encontramos frecuentemente que, cuando ninguno de los dos coeficientes se acercan a la unidad, r_s es aproximadamente un 50 por 100 mayor que τ en valor absoluto, pero esta regla no es invariable». Es decir, $r_s \simeq (3/2)\tau$. Conviene, no obstante, advertir que «esta regla no es invariable».

NOTA. Para el caso de empates existe una fórmula especial que no proponemos porque sólo aplicaremos el coeficiente de correlación de Kendall en situaciones en las que no aparezcan empates. Para un estudio más detallado de problemas relacionados con este coeficiente, véase Kendall (1970).

14.4. Coeficiente de correlación de Goodman y Kruskal

14.4.1. Introducción

Cuando la muestra consta de muchas observaciones y son muy pocos los valores ordinales alcanzables por ellas, será muy grande el número de empates. En este caso es recomendable la gamma de Goodman y Kruskal.

EJEMPLO 14.6. Supongamos dos variables: «nivel social» (X) y «nivel económico» (Y). Consideremos divididas ambas variables en tres categorías: «nivel bajo» (B), «nivel medio» (M), «nivel alto» (A). Esto significa que tanto en X como en Y sólo son posibles tres valores ordinales distintos. Supongamos que 45 personas se encuentran distribuidas, según la tabla 14.1.

TABLA 14.1

Y \ X	B	M	A	
A	1 ^{g}	4 ^{h}	8 ^{i}	13
M	6 ^{d}	9 ^{e}	5 ^{f}	20
B	7 ^{a}	3 ^{b}	2 ^{c}	12
	14	16	15	45

Con estas 45 personas podemos formar (44)(45)/2 = 990 pares que difieran en uno, al menos, de sus elementos. Diremos que un par es «semejante» o «no inverso» si la primera persona es superior a la segunda tanto en X como en Y o si es inferior a la segunda tanto en X como en Y. Diremos que un par es «desemejante» o «inverso» si la primera persona es superior a la segunda en X e inferior a ella en Y o si es inferior a la segunda en X y superior a ella en Y. Diremos que un par está «empatado» o es un «empate» si la primera persona es igual que la segunda bien sólo en X, bien sólo en Y, bien simultáneamente en X y en Y.

Veamos ahora cuántos de los 990 pares son «semejantes». Para ello consideremos las tablas siguientes, todas ellas extraídas de la tabla 14.1.

TABLA 14.2

Y \ X	B	M	A
A		4 ^{h}	8 ^{i}
M		9 ^{e}	5 ^{f}
B	7 ^{a}		

TABLA 14.3

Y \ X	B	M	A
A			8 ^{i}
M			5 ^{f}
B		3 ^{b}	

TABLA 14.4

Y \ X	B	M	A
A		4 ^{h}	8 ^{i}
M	6 ^{d}		
B			

TABLA 14.5

Y \ X	B	M	A
A			8 ^{i}
M		9 ^{e}	
B			

En la tabla 14.2 son «semejantes» todos los pares en los que una de las dos personas del par pertenezca a la casilla «a» y la otra a una cualquiera de las cuatro casillas «e», «f», «h», «i». En efecto, consideremos las casillas «a» y «f», por ejemplo. Toda persona P_a, perteneciente a «a», es «baja» en X y «baja» en Y. Toda persona P_f, perteneciente a «f» es «alta» en X y «media» en Y. Es decir, P_a es inferior a P_f tanto en X como en Y. Por análogo razonamiento, toda persona P_a es inferior a toda P_e, a toda P_h y a toda P_i tanto en X como en Y.

En la tabla 14.3 son «semejantes» todos los pares en los que una de las dos personas del par pertenezca a la casilla «b» y la otra a una cualquiera de las dos casillas «f» o «i». En efecto, toda persona P_b es inferior a toda persona P_f y a toda persona P_i tanto en X como en Y.

En la tabla 14.4 son «semejantes» todos los pares en los que una de las dos personas del par pertenezca a la casilla «d» y la otra a una cualquiera de las dos casillas «h» o «i». En efecto, toda persona P_d es inferior a toda persona P_h y P_i tanto en X como en Y.

En la tabla 14.5 son «semejantes» todos los pares en los que una de las dos personas pertenezca a la casilla «e» y la otra a la casilla «i». En efecto, toda persona P_e es inferior a toda persona P_i tanto en X como en Y.

Por consiguiente, el número de pares «semejantes» será

$$\begin{aligned} (7)(9+5+4+8) &= 182 \quad \text{(tabla 14.2)} \\ (3)(5+8) &= 39 \quad \text{(tabla 14.3)} \\ (6)(4+8) &= 72 \quad \text{(tabla 14.4)} \\ (9)(8) &= 72 \quad \text{(tabla 14.5)} \\ & \overline{365} \end{aligned}$$

De modo análogo, consideremos ahora las tablas siguientes, todas ellas extraídas de la tabla 14.1.

TABLA 14.6

		X		
		B	M	A
Y	A	1^g	4^h	
	M	6^d	9^e	
	B			2^c

TABLA 14.7

		X		
		B	M	A
Y	A	1^g		
	M	6^d		
	B		3^b	

TABLA 14.8

		X		
		B	M	A
Y	A	1^g	4^h	
	M			5^f
	B			

TABLA 14.9

		X		
		B	M	A
Y	A	1^g		
	M		9^e	
	B			

En la tabla 14.6 son «desemejantes» o «inversos» todos los pares en los que una de las dos personas del par pertenezca a la casilla «c» y la otra a una cualquiera de las cuatro casillas «d», «e», «g», «h». En efecto, consideremos las casillas «c» y «h», por ejemplo. Toda persona P_c, perteneciente a la casilla «c», es «alta» en X y «baja» en Y. Toda persona P_h, perteneciente a la casilla «h», es «media» en X y «alta» en Y. Es decir, P_c es superior en X a toda persona P_h y es inferior a ella en Y. Por análogo razonamiento, toda persona P_c es superior en X a toda persona P_d, P_e y P_g, y es inferior en Y a ellas.

En la tabla 14.7 son «desemejantes» o «inversos» todos los pares en los que una de las dos personas del par pertenezca a la casilla «b» y la otra a una cualquiera de las dos casillas «d» o «g». En efecto, toda persona P_b es superior en X a toda persona P_d y P_g, y es inferior en Y a ellas.

En la tabla 14.8 son «desemejantes» o «inversos» todos los pares en los que una de las dos personas del par pertenezca a la casilla «f» y la otra a una cualquiera de las dos casillas «g» o «h». En efecto, toda persona P_f es superior en X a toda persona P_g y P_h, y es inferior en Y a ellas.

En la tabla 14.9 son «desemejantes» o «inversos» todos los pares en los que una de las dos personas del par pertenezca a la casilla «e» y la otra a la casilla «g». En efecto, toda persona P_e es superior en X a toda persona P_g y es inferior en Y a la misma.

Por consiguiente, el número de pares «desemejantes» o «inversos» será:

$$\begin{aligned} (2)(6 + 9 + 1 + 4) &= 40 \quad \text{(tabla 14.6)} \\ (3)(6 + 1) &= 21 \quad \text{(tabla 14.7)} \\ (5)(1 + 4) &= 25 \quad \text{(tabla 14.8)} \\ (9)(1) &= 9 \quad \text{(tabla 14.9)} \\ &\overline{95} \end{aligned}$$

Veamos, finalmente, cuántos son los pares «empatados»:

a) Las dos personas del par pertenecen a un mismo nivel en X:

$$(7)(6 + 1) + (6)(1) + (3)(9 + 4) + (9)(4) + (2)(5 + 8) + (5)(8) = 196$$

b) Las dos personas del par pertenecen a un mismo nivel en Y:

$$(7)(3 + 2) + (3)(2) + (6)(9 + 5) + (9)(5) + (1)(4 + 8) + (4)(8) = 214$$

c) Las dos personas del par pertenecen a un mismo nivel en X y en Y. Se trata de todas combinaciones binarias formadas con las personas de cada una de las nueve casillas:

$$\binom{7}{2} + \binom{3}{2} + \binom{2}{2} + \binom{6}{2} + \binom{9}{2} + \binom{5}{2} + \binom{1}{2} + \binom{4}{2} + \binom{8}{2} =$$

$$21 + 3 + 1 + 15 + 36 + 10 + 0 + 6 + 28 = 120$$

Por tanto,

$$\begin{aligned} &\text{pares «semejantes»} && = 365 \\ &\text{pares «desemejantes» o «inversos»} && = 95 \\ &\text{pares «empatados»} \begin{cases} \text{(sólo en } X\text{)} \\ \text{(sólo en } Y\text{)} \\ \text{(en } X \text{ y en } Y\text{)} \end{cases} && \begin{matrix} = 196 \\ = 214 \\ = 120 \end{matrix} \\ & && \overline{990} \end{aligned}$$

Como era de esperar, la suma total (990) equivale al número de pares posibles distintos que podíamos formar a base de la tabla 14.1.

En conclusión, obtendremos los pares «semejantes» haciendo que la segunda persona de cada par pertenezca a una casilla cualquiera situada por encima y a la derecha de la casilla a la que pertenece la primera persona. Obtendremos los pares «desemejantes» o «inversos» haciendo que la segunda persona de cada par pertenezca a una casilla situada por encima y a la izquierda de la casilla a la que pertenece la primera persona.

Por supuesto, esto es válido si las categorías o niveles se encuentran ordenados de acuerdo con la tabla 14.1, es decir, con el nivel B en X a la izquierda de la tabla y con el nivel B en Y en la parte baja de la misma.

14.4.2. Definición

Llamemos n_s, n_d y n_e al número de pares «semejantes» o «no inversos», «desemejantes» o «inversos» y «empatados», con $n_s + n_d + n_e = n$. Supuesto esto, $n_s/(n - n_e) = n_s/(n_s + n_d)$ será la proporción de pares «semejantes» dentro de

los pares no «empatados». A su vez, $n_d/(n - n_e) = n_d/(n_s + n_d)$ será la proporción de pares «desemejantes» dentro de los pares no «empatados». Pues bien, por definición,

$$\gamma = \frac{n_s}{n_s + n_d} - \frac{n_d}{n_s + n_d} = \frac{n_s - n_d}{n_s + n_d}$$

14.4.3. Cálculo

Basta con aplicar la fórmula anterior al número de pares «semejantes» y al de pares «desemejantes» o «inversos», obtenidos de acuerdo con las indicaciones expuestas en 14.4.1. Así, con los datos del ejemplo 14.6 (tabla 14.1) hemos obtenido

$$n_s = 365 \quad , \quad n_d = 95$$

Por tanto,

$$\gamma = \frac{365 - 95}{365 + 95} = \frac{270}{460} = 0{,}587$$

EJEMPLO 14.7. Amón (1969) estudió la relación entre el prejuicio antiprotestante y la religiosidad utilitaria (es decir, usada como instrumento para conseguir beneficios bien materiales, bien espirituales) en un grupo de 223 estudiantes varones de edades entre quince y veinte años. Los resultados obtenidos fueron los siguientes:

		Prejuicio a.p.		
		Bajo	Alto	
Rel. util.	Alta	73	69	142
	Baja	64	17	81
		137	86	223

$$\gamma = \frac{(64)(69) - (17)(73)}{(64)(69) + (17)(73)} = \frac{4.416 - 1.241}{4.416 + 1.241} = \frac{3.175}{5.657} = 0{,}56$$

Una $\gamma = 0{,}56$ es bastante alta, si nos atenemos a los resultados obtenidos ordinariamente con variables de este tipo.

14.4.4. Propiedades

a) Si la proporción de observaciones dentro de cada casilla se mantiene constante, la gamma de Goodman y Kruskal se mantiene, también, constante, sea cual sea el tamaño de la muestra.

b) La gamma de Goodman y Kruskal es igual o mayor que -1 e igual o menor que 1. En efecto, si todos los pares «no-empatados» son semejantes o no-inversos, $\gamma = \frac{(n_s - 0)}{(n_s - 0)} = 1$. Si todos los pares «no-empatados» son desemejantes o inversos, $\gamma = \frac{(0 - n_d)}{(0 + n_d)} = -1$. En cualquier caso, $(n_s - n_d) < (n_s + n_d)$, ya que n_s y n_d son números esencialmente no-negativos y, por tanto, $(n_s - n_d)/(n_s + n_d)$ será (en valor absoluto) menor que 1. Su signo dependerá de que n_s sea mayor o menor que n_d. Para $n_s = n_d$, $\gamma = 0$.

14.5. Interpretación de los coeficientes de correlación ordinal

Ver los conceptos expuestos en el apartado 10.5.

14.6. Apéndice: Deducción del coeficiente de correlación de Spearman

Según sabemos, el coeficiente de correlación de Spearman no es más que el coeficiente de correlación de Pearson aplicado a dos series de valores ordinales, cada una de ellas compuesta por los n primeros números naturales.

También sabemos que el coeficiente de correlación de Pearson, r_{xy}, viene dado por

$$r_{xy} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - \bar{X})^2}\sqrt{\Sigma (Y - \bar{Y})^2}} \qquad (1)$$

Por otra parte,

$$\Sigma (X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n} \qquad (2)$$

$$\Sigma (Y - \bar{Y})^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n} \qquad (3)$$

Puesto que los valores de X (y de Y) son 1, 2, 3, ..., n, calcular ΣX (ΣY) y ΣX^2 (ΣY^2) equivale a calcular la suma de los n primeros números naturales y la suma de sus cuadrados. Ahora bien, se demuestra que:

$$1 + 2 + 3 + \cdots + n = n(n + 1)/2$$
$$1^2 + 2^2 + 3^2 + \cdots + n^2 = n(n + 1)(2n + 1)/6$$

Por tanto,

$$\Sigma (X - \bar{X})^2 = \Sigma X^2 - (\Sigma X)^2/n = \frac{n(n + 1)(2n + 1)}{6} - \frac{n^2(n + 1)^2}{4n} =$$
$$= \frac{4n^3 + 2n^2 + 4n^2 + 2n - 3n^3 - 6n^2 - 3n}{12} =$$
$$= \frac{n^3 - n}{12} = \frac{n(n^2 - 1)}{12} \qquad (4)$$

Por análoga razón,

$$\Sigma (Y - \bar{Y})^2 = \frac{n(n^2 - 1)}{12} \qquad (5)$$

Llamemos $d = X - Y$ a la diferencia entre el valor ordinal en X y el valor ordinal en Y de la persona i. Advirtiendo que $\bar{X} = \bar{Y}$ es decir, $\bar{X} - \bar{Y} = 0$, y, teniendo en cuenta (4), nos queda:

$$\Sigma d^2 = \Sigma [(X - Y) - (\bar{X} - \bar{Y})]^2 = \Sigma [(X - \bar{X}) - (Y - \bar{Y})]^2 =$$
$$= \Sigma (X - \bar{X})^2 - 2\Sigma (X - \bar{X})(Y - \bar{Y}) + \Sigma (Y - \bar{Y})^2 =$$
$$= 2\frac{n(n^2 - 1)}{12} - 2\Sigma (X - \bar{X})(Y - \bar{Y})$$

De donde,

$$\Sigma (X - \bar{X})(Y - \bar{Y}) = \frac{n(n^2 - 1)}{12} - \frac{\Sigma d^2}{2} \qquad (6)$$

En conclusión, teniendo en cuenta (1), (4), (5) y (6), nos queda:

$$r_s = \frac{\frac{n(n^2 - 1)}{12} - \frac{\Sigma d^2}{2}}{\sqrt{\frac{n(n^2 - 1)}{12}}\sqrt{\frac{n(n^2 - 1)}{12}}} = \frac{\frac{n(n^2 - 1) - 6\Sigma d^2}{12}}{\frac{n(n^2 - 1)}{12}} = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)}$$

14.7. Resumen: Definiciones y fórmulas

Coeficiente de correlación de Spearman: Coeficiente de correlación de Pearson aplicado a dos sucesiones de valores ordinales, considerados estos últimos como auténticas puntuaciones:

$$r_s = 1 - \frac{6\Sigma d_i^2}{(n)(n^2 - 1)}$$

Coeficiente de correlación de Kendall: Índice que nos mide el grado de semejanza entre los valores de dos sucesiones ordinales. Más concretamente, siendo P el número de no-inversiones y Q el número de inversiones, por definición,

$$\tau = \frac{P - Q}{P + Q}$$

Gamma de Goodman y Kruskal: Proporción de pares semejantes o no-inversos, dentro de los no empatados, menos proporción de pares desemejantes o inversos, dentro, también, de los no empatados:

$$\gamma = \frac{n_s}{n_s + n_d} - \frac{n_d}{n_s + n_d} = \frac{n_s - n_d}{n_s + n_d}$$

EJERCICIOS

14.1. En un colegio femenino socioeconómicamente medio-alto, siete alumnas fueron ordenadas de 1 a 7 según el mayor o menor grado en el que creían haber alcanzado las cosas deseadas. Cada alumna lleva asignados los ingresos familiares mensuales. Los datos son los siguientes: J. F. (250.000 ptas. o más), L. P. (210.000), M. A. (115.000), P. E. (72.000), A. Q. (78.000), E. R. (143.000), V. G. (85.000). Calcular el coeficiente de correlación de Spearman.

14.2. A partir de los datos del ejercicio anterior, calcular el coeficiente de correlación de Kendall.

14.3. Dos profesores, A y B, ordenan los ensayos literarios sobre la democracia escritos por 8 estudiantes. A partir de los datos del cuadro, calcular el coeficiente de correlación de Spearman, siendo las dos ordenaciones:

A	B
1	3
2	2
3	5
4	1
5	4
6	8
7	7
8	6

14.4. A partir de los datos del ejercicio anterior, calcular el coeficiente de correlación de Kendall.

14.5. Calcular el coeficiente de correlación de Spearman entre X e Y a partir de los cuadros siguientes. (Los números dentro de cada cuadro son puntuaciones en las variables X e Y. Como se sabe, estas puntuaciones deben ser transformadas en sus correspondientes valores ordinales.)

a)

X	Y
6	10
1	2
4	7
2	9
5	8

b)

X	Y
48	28
33	37
51	34
12	46
43	42

c)

X	Y
25	18
24	15
20	14
20	10
18	12
20	7

d)

X	Y
18	15
14	15
24	19
28	24
26	15
14	30
28	32

14.6. Calcular el coeficiente de correlación de Kendall entre X e Y a partir de los cuadros siguientes. (Los números dentro de cada cuadro son puntuaciones en X y en Y.)

a)

X	Y
2	4
5	3
6	5
1	2

b)

X	Y
18	14
35	7
42	9
6	13
24	15

c)

X	Y
74	27
80	32
65	28
82	29
69	35
81	34

d)

X	Y
72	31
62	28
54	32
42	35
70	33
48	34
57	30

14.7. En un grupo de 20 personas valió 0,60 el coeficiente de correlación de Kendall entre «responsabilidad» y «rendimiento escolar». Se pide calcular el número de inversiones y de no inversiones que tuvieron lugar en dicho grupo.

14.8. Calcular la gamma de Goodman y Kruskal a partir de los cuadros siguientes. En todos ellos suponemos que las categorías están ordenadas, es decir, $X_1 < X_2 < X_3 < \ldots$, $Y_1 < Y_2 < Y_3 < \ldots$

a)

	X_1	X_2
Y_2	6	16
Y_1	20	8

b)

	X_1	X_2	X_3
Y_2	10	6	1
Y_1	2	4	8

c)

	X_1	X_2	X_3
Y_3	8	4	1
Y_2	3	5	2
Y_1	2	4	6

d)

	X_1	X_2	X_3
Y_4	2	10	12
Y_3	3	6	4
Y_2	5	5	2
Y_1	10	6	1

e)

	X_1	X_2
Y_3	3	7
Y_2	6	5
Y_1	8	4

f)

	X_1	X_2	X_3	X_4
Y_4	6	4	3	3
Y_3	3	4	2	2
Y_2	2	1	4	1
Y_1	2	2	3	5

15

Relación entre variables nominales

15.1. Idea previa

Según sabemos, una variable es nominal, cuando a lo largo de ella sólo es posible establecer categorías no ordenadas, es decir, categorías cuyas posiciones pueden ser intercambiadas arbitrariamente.

15.2. Coeficiente *Q* de Yule

15.2.1. Fundamento y fórmula

Este coeficiente es aplicable cuando tenemos dos variables nominales, cada una de ellas con sólo dos categorías. Por ejemplo, las variables nominales «religión» (con las dos categorías: católico-protestante) y «raza» (con las dos categorías: latino-sajón). Consideremos el cuadro siguiente:

		RELIGION		
		A_1 (católicos)	A_2 (protestantes)	
RAZA	B_2 (saj.)	(A_1B_2)	(A_2B_2)	(B_2)
	B_1 (lat.)	(A_1B_1)	(A_2B_1)	(B_1)
		(A_1)	(A_2)	n

Según la terminología de Yule (1960):

(A_1B_1): número de católicos latinos.

(A_1B_2): número de católicos sajones.

(A_2B_1): número de protestantes latinos.

(A_2B_2): número de protestantes sajones.

(A_1): número de católicos.

(A_2): número de protestantes.

(B_1): número de latinos.

(B_2): número de sajones.

n: número total de personas.

Supuesto esto, no existirá relación entre raza y religión, si la proporción de católicos es la misma entre los latinos que entre los sajones. Es decir, si

$$\frac{(A_1B_1)}{(B_1)} = \frac{(A_1B_2)}{(B_2)} \tag{15.1}$$

Ahora bien,

$$\frac{(A_1B_1)}{(B_1)} + \frac{(A_2B_1)}{(B_1)} = \frac{(B_1)}{(B_1)}, \quad \text{de donde,} \quad \frac{(A_2B_1)}{(B_1)} = 1 - \frac{(A_1B_1)}{(B_1)}$$

$$\frac{(A_1B_2)}{(B_2)} + \frac{(A_2B_2)}{(B_2)} = \frac{(B_2)}{(B_2)}, \quad \text{de donde,} \quad \frac{(A_2B_2)}{(B_2)} = 1 - \frac{(A_1B_2)}{(B_2)}$$

Por tanto, si

$$\frac{(A_1B_1)}{(B_1)} = \frac{(A_1B_2)}{(B_2)} \quad , \quad \frac{(A_2B_1)}{(B_1)} = \frac{(A_2B_2)}{(B_2)}$$

Es decir, si no existe relación entre raza y religión, debe verificarse, también,

$$\frac{(A_2B_1)}{(B_1)} = \frac{(A_2B_2)}{(B_2)} \tag{15.2}$$

Esto es obvio, pues si no existe relación entre raza y religión, la proporción de protestantes entre los latinos debe ser la misma que entre los sajones. Las relaciones (15.1) y (15.2) se implican mutuamente. Si es verdadera una cualquiera de las dos, tiene que serlo, también, la otra.

Si, además, tenemos en cuenta una conocida propiedad de las proporciones, (15.1) y (15.2) llevan consigo

$$\frac{(A_1B_1)}{(B_1)} = \frac{(A_1B_2)}{(B_2)} = \frac{(A_1B_1) + (A_1B_2)}{(B_1) + (B_2)} = \frac{(A_1)}{n} \tag{15.3}$$

$$\frac{(A_2B_1)}{(B_1)} = \frac{(A_2B_2)}{(B_2)} = \frac{(A_2B_1) + (A_2B_2)}{(B_1) + (B_2)} = \frac{(A_2)}{n} \tag{15.4}$$

En conclusión, de acuerdo con (15.3) y (15.4), la independencia entre raza y religión implica (y es implicada por) las siguientes relaciones:

$$(A_1B_1) = \frac{(A_1)(B_1)}{n} \quad (15.5a), \quad \text{o sea,} \quad \frac{(A_1B_1)}{n} = \frac{(A_1)}{n}\frac{(B_1)}{n} \quad (15.5b)$$

$$(A_1B_2) = \frac{(A_1)(B_2)}{n} \quad (15.6a), \quad \text{o sea,} \quad \frac{(A_1B_2)}{n} = \frac{(A_1)}{n}\frac{(B_2)}{n} \quad (15.6b)$$

$$(A_2B_1) = \frac{(A_2)(B_1)}{n} \quad (15.7a), \quad \text{o sea,} \quad \frac{(A_2B_1)}{n} = \frac{(A_2)}{n}\frac{(B_1)}{n} \quad (15.7b)$$

$$(A_2B_2) = \frac{(A_2)(B_2)}{n} \quad (15.8a), \quad \text{o sea,} \quad \frac{(A_2B_2)}{n} = \frac{(A_2)}{n}\frac{(B_2)}{n} \quad (15.8b)$$

Si es cierta una cualquiera de las relaciones (15.5*a*), (15.6*a*), (15.7*a*) y (15.8*a*), lo son, también, las tres restantes. Consiguientemente, si es verdadera una cualquiera de las relaciones (15.5*b*), (15.6*b*), (15.7*b*) y (15.8*b*), lo son, también, las tres restantes. En otras palabras, la independencia entre raza y religión puede venir expresada por una de las cuatro proposiciones: «la proporción de católicos latinos es igual a la proporción de católicos por la proporción de latinos», «la proporción de protestantes latinos es igual a la proporción de protestantes por la proporción de latinos», «la proporción de católicos sajones es igual a la proporción de católicos por la proporción de sajones», «la proporción de protestantes sajones es igual a la proporción de protestantes por la proporción de sajones».

Por tanto, la independencia puede ser introducida mediante (15.5*b*) o, lo que es equivalente, mediante (15.5*a*). Ahora bien, esto es lo mismo que aceptar como criterio de independencia,

$$d = (A_1B_1) - \frac{(A_1)(B_1)}{n} = 0$$

Si $d = 0$, tendremos independencia. En cambio, si $d \neq 0$, existirá alguna dependencia o relación entre raza y religión. En otras palabras,

$$d = (A_1B_1) - \frac{(A_1)(B_1)}{n} = \frac{n(A_1B_1) - (A_1)(B_1)}{n} \quad (15.9)$$

puede ser introducido como índice de correlación.

Si advertimos que

$$(A_1) = (A_1B_1) + (A_1B_2)$$
$$(B_1) = (A_1B_1) + (A_2B_1)$$
$$n = (A_1B_1) + (A_1B_2) + (A_2B_1) + (A_2B_2)$$

y sustituimos en (15.9), (A_1), (B_1) y n por sus correspondientes valores, llegamos fácilmente a

$$d = \frac{(A_1B_1)(A_2B_2) - (A_1B_2)(A_2B_1)}{n} \quad (15.10)$$

que valdrá cero en caso de independencia y que será distinto de cero si existe alguna dependencia o relación entre las dos variables de que se trate, según lo acabado de ver hace unos instantes.

Así, por ejemplo, en la tabla 15.1, existe independencia entre raza y religión. Por el contrario, en la tabla 15.2, existe una gran relación entre ambas variables, pues la mayoría de los católicos tienden a ser latinos, la mayoría de los latinos tienden a ser católicos, la mayoría de los protestantes tienden a ser sajones y la mayoría de los sajones tienden a ser protestantes.

TABLA 15.1

	Católicos	Protestantes	
Sajones	78	52	130
Latinos	42	28	70
	120	80	200

$$d = \frac{(42)(52) - (78)(28)}{200} = \frac{2.184 - 2.184}{200} = 0$$

TABLA 15.2

	Católicos	Protestantes	
Sajones	10	75	85
Latinos	100	15	115
	110	90	200

$$d = \frac{(100)(75) - (10)(15)}{200} = \frac{7.500 - 150}{200} = 36{,}75$$

No obstante, este índice d tal como ha sido definido, presenta dos inconvenientes notables:

a) No existe un valor numérico máximo al cual podamos referir el d obtenido en un caso concreto. Por consiguiente, no podemos hacernos una idea aproximada sobre la magnitud del d encontrado en una situación concreta. Más aún, manteniendo constantes los porcentajes de las cuatro casillas interiores, podemos hacer que d aumente tanto como queramos, con tal de aumentar el tamaño de la muestra. Así, por ejemplo, si en la tabla 15.2 multiplicamos por 10 todas las casillas, d pasará a valer 367,5; y si las multiplicamos por 100, d pasará a valer 3.675. Y, sin embargo, la contextura intrínseca del cuadro es la misma en los tres casos:

	Católicos	Protestantes	
Sajones	5 %	37,5 %	42,5 %
Latinos	50 %	7,5 %	57,5 %
	55 %	45,0 %	100,0 %

b) Existiendo relación perfecta entre las dos variables (por ejemplo, todo católico es latino, todo latino es católico, todo protestante es sajón y todo sajón es protestante), d puede tomar valores muy distintos. Así, por ejemplo, consideremos los dos casos siguientes:

	Católicos	Protestantes	
Sajones	0	1	1
Latinos	99	0	99
	99	1	100

$$d = \frac{(99)(1) - 0}{100} = \frac{99}{100} = 0{,}99$$

	Católicos	Protestantes	
Sajones	0	50	50
Latinos	50	0	50
	50	50	100

$$d = \frac{(50)(50) - 0}{100} = \frac{2.500}{100} = 25$$

Para evitar estas dificultades, Yule (1965) propone el coeficiente Q definido así:

$$Q = \frac{(n)(d)}{(A_1B_1)(A_2B_2) + (A_1B_2)(A_2B_1)} = \frac{(A_1B_1)(A_2B_2) - (A_1B_2)(A_2B_1)}{(A_1B_1)(A_2B_2) + (A_1B_2)(A_2B_1)} \quad (15.11)$$

Si la relación es nula, $d = 0$ y, por tanto, $Q = 0$.

Si la relación es perfecta, ó $(A_1B_2) = (A_2B_1) = 0$, ó, $(A_1B_1) = (A_2B_2) = 0$. En el primer caso, $Q = 1$. En el segundo, $Q = -1$.

Sin embargo, aunque $Q = 1$ ó $Q = -1$, la relación no es necesariamente perfecta. Por ejemplo, en el cuadro siguiente, ni todo católico es latino, ni todo sajón es protestante. A pesar de ello, $Q = 1$.

	Católicos	Protestantes	
Sajones	30	40	70
Latinos	75	0	75
	105	40	145

$$Q = \frac{(75)(40) - (30)(0)}{(75)(40) + (30)(0)} = \frac{3.000}{3.000} = 1$$

15.2.2. Cálculo

Basta con aplicar la definición dada.

EJEMPLO 15.1. Calculemos Q entre raza y religión a partir de la tabla 15.3.

TABLA 15.3

	Católicos	Protestantes	
Sajones	20	70	90
Latinos	100	10	110
	120	80	200

$$Q = \frac{(100)(70) - (20)(10)}{(100)(70) + (20)(10)} = \frac{7.000 - 200}{7.000 + 200} = \frac{6.800}{7.200} = 0{,}944$$

El signo de Q es función de la organización del cuadro de frecuencias, supuestos unos mismos datos. Por ello, la táctica más sensata para interpretar dicho signo, es observar el cuadro de frecuencias. Cualquier otra norma rígida puede llevarnos a interpretaciones equivocadas. Así, por ejemplo, los mismos datos de la tabla 15.3 pueden legítimamente aparecer organizados del modo siguiente:

TABLA 15.4

	Católicos	Protestantes	
Latinos	100	10	110
Sajones	20	70	90
	120	80	200

$$Q = \frac{(20)(10) - (100)(70)}{(20)(10) + (100)(70)} = \frac{200 - 7.000}{200 + 7.000} = \frac{-6.800}{7.200} = -0{,}944$$

Es evidente que en ambos casos la relación entre raza y religión es la misma. El signo ha cambiado por el mero hecho de haber trastocado el orden de las filas, colocando la fila de los latinos por encima de la de los sajones (tabla 15.4) en vez de dejar la fila de los latinos por debajo de la de los sajones (tabla 15.3). Sin embargo, ambas tablas manifiestan una misma relación entre raza y religión, a saber: aparece una clara tendencia entre los católicos a ser latinos y entre los latinos a ser católicos; entre los protestantes a ser sajones y entre los sajones a ser protestantes. En otras palabras, existe una clara relación positiva entre ser católico y ser latino y entre ser latino y ser católico; entre ser protestante y ser sajón y entre ser sajón y ser protestante. Es claro que esta misma relación puede ser traducida así: existe clara relación negativa entre ser católico y ser sajón y entre ser sajón y ser protestante; entre ser protestante y ser latino y entre ser latino y ser protestante.

EJEMPLO 15.2. Getzels y Jackson (1962) distinguieron entre los adolescentes dos tipos de talento: inteligentes y creativos. Estudiando la relación entre el tipo de talento y la profesión del padre de los adolescentes considerados, llegaron al siguiente resultado:

		Profesión del padre		
		Prof. univers.	Negocios	
Tipo de talento	Creativos	7	11	18
	Inteligentes	15	4	19
		22	15	37

$$Q = \frac{(15)(11) - (7)(4)}{(15)(11) + (7)(4)} = \frac{137}{193} = 0{,}71$$

15.2.3. Propiedades

a) El coeficiente de correlación Q no puede valer menos que -1 ni más que 1. Es decir, $-1 \leq Q \leq 1$.

Ha quedado demostrado en el párrafo anterior.

b) Multiplicando los dos elementos de la primera fila de la tabla de frecuencias por una constante cualquiera, los dos de la segunda fila por otra, los dos de la primera columna por otra distinta y los de la segunda por otra cualquiera, Q no varía. La razón está en que en cada uno de estos productos multiplicamos numerador y denominador de la fórmula de Q por una misma constante y, por tanto, el cociente, Q, no varía.

Así, por ejemplo, de la tabla 15.5 pasamos a la 15.6 multiplicando 4 y 3 por 2. De la tabla 15.6 pasamos a la 15.7 multiplicando 1 y 2 por 4. De la tabla 15.7 pasamos a la 15.8 multiplicando 8 y 4 por 5. De la tabla 15.8 pasamos a la 15.9 multiplicando 6 y 8 por 3. El lector puede comprobar cómo en las cuatro tablas $Q = -5/11$.

TABLA 15.5

4	3
1	2

TABLA 15.6

8	6
1	2

TABLA 15.7

8	6
4	8

TABLA 15.8

40	6
20	8

TABLA 15.9

40	18
20	24

15.3. Coeficiente χ^2

15.3.1. Fundamento y fórmula

Mejor sería llamarle X^2 u otro símbolo distinto de χ^2. En realidad, lo que aquí vamos a llamar χ^2 no es más que un estadístico cuya distribución de probabilidad se aproxima a la distribución de probabilidad llamada χ^2, a medida que aumenta más y más el tamaño de la muestra. Sin embargo, le llamaremos χ^2, dada la universalidad de esta designación en los libros de Estadística para psicólogos.

El coeficiente χ^2 lo proponemos aquí principalmente como medio para poder calcular el coeficiente de contingencia C, del que hablaremos inmediatamente después.

Recordemos que el coeficiente Q exigía que las dos variables constaran de sólo dos categorías. Ahora bien, cada una de las dos variables puede constar de dos o más categorías. No obstante, el modo de razonar es el análogo al seguido al tratar del coeficiente Q. Supongamos, por ejemplo, la variable «religión» (con tres categorías) y la variable «raza» (con dos), según el cuadro de la página siguiente.

(A_i) y (B_j) representan el número de personas dentro de las categorías A_i y B_j, respectivamente. (A_iB_j) representa el número de personas que pertenecen simultáneamente a la categoría A_i (de la variable A: religión) y a la categoría B_j (de la variable B: raza).

	A_1 (católicos)	A_2 (protestantes)	A_3 (judíos)	
B_2 (sajones)	(A_1B_2)	(A_2B_2)	(A_3B_2)	(B_2)
B_1 (latinos)	(A_1B_1)	(A_2B_1)	(A_3B_1)	(B_1)
	(A_1)	(A_2)	(A_3)	n

Si no hubiera relación entre A y B, encontraríamos la misma proporción de católicos entre latinos y sajones, de protestantes entre latinos y sajones, de judíos entre latinos y sajones. Es decir,

$$\frac{(A_1B_1)}{(B_1)} = \frac{(A_1B_2)}{(B_2)} = \frac{(A_1B_1) + (A_1B_2)}{(B_1) + (B_2)} = \frac{(A_1)}{n}$$

Por consiguiente,

$$(A_1B_1) = \frac{(A_1)(B_1)}{n} \quad , \quad (A_1B_2) = \frac{(A_1)(B_2)}{n}$$

$$\frac{(A_2B_1)}{(B_1)} = \frac{(A_2B_2)}{(B_2)} = \frac{(A_2B_1) + (A_2B_2)}{(B_1) + (B_2)} = \frac{(A_2)}{n}$$

Por tanto,

$$(A_2B_1) = \frac{(A_2)(B_1)}{n} \quad , \quad (A_2B_2) = \frac{(A_2)(B_2)}{n}$$

$$\frac{(A_3B_1)}{(B_1)} = \frac{(A_3B_2)}{(B_2)} = \frac{(A_3B_1) + (A_3B_2)}{(B_1) + (B_2)} = \frac{(A_3)}{n}$$

Por consiguiente,

$$(A_3B_1) = \frac{(A_3)(B_1)}{n} \quad , \quad (A_3B_2) = \frac{(A_3)(B_2)}{n}$$

Nótese que la independencia implica que la frecuencia de cada casilla sea igual al producto de sus dos frecuencias marginales dividido por el número total de personas. Llamemos frecuencia teórica a esta frecuencia que debería aparecer en cada casilla en caso de independencia. Si observando un grupo de personas, las frecuencias encontradas, de hecho (que llamaremos empíricas) coinciden con las teóricas, diremos que no existe relación entre las dos variables en cuestión, A y B, para ese grupo de personas. Por el contrario, si no coinciden, diremos que

A y B no son independientes para ese grupo, es decir, que existe cierta relación entre A y B. Esta no independencia (o relación) tenderá a ser mayor a medida que las frecuencias empíricas se alejen más y más de las teóricas. Pues bien, llamando f_t a las frecuencias teóricas y f_e a las frecuencias empíricas, definiremos χ^2 del modo siguiente:

$$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t} \tag{15.12}$$

El signo de sumar, Σ, va de 1 a $(r)(s)$, siendo r el número de categorías en la variable A y s el número de categorías en la variable B, o sea, siendo $(r)(s)$ el número total de casillas en la tabla de frecuencias.

15.3.2. Cálculo

Es una mera aplicación de la fórmula anterior.

EJEMPLO 15.3. Con el fin de estudiar el influjo de las condiciones ambientales en que se desarrollan los gatos sobre su tendencia a cazar ratones, Kuo (1930) crió gatitos en distintas condiciones: 20 fueron criados en aislamiento; 21 con sus madres, viéndolas cazar ratones y 18 fueron criados con ratones. Después del período de crianza, a cada gatito se le presentaba un ratón para ver si lo mataba o no. Pues bien, encontró los siguientes resultados:

		Forma de crianza			
		En aislamiento	Con sus madres	Con ratones	
Respuesta ante el roedor	Matarlo	9 (10,17)	18 (10,68)	3 (9,15)	30
	No matarlo	11 (9,83)	3 (10,32)	15 (8,85)	29
		20	21	18	59

En cada una de las seis casillas hay dos números, uno sin paréntesis (la frecuencia empírica, la encontrada, de hecho), otro dentro de un paréntesis (la frecuencia teórica, la que debería aparecer en caso de independencia). Según lo dicho, estas frecuencias teóricas se calculan así:

$$\frac{(20)(30)}{59} = 10{,}17 \qquad \frac{(21)(30)}{59} = 10{,}68 \qquad \frac{(18)(30)}{59} = 9{,}15$$

$$\frac{(20)(29)}{59} = 9{,}83 \qquad \frac{(21)(29)}{59} = 10{,}32 \qquad \frac{(18)(29)}{59} = 8{,}85$$

Por consiguiente,

$$\chi^2 = \frac{(9 - 10{,}17)^2}{10{,}17} + \frac{(18 - 10{,}68)^2}{10{,}68} + \frac{(3 - 9{,}15)^2}{9{,}15} + \frac{(11 - 9{,}83)^2}{9{,}83} + \frac{(3 - 10{,}32)^2}{10{,}32} +$$

$$+ \frac{(15 - 8{,}85)^2}{8{,}85} = \frac{1{,}37}{10{,}17} + \frac{53{,}58}{10{,}68} + \frac{37{,}82}{9{,}15} + \frac{1{,}37}{9{,}83} + \frac{53{,}58}{10{,}32} + \frac{37{,}82}{8{,}85} = 0{,}13 +$$

$$+ 5{,}02 + 4{,}13 + 0{,}14 + 5{,}19 + 4{,}27 = 18{,}88$$

En el caso en que cada una de las dos variables conste de sólo dos categorías, tendremos la tabla siguiente:

a	b	$(a + b)$
c	d	$(c + d)$
$(a + c)$	$(b + d)$	n

En él, a, b, c y d son las cuatro frecuencias empíricas y $n = a + b + c + d$. Las cuatro frecuencias teóricas serán:

$$\frac{(a + b)(a + c)}{n}, \quad \frac{(a + b)(b + d)}{n}, \quad \frac{(c + d)(a + c)}{n}, \quad \frac{(c + d)(b + d)}{n}$$

Pues bien, es fácil probar que, bajo estas circunstancias,

$$\chi^2 = \frac{(n)(cb - ad)^2}{(a + b)(c + d)(a + c)(b + d)} \tag{15.13}$$

EJEMPLO 15.4. A partir de la tabla siguiente calculemos χ^2, aplicando (15.12) y (15.13).

2 (3,6)	10 (8,4)	12
4 (2,4)	4 (5,6)	8
6	14	20

Según (15.12):

$$\chi^2 = \frac{(-1{,}6)^2}{3{,}6} + \frac{(1{,}6)^2}{8{,}4} + \frac{(1{,}6)^2}{2{,}4} + \frac{(-1{,}6)^2}{5{,}6} =$$

$$= (1{,}6)^2(0{,}2778 + 0{,}1191 + 0{,}4167 + 0{,}1786) = (2{,}56)(0{,}9922) = 2{,}54$$

Según (15.13):

$$\chi^2 = \frac{(20)(40 - 8)^2}{(12)(8)(6)(14)} = \frac{20.480}{8.064} = 2{,}54$$

Como vemos, tanto (15.12) como (15.13) nos llevan al mismo resultado.

Con tablas de frecuencias de dos filas y dos columnas es recomendable la corrección de Yates, especialmente cuando una o más de las frecuencias *teóricas* en las cuatro casillas es pequeña. (Por pequeña se suele entender menor que 5 y aun menor que 10.) Mediante ella,

$$\chi^2 = \sum \frac{(|f_e - f_t| - 0{,}5)^2}{f_t} = \frac{n(|cb - ad| - n/2)^2}{(a + b)(c + d)(a + c)(b + d)}$$

Calculemos en el ejemplo siguiente χ^2, primero sin la corrección y luego con ella.

1 (1,2)	2 (1,8)	3
3 (2,8)	4 (4,2)	7
4	6	10

$$\chi^2 = \frac{(-0{,}2)^2}{1{,}2} + \frac{(0{,}2)^2}{1{,}8} + \frac{(0{,}2)^2}{2{,}8} + \frac{(-0{,}2)^2}{4{,}2} =$$

$$= (0{,}2)^2(0{,}8333 + 0{,}5556 + 0{,}3571 + 0{,}2381) =$$

$$= (0{,}04)(1{,}9841) = 0{,}079364$$

Con corrección y usando la primera fórmula:

$$\chi^2 = \frac{(0{,}2 - 0{,}5)^2}{1{,}2} + \frac{(0{,}2 - 0{,}5)^2}{1{,}8} + \frac{(0{,}2 - 0{,}5)^2}{4{,}2} + \frac{(0{,}2 - 0{,}5)^2}{2{,}8} =$$

$$= (0{,}3)^2(1{,}9841) = 0{,}178569$$

Con corrección y usando la segunda fórmula:

$$\chi^2 = \frac{(10)(|6-4|-10/2)^2}{(3)(7)(4)(6)} = \frac{(10)(-3)^2}{504} = \frac{90}{504} = 0{,}178570$$

Como se ve, el valor de χ^2 varía al usar la fórmula sin corrección y la fórmula corregida. Es debido a lo pequeñas que son las frecuencias teóricas. La importancia de esta diferencia se verá mejor al estudiar Estadística Inferencial.

Cuando la tabla es de dos filas y dos columnas, la diferencia $|f_e - f_t|$ (es decir, tomada en valor absoluto) es la misma en las cuatro casillas. El lector se encargará de demostrar esta afirmación en el ejercicio 15.7.

15.3.3. Propiedades

a) Al multiplicar por una constante k las frecuencias de todas las casillas de un cuadro de frecuencias (con cualquier número de filas y cualquier número de columnas) χ^2 queda multiplicado por esa constante.

En efecto, llamemos χ_1^2 al valor de χ^2 antes de la multiplicación y χ_k^2 al valor de χ^2 después de multiplicar todas las frecuencias por k. Tendremos,

$$\chi_k^2 = \sum \frac{(k f_e - k f_t)^2}{k f_t} = \frac{k^2}{k} \sum \frac{(f_e - f_t)^2}{f_t} = k \chi_1^2$$

Compruebe el alumno cómo, en efecto, el valor de χ^2 es 10/3 a partir de la tabla 15.10 y 40/3 a partir de la tabla 15.11.

TABLA 15.10

A)

8	4	12
2	6	8
10	10	20

TABLA 15.11

B)

32	16	48
8	24	32
40	40	80

NOTA. En bastantes textos las tablas de frecuencias de que hemos venido hablando suelen ser denominadas *tablas de contingencia* («contingency tables»).

15.4. Coeficiente de contingencia, *C*

15.4.1. Fundamento y fórmula

Acabamos de ver que χ^2 es función de n. Hemos visto, en efecto, que multiplicando las frecuencias de todas las casillas por una constante suficientemente alta, podemos hace que χ^2 aumente, a pesar de que las proporciones de todas las casillas sean las mismas antes y después de dicha multiplicación. Pues bien, una manera de evitar este inconveniente es utilizar C, definido de la siguiente manera:

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

15.4.2. Cálculo

Es una mera aplicación de la fórmula anterior.

EJEMPLO 15.5. Calculemos C a partir de $\chi^2 = 18{,}88$ obtenido en el ejemplo 15.3.

$$C = \sqrt{\frac{18{,}88}{59 + 18{,}88}} = \sqrt{0{,}2424} = 0{,}49$$

15.4.3. Propiedades

a) El coeficiente de contingencia C no valdrá menos que cero ni valdrá uno o más. Es decir, $0 \leq C < 1$.

Basta con mirar la fórmula de C para convencerse que nunca podrá llegar a valer uno, pues el numerador será siempre menor que el denominador, para todo valor finito de n. Por otra parte, de las dos raíces posibles aceptamos para C la positiva, con lo cual no tiene sentido hablar de valores negativos de C. Por consiguiente, C nos indica la magnitud de la relación entre dos variables. El significado positivo o negativo de la misma lo tendremos que inferir a base de la tabla de contingencia, considerando cómo se encuentran distribuidas las frecuencias en las diversas casillas. Incluso a veces, ni tendrá sentido hablar del signo positivo o negativo de C, sino meramente de la intensidad de la relación entre las dos variables en cuestión. Esto sucederá cuando la tabla de contingencia tenga más de dos filas o/y más de dos columnas.

b) Es función del número de filas y columnas. De aquí que dos *C's* (entre las mismas variables, X e Y) sólo son comparables si la primera C y la segunda, ambas han sido calculadas a partir de dos tablas de contingencia tales que la primera tiene el mismo número de filas que la segunda y la primera tiene el mismo número de columnas que la segunda. Así, por ejemplo, serán comparables dos *C's* si la primera ha sido calculada a partir de una tabla de contingencia con cinco filas y dos columnas y la segunda a partir de otra tabla, también, con cinco filas y dos columnas. (Suponemos, naturalmente, que son las mismas las variables estudiadas en ambos casos.) En cambio, no serán comparables dos *C's*, si una tabla tiene cuatro filas y dos columnas y la otra tiene siete filas y tres columnas.

c) Si la tabla de contingencia, a partir de la que obtenemos C, tiene igual número de filas que de columnas (llamemos a ese número, k) se demuestra que el valor máximo que puede alcanzar C viene dado por $\sqrt{(k-1)/k}$. Así,

$$\text{para} \quad k = 2 \qquad C_{\text{máx}} = \sqrt{(2-1)/2} = \sqrt{1/2} = 0{,}707$$
$$k = 3 \qquad C_{\text{máx}} = \sqrt{(3-1)/3} = \sqrt{2/3} = 0{,}816$$
$$k = 4 \qquad C_{\text{máx}} = \sqrt{(4-1)/4} = \sqrt{3/4} = 0{,}866$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$k = 10 \qquad C_{\text{máx}} = \sqrt{(10-1)/10} = \sqrt{9/10} = 0{,}949$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

Esto quiere decir, que el $C_{\text{máx}}$ no alcanzará el valor 1, para valores finitos de k por grandes que sean. Solamente en el límite alcanzaría ese valor. Es decir,

$$\lim_{k\to\infty} \sqrt{(k-1)/k} = \lim_{k\to\infty} \sqrt{1-(1/k)} = 1$$

Vamos a comprobar el valor del $C_{\text{máx}}$ para el caso de una tabla de contingencia de dos filas y dos columnas.

Las dos tablas de contingencia siguientes implican máxima correlación. Comprobemos cómo en ambos casos $C = \sqrt{1/2} = 0{,}707$.

1 (0,1)	0 (0,9)	1
0 (0,9)	9 (8,1)	9
1	9	10

$$\chi^2 = \frac{(0{,}9)^2}{0{,}1} + \frac{(0{,}9)^2}{0{,}9} + \frac{(0{,}9)^2}{0{,}9} + \frac{(0{,}9)^2}{8{,}1} = 8{,}1 + 0{,}9 + 0{,}9 + 0{,}1 = 10$$

Por consiguiente,

$$C = \sqrt{10/(10+10)} = \sqrt{1/2} = 0{,}707$$

0 (3,2)	4 (0,8)	4
16 (12,8)	0 (3,2)	16
16	4	20

$$\chi^2 = \frac{(3{,}2)^2}{3{,}2} + \frac{(3{,}2)^2}{0{,}8} + \frac{(3{,}2)^2}{12{,}8} + \frac{(3{,}2)^2}{3{,}2} = 3{,}2 + 12{,}8 + 0{,}8 + 3{,}2 = 20$$

Por consiguiente,

$$C = \sqrt{20/(20+20)} = \sqrt{1/2} = 0{,}707$$

d) Las distribuciones de ambas variables pueden ser de cualquier forma. C puede ser calculado y tiene sentido sea cual sea la forma de ambas distribuciones.

e) Puede ser calculado a cualquier nivel de medida. Si las variables son cuantitativas, C puede calcularse y tiene sentido, sea cual sea el diagrama de dispersión.

f) No es comparable directamente con otros índices de correlación como r_{xy}, r_s, etc. De aquí que no parezca muy recomendable interpretar como r_{xy} el cociente $C/C_{\text{máx}}$.

15.5. Interpretación de *Q* y *C*

Como en ocasiones anteriores, la única interpretación razonable de un valor determinado de Q y de C es compararlo con los valores encontrados por otros investigadores trabajando con las mismas variables o variables semejantes. Además, en el caso de C, para que sea sensata la comparación, es necesario que nuestro coeficiente C haya sido calculado a partir de una tabla de contingencia con el mismo número de filas y de columnas que tenían las tablas de contingencia de los otros investigadores o con un número muy parecido.

15.6. Resumen: Definiciones y fórmulas

Coeficiente Q

$$Q = \frac{(A_1B_1)(A_2B_2) - (A_1B_2)(A_2B_1)}{(A_1B_1)(A_2B_2) + (A_1B_2)(A_2B_1)}$$

Coeficiente χ^2

$$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$$

Coeficiente C

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

EJERCICIOS

15.1. Calcular el coeficiente Q de Yule a partir de los cuadros siguientes:

a)

15	30
10	4

b)

20	40
30	15

c)

10	5
8	14

d)

25	20
40	18

15.2. Calcular el coeficiente χ^2 a partir de los cuadros siguientes:

a)

15	45
30	10

b)

6	14	10
10	26	14

c)

9	5	2
1	10	13

d)

0	5
3	3
7	2

e)

10	6	4
5	12	3
0	2	8

f)

22	19	7
3	3	2
0	3	41

15.3. Calcular el coeficiente de contingencia C, a partir de los cuadros expuestos en 15.2.

15.4. Sabiendo que vale 0,4 el coeficiente de contingencia respecto a un grupo de 126 personas, ¿cuánto valdrá χ^2 para ese mismo grupo?

15.5. Sabiendo que $\left(\frac{C}{C_{\text{máx}}}\right)^2 = 0{,}90$ y que la tabla de contingencia es de 5 por 5, calcular el valor de C.

15.6. Demostrar que en cualquier tabla de contingencia la suma de las frecuencias teóricas tiene que ser igual que la suma de las frecuencias empíricas.

15.7. Demostrar que en una tabla de contingencia de 2 por 2, la diferencia $|f_e - f_t|$ (es decir, tomada en valor absoluto) tiene que valer lo mismo en las cuatro casillas.

15.8. Deducir la fórmula (15.13) a partir de la (15.12).

16

Relación entre variables dicotómicas o dicotomizadas

16.1. Conceptos previos

16.1.1. Variables dicotómicas

Aquellas que por su propia naturaleza sólo pueden manifestarse según dos modalidades: sexo (varón-mujer), nacionalidad (nacional-extranjero), estado vital (vivo-muerto), etc.

16.1.2. Variables dicotomizadas

Aquellas que por su propia naturaleza pueden manifestarse según muchas modalidades (al menos, según tres), pero que, de hecho, sólo se las permite manifestarse según dos: aprovechamiento escolar (aprobado-suspenso), altura (alto- bajo), número de hijos (menos de tres hijos-tres o más hijos), etc.

16.2. Coeficientes de correlación que son mera aplicación de r_{xy}

16.2.1. Coeficiente de correlación biserial puntual r_{bp}

a) *Fundamento y fórmula*

Introduzcamos una variable X, cuantitativa continua (eventualmente, discreta) y otra variable Y, dicotómica. Ésta sólo admite dos modalidades a las cuales vamos a atribuir dos números cualesquiera distintos que, por sencillez, suelen ser el 0 y el 1. Consideremos estos dos valores como auténticos números. Supuesto esto, tendremos dos columnas de números. La primera, referida a la variable X, que constará de bastantes números distintos y la segunda, referida a la variable Y, que sólo

constará de ceros y unos. Pues bien, la fórmula del coeficiente de correlación biserial puntual no es más que la fórmula del coeficiente de correlación de Pearson aplicada a estas dos columnas de números.

Se demuestra (véase 16.7. Apéndice), que el coeficiente de correlación de Pearson bajo estas condiciones, es decir, el coeficiente de correlación biserial puntual, r_{bp}, viene dado por

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{s_x}\sqrt{pq} \qquad (16.1)$$

$$= \frac{\bar{X}_p - \bar{X}}{s_x}\sqrt{\frac{p}{q}} \qquad (16.2)$$

X es la variable cuantitativa continua e Y es la dicotómica.
p es la proporción de personas con una de las dos modalidades posibles en la variable Y.
q es la proporción de personas con la otra modalidad.
$\bar{X}_p$ es la media en X de las personas cuya proporción es p.
$\bar{X}_q$ es la media en X de las personas cuya proporción es q.
$\bar{X}$ es la media en X de todas las personas.
s_x es la desviación típica en X de todas las personas.

Nótese que (16.1) admite una interpretación muy razonable. Supongamos, en efecto, que X representa agresividad, Y representa sexo, p es la proporción de varones, q la de mujeres y que los varones de nuestra muestra son más agresivos que las mujeres. Bajo estas condiciones, $\bar{X}_p$ (agresividad media de los varones) será mayor que $\bar{X}_q$ (agresividad media de las mujeres) y $\bar{X}_p - \bar{X}_q$ será tanto mayor cuanto más difieran en agresividad los varones de las mujeres, cuanto más estrecha concomitancia o correlación exista entre ser varón y alta agresividad y entre ser mujer y baja agresividad. Ahora bien, según (16.1), r_{bp} tenderá a crecer con $\bar{X}_p - \bar{X}_q$, es decir, con la correlación entre el sexo y la agresividad. Por consiguiente, r_{bp} aparece como índice idóneo de correlación. Algo parecido podía decirse sobre (16.2).

b) *Cálculo*

Apliquemos (16.1) y (16.2) al siguiente ejemplo.

EJEMPLO 16.1. Un grupo compuesto por 22 varones y 18 mujeres han obtenido las puntuaciones que se muestran en el cuadro de la página siguiente, en un test de inteligencia espacial.

En este cuadro la variable X, inteligencia espacial, es continua. La variable Y, sexo, es dicotómica.

Por su parte, n_p representa el número de varones (a los que atribuimos un 1 en Y); n_q representa el número de mujeres (a las que atribuimos un 0 en Y); n_t representa el número total de personas. Por tanto, las siete primeras personas (comenzando por arriba) han obtenido en X y en Y los siguientes pares de puntuaciones:

X	n_p	n_q	n_t	x'	x'^2	$n_t x'$	$n_t x'^2$	$n_p x'$	$n_q x'$
24-28	6	1	7	2	4	14	28	12	2
19-23	6	2	8	1	1	8	8	6	2
14-18	5	4	9	0	0	0	0	0	0
9-13	4	4	8	−1	1	−8	8	−4	−4
4-8	1	7	8	−2	4	−16	32	−2	−14
	22	18	40			−2	76	12	−14

(26,1), (26,1), (26,1), (26,1), (26,1), (26,1), (26,0); las ocho siguientes: (21,1), (21,1), (21,1), (21,1), (21,1), (21,1), (21,0), (21,0);. . . ; las ocho últimas: (6,1), (6,0), (6,0), (6,0), (6,0), (6,0), (6,0), (6,0). Pues bien, r_{bp} no es más que r_{xy} aplicado a estos cuarenta pares de números.

Calculemos previamente los elementos que aparecen en las dos fórmulas de r_{bp}.

$$\bar{X} = 16 + (5)(-2/40) = 16 - 0{,}25 = 15{,}75$$
$$\bar{X}_p = 16 + (5)(12/22) = 16 + 2{,}73 = 18{,}73$$
$$\bar{X}_q = 16 + (5)(-14/18) = 16 - 3{,}89 = 12{,}11$$
$$s_x = (5)\sqrt{76/40 - 4/1.600} = (5)\sqrt{1{,}9 - 0{,}0025} = (5)\sqrt{1{,}8975} = (5)(1{,}378) = 6{,}89$$
$$p = 22/40 = 0{,}55 \quad , \quad q = 18/40 = 0{,}45$$

De acuerdo con la tabla B (Apéndice III, al final del libro):

$$\sqrt{pq} = \sqrt{(0{,}55)(0{,}45)} = 0{,}4975 \quad , \quad \sqrt{p/q} = \sqrt{(0{,}55)/(0{,}45)} = 1{,}106$$

Entonces, aplicando (16.1),

$$r_{bp} = \frac{18{,}73 - 12{,}11}{6{,}89}\,0{,}4975 = \frac{6{,}62}{6{,}89}\,0{,}4975 = \frac{3{,}29345}{6{,}89} = 0{,}48$$

Y aplicando (16.2)

$$r_{bp} = \frac{18{,}73 - 15{,}75}{6{,}89}\,1{,}106 = \frac{2{,}98}{6{,}89}\,1{,}106 = \frac{3{,}29588}{6{,}89} = 0{,}48$$

Usando la fórmula y el esquema anterior, es claro que r_{bp} será positivo (negativo) siempre que $\bar{X}_p$ sea mayor (menor) que $\bar{X}_q$. Esto significa que si r_{bp} es positivo, a ser alto en X corresponde pertenecer a la categoría cuya proporción es p y a ser bajo en X corresponde pertenecer a la categoría cuya proporción es q. Si r_{bp} es ne-

gativo, a ser alto en X corresponde pertenecer a la categoría cuya proporción es q y a ser bajo en X corresponde pertenecer a la categoría cuya proporción es p.

El esquema anterior es arbitrario. Otros serían posibles. Sin embargo, le seguiremos siempre para evitar confusiones. No obstante, conviene que el alumno se acostumbre a considerar la tabla de frecuencias y, a partir de la misma, interprete oportunamente el valor positivo o negativo ofrecido por la fórmula. En nuestro ejemplo, $r_{bp} = 0{,}48$ es positivo. Esto quiere decir que a ser alto en X corresponde pertenecer a la categoría de los varones y a ser bajo en X corresponde pertenecer a la categoría de las mujeres.

Desde luego, no es necesario atribuir unos y ceros. Bastaría con llamar $\bar{X}_p$ a la media del grupo de personas cuya proporción es p y $\bar{X}_q$ a la media del grupo cuya proporción es q. Hemos atribuido ceros y unos para hacer ver cómo r_{bp} no es más que una aplicación de r_{xy} al caso de dos variables, una cuantitativa continua y la otra dicotómica. Compruebe el alumno cómo r_{xy}, aplicado a los 40 pares de que hemos hablado más arriba, vale 0,48.

16.2.2. Coeficiente de correlación, φ

a) *Fundamento y fórmula*

Introduzcamos dos variables dicotómicas, X e Y. Cada una de ellas admite sólo dos modalidades a las que vamos a atribuir dos números distintos que, por sencillez suelen ser el 0 y el 1. Consideremos estos dos valores como auténticos números. Esto supuesto, tendremos dos columnas de números, cada una de las cuales consta únicamente de ceros y unos. Pues bien, la fórmula del coeficiente φ no es más que la fórmula del coeficiente de correlación de Pearson aplicada a estas dos columnas de números.

Se demuestra (véase 16.7. Apéndice), que el coeficiente de correlación de Pearson bajo estas condiciones, es decir, el coeficiente φ, viene dado por:

$$\varphi = \frac{cb - ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} \qquad (16.3)$$

Donde a, b, c, y d tienen el significado expuesto en el cuadro adjunto:

		X		
		0	1	
Y	1	a	b	$(a + b)$
	0	c	d	$(c + d)$
		$(a + c)$	$(b + d)$	n

Es decir, a, b, c y d representan el número de personas cuyas puntuaciones en X e Y son respectivamente: (0,1), (1,1), (0,0), (1,0). Asimismo, $(a + b)$ representa el número de personas con puntuación igual a 1 en Y (o, mejor aún, el número de personas pertenecientes a la categoría de Y a la que hemos atribuido el valor 1), $(c + d)$ representa el número de personas con puntuación igual a 0 en Y, $(a + c)$ el de las personas con puntuación igual a 0 en X, $(b + d)$ el de las personas con puntuación igual a 1 en X.

Usando el esquema anterior, si φ es positivo (negativo) quiere decir que es positiva (negativa) la relación entre la categoría 0 en X y la categoría 0 en Y, o entre la categoría 1 en X y la categoría 1 en Y. Con todo, lo mejor es estudiar el cuadro de frecuencias para asegurarse del signo de la relación.

b) *Cálculo*

Apliquemos (16.3) al ejemplo siguiente:

EJEMPLO 16.2. Calculemos la relación posible entre poseer coche y poseer piso a partir de los datos que proponemos a continuación. Se trata de dos variables dicotómicas: o se tiene o no se tiene coche, o se tiene o no se tiene piso. Atribuyamos un 0 a las categorías «no coche» y «no piso»; atribuyamos un 1 a las categorías «coche» y «piso». Supongamos que tenemos la siguiente tabla de frecuencias:

		X		
		No coche 0	Coche 1	
Y	Piso 1	30	60	90
	No piso 0	70	40	110
		100	100	200

Aplicando la fórmula (16.3):

$$\varphi = \frac{(70)(60) - (30)(40)}{\sqrt{(90)(110)(100)(100)}} = \frac{4.200 - 1.200}{\sqrt{99.000.000}} = \frac{3.000}{9.950} = 0{,}30$$

$\varphi = 0{,}30$ es positivo. Ello quiere decir que es positiva la relación entre «tener coche» y «tener piso» o entre «no tener coche» y «no tener piso». A esta misma

conclusión llegamos observando el cuadro de frecuencias. En él son mayoría los que poseen simultáneamente ambas cosas o carecen simultáneamente de ellas; en cambio, son minoría los que poseyendo una de ellas, carecen de la otra.

No es necesario atribuir expresamente los unos y ceros. Basta con poner las frecuencias. Hemos atribuido los unos y ceros para que el alumno vea cómo φ es un caso particular de r_{xy}. En este ejemplo tenemos 30 personas cuyas puntuaciones en X e Y serían (0,1), 60 cuyas puntuaciones serían (1,1), 70 cuyas puntuaciones serían (0,0) y 40 cuyas puntuaciones serían (1,0). Tendríamos, por tanto, 200 pares de puntuaciones. Compruebe el lector cómo, de hecho, al aplicar el coeficiente de correlación de Pearson a estos 200 pares de números se obtiene $r_{xy} = 0{,}30$.

EJEMPLO 16.3. Dement y Kleitman (1957) se plantearon la hipótesis de que cuando una persona está soñando, sus ojos se mueven rápidamente. Para verificarla, hicieron que un grupo de personas durmieran en un laboratorio y durante la noche les despertaban varias veces. Unas veces, cuando presentaban movimientos oculares rápidos (M. O. R.), y otras, cuando no los presentaban. En ambas ocasiones les preguntaban si recordaban algún sueño. Pues bien, encontraron los siguientes resultados.

		X		
		Despertados sin M.O.R.	Despertados con M.O.R.	
		0	1	
Y	Recuerdan algún sueño 1	11	152	163
	No recuerdan sueño alguno 0	149	39	188
		160	191	351

Aplicando la fórmula (16.3),

$$\varphi = \frac{(149)(152) - (11)(39)}{\sqrt{(163)(188)(160)(191)}} = \frac{22.648 - 429}{\sqrt{936.480.640}} = \frac{22.219}{30.601{,}97} = 0{,}73$$

El resultado obtenido apoya la hipótesis propuesta.

16.2.3. Propiedades de r_{bp} y φ

a) Ninguno de los dos coeficientes puede valer menos que -1, ni más que 1. Es decir, $-1 \leq r_{bp} \leq 1$, $-1 \leq \varphi \leq 1$.

En efecto, basta con tener en cuenta que ambos son aplicaciones particulares de r_{xy}.

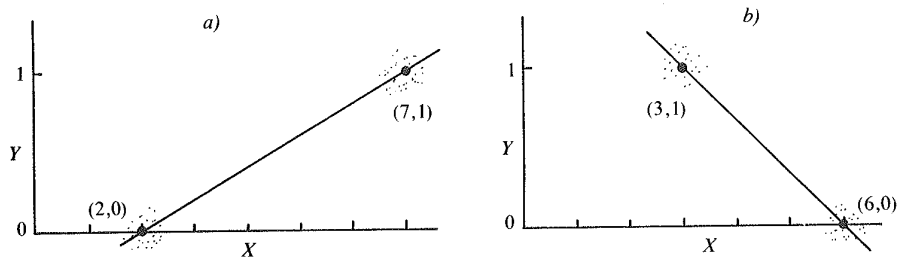
$r_{bp} = \pm 1$ cuando todas las personas con 1 en Y (variable dicotómica), obtengan la misma puntuación en X (variable cuantitativa continua) y cuando todas las personas con 0 en Y, obtengan la misma puntuación en X.

En particular, valdrá 1, cuando la única puntuación en X de los que obtienen 1 en Y sea mayor que la única puntuación en X de los que obtienen 0 en Y. Valdrá -1, cuando la única puntuación en X de los que obtienen 1 en Y sea menor que la de los que obtienen 0. Nótese que en ambos casos los puntos representativos de las personas están sobre una línea recta. Veamos dos ejemplos:

a') Todos los varones (1 en Y) obtienen 7 en X; todas las mujeres (0 en Y) obtienen 2 en X; r_{bp} será positivo.

a'') Todos los varones (1 en Y) obtienen 3 en X; todas las mujeres (0 en Y) obtienen 6 en X; r_{bp} será negativo.

He aquí las correspondientes representaciones gráficas:



$\varphi = 1$ cuando todas las personas con 1 en X obtienen 1 en Y, y todas con 0 en X obtienen 0 en Y.

$\varphi = -1$ cuando todas las personas con 1 en X obtienen 0 en Y, y todas con 0 en X obtienen 1 en Y.

He aquí los cuadros que representan estas dos situaciones extremas:

			X	
			0	1	
Y	1	0	b	b
	0	c	0	c
		c	b	n

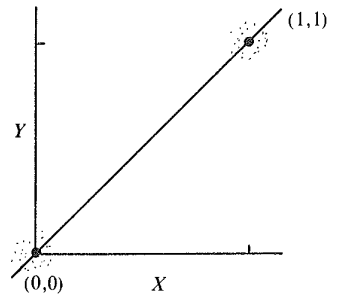
			X	
			0	1	
Y	1	a	0	a
	0	0	d	d
		a	d	n

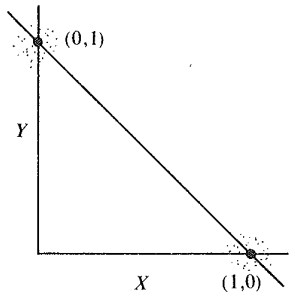
$$\varphi = \frac{cb - 0}{\sqrt{(b)(c)(b)(c)}} = \frac{cb}{cb} = 1$$

$$\varphi = \frac{0 - ad}{\sqrt{(a)(d)(a)(d)}}$$

$$= \frac{-ad}{ad} = -1$$

Estas dos situaciones extremas quedan representadas gráficamente así:





b) Dada una tabla con las cuatro frecuencias marginales fijas, el valor máximo alcanzable por φ es función de ellas.

Acabamos de ver que $\varphi = \pm 1$ sólo cuando las cuatro frecuencias marginales son (*b*, *c*, *b*, *c*) o (*a*, *d*, *a*, *d*). En cualquier otro caso $|\varphi| < 1$. Ahora bien, fijadas las cuatro frecuencias marginales, pueden variar las cuatro frecuencias interiores y, consiguientemente, φ puede ir tomando diversos valores. Lo que ahora nos interesa es determinar cuál de esos valores posibles es el máximo. Se demuestra que, fijadas las cuatro frecuencias marginales, ese valor máximo viene dado por la siguiente expresión:

$$\varphi_{\max} = \sqrt{\frac{n_j}{n_i}\frac{n'_i}{n'_j}} \quad \text{(en valor absoluto)}$$

Donde:

n_i es la frecuencia marginal máxima de las cuatro.
n_j es su complemento, es decir, $n - n_i$.
n'_i es la frecuencia mayor de las dos restantes.
n'_j es su complemento, es decir, $n - n'_i$.

Por ejemplo, el valor máximo para

	30 20
22 28	50

y

	20 30
22 28	50

es

$$\sqrt{\frac{20}{30}\frac{28}{22}} = 0{,}92$$

Este valor máximo es alcanzable, por ejemplo, con los dos cuadros siguientes:

2 28	30 20
20 0	
22 28	50

con $\varphi = 0{,}92$

20 0	20 30
2 28	
22 28	50

con $\varphi = -0{,}92$

c) Para una tabla de contingencia de dos filas y dos columnas, $\chi^2 = n\varphi^2$.

En efecto, basta con comparar la fórmula (15.13) de χ^2, con la fórmula (16.3) de φ (elevada al cuadrado).

16.2.4. Interpretación de r_{bp} y de φ

Según lo dicho en ocasiones semejantes, el único modo razonable de interpretar un valor determinado de r_{bp} o de φ es compararlo con los valores de r_{bp} y φ obtenidos por otros investigadores trabajando con las mismas o parecidas variables. Conviene, además, que sean iguales o parecidas las proporciones marginales de X en sus tablas de contingencia y en la nuestra, y que ocurra lo mismo con las proporciones marginales de Y.

16.3. Coeficientes de correlación que son estimación de r_{xy}

16.3.1. Coeficiente de correlación biserial, r_b

a) *Fundamento y fórmula*

Supongamos dos variables, X e Y, ambas continuas. Una de ellas (X) aparece como continua. La otra (Y) aparece dicotomizada artificialmente. Por tanto, en Y sólo tenemos, de hecho, dos categorías; es decir, unos y ceros. Pues bien, el coeficiente de correlación biserial, r_b, es una estimación de r_{xy} si Y se hubiera mantenido continua y se hubieran cumplido las dos condiciones siguientes:

a) La distribución de Y, considerada como continua, es normal.

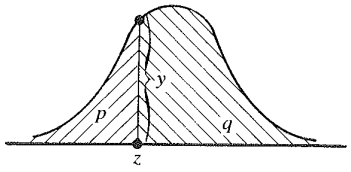
b) La relación entre X e Y (consideradas ambas como continuas) es lineal.

Se demuestra que esta estimación, bajo las dos condiciones acabadas de exponer, viene dada por:

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{s_x}\frac{pq}{y} = \tag{16.4}$$

$$= \frac{\bar{X}_p - \bar{X}}{s_x}\frac{p}{y} \tag{16.5}$$

En estas fórmulas $\bar{X}_p$, $\bar{X}_q$, $\bar{X}$, p, q tienen el mismo significado que en 16.2.1. Por otro lado, el significado de y es el siguiente: Supongamos una distribución normal con área unidad. Sea z la abscisa que deja a uno de sus lados un área igual a p y al otro un área igual a q. Pues bien, y es la ordenada correspondiente a dicha abscisa z.



No tenemos que preocuparnos por el cálculo de y, pues en la tabla B (Apéndice III) podemos encontrar directamente lo que vale p/y y pq/y, sabiendo lo que vale p.

La deducción de las fórmulas anteriores puede verse en Magnusson (1967) o en Lord y Novick (1968).

Por no alargarnos no repetimos las observaciones hechas al final de 16.2.1, al tratar de r_{bp}. Dichas observaciones, *mutatis mutandis*, pueden ser útiles, también, aquí.

b) *Cálculo*

Apliquemos (16.4) y (16.5) al siguiente ejemplo.

EJEMPLO 16.4. Llamemos X a la agudeza visual e Y a la habilidad para el oficio de relojero. Un comité de expertos califica como «buenos» o como «malos» profesionales a un grupo de relojeros. Éstos, a su vez, realizan una prueba de agudeza visual que les permite alcanzar puntuaciones que van de 0 a 30. X e Y son variables continuas, pero la segunda ha quedado dicotomizada artificialmente, ya que en ella sólo son posibles dos puntuaciones: una correspondiente a «buen» relojero y la otra correspondiente a «mal» relojero. En estas circunstancias parece muy oportuno el uso de la correlación biserial. Supongamos ahora que los datos numéricos obtenidos coinciden con los datos del ejemplo 16.1 a partir de los cuales hemos calculado r_{bp}. Lo hacemos así para poder comparar luego los resultados numéricos alcanzados en uno y otro caso. Llamemos n_p al número de «buenos» relojeros, n_q al de «malos» relojeros y n al número total de relojeros de nuestra muestra. Supuesto esto, apliquemos (16.4) y (16.5) al cuadro de la página anterior.

X	n_p	n_q	n
24-28	6	1	7
19-23	6	2	8
14-18	5	4	9
9-13	4	4	8
4-8	1	7	8
	22	18	40

Recordemos los resultados previos obtenidos en 16.2.1.

$$\bar{X} = 15{,}75 \quad , \quad \bar{X}_p = 18{,}73 \quad , \quad \bar{X}_q = 12{,}11 \quad , \quad s_x = 6{,}89$$
$$p = 0{,}55 \quad , \quad q = 0{,}45$$

De acuerdo con la tabla B (Apéndice III) $pq/y = 0{,}6253$, $p/y = 1{,}390$. Supuesto esto, aplicando (16.4)

$$r_b = \frac{18{,}73 - 12{,}11}{6{,}89}\,0{,}6253 = \frac{6{,}62}{6{,}89}\,0{,}6253 = \frac{4{,}1395}{6{,}89} = 0{,}60$$

y aplicando (16.5)

$$r_b = \frac{18{,}73 - 15{,}75}{6{,}89}\,1{,}390 = \frac{2{,}98}{6{,}89}\,1{,}390 = \frac{4{,}1422}{6{,}89} = 0{,}60$$

Esto nos indica que a ser altos en la prueba de agudeza visual, corresponde ser buenos relojeros; y a ser bajos en la misma, corresponde ser malos relojeros. No todo el que es alto es buen relojero, ni todo el que es bajo es mal relojero, pero, en general, los altos en la prueba, tienden a ser buenos relojeros y los bajos en la misma, tienden a ser malos relojeros.

Recuérdese aquí las observaciones hechas en este mismo contexto al tratar del cálculo de r_{bp} al final de 16.2.1.

16.3.2. Coeficiente de correlación tetracórica, r_t

a) *Fundamento y fórmulas*

Supongamos dos variables, X e Y, ambas continuas. Las dos aparecen dicotomizadas artificialmente. Por tanto, en X y en Y sólo tenemos, de hecho, dos categorías. Pues bien, el coeficiente de correlación tetracórica, r_t, es una estimación de r_{xy} si tanto X como Y se hubieran mantenido continuas y se hubieran cumplido las dos condiciones siguientes:

a') Las distribuciones de X e Y, consideradas como continuas, son normales.

b') La relación entre X e Y, consideradas ambas como continuas, es lineal.

Se demuestra que esta estimación viene dada por el siguiente desarrollo en serie de potencias en r_t que, por sencillez, llamaremos r en este desarrollo.

$$\frac{cb - ad}{n^2 yy'} = r + zz'\frac{r^2}{2} + (z^2 - 1)(z'^2 - 1)\frac{r^3}{6} + (z^3 - 3z)(z'^3 - 3z')\frac{r^4}{24} + \cdots$$

En este desarrollo a, b, c y d tienen el mismo significado que el presentado al tratar de φ (véase 16.2.2):

z: es una puntuación típica tal, que divide el área (de valor unidad) bajo la curva normal en dos áreas iguales a las dos proporciones $(a + c)/n$ y $(b + d)/n$.

z': es una puntuación típica tal, que divide el área (de valor unidad) bajo la curva normal en dos áreas iguales a las dos proporciones $(c + d)/n$ y $(a + b)/n$.

y e y': son las ordenadas que la curva normal hace corresponder respectivamente a las puntuaciones típicas z y z'.

El cálculo de r_t mediante el anterior desarrollo en serie de potencias es laborioso, especialmente cuando r_t es alto. En este caso no será muy aceptable la aproximación obtenida valiéndonos sólo de los dos primeros términos, es decir, mediante una ecuación de segundo grado en r_t. Necesitaremos tres o más términos, deberemos resolver una ecuación de grado superior al segundo. Debido a la dificultad de estos cálculos se han buscado otras fórmulas que ofrezcan buenas aproximaciones de r_t y cuya aplicación sea sencilla. Afortunadamente, no tenemos que preocuparnos ni por el desarrollo en serie de potencias, ni por las otras fórmulas subsidiarias. Existen tablas y diagramas que nos permiten calcular r_t de modo muy sencillo y, a la vez, bastante fiable en la mayoría de los casos. Entre ellos son muy conocidos los diagramas de Chesire, Saffir y Thurstone. Nosotros utilizaremos la tabla *C* (Apéndice III) que nos permitirá calcular r_t en unos momentos, conocidas las cuatro frecuencias a, b, c y d.

b) *Cálculo*

Llamemos X a la variable «rendimiento en aritmética» e Y a la variable «rendimiento en gramática». Atribuyamos un 1 a las personas que en aritmética y en gramática obtienen puntuaciones iguales o superiores a dos valores, uno en aritmética y otro en gramática. Son los «aprobados». Atribuyamos un 0 a las personas que en ambas asignaturas obtienen puntuaciones inferiores a los dos valores anteriores. Son los «suspensos». Por consiguiente, las dos variables, en sí continuas, aparecen dicotomizadas.

EJEMPLO 16.5. Supongamos que 200 personas se encuentran repartidas según el cuadro siguiente:

		X		
		0	1	
Y	1	30 (a)	60 (b)	90
	0	70 (c)	40 (d)	110
		100	100	200

Vamos a calcular r_t valiéndonos del desarrollo en serie de potencias, y de la tabla *C* (Apéndice III).

Utilizaremos sólo los dos primeros términos. Para ello necesitamos calcular z, z', y e y'.

En primer lugar,

$$\begin{aligned}(a + c)/n &= 100/200 = 0{,}50\\ (b + d)/n &= 100/200 = 0{,}50\\ (c + d)/n &= 110/200 = 0{,}55\\ (a + b)/n &= 90/200 = 0{,}45\end{aligned}$$

$z = 0$, pues esa puntuación típica divide el área (de valor unidad) bajo la curva normal en dos áreas, ambas iguales a 0,50.

$z' = 0{,}1257$, pues esa puntuación típica (de acuerdo con la tabla de las áreas bajo la curva normal) divide el área (de valor unidad) en dos áreas iguales a 0,45 y 0,55.

Las ordenadas correspondientes a z y z' son, respectivamente, 0,3989 y 0,3958. Véase la tabla *B*, columna F (Apéndice III). Allí 0,3989 corresponde a la proporción 0,50 y 0,3958 corresponde a la proporción 0,55.

En este supuesto,

$$\frac{(70)(60) - (30)(40)}{(200)^2(0{,}3989)(0{,}3958)} = r_t + (0)(0{,}1257)\frac{r_t^2}{2} = r_t + 0$$

Es decir:

$$\frac{3.000}{6.315{,}4} = 0{,}47 = r_t$$

o sea:

$$r_t = 0{,}47$$

Usando la tabla C, tenemos:

$$\frac{(c)(b)}{(a)(d)} = \frac{(70)(60)}{(30)(40)} = \frac{4.200}{1.200} = 3{,}5$$

En la tabla C vemos que a todos los valores comprendidos entre 3,461 y 3,571 les corresponde $r_t = 0{,}46$.

La tabla de frecuencias nos dirá qué signo debemos atribuir a un r_t determinado. Recuérdese a este respecto las consideraciones hechas al final de 16.2.1.

EJEMPLO 16.6. Vamos a analizar los datos presentados por Woodworth (1941) sobre 19 pares de gemelos univitelinos. Los dos elementos de cada par son educados en ambientes diferentes. Consideremos la diferencia en educación (X) entre los dos elementos de cada par y su diferencia en cociente intelectual (Y). Tendremos 19 diferencias en educación y 19 diferencias en cociente intelectual. Calculemos r_t, habiendo dicotomizado X e Y del modo siguiente:

		X: Diferencia en educación		
		Baja 0	Alta 1	
X: Diferencia en CI	Alta 1	1 (a)	8 (b)	9
	Baja 0	8 (c)	2 (d)	10
		9	10	19

$\frac{(8)(8)}{(1)(2)} = 32$. En la tabla C (Apéndice III) vemos que a 32 le corresponde $r_t = 0{,}89$.

Esto significa que el cociente intelectual es función de la educación recibida, ya que los elementos de cada par no difieren entre sí genéticamente, al ser univitelinos, sino sólo por razón del entorno distinto dentro del cual uno y otro han sido educados.

16.3.3. Propiedades de r_b y de r_t

a) El coeficiente de correlación biserial, r_b, puede valer más que uno (o menos que -1).

b) Para unos mismos datos, $|r_b|$ es siempre mayor que $|r_{bp}|$. En efecto:

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{s_x}\frac{pq}{y} = \frac{\bar{X}_p - \bar{X}_q}{s_x}\sqrt{pq}\frac{\sqrt{pq}}{y} = r_{bp}\frac{\sqrt{pq}}{y}$$

Ahora bien, de acuerdo con la tabla B, (Apéndice II), $\sqrt{pq}/y$ va de 3,733, para $p = 0{,}99$ (ó $p = 0{,}01$), a 1,253, para $p = 0{,}50$. Por consiguiente $\sqrt{pq}/y$ es siempre mayor que uno y, en consecuencia, r_b será siempre mayor que r_{bp} para unos mismos datos.

Recordemos cómo para los mismos datos del ejemplo 16.1, r_{bp} valía 0,48 y r_b valía 0,60 (ejemplo 16.4).

c) Para unos mismos datos, $r_t \simeq (3/2)(\varphi)$ (véase Lord y Novick, 1968). Esta aproximación es tanto mejor, cuanto más se aproximen a la mediana los puntos de dicotomización en ambas variables y cuando r_t es igual o menor que 0,50. Así, por ejemplo, compruebe el alumno cómo se verifica esta relación en el cuadro siguiente:

10	20	30
15	5	20
25	25	50

$r_t = 0{,}61$ (consultando tabla C del Apéndice II)
$\varphi = 0{,}408$ (mediante fórmula)
$(3/2)(\varphi) = (3/2)(0{,}408) = 0{,}612 \simeq 0{,}61 = r_t$

d) Siempre que una de las cuatro frecuencias interiores sea nula, $r_t = \pm 1$. En efecto, si uno de los cuatro valores a, b, c o d es nulo, claramente ad ó cb serán nulos. Consiguientemente valdrá cero uno de los dos cocientes cb/ad ó ad/cb. En ambos casos, al cociente nulo le corresponde en la tabla C, (Apéndice III) $r_t = 1$. El signo positivo o negativo dependerá de cuál de los cuatro valores es nulo.

16.3.4. Interpretación de r_b y de r_t

La interpretación razonable es compararlos con los valores de r_b y r_t obtenidos por otros investigadores trabajando con iguales o parecidas variables. Conviene, además, tener en cuenta los puntos de dicotomización usados por nosotros y por ellos.

16.4. Comparación de r_{bp} y de r_b

a) El coeficiente r_{bp} es una aplicación de r_{xy}. El coeficiente r_b es una estimación de r_{xy}, bajo ciertas condiciones.

b) El coeficiente $|r_{bp}|$ es siempre menor que el coeficiente $|r_b|$, para unos mismos datos.

c) El coeficiente $|r_{bp}|$ es siempre igual o menor que 1. El coeficiente $|r_b|$ puede superar ese valor.

d) Ambos coeficientes son usados en teoría de tests. Cada uno tiene sus ven-

tajas e inconvenientes, según sean las situaciones concretas en que van a ser usados y el aspecto que desea ser considerado. Esta discusión, en sí muy interesante, no la creemos oportuna ahora. El lector preocupado por este tema puede consultar Lord y Novick (1968).

16.5. Comparación de φ y de r_t

a) El coeficiente φ es una aplicación de r_{xy}. El coeficiente r_t es una estimación de r_{xy}, bajo ciertas condiciones.

b) El coeficiente $|\varphi|$ es igual o menor que $|r_t|$ para unos mismos datos. La diferencia tiende a aumentar a medida que los puntos de dicotomización en una o en las dos variables se van alejando de la mediana.

c) El valor máximo de $|\varphi|$ es función de las frecuencias marginales y sólo *puede* valer 1, cuando las frecuencias marginales en X son iguales a las frecuencias marginales en Y. Por el contrario, $|r_t|$ puede valer 1 con cualquier cuaterna de frecuencias marginales. En efecto, $r_t = 1$ con tal que sea 0 una cualquiera de las cuatro frecuencias interiores, lo cual es siempre posible, sean cualesquiera las cuatro frecuencias marginales. (Naturalmente, suponemos que estas frecuencias marginales son todas distintas de cero.)

d) El coeficiente φ es el más apropiado cuando las variables son estrictamente dicotómicas. El coeficiente r_t cuando, siendo continuas, se encuentran dicotomizadas.

e) Si las variables continuas aparecen como tales, calcúlese r_{xy} y no se introduzca dicotomización alguna. Es verdad que dicotomizando simplificamos mucho los cálculos, pero, a la vez, desaprovechamos mucha información que teníamos en nuestras manos. Utilicemos esta información. Hoy disponemos de pequeñas máquinas electrónicas de bolsillo que nos permiten calcular r_{xy} de modo muy sencillo.

g) Ambos coeficientes son usados en teoría de tests.

h) No parece recomendable calcular el cociente entre el φ obtenido y el φ máximo e interpretar dicho cociente como r_{xy}.

16.6. Resumen: Definiciones y fórmulas

Variable dicotómica: Aquella que sólo puede manifestarse según dos únicas modalidades.

Variable dicotomizada: Aquella que puede manifestarse según tres o más modalidades, pero a la que, de hecho, sólo se le permite manifestarse según dos modalidades.

Coeficiente de correlación biserial puntual: Aplicación de r_{xy} a dos variables, una continua y la otra dicotómica.

$$r_{bp} = \frac{\bar{X}_p - \bar{X}_q}{s_x}\sqrt{pq} = \frac{\bar{X}_p - \bar{X}}{s_x}\sqrt{\frac{p}{q}}$$

Coeficiente de correlación φ: Aplicación de r_{xy} a dos variables, ambas dicotómicas.

$$\varphi = \frac{cb - ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

Coeficiente de correlación biserial: Estimación de lo que habría valido r_{xy} entre dos variables continuas, obtenidas a partir de dichas variables manteniéndose una de ellas continua y habiendo sido dicotomizada la otra.

$$r_b = \frac{\bar{X}_p - \bar{X}_q}{s_x}\frac{pq}{y} = \frac{\bar{X}_p - \bar{X}}{s_x}\frac{p}{y}$$

Coeficiente de correlación tetracórica: Estimación de lo que habría valido r_{xy} entre dos variables continuas, obtenidas a partir de dichas variables dicotomizadas ambas artificialmente.

$$\frac{cb - ad}{n^2 yy'} = r_t + zz'\frac{r_t^2}{2} + (z^2 - 1)(z'^2 - 1)\frac{r_t^3}{6} + \cdots$$

16.7. Apéndice: Deducción de las fórmulas de r_{bp} y de φ a partir de r_{xy}

16.7.1. Deducción de la fórmula de r_{bp} a partir de r_{xy}

Sea Y la variable dicotómica. Sólo son posibles dos puntuaciones: 0 y 1. (Tan legítimo hubiera sido otro par cualquiera de números distintos.)

Llamemos p a la proporción de personas con 1 en Y, q a la proporción de personas con 0 en Y, n al número total de personas. Por consiguiente, tendremos np personas con 1 en Y, y nq personas con 0 en Y. De aquí son inmediatas las siguientes relaciones:

$$p + q = 1 \quad , \quad q = 1 - p \quad , \quad np + nq = n(p + q) = n$$

Calculemos $\bar{Y}$ y s_y.

$$\bar{Y} = \frac{(np)(1) + (nq)(0)}{n} = \frac{np}{n} = p$$

$$s_y^2 = \frac{(np)(1 - p)^2 + (nq)(0 - p)^2}{n} = \frac{npq^2 + nqp^2}{n} = pq^2 + qp^2 = pq(p + q) = pq$$

En conclusión:

$$\bar{Y} = p \quad , \quad s_y = \sqrt{pq} \tag{1}$$

Sea $\bar{X}_p$ la media en X de las np personas con 1 en Y.
Sea $\bar{X}_q$ la media en X de las nq personas con 0 en Y.
Sea s_x la desviación típica en X de las n personas.

Es claro que:

$$\Sigma XY = (X_1)(1) + \cdots + (X_{np})(1) + (X_{np+1})(0) + \cdots + (X_n)(0) = \Sigma X + 0 = \\ = \Sigma X, \text{ donde } \Sigma \text{ va de 1 a } np$$

Por consiguiente,

$$\frac{\Sigma XY}{np} = \frac{\Sigma X}{np} = \bar{X}_p$$

o sea,

$$\Sigma XY = np\bar{X}_p \tag{2}$$

Con estas condiciones y teniendo en cuenta (1) y (2),

$$r_{xy} = \frac{\Sigma xy}{ns_x s_y} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{ns_x s_y} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{ns_x s_y} = \frac{np\bar{X}_p - np\bar{X}}{ns_x\sqrt{pq}} =$$

$$= \frac{p\bar{X}_p - p\bar{X}}{s_x\sqrt{pq}} = \frac{\bar{X}_p - \bar{X}}{s_x}\,\frac{p}{\sqrt{pq}} = \frac{\bar{X}_p - \bar{X}}{s_x}\sqrt{p/q}$$

Observamos que esta fórmula es la (16.2).

Por otra parte:

$$\bar{X} = \frac{np\bar{X}_p + nq\bar{X}_q}{n} = p\bar{X}_p + q\bar{X}_q$$

Consiguientemente:

$$\bar{X}_p - \bar{X} = \bar{X}_p - p\bar{X}_p - q\bar{X}_q = (1 - p)\bar{X}_p - q\bar{X}_q = q\bar{X}_p - q\bar{X}_q = q(\bar{X}_p - \bar{X}_q)$$

Sustituyendo $q(\bar{X}_p - \bar{X}_q)$ por $\bar{X}_p - \bar{X}$ en la fórmula de r_{bp} acabada de obtener, nos queda:

$$r_{xy} = \frac{\bar{X}_p - \bar{X}}{s_x}\sqrt{p/q} = \frac{q(\bar{X}_p - \bar{X}_q)}{s_x}\sqrt{p/q} = \frac{\bar{X}_p - \bar{X}_q}{s_x}\sqrt{pq}$$

Observamos que esta fórmula es la (16.1).

16.7.2. Deducción de la fórmula de φ a partir de r_{xy}

Las dos variables, X e Y, son dicotómicas.

Sea p_x la proporción de personas con puntuación 1 en X.
Sea q_x la proporción de personas con puntuación 0 en X.
Sea p_y la proporción de personas con puntuación 1 en Y.
Sea q_y la proporción de personas con puntuación 0 en Y.

Hemos visto en 16.7.1 que la media y la desviación típica de la variable dicotómica, Y, valían $\bar{Y} = p$, $s_y = \sqrt{pq}$. Como ahora X e Y son dicotómicas,

$$\bar{X} = p_x \quad , \quad s_x = \sqrt{p_x q_x} \quad , \quad \bar{Y} = p_y \quad , \quad s_y = \sqrt{p_y q_y}$$

Es claro que, siendo np_{xy} el número de personas que obtienen 1 en X y 1 en Y,

$$\Sigma XY = \{(1)(1) + \cdots + (1)(1)\} + \{(1)(0) + \cdots + (1)(0)\} + \{(0)(1) + \cdots + \\ + (0)(1)\} + \{(0)(0) + \cdots + (0)(0)\} = (1)(1) + \cdots + (1)(1) + 0 + 0 + \\ + 0 = np_{xy}$$

Bajo este supuesto,

$$r_{xy} = \frac{\Sigma xy}{ns_x s_y} = \frac{\Sigma XY - n\bar{X}\bar{Y}}{ns_x s_y} = \frac{np_{xy} - np_x p_y}{n\sqrt{p_x q_x}\sqrt{p_y q_y}} = \frac{p_{xy} - p_x p_y}{\sqrt{p_x q_x p_y q_y}} \tag{3}$$

Recordando la tabla en 16.2.2.*a*,

$$p_x = \frac{b + d}{n}, \quad q_x = \frac{a + c}{n}, \quad p_y = \frac{a + b}{n}, \quad q_y = \frac{c + d}{n}, \quad p_{xy} = \frac{b}{n}$$

Sustituyendo estos valores en (3),

$$r_{xy} = \frac{\frac{b}{n} - \frac{(b + d)}{n}\,\frac{(a + b)}{n}}{\sqrt{\frac{(b + d)}{n}\,\frac{(a + c)}{n}\,\frac{(a + b)}{n}\,\frac{(c + d)}{n}}} = \frac{nb - b^2 - ab - ad - bd}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} =$$

$$= \frac{(a + b + c + d)(b) - b^2 - ab - ad - bd}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} =$$

$$= \frac{ab + b^2 + cb + bd - b^2 - ab - ad - bd}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} =$$

$$= \frac{cb - ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}}$$

que es precisamente la fórmula (16.3)

EJERCICIOS

16.1. Calcular el coeficiente de correlación biserial puntual, a partir de los siguientes datos *no agrupados* en intervalos (en la variable continua X).

a)

X	n_p	n_q
4	1	4
3	3	4
2	5	3
1	4	1

b)

X	n_p	n_q
5	6	2
4	6	4
3	8	10
2	2	6
1	2	4

c)

X	n_p	n_q
5	3	0
4	3	1
3	2	3
2	0	4
1	0	4

d)

X	n_p	n_q
5	4	0
4	6	2
3	6	6
2	2	8
1	2	4

16.2. Calcular el coeficiente de correlación biserial puntual, a partir de los siguientes datos *agrupados* en intervalos (en la variable continua X).

a)

X	n_p	n_q
45-51	2	0
38-44	5	3
31-37	6	4
24-30	5	7
17-23	5	8
10-16	1	4

b)

X	n_p	n_q
15-18	3	1
11-14	5	2
7-10	1	3

c)

X	n_p	n_q
15-17	4	2
12-14	6	4
9-11	7	5
6-8	5	5
3-5	3	5
0-2	0	4

d)

X	n_p	n_q
17-21	0	4
12-16	4	12
7-11	6	8
2-6	4	2

16.3. Teniendo en cuenta lo expuesto en 16.2.1, compruebe el lector cómo, en efecto, los valores de r_{bp} obtenidos en todos los casos de los dos ejercicios anteriores, no son más que una aplicación del coeficiente de correlación de Pearson.

16.4. Calcular el coeficiente φ a partir de las tablas de frecuencias siguientes:

a)

50	60
80	10

b)

36	30
24	30

c)

15	10
10	15

d)

70	100
130	100

e)

50	60
60	30

16.5. Teniendo en cuenta lo expuesto en 16.2.2, compruebe el lector cómo, en efectos, los valores de φ obtenidos en todos los casos del ejercicio anterior no son más que una aplicación del coeficiente de correlación de Pearson.

16.6. Calcular el coeficiente de correlación biserial, teniendo en cuenta los mismos cuadros a partir de los cuales ha calculado r_{bp} en el ejercicio 16.1. Ahora suponemos que X e Y son continuas, aunque Y aparece dicotomizada. Nótese cómo r_b es mayor (en valor absoluto) que r_{bp} para unos mismos datos numéricos.

16.7. Calcular el coeficiente de correlación biserial, teniendo en cuenta los mismos cuadros a partir de los cuales ha calculado r_{bp} en el ejercicio 16.2. Nótese, de nuevo, cómo r_b es mayor (en valor absoluto) que r_{bp} para unos mismos datos numéricos.

16.8. Calcular el coeficiente de correlación tetracórica, teniendo en cuenta los mismos cuadros a partir de los cuales ha calculado φ en el ejercicio 16.4. Ahora suponemos que X e Y son continuas, aunque ambas aparecen dicotomizadas. Compruébese cómo r_t es mayor (en valor absoluto) que φ para unos valores numéricos y, además, cómo se verifica la relación $|r_t| \simeq \frac{3}{2}|\varphi|$.

IV

Estudio conjunto de tres variables

17

Correlación y regresión

17.1. Introducción

Desde ahora estudiaremos conjuntamente tres variables. En el presente capítulo vamos a exponer tres temas:

a) Correlación existente entre dos variables, eliminando el influjo de la tercera (correlación parcial).

b) Pronóstico de las puntuaciones en una de ellas, conociendo las puntuaciones en las otras dos (regresión múltiple).

c) Correlación de una de ellas con las otras dos consideradas conjuntamente (correlación múltiple).

17.2. Correlación parcial

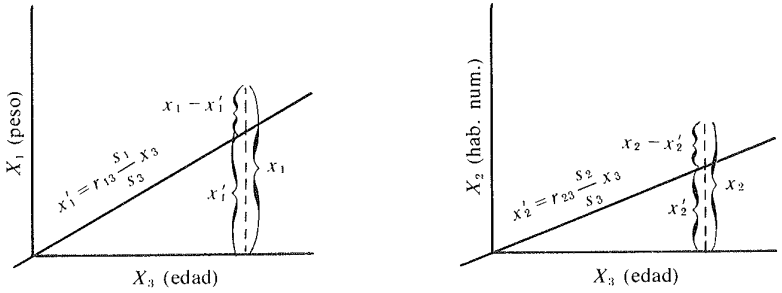
17.2.1. Fundamento y fórmula

En el capítulo 10, apartado 10.2.5*b*, considerábamos un grupo de niños cuyas edades oscilaban entre cuatro y doce años, planteándonos el problema de calcular el coeficiente de correlación de Pearson, respecto a dicho grupo, entre el peso y la habilidad en realizar operaciones aritméticas. Es claro que al ir aumentando su edad, van aumentando, simultáneamente, el peso y la habilidad numérica. En general, los niños de corta edad (de cuatro a seis años, por ejemplo) pesarán poco y mostrarán escasa habilidad en el manejo de números. En cambio, los más mayores (de diez a doce años, por ejemplo) pesarán más, ordinariamente, que los primeros y, a la vez, serán más hábiles que ellos en operar aritméticamente. Ello hace que el coeficiente de correlación de Pearson entre peso y habilidad numérica sea alto y positivo por culpa de la edad. Es, pues, necesario eliminar el influjo de ésta si deseamos captar la verdadera relación entre el peso y la habilidad en realizar operaciones aritméticas. ¿Cómo llevar a cabo esta eliminación? Vamos a presentar dos caminos posibles:

a) Eliminación empírica. Es decir, empíricamente formamos subgrupos, cada uno de ellos con niños de la misma o parecida edad y calculamos r_{xy} dentro de cada grupo. Así, la correlación entre peso y cálculo numérico quedará libre del influjo de la edad. Esta eliminación, no obstante, presenta dos inconvenientes. En primer lugar, deberemos calcular varios coeficientes de correlación (uno para cada subgrupo). En segundo lugar, cada uno de estos coeficientes habrá sido calculado con muy pocas personas y será poco fiable.

b) Eliminación estadística. Es decir, utilizamos los datos del grupo total de niños y eliminamos el influjo de la edad valiéndonos del siguiente razonamiento «estadístico».

Dentro de las edades propuestas, podemos suponer razonablemente que son lineales las relaciones entre peso y edad y entre habilidad numérica y edad. Por tanto, utilizando puntuaciones diferenciales, tendremos:



$$x_1' = r_{13}\frac{s_1}{s_3}x_3 \quad \text{(pronóstico del peso, mediante la edad)}$$

$$x_2' = r_{23}\frac{s_2}{s_3}x_3 \quad \text{(pronóstico de la habilidad numérica, mediante la edad).}$$

Donde s_1, s_2, s_3 son las desviaciones típicas del grupo total de niños en peso, habilidad numérica y edad, respectivamente; r_{13} y r_{23} son los coeficientes de correlación de Pearson obtenidos a partir del grupo total entre peso y edad, y entre habilidad numérica y edad, respectivamente.

Según ya sabemos (véase 11.3), $x_1 - x_1'$ y $x_2 - x_2'$ (errores en los pronósticos) no dependen de la variable predictora (edad).

Pues bien, calcular la correlación entre peso y habilidad numérica, independientemente de la edad, equivaldrá a calcular la correlación entre aquella parte del peso y aquella parte de la habilidad numérica que no dependen de la edad, o sea, entre $x_1 - x_1'$ y $x_2 - x_2'$.

Según lo visto en 11.1, tendremos ahora:

a)

$$r_{13}^2 = 1 - \frac{\Sigma\,(x_1 - x_1')^2/n}{s_1^2}$$

de donde, var $(x_1 - x_1') = \dfrac{\Sigma\,(x_1 - x_1')^2}{n} = s_1^2(1 - r_{13}^2)$ que designaremos por $s_{1.3}^2$

b)

$$r_{23}^2 = 1 - \frac{\Sigma\,(x_2 - x_2')^2/n}{s_2^2}$$

de donde, var $(x_2 - x_2') = \dfrac{\Sigma\,(x_2 - x_2')^2}{n} = s_2^2(1 - r_{23}^2)$ que designaremos por $s_{2.3}^2$

Teniendo en cuenta estas consideraciones previas, el coeficiente de correlación de Pearson entre el peso (X_1) y el cálculo numérico (X_2), eliminado el influjo de la edad (X_3), (y que designaremos por $r_{12.3}$), vendrá dado por

$$r_{12.3} = r_{(x_1 - x_1')(x_2 - x_2')} = \frac{\Sigma\,(x_1 - x_1')(x_2 - x_2')}{n\,s_{1.3}\,s_{2.3}} = \frac{\Sigma\left(x_1 - r_{13}\dfrac{s_1}{s_3}x_3\right)\left(x_2 - r_{23}\dfrac{s_2}{s_3}x_3\right)}{n\,s_{1.3}\,s_{2.3}} =$$

$$= \frac{\Sigma\,x_1x_2 - r_{23}\dfrac{s_2}{s_3}\Sigma\,x_1x_3 - r_{13}\dfrac{s_1}{s_3}\Sigma\,x_2x_3 + r_{13}r_{23}\dfrac{s_1s_2}{s_3s_3}\Sigma\,x_3^2}{n\,s_1s_2\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} =$$

$$= \frac{r_{12} - r_{23}r_{13} - r_{13}r_{23} + r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}} \qquad (17.1)$$

17.2.2. Cálculo

Es una mera aplicación de (17.1).

EJEMPLO 17.1. Supongamos que hemos calculado, de modo ordinario, r_{12} entre el peso (X_1) de un grupo de niños y su habilidad en el manejo de símbolos numéricos (X_2), obteniendo $r_{12} = 0{,}50$. Calculemos, ahora, r_{13} entre peso (X_1) y edad (X_3) y r_{23} entre habilidad numérica (X_2) y edad (X_3). Los siguientes valores hipotéticos son muy razonables: $r_{13} = 0{,}80$, $r_{23} = 0{,}60$. Esto supuesto, veamos lo que sucede aplicando (17.1).

$$r_{12.3} = \frac{0{,}50 - (0{,}80)(0{,}60)}{\sqrt{1 - 0{,}64}\sqrt{1 - 0{,}36}} = \frac{0{,}02}{(0{,}60)(0{,}80)} = 0{,}042$$

En conclusión, el coeficiente de correlación de Pearson entre peso y habilidad numérica, en vez de valer 0,50, es prácticamente nulo, en cuanto hemos eliminado el influjo de la edad.

Este y otros muchos ejemplos nos enseñan con qué cautela debemos proceder en algunos casos al analizar el valor de r_{xy} encontrado entre dos variables. Es posible una relación verdadera entre las mismas. Pero, también, es posible que esa relación sea sólo aparente y se deba totalmente o, al menos, en gran parte, al influjo de otra u otras variables.

17.2.3. Propiedades

Señalaremos como única propiedad del coeficiente de correlación parcial, que no puede valer menos que -1, ni más que 1. En efecto, dicho coeficiente no es más que un coeficiente de correlación de Pearson.

17.3. Regresión múltiple

17.3.1. Introducción

Recordemos el caso de dos variables. Allí teníamos una variable predictora, X, y un criterio, Y, que ahora nos conviene designar por X_2 y X_1, respectivamente. La recta de regresión de X_1 sobre X_2 era una recta tal que hacía mínima la suma de errores cuadráticos, $\Sigma (X_1 - X_1')^2$, es decir, la suma de las diferencias cuadráticas entre las puntuaciones obtenidas, de hecho, en el criterio y las pronosticadas en el mismo mediante dicha recta. Ahora tenemos dos variables predictoras, X_2 y X_3, y un criterio, X_1. Pues bien, en vez de la recta de regresión de X_1 sobre X_2, tendremos un plano de regresión de X_1 sobre X_2 y X_3 tal que haga mínima la suma de errores cuadráticos, $\Sigma (X_1 - X_1')^2$, es decir, la suma de las diferencias cuadráticas entre las puntuaciones obtenidas, de hecho, en el criterio y las puntuaciones pronosticadas en el mismo mediante dicho plano.

Hablaremos de regresión múltiple, siempre que aparezcan dos o más variables predictoras. Por ahora nos limitaremos al caso de sólo dos variables.

17.3.2. Ecuación del plano en un espacio tridimensional

La ecuación de la recta en el plano (espacio bidimensional) era $Y = a + bX$ (ó $X_1 = a + b_2X_2$). La ecuación del plano en un espacio tridimensional viene dada por $X_1 = a + b_2X_2 + b_3X_3$, donde a, b_2 y b_3 son tres constantes características de cada plano. La constante a nos determina el punto en que el plano corta el eje OX_1, o sea, el punto de coordenadas $(a, 0,0)$. Cuando $a = 0$, el plano pasará por el origen de coordenadas (0,0,0) y recíprocamente. Las constantes b_2 y b_3 nos determinan la inclinación del plano.

17.3.3. Ecuaciones de los planos de regresión de X_1 sobre X_2 y X_3 según el criterio de mínimos cuadrados

Supongamos un grupo de n personas de las que conocemos sus puntuaciones en las dos variables predictoras, X_2 y X_3, y en el criterio, X_1. Por ejemplo, en dos tests de aptitud y en el examen final de curso. Consideremos tres ejes ortogonales, OX_1, OX_2 y OX_3. (Véase figura 17.1.) Cada persona i viene representada por un punto

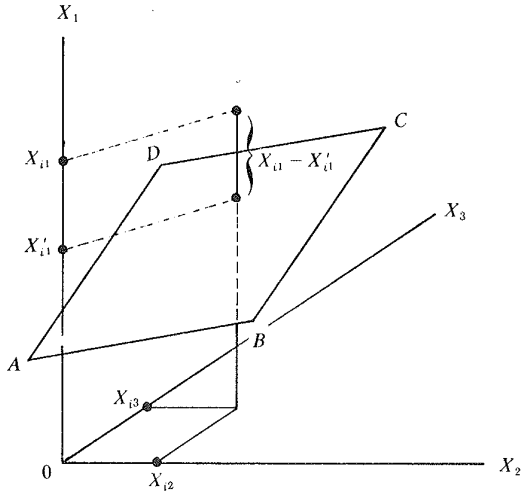


Figura 17.1

cuyas coordenadas (X_{i1}, X_{i2}, X_{i3}) no son más que las puntuaciones obtenidas en las dos variables predictoras y en el criterio. El grupo total de las n personas vendrá representado por una nube de puntos que será de forma elipsoidal (algo así como un balón de rugby) si existe correlación entre las tres variables. Sea un plano $ABCD$. Hagamos pasar por el punto (X_{i1}, X_{i2}, X_{i3}) una perpendicular al plano determinado por los ejes OX_2 y OX_3. Esta perpendicular cortará al plano $ABCD$ en un un punto $(X'_{i1}, X_{i2}, X_{i3})$. Si a la persona i que ha obtenido, de hecho, la puntuación X_{i1} en el criterio le atribuimos la puntuación X'_{i1} (dada por el plano $ABCD$), cometeremos el error $(X_{i1} - X'_{i1})$ y, consiguientemente, el error cuadrático $(X_{i1} - X'_{i1})^2$. Haciendo lo mismo con las restantes personas, tendremos n errores cuadráticos y, por tanto, una suma de errores cuadráticos expresada por $\Sigma (X_{i1} - X'_{i1})^2$.

Con cada plano del espacio irá asociada una suma de errores cuadráticos. Pues bien, el plano de regresión de X_1 sobre X_2 y X_3 es aquel plano cuya suma de errores cuadráticos asociada, $\Sigma (X_{i1} - X'_{i1})^2$, es más pequeña que la suma de errores cuadráticos asociada a cualquier otro de los infinitos planos del espacio. Más brevemente, es aquel plano que hace mínima la suma de errores cuadráticos $\Sigma (X_{i1} - X'_{i1})^2$.

Nótese que para construir el plano de regresión necesitamos un grupo de personas cuyas puntuaciones en X_1, en X_2 y en X_3 debemos conocer. En cambio, lo aplicaremos a otras personas, *semejantes* a las anteriores, de las que sólo conoceremos sus puntuaciones en X_2 y en X_3. Supongamos que X_2 y X_3 son dos tests de aptitud para la Estadística y X_1 el rendimiento en la misma, manifestado mediante un examen. Queremos construir la ecuación del plano que nos permita pronosticar del mejor modo posible el rendimiento, conocido el resultado en ambos tests. Pues bien, para construir ese plano, necesitamos unas personas cuyas puntuaciones en los dos tests y en el examen nos sean conocidas. Una vez construido, lo aplicamos a otras personas, *semejantes* a las anteriores, de las que sólo conoceremos sus puntuaciones en los dos tests de aptitud.

Dada la semejanza entre los dos grupos de personas, es de esperar que el plano de regresión que fue óptimo en la reducción de los errores cuadráticos respecto al primer grupo, será, también, razonablemente bueno en la reducción de los errores cuadráticos respecto al segundo.

Expuestas estas consideraciones previas, veamos cuál es el plano de regresión de X_1 sobre X_2 y X_3.

a) *Expresado en puntuaciones directas*

Comenzamos con la ecuación

$$X'_1 = A + B_2X_2 + B_3X_3 \tag{17.2}$$

Nuestro propósito es determinar A, B_2 y B_3 de modo que $\Phi \equiv \Sigma (X_1 - A - B_2X_2 - B_3X_3)^2$ sea mínima. Según se demuestra en Cálculo, ello equivale a

resolver las tres ecuaciones $\frac{\partial \Phi}{\partial A} = 0$, $\frac{\partial \Phi}{\partial B_2} = 0$, $\frac{\partial \Phi}{\partial B_3} = 0$ donde $\frac{\partial \Phi}{\partial A}$, $\frac{\partial \Phi}{\partial B_2}$, $\frac{\partial \Phi}{\partial B_3}$ son las derivadas parciales de Φ respecto a A, a B_2 y a B_3. Es decir,

$$\frac{\partial \Sigma (X_1 - A - B_2X_2 - B_3X_3)^2}{\partial A} = -2 \Sigma (X_1 - A - B_2X_2 - B_3X_3) = 0$$

$$\frac{\partial \Sigma (X_1 - A - B_2X_2 - B_3X_3)^2}{\partial B_2} = -2 \Sigma (X_1 - A - B_2X_2 - B_3X_3)X_2 = 0$$

$$\frac{\partial \Sigma (X_1 - A - B_2X_2 - B_3X_3)^2}{\partial B_3} = -2 \Sigma (X_1 - A - B_2X_2 - B_3X_3)X_3 = 0$$

O, lo que es equivalente:

$$\Sigma (X_1 - A - B_2X_2 - B_3X_3) = 0, \text{ de donde,}$$
$$\Sigma X_1 = nA + B_2 \Sigma X_2 + B_3 \Sigma X_3 \tag{17.3}$$

$$\Sigma (X_1 - A - B_2X_2 - B_3X_3)X_2 = 0, \text{ de donde,}$$
$$\Sigma X_1X_2 = A \Sigma X_2 + B_2 \Sigma X_2^2 + B_3 \Sigma X_2X_3 \tag{17.4}$$

$$\Sigma (X_1 - A - B_2X_2 - B_3X_3)X_3 = 0, \text{ de donde,}$$
$$\Sigma X_1X_3 = A \Sigma X_3 + B_2 \Sigma X_2X_3 + B_3 \Sigma X_3^2 \tag{17.5}$$

Las ecuaciones (17.3), (17.4) y (17.5) suelen ser llamadas normales. Ellas nos permiten despejar A, B_2 y B_3 (véase NOTA 3, al final de este capítulo). Sus valores son:

$$A = \bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3 \tag{17.6}$$

$$B_2 = \frac{[n \Sigma X_1X_2 - \Sigma X_1 \Sigma X_2][n \Sigma X_3^2 - (\Sigma X_3)^2]}{[n \Sigma X_2^2 - (\Sigma X_2)^2][n \Sigma X_3^2 - (\Sigma X_3)^2] - [n \Sigma X_2X_3 - \Sigma X_2 \Sigma X_3]^2} -$$
$$- \frac{[n \Sigma X_1X_3 - \Sigma X_1 \Sigma X_3][n \Sigma X_2X_3 - \Sigma X_2 \Sigma X_3]}{[n \Sigma X_2^2 - (\Sigma X_2)^2][n \Sigma X_3^2 - (\Sigma X_3)^2] - [n \Sigma X_2X_3 - \Sigma X_2 \Sigma X_3]^2} \tag{17.7}$$

$$B_3 = \frac{[n \Sigma X_1X_3 - \Sigma X_1 \Sigma X_3][n \Sigma X_2^2 - (\Sigma X_2)^2]}{[n \Sigma X_3^2 - (\Sigma X_3)^2][n \Sigma X_2^2 - (\Sigma X_2)^2] - [n \Sigma X_3X_2 - \Sigma X_3 \Sigma X_2]^2} -$$
$$- \frac{[n \Sigma X_1X_2 - \Sigma X_1 \Sigma X_2][n \Sigma X_3X_2 - \Sigma X_3 \Sigma X_2]}{[n \Sigma X_3^2 - (\Sigma X_3)^2][n \Sigma X_2^2 - (\Sigma X_2)^2] - [n \Sigma X_3X_2 - \Sigma X_3 \Sigma X_2]^2} \tag{17.8}$$

Teniendo en cuenta (17.2) y (17.6), nos queda

$$X'_1 = A + B_2X_2 + B_3X_3 = (\bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3) + B_2X_2 + B_3X_3 \tag{17.9}$$

De (17.9) se infieren inmediatamente las siguientes consecuencias:

1) $$\bar{X}'_1 = \frac{\Sigma X'_1}{n} = (\bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3) + B_2 \frac{\Sigma X_2}{n} + B_3 \frac{\Sigma X_3}{n} =$$
$$= \bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3 + B_2\bar{X}_2 + B_3\bar{X}_3 = \bar{X}_1 \tag{17.10}$$

Es decir, son iguales la media de las puntuaciones directas pronosticadas, $\bar{X}'_1$, y la media de las puntuaciones directas obtenidas, $\bar{X}_1$.

2) Sustituyendo X_2 por $\bar{X}_2$ y X_3 por $\bar{X}_3$ en (17.9),

$$X'_1 = \bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3 + B_2\bar{X}_2 + B_3\bar{X}_3 = \bar{X}_1 \qquad (17.11)$$

Es decir, el plano de regresión de X_1 sobre X_2 y X_3, en puntuaciones directas, pasa por el punto $(\bar{X}_1, \bar{X}_2, \bar{X}_3)$. (Lo mismo les sucede al plano de regresión de X_2 sobre X_1 y X_3 y al plano de regresión de X_3 sobre X_1 y X_2.)

b) *Expresado en puntuaciones diferenciales*

Comenzamos con la ecuación

$$x'_1 = a + b_2x_2 + b_3x_3 \qquad (17.12)$$

Nuestro propósito es determinar a, b_2 y b_3 de modo que $\Sigma\,(x_1 - a - b_2x_2 - b_3x_3)^2$ sea mínima. Según un razonamiento análogo al seguido en el caso de las puntuaciones directas, llegamos a:

$$\Sigma\, x_1 = na + b_2\,\Sigma\, x_2 + b_3\,\Sigma\, x_3 \qquad (17.13)$$
$$\Sigma\, x_1x_2 = a\,\Sigma\, x_2 + b_2\,\Sigma\, x_2^2 + b_3\,\Sigma\, x_2x_3 \qquad (17.14)$$
$$\Sigma\, x_1x_3 = a\,\Sigma\, x_3 + b_2\,\Sigma\, x_2x_3 + b_3\,\Sigma\, x_3^2 \qquad (17.15)$$

Las ecuaciones (17.13), (17.14) y (17.15) suelen ser llamadas normales. Ellas nos permiten despejar a, b_2 y b_3 (véase NOTA 1 al final de este capítulo). Sus valores son:

$$a = 0 \qquad (17.16)$$

$$b_2 = \frac{\Sigma\, x_1x_2\,\Sigma\, x_3^2 - \Sigma\, x_1x_3\,\Sigma\, x_2x_3}{\Sigma\, x_2^2\,\Sigma\, x_3^3 - (\Sigma\, x_2x_3)^2} \qquad (17.17)$$

$$b_3 = \frac{\Sigma\, x_1x_3\,\Sigma\, x_2^2 - \Sigma\, x_1x_2\,\Sigma\, x_3x_2}{\Sigma\, x_3^2\,\Sigma\, x_2^2 - (\Sigma\, x_3x_2)^2} \qquad (17.18)$$

Recordando que $\Sigma\, x_ix_j = nr_{ij}s_is_j$ y que $r_{ii} = 1$, nos quedará:

$$b_2 = \frac{nr_{12}s_1s_2ns_3^2 - nr_{13}s_1s_3nr_{23}s_2s_3}{ns_2^2ns_3^2 - (nr_{23}s_2s_3)^2} = \frac{s_1}{s_2}\left(\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}\right) \qquad (17.19)$$

$$b_3 = \frac{nr_{13}s_1s_3ns_2^2 - nr_{12}s_1s_2nr_{23}s_2s_3}{ns_3^2ns_2^2 - (nr_{23}s_2s_3)^2} = \frac{s_1}{s_3}\left(\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}\right) \qquad (17.20)$$

Teniendo en cuenta (17.12) y (17.16), nos queda

$$x'_1 = a + b_2x_2 + b_3x_3 = 0 + b_2x_2 + b_3x_3 \qquad (17.21)$$

De (17.21) se deducen inmediatamente las siguientes consecuencias

1) $$\bar{x}'_1 = b_2\frac{\Sigma\, x_2}{n} + b_3\frac{\Sigma\, x_3}{n} = (b_2)(0) + (b_3)(0) = 0 \qquad (17.22)$$

Es decir, la media de las puntuaciones diferenciales pronosticadas, $\bar{x}'_1$, vale 0, lo mismo que la media de las puntuaciones diferenciales obtenidas, $\bar{x}_1 = 0$.

2) Sustituyendo x_2 por $\bar{x}_2 = 0$ y x_3 por $\bar{x}_3 = 0$ en (17.21), $x'_1 = 0$ (17.23)

Es decir, el plano de regresión de X_1 sobre X_2 y X_3, en puntuaciones diferenciales, pasa por el origen (0,0,0). (Lo mismo les sucede al plano de regresión de X_2 sobre X_1 y X_3 y al plano de regresión de X_3 sobre X_1 y X_2.)

Nótese que:

$$n\,\Sigma\, x_ix_j = n\,\Sigma\,(X_i - \bar{X})(X_j - \bar{X}) = n\,\Sigma\, X_iX_j - n^2\bar{X}_i\bar{X}_j = n\,\Sigma\, X_iX_j - \Sigma\, X_i\,\Sigma\, X_j$$

De donde,

$$n\,\Sigma\, x_i^2 = n\,\Sigma\, X_i^2 - (\Sigma\, X_i)^2$$

Por tanto, teniendo en cuenta estas relaciones en (17.7) y en (17.8), y recordando (17.17) y (17.18), nos queda

$$B_2 = \frac{\not{n}\,\Sigma\, x_1x_2\not{n}\,\Sigma\, x_3^2 - \not{n}\,\Sigma\, x_1x_3\not{n}\,\Sigma\, x_2x_3}{\not{n}\,\Sigma\, x_2^2\not{n}\,\Sigma\, x_3^2 - \not{n}^2\,(\Sigma\, x_2x_3)^2} = b_2 \qquad (17.24)$$

$$B_3 = \frac{\not{n}\,\Sigma\, x_1x_3\not{n}\,\Sigma\, x_2^2 - \not{n}\,\Sigma\, x_1x_2\not{n}\,\Sigma\, x_3x_2}{\not{n}\,\Sigma\, x_3^2\not{n}\,\Sigma\, x_2^2 - \not{n}^2(\Sigma\, x_3x_2)^2} = b_3 \qquad (17.25)$$

Las igualdades (17.24) y (17.25) nos indican que son paralelos el plano de regresión en puntuaciones diferenciales y el correspondiente plano de regresión en puntuaciones directas.

c) *Expresado en puntuaciones típicas*

Comenzamos con la ecuación

$$z'_1 = a^* + b_2^*z_2 + b_3^*z_3 \qquad (17.26)$$

Nuestro propósito es determinar a^*, b_2^* y b_3^* de modo que $\Sigma (z_1 - a^* - b_2^* z_2 - b_3^* z_3)^2$ sea mínima. Según un razonamiento análogo al seguido en el caso de las puntuaciones directas llegamos a

$$\Sigma z_1 = na^* + b_2^* \Sigma z_2 + b_3^* \Sigma z_3 \tag{17.27}$$

$$\Sigma z_1 z_2 = a^* \Sigma z_2 + b_2^* \Sigma z_2^2 + b_3^* \Sigma z_2 z_3 \tag{17.28}$$

$$\Sigma z_1 z_3 = a^* \Sigma z_3 + b_2^* \Sigma z_2 z_3 + b_3^* \Sigma z_3^2 \tag{17.29}$$

Las ecuaciones (17.27), (17.28) y (17.29) suelen ser llamados normales. Ellas nos permiten despejar a^*, b_2^* y b_3^*. (Véase NOTA 1 al final de este capítulo.) Sus valores son:

$$a^* = 0 \tag{17.30}$$

$$b_2^* = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \tag{17.31}$$

$$b_3^* = \frac{r_{13} - r_{12} r_{23}}{1 - r_{23}^2} \tag{17.32}$$

Teniendo en cuenta (17.26) y (17.30), nos queda

$$z_1' = a^* + b_2^* z_2 + b_3^* z_3 = 0 + b_2^* z_2 + b_3^* z_3 \tag{17.33}$$

De (17.33) se deducen inmediatamente las siguientes consecuencias:

1) $$\bar{z}_1' = \frac{\Sigma z_1'}{n} = b_2^* \frac{\Sigma z_2}{n} + b_3^* \frac{\Sigma z_3}{n} = (b_2^*)(0) + (b_3^*)(0) = 0 \tag{17.34}$$

Es decir, la media de las puntuaciones típicas pronosticadas, $\bar{z}_1'$, vale 0, lo mismo que la media de las puntuaciones típicas obtenidas, $\bar{z}_1 = 0$.

2) Sustituyendo z_2 por $\bar{z}_2 = 0$ y z_3 por $\bar{z}_3 = 0$, nos queda $z_1' = 0$.

Es decir, el plano de regresión de X_1 sobre X_2 y X_3, en puntuaciones típicas, pasa por el origen (0,0,0). (Lo mismo les sucede a los planos de regresión de X_2 sobre X_1 y X_3 y de X_3 sobre X_1 y X_2.)

Nótese que, según (17.19), (17.20), (17.31) y (17.32),

$$b_2 = \frac{s_1}{s_2} b_2^* \tag{17.35}$$

$$b_3 = \frac{s_1}{s_3} b_3^* \tag{17.36}$$

En el capítulo siguiente veremos que $s_{z_1'}^2 = \frac{\Sigma z_1'^2}{n} \leq 1$. En otras palabras, las puntuaciones z_1' no cumplen con una de las condiciones necesarias con las que cumplen las auténticas puntuaciones típicas, a saber, que su varianza vale siempre 1. Por esta razón, las puntuaciones z_1' deberían ser llamadas pseudotípicas.

EJEMPLO 17.2. Comenzaremos introduciendo un miniejemplo que ayude al lector a comprender mejor la aplicación de las fórmulas anteriores. Después ofreceremos otro ejemplo algo más largo, con unos datos obtenidos en la vida real.

Supongamos cinco personas con puntuaciones en dos variables predictoras, X_2 y X_3, y un criterio X_1, según las tres tablas siguientes:

TABLA 17.1
(*puntuaciones directas*)

X_1	X_2	X_3	X_1X_2	X_1X_3	X_2X_3	X_1^2	X_2^2	X_3^2	X_1'
5	9	8	45	40	72	25	81	64	4,2
1	0	4	0	4	0	1	0	16	1,4
3	6	0	18	0	0	9	36	0	2,6
7	18	12	126	84	216	49	324	144	7,0
4	12	6	48	24	72	16	144	36	4,8
20	45	30	237	152	360	100	585	260	20,0

TABLA 17.2
(*puntuaciones diferenciales*)

x_1	x_2	x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_3^2			x_1'
1	0	2	0	2	0	1	0	4	$s_1^2 = 4$	$s_1 = 2$	0,2
−3	−9	−2	27	6	18	9	81	4	$s_2^2 = 36$	$s_2 = 6$	−2,6
−1	−3	−6	3	6	18	1	9	36	$s_3^2 = 16$	$s_3 = 4$	−1,4
3	9	6	27	18	54	9	81	36			3,0
0	3	0	0	0	0	0	9	0			0,8
0	0	0	57	32	90	20	180	80			0,0

TABLA 17.3
(*puntuaciones típicas*)

z_1	z_2	z_3	z_1z_2	z_1z_3	z_2z_3	z_1'	$z_1'^2$
0,5	0,0	0,5	0,00	0,25	0,00	0,1	0,01
−1,5	−1,5	−0,5	2,25	0,75	0,75	−1,3	1,69
−0,5	−0,5	−1,5	0,25	0,75	0,75	−0,7	0,49
1,5	1,5	1,5	2,25	2,25	2,25	1,5	2,25
0,0	0,5	0,0	0,00	0,00	0,00	0,4	0,16
0,0	0,0	0,0	4,75	4,00	3,75	0,0	4,60

$$r_{12} = \frac{4,75}{5} = 0,95, \qquad r_{13} = \frac{4}{5} = 0,80, \qquad r_{23} = \frac{3,75}{5} = 0,75$$

a) *Plano de regresión a partir de puntuaciones directas*

Según (17.7):

$$B_2 = \frac{[(5)(237) - (20)(45)][(5)(260) - (30)^2]}{[(5)(585) - (45)^2][(5)(260) - (30)^2] - [(5)(360) - (45)(30)]^2} - \frac{[(5)(152) - (20)(30)][(5)(360) - (45)(30)]^2}{[(5)(585) - (45)^2][(5)(260) - (30)^2] - [(5)(360) - (45)(30)]^2} = \frac{42.000}{157.500} = 0,267$$

Según (17.8):

$$B_3 = \frac{[(5)(152) - (20)(30)][(5)(585) - (45)^2]}{157.500} - \frac{[(5)(237) - (20)(45)][(5)(360) - (30)(45)]}{157.500} = \frac{15.750}{157.500} = 0,100$$

Según (17.6):

$$A = 4 - (0,267)(9) - (0,1)(6) = 0,997$$

Por consiguiente,

$$X_1' = 0,997 + (0,267)\,X_2 + (0,100)\,X_3$$

Aplicando esta ecuación a los correspondientes valores X_2 y X_3 obtenemos la columna encabezada por X_1' en la tabla 17.1.

Nótese cómo:

$$\Sigma X_1 = \Sigma X_1' = 20, \quad \text{ó,} \quad \bar{X}_1 = \bar{X}_1' = 4$$

y cómo:

$$X_1' = 0,997 + (0,267)(\bar{X}_2) + (0,1)(\bar{X}_3) = 0,997 + (0,267)(9) + (0,1)(6) = 4 = \bar{X}_1$$

b) *Plano de regresión a partir de puntuaciones diferenciales*

Según (17.24): $b_2 = B_2 = 0,267$.
Según (17.25): $b_3 = B_3 = 0,100$.
Según (17.16): $a = 0$.

Por consiguiente,

$$x_1' = (0,267)\,x_2 + (0,100)\,x_3$$

Aplicando esta ecuación a los correspondientes valores de x_2 y x_3, obtenemos la columna encabezada por x_1' en la tabla 17.2.
Nótese cómo:

$$\Sigma x_1 = \Sigma x_1' = 0, \quad \text{ó,} \quad \bar{x}_1 = \bar{x}_1' = 0$$

y cómo

$$x_1' = (0,267)(\bar{x}_2) + (0,1)(\bar{x}_3) = (0,267)(0) + (0,1)(0) = 0 = \bar{x}_1$$

c) *Plano de regresión a partir de puntuaciones típicas*

Según (17.31): $b_2^* = \dfrac{0,95 - (0,80)(0,75)}{1 - (0,75)^2} = \dfrac{0,35}{0,4375} = 0,8$

Según (17.32): $b_3^* = \dfrac{0,80 - (0,95)(0,75)}{1 - (0,75)^2} = \dfrac{0,0875}{0,4375} = 0,2$

Según (17.30): $a^* = 0$.

Por consiguiente:

$$z_1' = (0,8)\,z_2 + (0,2)\,z_3$$

Aplicando esta ecuación a los correspondientes valores z_2 y z_3, obtenemos la columna encabezada por z_1' en la tabla 17.3.
Nótese cómo:

$$\Sigma z_1 = \Sigma z_1' = 0, \quad \text{ó,} \quad \bar{z}_1 = \bar{z}_1' = 0$$

y cómo:

$$z_1' = (0,8)(\bar{z}_2) + (0,2)(\bar{z}_3) = 0 = \bar{z}_1$$

EJEMPLO 17.3. En el ejemplo 11.2 hemos construido las rectas de regresión de X_1 (rendimiento escolar) sobre X_2 (razonamiento abstracto) a partir de las puntuaciones obtenidas en X_1 y X_2 por 22 alumnos de Enseñanza General Básica. Vamos ahora a considerar como nueva variable predictora la comprensión verbal (X_3). La tabla 17.4 nos presenta las puntuaciones obtenidas por dichos 22 alumnos en X_1, X_2 y X_3. Construyamos los planos de regresión de X_1 sobre X_2 y X_3.

TABLA 17.4

X_1	X_2	X_3	X'_1	x'_1	z'_1
4,3	2	16	2,8228	−3,2499	−1,9633
7,5	20	27	8,1876	2,1149	1,2777
5,0	12	17	4,9640	−1,1087	−0,6698
6,9	21	20	7,2359	1,1632	−0,2305
5,2	19	13	5,6911	−0,3816	0,7028
4,4	17	18	6,1167	0,0440	0,0266
6,3	14	17	5,3594	−0,7133	−0,4309
5,6	13	23	6,1469	0,0742	0,0448
8,7	21	22	7,5643	1,4916	0,9011
7,6	19	18	6,5121	0,4394	0,2654
7,0	14	18	5,5236	−0,5491	−0,3317
9,2	20	25	7,8592	1,7865	1,0793
6,5	16	18	5,9190	−0,1537	−0,0929
6,3	17	20	6,4451	0,3724	0,2250
8,8	18	30	8,2848	2,2121	1,3363
7,9	21	14	6,2507	0,1780	0,1075
5,0	11	23	5,7515	−0,3212	−0,1940
4,2	17	10	4,8031	−1,2696	−0,7670
4,1	16	20	6,2474	0,1747	0,1055
5,2	13	16	4,9975	−1,0752	−0,6496
3,6	13	19	5,4901	−0,5826	−0,3519
4,3	16	15	5,4264	−0,6463	−0,3905
133,6	350	419	133,5992	−0,0002	−0,0001

$$\bar{X}_1 = \frac{133,6}{22} = 6,0727$$

$$\bar{X}_2 = \frac{350}{22} = 15,9091$$

$$\bar{X}_3 = \frac{419}{22} = 19,0455$$

$$\Sigma X_1X_2 = 2.217,8$$

$$\Sigma X_1X_3 = 2.633,9$$

$$\Sigma X_2X_3 = 6.742$$

$$\Sigma X_1^2 = 871,62$$

$$\Sigma X_2^2 = 5.972$$

$$\Sigma X_3^2 = 8.433$$

$$s_1 = 1,6556, \quad s_2 = 4,2843$$

$$s_3 = 4,5375$$

$$r_{12} = 0,5918, \quad r_{13} = 0,5411$$

$$r_{23} = 0,1779$$

a) *Plano de regresión a partir de puntuaciones directas*

Según (17.7).

$$B_2 = \frac{[(22)(2.217,8) - (133,6)(350)][(22)(8.433) - (419)^2]}{[(22)(5.972) - (350)^2][(22)(8.433) - (419)^2] - [(22)(6.742) - (350)(419)]^2} -$$

$$- \frac{[(22)(2.633,9) - (133,6)(419)][(22)(6.742) - (350)(419)]}{[(22)(5.972) - (350)^2][(22)(8.433) - (419)^2] - [(22)(6.742) - (350)(419)]^2} =$$

$$= \frac{16.951.466,4}{85.726.784} = 0,1977$$

Según (17.8).

$$B_3 = \frac{[(22)(2.633,9) - (133,6)(419)][(22)(5.972) - (350)^2]}{85.726.784} -$$

$$- \frac{[(22)(2.217,8) - (133,6)(350)][(22)(6.742) - (419)(350)]}{85.726.784} =$$

$$= \frac{14.077.483,2}{85.726.784} = 0,1642$$

Según (17.6).

$$A = 6,0727 - (0,1977)(15,9091) - (0,1642)(19,0455) = -0,1998$$

Por consiguiente:

$$X'_1 = -0,1998 + 0,1977\,X_2 + 0,1642\,X_3$$

Aplicando esta ecuación a los correspondientes valores X_2 y X_3, obtenemos la columna encabezada por X'_1 en la tabla 17.4.

Nótese cómo:

$$\Sigma X_1 = \Sigma X'_1 = 133,6, \quad \text{ó}, \quad \bar{X}_1 = \bar{X}'_1 = 6,0727$$

y cómo:

$$X'_1 = -0,1998 + (0,1977)(15,9091) + (0,1642)(19,0455) = 6,0727 = \bar{X}_1$$

b) *Plano de regresión a partir de puntuaciones diferenciales*

Según (17.24): $b_2 = B_2 = 0,1977$
Según (17.25): $b_3 = B_3 = 0,1642$
Según (17.16): $a = 0$

Por consiguiente,

$$x'_1 = 0,1977\,x_2 + 0,1642\,x_3$$

Aplicando esta ecuación a los correspondientes valores x_2 y x_3, obtenemos la columna encabezada por x'_1 en la tabla 17.4.

Nótese cómo:

$$\Sigma x_1 = \Sigma x'_1 = 0, \quad \text{ó}, \quad \bar{x}_1 = \bar{x}'_1 = 0$$

y cómo:

$$x'_1 = (0{,}1977)(0) + (0{,}1642)(0) = 0 = \bar{x}_1$$

c) *Plano de regresión a partir de puntuaciones típicas*

Según (17.31): $b_2^* = \dfrac{0{,}5918 - (0{,}5411)(0{,}1779)}{1 - (0{,}1779)^2} = 0{,}5117$

Según (17.32): $b_3^* = \dfrac{0{,}5411 - (0{,}5918)(0{,}1779)}{1 - (0{,}1779)^2} = 0{,}4501$

Según (17.30): $a^* = 0$

Por consiguiente,

$$z'_1 = (0{,}5117)\, z_2 + (0{,}4501)\, z_3$$

Aplicando esta ecuación a los correspondientes valores z_2 y z_3, obtenemos la columna encabezada por z'_1 en la tabla 17.4.

Nótese cómo:

$$\Sigma z_1 = \Sigma z'_1 = 0, \quad \text{ó}, \quad \bar{z}_1 = \bar{z}'_1 = 0$$

y cómo:

$$z'_1 = (0{,}5117)(0) + (0{,}4501)(0) = 0 = \bar{z}_1$$

17.3.4. Aplicación de los planos de regresión

Una vez construidos los planos de regresión, los podemos aplicar a otras personas con tal que sean semejantes a aquellas con las que los hemos construido. En realidad, suponemos que tanto el grupo con el que hemos construido los planos como el grupo al que se los aplicamos, no son más que dos muestras de la misma población.

Sean X_2 y X_3 dos tests de aptitud y sea X_1 el aprovechamiento escolar o notas en el examen de fin de curso. Supongamos que para las personas del grupo primero (mediante las cuales hemos construido los planos) $\bar{X}_1 = 24$, $\bar{X}_2 = 25$, $\bar{X}_3 = 15$, $s_1 = 4$, $s_2 = 5$, $s_3 = -2$, $r_{12} = 0{,}80$, $r_{13} = 0{,}40$, $r_{23} = 0{,}20$. A partir de estos datos tendremos las siguientes ecuaciones de regresión:

$$X'_1 = 1{,}50 + (0{,}60)\, X_2 + (0{,}50)\, X_3 \qquad (17.37)$$
$$x'_1 = (0{,}60)\, x_2 + (0{,}50)\, x_3 \qquad (17.38)$$
$$z'_1 = (0{,}75)\, z_2 + (0{,}25)\, z_3 \qquad (17.39)$$

Un nuevo alumno (semejante a los primeros) hace los dos tests de aptitud y obtiene las puntuaciones directas $X_2 = 30$, $x_3 = 20$. ¿Qué puntuación directa, diferencial y típica le pronosticaremos como nota de fin de curso?

Para obtener la puntuación directa pronosticada, basta con aplicar (17.37). Es decir:

$$X'_1 = 1{,}50 + (0{,}60)(30) + (0{,}50)(20) = 1{,}50 + 18 + 10 = 29{,}50$$

Para obtener la puntuación diferencial pronosticada podemos seguir dos caminos:

a) Transformar 30 y 20 en diferenciales y aplicar (17.38). Es decir:

$$x'_1 = (0{,}60)(30 - 25) + (0{,}50)(20 - 15) = 3 + 2{,}5 = 5{,}5$$

b) Transformar la puntuación directa pronosticada, 29,50, en diferencial pronosticada. Es decir:

$$x'_1 = 29{,}5 - 24 = 5{,}5$$

el mismo resultado que en *a*).

Para obtener la puntuación típica pronosticada podemos seguir dos caminos:

a) Transformar 30 y 20 en típicas y aplicar (17.39). Es decir:

$$z'_1 = (0{,}75)(30 - 25)/(5) + (0{,}25)(20 - 15)/(2) = 0{,}75 + 0{,}625 = 1{,}375$$

b) Transformar la puntuación directa (o diferencial) pronosticada en típica pronosticada. Es decir:

$$z'_1 = (29{,}5 - 24)/(4) = (5{,}5)/(4) = 1{,}375$$

el mismo resultado que en *a*).

Evidentemente, este nuevo alumno no pertenece al grupo primero mediante el cual hemos calculado $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, s_1, s_2, s_3. Sin embargo, podemos referir sus puntuaciones directas a esas medias y desviaciones típicas (para calcular sus puntuaciones diferenciales y típicas) porque suponemos que pertenece a la misma población a la que pertenecía el grupo primero.

17.4. Correlación múltiple

17.4.1. Definición

El coeficiente de correlación múltiple es un índice que nos mide la relación existente entre una variable, X_1, y otras variables, X_2, X_3, X_4, . . . , consideradas conjuntamente. Por ahora sólo consideraremos tres variables. Por tanto, será un índice que nos mide la relación existente entre X_1 y las dos variables X_2 y X_3 consideradas conjuntamente. Definamos este índice de un modo más preciso. De las varias definiciones posibles (equivalentes entre sí) elegimos la siguiente:

Sean X_{11}, X_{21}, . . . , X_{n1} las puntuaciones directas obtenidas por n personas en el criterio, X_1.

Sean X'_{11}, X'_{21}, . . . , X'_{n1} las puntuaciones directas pronosticadas a esas n personas mediante el plano de regresión de X_1 sobre X_2 y X_3.

Pues bien, definimos el coeficiente de correlación múltiple de X_1 con X_2 y X_3 (designándolo por $R_{1.23}$) como el coeficiente de correlación de Pearson entre (X_{11}, X_{21}, . . . , X_{n1}) y (X'_{11}, X'_{21}, . . . , X'_{n1}). Es decir:

$$R_{1.23} = r_{x_1x'_1} \tag{17.40}$$

Aceptamos esta definición de $R_{1.23}$ por dos razones. En primer lugar, X'_1 es función conjunta de X_2 y X_3 (recuérdense las ecuaciones de regresión). Por tanto, correlacionar X_1 con X'_1 es correlacionar X_1 con una función conjunta de X_2 y X_3, es decir, es correlacionar X_1 con X_2 y X_3 consideradas conjuntamente. En segundo lugar esta definición es muy parecida a la del coeficiente de correlación de Pearson entre dos variables, X_1 y X_2. En efecto, sabemos que el coeficiente de correlación de Pearson entre un criterio, X_1 y la variable predictora, X_2, no es más que el mismo coeficiente entre las puntuaciones obtenidas en el criterio (X_1) y las puntuaciones pronosticadas en el mismo (X'_1), es decir, $r_{x_1x_2} = r_{x_1x'_1}$.

Dada esta definición, como coeficiente de correlación de Pearson, es claro que $R_{1.23}$ valdrá lo mismo tanto si lo calculamos a partir de puntuaciones directas, como si lo hacemos a partir de puntuaciones diferenciales o típicas. Más aún, no tienen por qué ser del mismo tipo las puntuaciones obtenidas y las pronosticadas. Así, por ejemplo, las obtenidas en el criterio pueden ser directas y las pronosticadas en el mismo, diferenciales.

De la definición expuesta, $R_{1.23} = r_{x_1x'_1}$, se deducen otras fórmulas más aptas, tal vez, que ella misma para el cálculo efectivo de $R_{1.23}$. Vamos a deducir dos, una de las cuales suele ser presentada, con frecuencia, como definición de $R_{1.23}$. Pero antes hagamos algunas aclaraciones introductorias que creemos necesarias.

En primer lugar, dividiendo por n (17.28) y (17.29), nos queda

$$b_2^* + b_3^* r_{23} = r_{12} \tag{17.41}$$
$$b_3^* + b_2^* r_{23} = r_{13} \tag{17.42}$$

En segundo lugar, veamos lo que vale $s^2_{z'_1}$, es decir, la varianza de las puntuaciones pseudotípicas pronosticadas, z'_1.

$$s^2_{z'_1} = \frac{\Sigma z'^2_1}{n} = \frac{\Sigma (b_2^* z_2 + b_3^* z_3)^2}{n} = b_2^{*2} \frac{\Sigma z_2^2}{n} + 2b_2^* b_3^* \frac{\Sigma z_2 z_3}{n} + b_3^{*2} \frac{\Sigma z_3^2}{n} =$$
$$= b_2^{*2} + 2b_2^* b_3^* r_{23} + b_3^{*2} = b_2^{*2} + b_2^* b_3^* r_{23} + b_3^{*2} + b_2^* b_3^* r_{23} =$$
$$= b_2^*(b_2^* + b_3^* r_{23}) + b_3^*(b_3^* + b_2^* r_{23}) = b_2^* r_{12} + b_3^* r_{13}$$

(de acuerdo con (17.41) y (17.42).

Por consiguiente,

$$s_{z'_1} = \sqrt{b_2^* r_{12} + b_3^* r_{13}} \tag{17.43}$$

En este supuesto, aplicando la definición de correlación propuesta:

$$R_{1.23} = r_{z_1z'_1} = \frac{\Sigma z_1 z'_1}{n s_{z_1} s_{z'_1}} = \frac{\Sigma z_1 z'_1}{n s_{z'_1}} = \frac{\Sigma z_1 (b_2^* z_2 + b_3^* z_3)}{n s_{z'_1}} =$$
$$= \frac{1}{s_{z'_1}} \left(b_2^* \frac{\Sigma z_1 z_2}{n} + b_3^* \frac{\Sigma z_1 z_3}{n} \right) = \frac{b_2^* r_{12} + b_3^* r_{13}}{\sqrt{b_2^* r_{12} + b_3^* r_{13}}} = \sqrt{b_2^* r_{12} + b_3^* r_{13}} \tag{17.44}$$

Por consiguiente, teniendo en cuenta (17.44) y (17.43).

$$R^2_{1.23} = s^2_{z'_1} \tag{17.45}$$

Si en (17.44) sustituimos b_2^* y b_3^* por sus valores en (17.31) y (17.32),

$$R^2_{1.23} = \frac{r_{12} - r_{13} r_{23}}{1 - r^2_{23}} r_{12} + \frac{r_{13} - r_{12} r_{23}}{1 - r^2_{23}} r_{13} = \frac{r^2_{12} - 2 r_{12} r_{13} r_{23} + r^2_{13}}{1 - r^2_{23}} \tag{17.46}$$

17.4.2. Cálculo

Es una aplicación de las fórmulas anteriores.

EJEMPLO 17.4. Calculemos $R_{1.23}$ a partir de los datos contenidos en la tabla 17.5 El lector se encargará de comprobar que:

$\bar{X}_1 = 3$; $\bar{X}_2 = 4$; $\bar{X}_3 = 6$; $s_1 = 2$; $s_2 = 2$; $s_3 = 4$; $r_{12} = 0{,}85$; $r_{13} = 0{,}55$; $r_{23} = 0{,}45$

$b_2^* = 0{,}755$; $b_3^* = 0{,}210$; $s_{z'_1} = \sqrt{3{,}7841/5} = \sqrt{0{,}75682} = 0{,}87$

TABLA 17.5

X_1	X_2	X_3	z_1	z'_1	z'^2_1	$z_1 z'_1$
0	1	4	−1,5	−1,2375	1,53140625	1,85625
4	3	6	0,5	−0,3775	0,14250625	−0,18875
3	5	0	0,0	0,0625	0,00390625	0,00000
2	4	8	−0,5	0,1050	0,01102500	−0,05250
6	7	12	1,5	1,4475	2,09525625	2,17125
15	20	30	0,0	0,0000	3,78410000	3,78625

Tendremos, por tanto:

según (17.40): $R_{1.23} = \dfrac{3,78625}{(5)(0,87)} = 0,87$

según (17.44): $R_{1.23} = \sqrt{(0,755)(0,85) + (0,210)(0,55)} = 0,87$

según (17.45): $R_{1.23} = \sqrt{3,7841/5} = 0,87$

según (17.46): $R_{1.23} = \sqrt{\dfrac{(0,85)^2 - (2)(0,85)(0,55)(0,45) + (0,55)^2}{1 - (0,45)^2}} = 0,87$

EJEMPLO 17.5. Calculemos $R_{1.23}$ a partir de la tabla 17.4. Recordemos que:

$$r_{12} = 0,5918,\ r_{13} = 0,5411,\ r_{23} = 0,1779,\ b^*_2 = 0,5117,\ b^*_3 = 0,4501$$

Compruebe, además, el lector que:

$$\Sigma z'^2_1 = 12,0201,\ s_{z'_1} = \sqrt{12,0201/22} = 0,74$$

$$\Sigma z_1 z'_1 = 12,0196$$

En este supuesto:

según (17.40): $R_{1.23} = \dfrac{12,0196}{(22)(0,74)} = 0,74$

según (17.44): $R_{1.23} = \sqrt{(0,5117)(0,5918) + (0,4501)(0,5411)} = 0,74$

según (17.45): $R_{1.23} = \sqrt{12,0201/22} = 0,74$

según (17.46) $R_{1.23} = \sqrt{\dfrac{(0,5918)^2 - (2)(0,5918)(0,5411)(0,1779) + (0,5411)^2}{1 - (0,1779)^2}} = 0,74$

17.4.3. Propiedades

a) El coeficiente de correlación múltiple, $R_{1.23}$, al cuadrado, no puede ser menor que cero ni mayor que uno. Es decir, $0 \leq R^2_{1.23} \leq 1$.

En efecto, será no negativo por ser una expresión elevada al cuadrado. Será igual o menor que uno por su propia definición como coeficiente de correlación de Pearson.

b) $R^2_{1.23}$ es igual o mayor que r^2_{12} y que r^2_{13}. Es decir, la correlación múltiple (al cuadrado) del criterio con las dos variables predictoras es igual o mayor que la correlación simple (al cuadrado) del mismo con cada una de las dos variables predictoras consideradas por separado. Así, en el ejemplo 17.4, $R^2_{1.23} = (0,87)^2 = 0,76$ es mayor que $r^2_{12} = (0,85)^2 = 0,72$ y que $r^2_{13} = (0,55)^2 = 0,30$. Lo mismo sucede en el ejemplo 17.5, donde $R^2_{1.23} = (0,74)^2 = 0,55$ es mayor que $(0,0918)^2 = 0,35$ y que $(0,5411)^2 = 0,29$. En otras palabras, es mayor la correlación entre el rendimiento escolar y los dos tests predictores («razonamiento abstracto» y «comprensión verbal») considerados conjuntamente, que la correlación del rendimiento escolar tanto con el «razonamiento abstracto», considerado solo, como con la «comprensión verbal», considerada sola.

c) El coeficiente $R_{1.23}$ es aceptado siempre como positivo. En realidad, casi ni tiene sentido hablar del signo de $R_{1.23}$ ya que es función de varias correlaciones simples con signos posiblemente distintos. Nos mide, sin más, la intensidad de la relación entre el criterio y una combinación lineal de variables predictoras.

d) En general, $R_{1.23}$ tiende a aumentar cuando aumentan r_{12} y r_{13} y disminuye r_{23}. Sin embargo, esta afirmación necesitaría ser matizada algo más. Las relaciones entre $R_{1.23}$, r_{12}, r_{13} y r_{23} no son tan sencillas como puede aparecer a primera vista.

e) $R_{1.23}$ será mayor para la muestra mediante la cual hemos construido las ecuaciones de regresión, que para otra muestra a la que aplicamos esas mismas ecuaciones.

NOTA. No hemos creído necesario hablar expresamente de $r_{13.2}$ y $r_{23.1}$; de los planos de regresión de X_2 sobre X_1 y X_3, y de X_3 sobre X_1 y X_2; de $R_{2.13}$ y $R_{3.12}$. No son nada nuevo y ni aun necesario. Basta con llamar siempre X_3 a la variable que deseamos mantener constante y calcular $r_{12.3}$. Basta con llamar siempre X_1 a la variable que actuará como criterio y calcular las ecuaciones de regresión de X_1 sobre X_2 y X_3. Es decir, llamar siempre X_1 a la variable que deseamos correlacionar con las otras dos tomadas conjuntamente y calcular $R_{1.23}$. Con todo, puede ser un ejercicio útil para el lector intentar calcular $r_{13.2}$, $r_{23.1}$, y las ecuaciones de regresión de X_2 sobre X_1 y X_3 y las de X_3 sobre X_1 y X_2 y $R_{2.13}$, $R_{3.12}$.

17.5. Resumen: Definiciones y fórmulas

Coeficiente de correlación parcial, $r_{12.3}$: Coeficiente de correlación de Pearson entre aquella parte de X_1 que no depende de X_3 y aquella parte de X_2 que tampoco depende de X_3

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}$$

Ecuaciones de regresión múltiple de X_1 sobre X_2 y X_3: Ecuaciones que nos permiten pronosticar las puntuaciones en X_1 a partir de las puntuaciones obtenidas en X_2 y en X_3.

a) *Supuestas puntuaciones directas*: $X'_1 = A + B_2X_2 + B_3X_3$

$$A = \bar{X}_1 - B_2\bar{X}_2 - B_3\bar{X}_3$$

$$B_2 = \frac{[n\Sigma X_1X_2 - \Sigma X_1 \Sigma X_2][n\Sigma X_3^2 - (\Sigma X_3)^2]}{H} - \frac{[n\Sigma X_1X_3 - \Sigma X_1 \Sigma X_3][n\Sigma X_2X_3 - \Sigma X_2 \Sigma X_3]}{H}$$

$$B_3 = \frac{[n\Sigma X_1X_3 - \Sigma X_1 \Sigma X_3][n\Sigma X_2^2 - (\Sigma X_2)^2]}{H} - \frac{[n\Sigma X_1X_2 - \Sigma X_1 \Sigma X_2][n\Sigma X_3X_2 - \Sigma X_3 \Sigma X_2]}{H}$$

siendo:

$$H = [n\Sigma X_2^2 - (\Sigma X_2)^2][n\Sigma X_3^2 - (\Sigma X_3)^2] - [n\Sigma X_2X_3 - \Sigma X_2 \Sigma X_3]^2$$

b) *Supuestas puntuaciones diferenciales*: $x'_1 = b_2x_2 + b_3x_3$

$$b_2 = B_2 \quad , \quad b_3 = B_3$$

c) *Supuestas puntuaciones típicas*: $z'_1 = b_2^*z_2 + b_3^*z_3$

$$b_2^* = \frac{s_2}{s_1}b_2 \quad , \quad b_3^* = \frac{s_3}{s_1}b_3$$

Coeficiente de correlación múltiple, $R_{1.23}$: Coeficiente de correlación de Pearson entre X_1, y X'_1, donde X'_1 es el pronóstico ofrecido por el plano de regresión de X_1 sobre X_2 y X_3.

$$R_{1.23} = r_{X_1X'_1} = \sqrt{b_2^*r_{12} + b_3^*r_{13}} = \sqrt{\frac{\Sigma z_1'^2}{n}} = \sqrt{\frac{r_{12}^2 - 2r_{12}r_{13}r_{23} + r_{13}^2}{1 - r_{23}^2}}$$

NOTA 1. Las puntuaciones diferenciales y típicas no son más que las directas sometidas a ciertas condiciones restrictivas. Por tanto, las relaciones válidas entre las directas, serán también válidas entre las diferenciales y típicas, con tal que impongamos a las directas las restricciones requeridas.

Ahora bien, introducir puntuaciones diferenciales equivale a imponer $\Sigma x_i = \bar{x}_i = 0$, $(i = 1, 2, 3)$. Consiguientemente, (17.6), (17.7) y (17.8) quedarán convertidas en

$$A = 0 + (B_2)(0) + (B_3)(0) = 0, \quad \text{o sea, (17.16)}$$

$$B_2 = \frac{(n\Sigma x_1x_2 - 0)(n\Sigma x_3^2 - 0) - (n\Sigma x_1x_3 - 0)(n\Sigma x_2x_3 - 0)}{(n\Sigma x_2^2 - 0)(n\Sigma x_3^2 - 0) - (n\Sigma x_2x_3 - 0)^2} = \frac{\cancel{n^2}(\Sigma x_1x_2 \Sigma x_3^2 - \Sigma x_1x_3 \Sigma x_2x_3)}{\cancel{n^2}(\Sigma x_2^2 \Sigma x_3^2 - (\Sigma x_2x_3)^2)}, \quad \text{o sea, (17.17)}$$

$$B_3 = \frac{(n\Sigma x_1x_3 - 0)(n\Sigma x_2^2 - 0) - (n\Sigma x_1x_2 - 0)(n\Sigma x_3x_2 - 0)}{(n\Sigma x_3^2 - 0)(n\Sigma x_2^2 - 0) - (n\Sigma x_3x_2 - 0)^2} = \frac{\cancel{n^2}(\Sigma x_1x_3 \Sigma x_2^2 - \Sigma x_1x_2 \Sigma x_3x_2)}{\cancel{n^2}(\Sigma x_3^2 \Sigma x_2^2 - (\Sigma x_3x_2)^2)}, \quad \text{o sea, (17.18)}$$

A su vez, introducir puntuaciones típicas equivale a imponer $\Sigma z_i = \bar{z}_i = 0$, $\Sigma z_iz_j = nr_{ij}$ $(i, j = 1, 2, 3)$, con $r_{ii} = 1$. Consiguientemente, (17.6), (17.7) y (17.8) quedarán convertidas en

$$A = 0 + (B_2)(0) + (B_3)(0) = 0, \quad \text{o sea, (17.30)}$$

$$B_2 = \frac{(nnr_{12} - 0)(nn - 0) - (nnr_{13} - 0)(nnr_{23} - 0)}{(nn - 0)(nn - 0) - (nnr_{23} - 0)^2} = \frac{\cancel{n^4}}{\cancel{n^4}}\frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2}, \quad \text{o sea, (17.31)}$$

$$B_3 = \frac{(nnr_{13} - 0)(nn - 0) - (nnr_{12} - 0)(nnr_{23} - 0)}{(nn - 0)(nn - 0) - (nnr_{23} - 0)^2} = \frac{\cancel{n^4}}{\cancel{n^4}}\frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2}, \quad \text{o sea, (17.32)}$$

NOTA 2. Según (17.21), (17.35) y (17.36):

$$x'_1 = \frac{s_1}{s_2}b_2^*x_2 + \frac{s_1}{s_3}b_3^*x_3$$

Dividiendo ambos miembros por s_1 y teniendo en cuenta (17.33), nos queda $\frac{x'_1}{s_1} = b_2^* \frac{x_2}{s_2} + b_3^* \frac{x_3}{s_3} = b_2^* z_2 + b_3^* z_3 = z'_1$. Por tanto, $z'_1 = \frac{x'_1}{s_1}$. Esta igualdad nos vuelve a recordar algo que ya conocíamos, al saber que z'_1 no es auténtica puntuación típica, por ser cociente entre una puntuación diferencial *pronosticada* (x'_1) y la desviación típica de las puntuaciones obtenidas (s_1). Es decir, las puntuaciones z'_1 son puntuaciones pseudotípicas y no puntuaciones auténticamente típicas.

NOTA 3. *Deducción de los coeficientes de los planos de regresión, a partir de puntuaciones directas.*

Recordemos las tres ecuaciones normales:

$$\Sigma X_1 = nA + B_2 \Sigma X_2 + B_3 \Sigma X_3 \tag{17.3}$$
$$\Sigma X_1 X_2 = A \Sigma X_2 + B_2 \Sigma X_2^2 + B_3 \Sigma X_2 X_3 \tag{17.4}$$
$$\Sigma X_1 X_3 = A \Sigma X_3 + B_2 \Sigma X_2 X_3 + B_3 \Sigma X_3^2 \tag{17.5}$$

Dividiendo (17.3) por n, nos queda

$$\bar{X}_1 = A + B_2 \bar{X}_2 + B_3 \bar{X}_3 \tag{17.47}$$
$$A = \bar{X}_1 - B_2 \bar{X}_2 - B_3 \bar{X}_3 \quad , \quad \text{que es} \quad (17.6)$$

Multiplicando (17.47) por $n\bar{X}_2$ y restando este resultado de (17.4),

$$\begin{aligned} \Sigma X_1 X_2 - n\bar{X}_1\bar{X}_2 &= (A \Sigma X_2 + B_2 \Sigma X_2^2 + B_3 \Sigma X_2 X_3) - \\ &- (An\bar{X}_2 + B_2 n\bar{X}_2^2 + B_3 n\bar{X}_2\bar{X}_3) = \\ &= A(\Sigma X_2 - n\bar{X}_2) + B_2(\Sigma X_2^2 - n\bar{X}_2^2) + B_3(\Sigma X_2 X_3 - n\bar{X}_2\bar{X}_3) \end{aligned}$$

Esta igualdad puede ser escrita del modo siguiente:

$$\frac{n \Sigma X_1 X_2 - \Sigma X_1 \Sigma X_2}{n} = \frac{(A)(0)}{n} + \frac{B_2(n \Sigma X_2^2 - (\Sigma X_2)^2)}{n} + \frac{B_3(n \Sigma X_2 X_3 - \Sigma X_2 \Sigma X_3)}{n} \tag{17.48}$$

De modo análogo, multiplicando (17.47) por $n\bar{X}_3$ y restando este resultado de (17.5),

$$\begin{aligned} n \Sigma X_1 X_3 - \Sigma X_1 \Sigma X_3 &= (A)(0) + B_2(n \Sigma X_2 X_3 - \Sigma X_2 \Sigma X_3) + \\ &+ B_3(n \Sigma X_3^2 - (\Sigma X_3)^2) \end{aligned} \tag{17.49}$$

Hagamos

$$n \Sigma X_i X_j - \Sigma X_i \Sigma X_j = H_{ij} \tag{17.50}$$

Bajo este supuesto, (17.48) y (17.49) vendrán expresadas así:

$$H_{12} = B_2 H_{22} + B_3 H_{23} \tag{17.51}$$
$$H_{13} = B_2 H_{23} + B_3 H_{33} \tag{17.52}$$

Multiplicando (17.51) por H_{33}, (17.52) por H_{23} y restando

$$H_{12}H_{33} - H_{13}H_{23} = B_2(H_{22}H_{33} - H_{23}^2)$$

De donde:

$$B_2 = \frac{H_{12}H_{33} - H_{13}H_{23}}{H_{22}H_{33} - H_{23}^2}$$

que, teniendo en cuenta (17.50), no es más que (17.7).

Multiplicando (17.51) por H_{23}, (17.52) por H_{22} y restando,

$$H_{12}H_{23} - H_{13}H_{22} = B_3(H_{23}^2 - H_{33}H_{22})$$

De donde:

$$B_3 = \frac{H_{13}H_{22} - H_{12}H_{23}}{H_{22}H_{33} - H_{23}^2}$$

que, teniendo en cuenta (17.50), no es más que (17.8).

Recordando la NOTA 1, podemos deducir inmediatamente los coeficientes correspondientes a las ecuaciones de regresión en puntuaciones *diferenciales* y en puntuaciones típicas, a partir de A_1, B_2 y B_3 imponiendo las condiciones restrictivas propias de las puntuaciones diferenciales y de las típicas, respectivamente.

EJERCICIOS

17.1. Calcular $r_{12.3}$ sabiendo que

a) $r_{12} = 0{,}80$; $r_{13} = 0{,}60$; $r_{23} = 0{,}50$.
b) $r_{12} = 0{,}30$; $r_{13} = 0{,}40$; $r_{23} = 0{,}60$.
c) $r_{12} = 0{,}80$; $r_{13} = 0{,}70$; $r_{23} = 0{,}60$.

17.2. Calcular el valor de r_{12}, sabiendo que $r_{12.3} = 0{,}60$; $r_{13} = 0{,}80$; $r_{23} = 0{,}60$.

17.3. ¿Cuánto valdrá $r_{12.3}$ si $r_{13} = 0$ y $r_{23} = 0$?

17.4. Supongamos que r_{12}, r_{13} y r_{23} tienen un mismo valor que llamaremos r. Con estas condiciones, demostrar que $r_{12.3} = r_{23.1} = r_{13.2} = r/(1 + r)$.

17.5. Comprobar lo demostrado en 17.4, para $r = 0{,}60$.

17.6. Demostrar que $r_{23} = 1$, tanto si $r_{12} = r_{13} = 1$, como si $r_{12} = r_{13} = -1$.

17.7. Demostrar que $r_{23} = -1$, tanto si $r_{12} = 1$ y $r_{13} = -1$, como si $r_{12} = -1$ y $r_{13} = 1$.

17.8. Demostrar que $r_{13.2} = r_{12.3}$ si $r_{13} = r_{12}$.

17.9. Comprobar lo demostrado en 17.8, para $r_{12} = r_{13} = 0{,}80$.

17.10. Suponiendo $s_1 = 6$, $s_2 = 3$, $s_3 = 3$, $r_{12} = 0{,}50$, $r_{13} = 0{,}40$, $r_{23} = 0{,}20$, calcular las ecuaciones de regresión de X_1 sobre X_2 y X_3, en puntuaciones diferenciales y típicas.

17.11. Suponiendo $s_1 = 4$, $s_2 = 5$, $s_3 = 2$, $r_{12} = 0{,}80$, $r_{13} = 0{,}40$, $r_{23} = 0{,}20$, calcular las ecuaciones de regresión de X_1 sobre X_2 y X_3, en puntuaciones diferenciales y típicas.

17.12. Calcular las ecuaciones de regresión de X_1 sobre X_2 y X_3 a partir de los datos siguientes:

a)

X_1	X_2	X_3
0	2	1
4	0	9
8	4	5
6	6	13
12	3	7

b)

X_1	X_2	X_3
2	1	1
6	3	7
10	4	5
8	5	3
14	7	4

c)

X_1	X_2	X_3
3	9	12
5	12	8
4	0	4
7	18	6
1	6	0

d)

X_1	X_2	X_3
1	1	0
1	0	1
0	1	1
2	2	2

17.13. Calcular las puntuaciones directas, diferenciales y pseudotípicas pronosticadas en el criterio X_1, mediante las ecuaciones de regresión de X_1 sobre X_2 y X_3, teniendo en cuenta los datos del ejercicio 17.12.

17.14. Sean X_{1i}, X'_{1i}, x_{1i}, x'_{1i} las puntuaciones directas (obtenida y pronosticada) y diferenciales (obtenida y pronosticada) de la persona i. Demostrar que para toda persona i se verifica: $X_{1i} - X'_{1i} = x_{1i} - x'_{1i}$.

17.15. Calcular el coeficiente de correlación múltiple, teniendo en cuenta los datos del ejercicio 17.12.

18

El coeficiente de correlación múltiple y los planos de regresión

18.1. $R^2_{1.23}$ como índice de reducción de error en los pronósticos

Vamos a considerar, en concreto, el plano de regresión de X_1 sobre X_2 y X_3. Algo parecido valdrá para los planos de regresión de X_2 sobre X_1 y X_3 y de X_3 sobre X_1 y X_2.

Sea X_1 la puntuación directa obtenida por una persona i en el criterio X_1 y sea X'_1 la puntuación directa pronosticada a esa misma persona mediante el plano de regresión de X_1 sobre X_2 y X_3. El error cuadrático cometido en ese pronóstico individual valdrá $(X_1 - X'_1)^2$ y $\Sigma (X_1 - X'_1)^2$ será la suma de errores cuadráticos respecto a todas las personas de la muestra. Por su parte, la suma $\Sigma(X_1 - \bar{X}_1)^2$ queda descompuesta en dos sumas de términos cuadráticos. En efecto:

$$\Sigma (X_1 - \bar{X}_1)^2 = \Sigma [(X_1 - X'_1) + (X'_1 - \bar{X}_1)]^2 = \Sigma (X_1 - X'_1)^2 + \Sigma (X'_1 - \bar{X}_1)^2 \quad (18.1)$$

pues:

$$\Sigma (X_1 - X'_1)(X'_1 - \bar{X}_1) = 0$$

según se encargará de comprobar el lector en el ejercicio 18.4.

Ahora bien:

$$\frac{\Sigma (X_1 - \bar{X}_1)^2}{n}$$

que solemos designar por s_1^2 ó $s_{x_1}^2$, es el error cuadrático medio (E_m^2) cometido al atribuir a cada persona, como puntuación, la media $\bar{X}_1$.

A su vez:

$$\frac{\Sigma (X_1 - X'_1)^2}{n}$$

que designaremos por $s^2_{1.23}$, es el E^2_m cometido al atribuir a cada persona la puntuación X'_1 obtenida mediante el plano de regresión.

Finalmente, según (18.1):

$$\frac{\Sigma (X'_1 - \bar{X}_1)^2}{n} = \frac{\Sigma (X_1 - \bar{X}_1)^2}{n} - \frac{\Sigma (X_1 - X'_1)^2}{n}$$

que designaremos por $s^2_{x'_1}$, no es más que la diferencia entre el E^2_m cometido valiéndonos de $\bar{X}_1$ y el E^2_m cometido valiéndonos de X'_1. En otras palabras, es aquella parte del E^2_m primitivo que dejamos de cometer por el hecho de atribuir X'_1 a cada persona en vez de atribuirle $\bar{X}_1$.

Por consiguiente:

$$\frac{s^2_{x'_1}}{s^2_{x_1}} = \frac{s^2_{x'_1}}{s^2_{x'_1} + s^2_{1.23}} = \frac{\text{parte de } E^2_m \text{ eliminada}}{\text{parte de } E^2_m \text{ eliminada} + \text{parte de } E^2_m \text{ no eliminada}} =$$

= proporción de E^2_m eliminado o, de otro modo, proporción en que reducimos el E^2_m primitivo, el que cometíamos valiéndonos de $\bar{X}_1$.

Si, por ejemplo, ese cociente fuera igual a 0,60, ello significaría que habíamos reducido el E^2_m en un 60 por 100, es decir, que, valiéndonos de X'_1 sólo cometeríamos un 40 por 100 del E^2_m que habríamos cometido valiéndonos de $\bar{X}_1$.

Pero

$$\frac{s^2_{x'_1}}{s^2_{x_1}} = \frac{\Sigma x'^2_1}{n}\frac{1}{s^2_{x_1}} = \frac{\Sigma (x'_1/s_{x_1})^2}{n} = \frac{\Sigma z'^2_1}{n} = s^2_{z'_1} = R^2_{1.23} \qquad (18.2)$$

según (17.44).

Por tanto, $R^2_{1.23}$ no es más que la proporción en que ha sido reducido el error cuadrático primitivo, el que habríamos cometido si nos hubiéramos valido de $\bar{X}_1$ como puntuación pronosticada.

En conclusión:

$$R^2_{1.23} = \frac{s^2_{x'_1}}{s^2_{x_1}} = \frac{s^2_{x_1} - s^2_{1.23}}{s^2_{x_1}} = 1 - \frac{s^2_{1.23}}{s^2_{x_1}} \qquad (18.3)$$

Consideremos los dos casos extremos posibles: $s^2_{1.23} = 0$ (equivalente a $s^2_{x'_1} = s^2_{x_1}$), y $s^2_{x'_1} = 0$ (equivalente a $s^2_{1.23} = s^2_{x_1}$).

El primero nos indica que, valiéndonos del plano de regresión, hemos eliminado o reducido por completo el E^2_m primitivo, el que habríamos cometido atribuyendo $\bar{X}_1$, como puntuación en X_1, a cada una de las personas de la muestra. Ahora bien, si $s^2_{1.23} = 0$, entonces, según (18.3), $R^2_{1.23} = 1$; luego a reducción total del E^2_m primitivo, corresponde $R^2_{1.23} = 1$. Recíprocamente, si $R^2_{1.23} = 1$, entonces según (18.3), $s^2_{1.23} = 0$; luego a $R^2_{1.23} = 1$, corresponde reducción total del E^2_m primitivo.

El segundo caso extremo nos indica que, valiéndonos del plano de regresión, no hemos reducido en nada el E^2_m que cometíamos atribuyendo $\bar{X}_1$, como puntuación en X_1, a cada una de las personas de la muestra. Ahora bien, si $s^2_{x'_1} = 0$, entonces, según (18.3), $R^2_{1.23} = 0$; luego a reducción nula del E^2_m primitivo, corresponde $R^2_{1.23} = 0$. Recíprocamente, si $R^2_{1.23} = 0$, entonces, según (18.3), $s^2_{x'_1} = 0$; luego a $R^2_{1.23} = 0$ corresponde reducción nula del E^2_m primitivo.

En conclusión, a:

$R^2_{1.23} = 1$, corresponde reducción total del E^2_m primitivo y recíprocamente.
$R^2_{1.23} = 0$, corresponde reducción nula del E^2_m primitivo y recíprocamente.

El valor $s_{1.23}$ suele ser llamado error típico de estimación.

Teniendo en cuenta (18.3), nos queda

$$s^2_{1.23} = s^2_{x_1}(1 - R^2_{1.23}) \qquad (18.4)$$

De acuerdo con lo dicho en el apartado 12.1, el paso de $R_{1.23}$ de 0,20 a 0,50, por ejemplo, significa mucho menos que el paso de 0,65 a 0,95, en cuanto a reducción de E^2_m.

Según (17.37), (17.44) y (18.2), vemos que, definiendo $R^2_{1.23}$ como $r^2_{z_1 z'_1} = s^2_{z'_1}$, llegamos a que $R^2_{1.23} = s^2_{x'_1}/s^2_{x_1}$ y vemos, también, que, definiendo $R^2_{1.23}$ como $s^2_{x'_1}/s^2_{x_1}$, llegamos a que $R^2_{1.23} = s^2_{z'_1} = r^2_{z_1 z'_1}$.

La definición de $R^2_{1.23}$ como $s^2_{x'_1}/s^2_{x_1} = \Sigma (X'_1 - \bar{X}_1)^2/\Sigma (X_1 - \bar{X}_1)^2 = 1 - \Sigma (X_1 - X'_1)^2/\Sigma (X_1 - \bar{X}_1)^2$ tiene la ventaja de su paralelismo con la definición de otros coeficientes de correlación como índices de reducción de error. Unos diferirán de otros en diversos aspectos accidentales, pero todos ellos coincidirán en ser índices de reducción de error.

En los ejercicios 18.12. y 18.13, al final del capítulo, el lector se encargará de probar que

$$\Sigma x^2_1 = \Sigma (x_1 - x'_1)^2 + \Sigma x'^2_1 \qquad (18.5)$$

$$\Sigma z^2_1 = \Sigma (z_1 - z'_1)^2 + \Sigma z'^2_1 \qquad (18.6)$$

EJEMPLO 18.1. Comprobemos las relaciones acabadas de exponer con los datos siguientes:

X_1	X_2	X_3	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3	z_1	z_2	z_3	z_1z_2	z_1z_3	z_2z_3
3	9	12	−1	0	6	1	0	36	0	−6	0	−0,5	0,0	1,5	0,00	−0,75	0,00
5	12	8	1	3	2	1	9	4	3	2	6	0,5	0,5	0,5	0,25	0,25	0,25
4	0	4	0	−9	−2	0	81	4	0	0	18	0,0	−1,5	−0,5	0,00	0,00	0,75
7	18	6	3	9	0	9	81	0	27	0	0	1,5	1,5	0,0	2,25	0,00	0,00
1	6	0	−3	−3	−6	9	9	36	9	18	18	−1,5	−0,5	−1,5	0,75	2,25	0,75
20	45	30	0	0	0	20	180	80	39	14	42	0,0	0,0	0,0	3,25	1,75	1,75

Del cuadro anterior es fácil construir las ecuaciones de regresión:

$$X'_1 = 1{,}77778 + 0{,}20038\,X_2 + 0{,}06980\,X_3 \quad , \quad x'_1 = 0{,}20038\,x_2 + 0{,}06980\,x_3$$
$$z'_1 = 0{,}60114\,z_2 + 0{,}13960\,z_3$$

Con estas ecuaciones podemos formar el cuadro siguiente:

$(X_1 - \bar{X}_1)^2 = x_1^2$	z_1^2	X'_1	x'_1	z'_1	$X_1 - X'_1 = x_1 - x'_1$	$z_1 - z'_1$	$(X_1 - X'_1)^2 = (x_1 - x'_1)^2$	$(z_1 - z'_1)^2$	$(X'_1 - \bar{X}_1)^2 = x'^2_1$	z'^2_1
1	0,25	4,41880	0,41880	0,20940	−1,41880	−0,70940	2,01299	0,50325	0,17539	0,04385
1	0,25	4,74074	0,74074	0,37037	0,25926	0,12963	0,06722	0,01680	0,54870	0,13717
0	0,00	2,05698	−1,94302	−0,97151	1,94302	0,97151	3,77533	0,94383	3,77533	0,94383
9	2,25	5,80324	1,80432	0,90171	1,19658	0,59829	1,43180	0,35795	3,25232	0,81308
9	2,25	2,98006	−1,01994	−0,50997	−1,98006	−0,99003	3,92064	0,98016	1,04028	0,26007
20	5,00	20,00000	0,00000	0,00000	0,00000	0,00000	11,20798	2,80199	8,79202	2,19800

Comprobación de (18.1):

$$\Sigma\,(X_1 - X'_1)^2 + \Sigma\,(X'_1 - \bar{X}_1)^2 = 11{,}20798 + 8{,}79202 = 20{,}00000 = \Sigma\,(X_1 - \bar{X}_1)^2$$

Comprobación de (18.5):

$$\Sigma\,(x_1 - x'_1)^2 + \Sigma\,x'^2_1 = 11{,}20798 + 8{,}79202 = 20{,}00000 = \Sigma\,x_1^2$$

Según sabemos,

$$R^2_{1.23} = s^2_{z'_1} = \frac{2{,}19800}{5} = 0{,}4396$$

Comprobación de (18.2):

$$s^2_{x'_1}/s^2_{x_1} = \frac{(8{,}79202)/(5)}{(20{,}0000)/(5)} = 0{,}4396 = R^2_{1.23}$$

Comprobación de (18.3):

$$1 - s^2_{1.23}/s^2_{x_1} = 1 - \frac{(11{,}20798)/(5)}{(20{,}00000)/(5)} = 1 - 0{,}5604 = 0{,}4396 = R^2_{1.23}$$

Comprobación de (18.4):

$$s^2_{x_1}(1 - R^2_{1.23}) = \frac{20}{5}(1 - 0{,}4396) = 2{,}2416 = \frac{11{,}20798}{5} = s^2_{1.23}$$

Comprobación de (18.6):

$$\Sigma\,(z_1 - z'_1)^2 + \Sigma\,z'^2_1 = 2{,}80199 + 2{,}19800 = 4{,}99999 \simeq 5 = \Sigma\,z_1^2$$

NOTA. A pesar de lo expuesto, es posible que $R^2_{1.23} = 0$, valiendo $s^2_{1.23} = 0$. Esto sucedería si todos los puntos estuvieran situados sobre el plano de regresión y éste fuera paralelo al plano X_2OX_3. Ciertamente, $s^2_{1.23} = 0$. Además, $s^2_{x_1} = 0$. Por tanto, $s^2_{x'_1} = 0 - 0 = 0$. Consiguientemente, $R^2_{1.23} = s^2_{x'_1}/s^2_{x_1} = 0/0$, quedaría indeterminado. Sin embargo, tiene sentido admitir $R_{1.23} = 0$, ya que en esta situación, tanto a puntuaciones altas, como medias y bajas en X_2 y X_3, corresponde una misma puntuación, $\bar{X}_1$, en X_1 y, por tanto, la covariación entre X_1 y X_2 y X_3 es nula.

18.2. $R^2_{1.23}$ como índice de aproximación de los puntos al plano de regresión

Hemos visto en el párrafo anterior que $R^2_{1.23} = 1 - \Sigma\,(X_1 - X'_1)^2/\Sigma\,(X_1 - \bar{X}_1)^2$. De aquí se deduce inmediatamente lo siguiente:

a) Si todos los puntos (representantes de las puntuaciones obtenidas) se encuentran sobre el plano de regresión de X_1 sobre X_2 y X_3, $X_1 = X'_1$ para toda persona de la muestra. Por tanto, $X_1 - X'_1 = 0$. Consiguientemente, $\Sigma\,(X_1 - X'_1)^2 = 0$. Es decir, $R^2_{1.23} = 1$.

b) Si $R^2_{1.23} = 1$, $\Sigma\,(X_1 - X'_1)^2 = 0$. Ahora bien, como los n sumandos son no negativos (por ser cuadráticos), si su suma es cero quiere decir que cada uno de ellos tiene que valer cero. Por consiguiente, $X_1 - X'_1 = 0$, es decir, $X_1 = X'_1$, para toda persona de la muestra. En otras palabras, todos los puntos se encuentran sobre el plano de regresión de X_2 y X_3 sobre X_1.

En conclusión:

Si todos los puntos se encuentran sobre el plano de regresión, $R^2_{1.23} = 1$.

Si $R^2_{1.23} = 1$, todos los puntos se encuentran sobre el plano de regresión.

Por consiguiente, $R^2_{1.23}$ nos mide la aproximación de los puntos al plano de regresión. Nótese, por tanto, que lo que mide $R_{1.23}$ es la aproximación de los puntos a un plano. Es posible que todos los puntos se encuentren sobre una superficie no lineal (una superficie esférica, por ejemplo), que exista una relación funcional perfecta entre X_2 y X_3 con X_1 que nos permita pronosticar un valor exacto de X_1 para cada par de valores dados de X_2 y X_3, y que $R_{1.23}$ valga 0. Así, por ejemplo, los seis puntos siguientes están situados sobre la superficie esférica de ecuación $X_1^2 + X_2^2 + X_3^2 = 4$. Sin embargo, $R_{1.23} = 0$.

X_1	X_2	X_3
0	0	2
0	2	0
0	0	−2
0	−2	0
2	0	0
−2	0	0

Nótese que $R^2_{1.23}$ no es sólo función de $s^2_{1.23}$, sino también de $s^2_{x_1}$. Esto quiere decir que para un mismo error cuadrático medio ($s^2_{1.23}$), $R^2_{1.23}$ puede ser mayor o menor según que sea mayor o menor $s^2_{x_1}$.

Por supuesto, si operamos con puntuaciones típicas:

$$s^2_{z_1} = 1$$

Por tanto,

$$R^2_{1.23} = 1 - \frac{s^2_{z_{1.23}}}{s^2_{z_1}} = 1 - s^2_{z_{1.23}} = 1 - \frac{\Sigma (z_1 - z'_1)^2}{n}$$

Es decir, operando con puntuaciones típicas, siempre que aumenta (disminuye) $\Sigma (z_1 - z'_1)^2$, disminuye (aumenta) $R^2_{1.23}$ y recíprocamente. Por consiguiente, a mayor (menor) aproximación de los puntos a la recta al plano de regresión, corresponde siempre mayor (menor) valor de $R^2_{1.23}$.

18.3. $R^2_{1.23}$ como proporción de la varianza de X_1 asociada a la variación de X_2 y de X_3

La relación

$$X_1 = X'_1 + (X_1 - X'_1)$$

es una pura identidad que puede ser interpretada aceptando que la puntuación directa obtenida es igual al pronóstico (X'_1) más el error en dicho pronóstico ($X_1 - X'_1$).

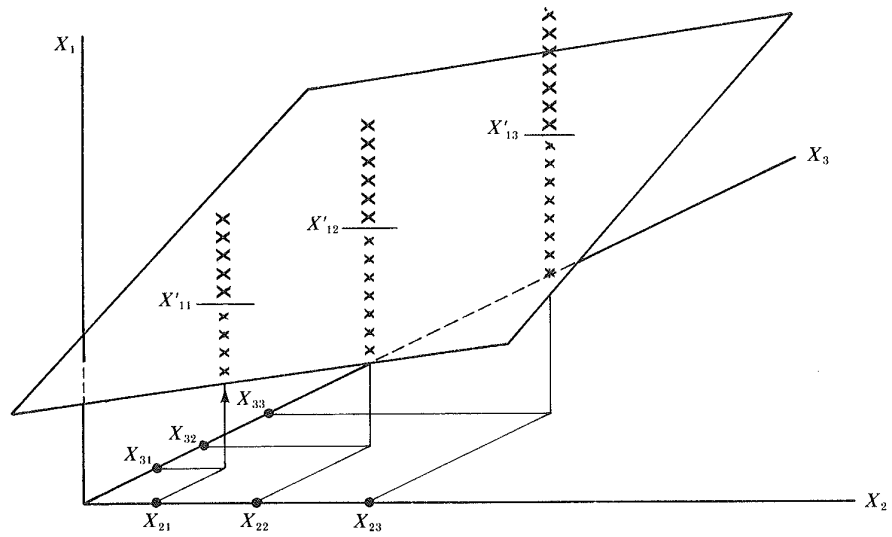
Sabemos que

$$X'_1 = a + b_2 X_2 + b_3 X_3$$

Esto nos indica que X'_1 depende de, es función de, está asociada a, X_2 y X_3.

Por el contrario, $X_1 - X'_1$ no depende, no es función de, no está asociada a, X_2 y X_3. De hecho, $r_{x_2(x_1 - x'_1)} = r_{x_3(x_1 - x'_1)} = 0$. A puntuación baja en X_2 y/o en X_3 puede corresponder error bajo, medio o alto en X_1. Como se ve en la figura, tanto entre las personas con $X_2 = X_{21}$ (puntuación baja en X_2) y $X_3 = X_{31}$ (puntuación baja en X_3), como entre las personas con $X_2 = X_{22}$ (puntuación media en X_2) y $X_3 = X_{32}$ (puntuación media en X_3), como entre las personas con $X_2 = X_{23}$ (puntuación alta en X_2) y $X_3 = X_{33}$ (puntuación alta en X_3), a unas les corresponden

errores grandes (a las que se alejan bastante del plano de regresión), a otras medios (a las que se alejan moderadamente del mismo) y a otras pequeños (a las que se alejan muy poco de dicho plano).



Por otra parte, hay variación en los errores, cuando no la hay en la variables predictoras. Es decir, a personas con una misma puntuación en X_2 y en X_3, les corresponden, en general, distintos errores en los pronósticos (a unas, pequeños; a otras, medios; y a otras, altos).

Sabemos, además, que:

$$\frac{\Sigma (X_1 - \bar{X}_1)^2}{n} = \frac{\Sigma (X'_1 - \bar{X}_1)^2}{n} + \frac{\Sigma (X_1 - X'_1)^2}{n}$$

$$s^2_{x_1} = s^2_{x'_1} + s^2_{1.23}$$

En otras palabras, la varianza total de X_1 ($s^2_{x_1}$) se descompone en dos partes aditivas. Una, $s^2_{x'_1}$, asociada a, dependiente de, explicada por, la variación de X_2 y X_3, ya que X'_1 estaba asociada a, dependía de, era explicada por, la variación de X_2 y X_3. Otra, $s^2_{1.23}$, no asociada a, no dependiente de, no explicada por, la variación de X_2 y X_3.

Por estas razones, llamaremos a $s^2_{x'_1}$ varianza asociada, y a $s^2_{1.23}$ la llamaremos varianza no asociada.

Ahora bien, sabemos que: $R^2_{1.23} = s^2_{x'_1}/s^2_{x_1}$. Pero $s^2_{x'_1}/s^2_{x_1}$, dentro de este con-

texto, no es más que la proporción de la varianza total de X_1 que está asociada a la variación de X_2 y X_3. Por tanto, $R^2_{1.23}$ representará esa proporción de varianza asociada. Si, por ejemplo, $R_{1.23} = 0{,}80$, diremos que 0,64 es la proporción de la varianza total de X_1 que está asociada a la variación de X_2 y X_3 y que 0,36 es la proporción que no está asociada. De otro modo, diremos que el 64 por 100 de las diferencias individuales en X_1 está asociado, es atribuible a, queda explicado por, ... las diferencias individuales en las dos variables predictoras, X_2 y X_3, y el 36 por 100 restante no está asociado, no es atribuible a, no queda explicado por, ... las diferencias individuales en las dos variables predictoras.

Igual que en el caso de dos variables, podemos exponer las ideas anteriores de un modo algo más intuitivo. De las n personas, unas son altas y otras son bajas en X_1 (criterio). Esa variabilidad es debida a la posesión en alto o bajo grado de ciertos factores (quienes los poseen en alto grado, obtienen altas puntuaciones en X_1; quienes los poseen en bajo grado, obtienen bajas puntuaciones en X_1). Parte de esos factores son tales que su posesión en alto o bajo grado hacen, también, que las personas obtengan puntuaciones altas o bajas, respectivamente, en X_2 y X_3 (variables predictoras). Pero el resto de los factores, aunque influyen en obtener altas o bajas puntuaciones en X_1, no influyen en obtener sistemáticamente puntuaciones altas o bajas en X_2 y X_3. Es decir, del hecho de poseer estos últimos factores en alto (bajo) grado, nada podemos concluir sobre si las puntuaciones en las variables predictoras serán, también, altas (bajas). Pues bien, los primeros factores darían razón de la varianza asociada y los restantes darían razón de la no asociada.

18.4. Resumen: Definiciones y fórmulas

$$\Sigma (X_1 - \bar{X}_1)^2 = \Sigma (X'_1 - \bar{X}_1)^2 + \Sigma (X_1 - X'_1)^2$$

$$s^2_1 = s^2_{x'_1} + s^2_{1.23}$$

$$R^2_{1.23} = \frac{s^2_{x'_1}}{s^2_{x_1}} = 1 - \frac{s^2_{1.23}}{s^2_{x_1}}$$

EJERCICIOS

18.1. A partir de los cuatro cuadros del ejercicio 17.12 (del capítulo anterior), calcular $s^2_{x'_1}$, $s^2_{1.23} = s^2_{x_1 - x'_1}$, $s^2_{z_1 - z'_1}$.

18.2. Demostrar que $r_{z'_1(z_1 - z'_1)} = 0$. Es decir, que es nulo el coeficiente de correlación de Pearson entre los pronósticos, z'_1, y los errores en dichos pronósticos $(z_1 - z'_1)$.

18.3. Demostrar que $r_{z_2(z_1 - z'_1)} = r_{z_3(z_1 - z'_1)} = 0$. Es decir, que es nulo el coeficiente de correlación de Pearson entre los errores cometidos en los pronósticos (en X_1) y las puntuaciones obtenidas en X_2 y entre los mismos errores y las puntuaciones obtenidas en X_3.

18.4. Demostrar que $\Sigma (X_1 - X'_1)(X'_1 - \bar{X}_1) = 0$.

18.5. Demostrar que $\Sigma (x_1 - x'_1)(x'_1) = 0$.

18.6. Demostrar que $\Sigma (z_1 - z'_1)(z'_1) = 0$.

18.7. Apoyándose en 18.6, demostrar que $r_{z'_1(z_1 - z'_1)} = 0$. Es decir, que es nulo el coeficiente de correlación de Pearson entre los pronósticos, z'_1, y los errores en dichos pronósticos, $z_1 - z'_1$.

18.8. Demostrar que $R_{2.13} = R_{3.12} = 1$, si $R_{1.23} = 1$.

18.9. Sean r_{12} y r_{13} las correlaciones simples entre un criterio, X_1, y dos variables predictoras, X_2 y X_3. Suponiendo que ambas correlaciones tienen un mismo valor (que llamaremos r), demostrar que la correlación simple, r_{23}, entre las dos variables predictoras es igual o mayor que $2r^2 - 1$.

Utilícese la fórmula $R^2_{1.23} = (r^2_{12} - 2r_{12}r_{13}r_{23} + r^2_{13})/(1 - r^2_{23})$ y del hecho de que $R^2_{1.23} \leq 1$.)

18.10. A partir de 18.9, demostrar que si $r_{12} = r_{13} = 1$, necesariamente $r_{23} = 1$.

18.11. Sean 10, 14, 6, 2, 8 las puntuaciones directas pronosticadas en un criterio X_1, mediante el plano de regresión de X_1 sobre X_2 y X_3. Siendo $s^2_{x_1} = 25$, decir qué parte de esta varianza no es atribuible a la variación de X_2 y X_3, y cuánto vale $R_{1.23}$.

18.12. Demostrar que $\Sigma x^2_1 = \Sigma (x_1 - x'_1)^2 + \Sigma x'^2_1$.

18.13. Demostrar que $\Sigma z^2_1 = \Sigma (z_1 - z'_1)^2 + \Sigma z'^2_1$.

V

Apéndices

Apéndice I

1. SIGNO (SIMPLE) DE SUMAR, Σ

1.1. Definición

Comencemos con algunos casos particulares.
Por definición:

$$\sum_{i=1}^{6} X_i = X_1 + X_2 + X_3 + X_4 + X_5 + X_6$$

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$$

Por definición:

$$\sum_{i=1}^{4} (X_i - 5)^3 = (X_1 - 5)^3 + (X_2 - 5)^3 + (X_3 - 5)^3 + (X_4 - 5)^3$$

$$\sum_{i=1}^{n} (X_i - 5)^3 = (X_1 - 5)^3 + (X_2 - 5)^3 + \cdots + (X_n - 5)^3$$

Por definición:

$$\sum_{i=1}^{3} \left(\frac{2X_i}{7} - 1\right)^2 = \left(\frac{2X_1}{7} - 1\right)^2 + \left(\frac{2X_2}{7} - 1\right)^2 + \left(\frac{2X_3}{7} - 1\right)^2$$

$$\sum_{i=1}^{n} \left(\frac{2X_i}{7} - 1\right)^2 = \left(\frac{2X_1}{7} - 1\right)^2 + \left(\frac{2X_2}{7} - 1\right)^2 + \cdots + \left(\frac{2X_n}{7} - 1\right)^2$$

De estos ejemplos particulares inferimos que el signo de sumar, Σ, viene a significar lo siguiente: «Sume los n términos obtenidos sustituyendo el subíndice i por $1, 2, \ldots, n$ en la expresión afectada por dicho signo».

1.2. Propiedades

a) $\sum_{i=1}^{n} (cX_i + k) = c \sum_{i=1}^{n} X_i + nk$, donde c y k son dos constantes arbitrarias.

En efecto:

$$\sum_{i=1}^{n} (cX_i + k) = (cX_1 + k) + (cX_2 + k) + \cdots + (cX_n + k) =$$
$$= (cX_1 + cX_2 + \cdots + cX_n) + (k + k \cdots + k) =$$
$$= c(X_1 + X_2 + \cdots + X_n) + nk =$$
$$= c \sum_{i=1}^{n} X_i + nk$$

b) $\sum_{i=1}^{n} (X_i + k) = \sum_{i=1}^{n} X_i + nk$. Es un caso particular de *a)*, para $c = 1$, $k \neq 0$.

c) $\sum_{i=1}^{n} cX_i = c \sum_{i=1}^{n} X_i$. Es un caso particular de *a)*, para $c \neq 0$, $k = 0$.

d) $\sum_{i=1}^{n} k = nk$. Es un caso particular de *a)*, para $c = 0$, $k \neq 0$.

e) $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$.

En efecto:

$$\sum_{i=1}^{n} (X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + \cdots + (X_n + Y_n) =$$
$$= (X_1 + X_1 + \cdots + X_n) + (Y_1 + Y_2 + \cdots + Y_n) =$$
$$= \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$$

f) $\sum_{i=1}^{n} (X_i - k)^2 = \sum_{i=1}^{n} X_i^2 - 2k \sum_{i=1}^{n} X_i + nk^2$.

En efecto:

$$\sum_{i=1}^{n} (X_i - k)^2 = \sum_{i=1}^{n} (X_i^2 - 2k\, X_i + k^2) = \sum_{i=1}^{n} X_i^2 - 2k \sum_{i=1}^{n} X_i + nk^2$$

g) En general, $\sum_{i=1}^{n} X_i^2 \neq \left(\sum_{i=1}^{n} X_i \right)^2$. Es decir, la suma de los cuadrados es distinta del cuadrado de la suma.

En efecto:

$$\left(\sum_{i=1}^{n} X_i \right)^2 = (X_1 + X_2 + \cdots + X_n)^2 = X_1^2 + X_2^2 + \cdots + X_n^2 + X_1X_2 +$$
$$+ X_1X_3 + \cdots + X_nX_{n-1} = \sum_{i=1}^{n} X_i^2 + X_1X_2 + X_1X_3 + \cdots +$$
$$+ X_nX_{n-1}$$

Así, por ejemplo, para:

$$X_1 = 2,\ X_2 = 3,\ X_3 = 4,\ \sum_{i=1}^{3} X_i^2 = 2^2 + 3^2 + 4^2 = 4 + 9 + 16 = 29$$
$$\left(\sum_{i=1}^{3} X_i \right)^2 = (2 + 3 + 4)^2 = 9^2 = 81$$

Sin embargo, es posible la igualdad, siempre que:

$$X_1X_2 + X_1X_3 + \cdots + X_nX_{n-1} = 0$$

Así, por ejemplo, para:

$$X_1 = -1,\ X_2 = 2,\ X_3 = 2,\ \sum_{i=1}^{3} X_i^2 = (-1)^2 + (2)^2 + (2)^2 = 1 + 4 + 4 = 9$$
$$\left(\sum_{i=1}^{3} X_i \right)^2 = (-1 + 2 + 2)^2 = (3)^2 = 9$$

En este último caso:

$$(-1)(2) + (-1)(2) + (2)(-1) + (2)(2) + (2)(-1) + (2)(2) = -8 + 8 = 0$$

h) En general, $\sum_{i=1}^{n} X_iY_i \neq \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i$. Es decir, la suma de los productos es distinta del producto de las sumas.

En efecto:

$$\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i = (X_1 + X_2 + \cdots + X_n)(Y_1 + Y_2 + \cdots + Y_n) = (X_1Y_1 +$$
$$+ X_2Y_2 + \cdots + X_nY_n) + (X_1Y_2 + X_1Y_3 + \cdots + X_nY_{n-1}) =$$
$$= \sum_{i=1}^{n} X_iY_i + (X_1Y_2 + X_1Y_3 + \cdots + X_nY_{n-1})$$

Así, por ejemplo, consideremos el cuadro siguiente:

X_i	Y_i	X_iY_i
2	3	(2)(3) = 6
4	6	(4)(6) = 24
5	1	(5)(1) = 5
$\sum_{i=1}^{3} X_i = 11$	$\sum_{i=1}^{3} Y_i = 10$	$\sum_{i=1}^{3} X_iY_i = 35$

$$\sum_{i=1}^{3} X_iY_i = 35 \neq 110 = (11)(10) = \sum_{i=1}^{3} X_i \sum_{i=1}^{3} Y_i$$

Sin embargo, es posible la igualdad, siempre que:

$$X_1Y_2 + X_1Y_3 + \cdots + X_nY_{n-1} = 0$$

Así, por ejemplo, consideremos el cuadro siguiente:

X_i	Y_i	X_iY_i
2	−2	(2)(−2) = −4
1	1	(1)(1) = 1
$\sum_{i=1}^{2} X_i = 3$	$\sum_{i=1}^{2} Y_i = -1$	$\sum_{i=1}^{2} X_iY_i = -3$

$$\sum_{i=1}^{2} X_iY_i = -3 = (3)(-1) = \sum_{i=1}^{2} X_i \sum_{i=1}^{2} Y_i$$

En este último caso, (2)(1) + (−2)(1) = 2 − 2 = 0.

NOTA. Ordinariamente, por sencillez, usaremos ΣX_i en vez de $\sum_{i=1}^{n} X_i$, y aun ΣX en vez de ΣX_i, pero entendiendo que existe un subíndice i que va de 1 a n (o a otro valor concreto, de acuerdo con el contexto de que se trate).

2. SIGNO (DOBLE) DE SUMAR, ΣΣ

2.1. Definición

Supongamos que un grupo total queda descompuesto en k subgrupos, con n_1, $n_2, \ldots, n_k$ personas respectivamente.

G. 1.º	G. 2.º	G. j	...	G. k
X_{11}	X_{12}	X_{1j}	...	X_{1k}
X_{21}	X_{22}	X_{2j}	...	X_{2k}
⋮	⋮	⋮		⋮
X_{i1}	X_{i2}	X_{ij}		X_{ik}
⋮	⋮	⋮		⋮
$X_{n_1 1}$	$X_{n_2 2}$	$X_{n_j j}$		$X_{n_k k}$

$$n_1 + n_2 + \cdots + n_k = n$$

X_{ij} representa la puntuación de la persona i perteneciente al grupo j.

La suma de las puntuaciones del grupo 1.º vendrá dada por $\sum_{i=1}^{n_1} X_{i1}$

La suma de las puntuaciones del grupo 2.º vendrá dada por $\sum_{i=1}^{n_2} X_{i2}$

. . . . . (A)

La suma de las puntuaciones del grupo j vendrá dada por $\sum_{i=1}^{n_j} X_{ij}$

. . . . . .

La suma de las puntuaciones del grupo k vendrá dada por $\sum_{i=1}^{n_k} X_{ik}$

La suma de las n puntuaciones del grupo total viene dada por $\sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}$. En efecto, consideremos $\sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij} = \sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij} \right)$, es decir, el signo simple de sumar (con j variando de 1 a k) aplicado a la expresión $\sum_{i=1}^{n_j} X_{ij}$. Sabemos que, por definición, $\sum_{j=1}^{k} \left(\sum_{i=1}^{n_j} X_{ij} \right) = \sum_{i=1}^{n_1} X_{i1} + \sum_{i=1}^{n_2} X_{i2} + \cdots + \sum_{i=1}^{n_k} X_{ik}$, que no es más que la suma de las n puntuaciones, si tenemos en cuenta (A).

Supongamos ahora que nos ofrecen las puntuaciones de n personas en k pruebas. Tendremos el cuadro siguiente:

		Pruebas					
		1	2	...	j	...	k
Personas	1	X_{11}	X_{12}	...	X_{1j}	...	X_{1k}
	2	X_{21}	X_{22}	...	X_{2j}	...	X_{2k}
	⋮	⋮	⋮		⋮		⋮
	i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ik}
	⋮	⋮	⋮		⋮		⋮
	n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{nk}
		$\sum_{i=1}^{n} X_{ij}$					

$\sum_{j=1}^{k} X_{ij}$

En este cuadro la fila i representa las puntuaciones obtenidas por la persona i en las k pruebas. La columna j representa las puntuaciones de las n personas en la

prueba j. La puntuación X_{ij} no es más que la puntuación obtenida por la persona i en la prueba j.

La suma de las puntuaciones de la persona i en las k pruebas vendrá dada por

$$X_{i1} + X_{i2} + \cdots + X_{ik} = \sum_{j=1}^{k} X_{ij}$$

La suma de las puntuaciones de las n personas en la prueba j vendrá dada por

$$X_{1j} + X_{2j} + \cdots + X_{nj} = \sum_{i=1}^{n} X_{ij}$$

La suma de las puntuaciones de las n personas en las k pruebas vendrá dada por $\sum_{j=1}^{k} \sum_{i=1}^{n} X_{ij}$. En efecto, consideremos $\sum_{j=1}^{k} \sum_{i=1}^{n} X_{ij} = \sum_{j=1}^{k} \left(\sum_{i=1}^{n} X_{ij} \right)$, es decir, el signo (simple) de sumar aplicado a la expresión $\sum_{i=1}^{n} X_{ij}$. Sabemos que, por definición,

$$\sum_{j=1}^{k} \left(\sum_{i=1}^{n} X_{ij} \right) = \sum_{i=1}^{n} X_{i1} + \sum_{i=1}^{n} X_{i2} + \cdots + \sum_{i=1}^{n} X_{ik} = (X_{11} + X_{21} + \cdots + X_{n1}) +$$

$+ (X_{12} + X_{22} + \cdots + X_{n2}) + \cdots + (X_{1k} + X_{2k} + \cdots + X_{nk})$ que no es más que la suma de las puntuaciones de las n personas en la prueba primera, más la suma de las puntuaciones de las n personas en la prueba segunda, . . . , más la suma de las puntuaciones de las n personas en la prueba k. Es decir, la suma de las puntuaciones de las n personas en las k pruebas.

Nótese que la suma de las puntuaciones de las n personas en las k pruebas puede venir dada de esta forma:

$$(X_{11} + X_{12} + \cdots + X_{1k}) + (X_{21} + X_{22} + \cdots + X_{2k}) + \cdots + (X_{n1} + X_{n2} + \cdots + X_{nk}) = \sum_{j=1}^{k} X_{1j} + \sum_{j=1}^{k} X_{2j} + \cdots + \sum_{j=1}^{k} X_{nj} = \sum_{i=1}^{n} \left(\sum_{j=1}^{k} X_{ij} \right) = \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}$$

En otras palabras:

$$\sum_{j=1}^{k} \sum_{i=1}^{n} X_{ij} = \sum_{i=1}^{n} \sum_{j=1}^{k} X_{ij}$$

De todo lo dicho inferimos que el doble signo de sumar viene a significar lo siguiente: «Sume todos los términos obtenidos sustituyendo los subíndices i y j por todos sus valores posibles en la expresión afectada por dicho signo».

EJEMPLO A.1

Grupo 1.º	Grupo 2.º	Grupo 3.º
4	5	10
2	4	8
8	2	6
5		7
		9

$n_1 = 4 \quad , \quad n_2 = 3 \quad , \quad n_3 = 5$

$$\sum_{j=1}^{3} \sum_{i=1}^{n_j} X_{ij} = \sum_{i=1}^{4} X_{i1} + \sum_{i=1}^{3} X_{i2} + \sum_{i=1}^{5} X_{i3} = (4 + 2 + 8 + 5) + (5 + 4 + 2) + (10 + 8 + 6 + 7 + 9) = 19 + 11 + 40 = 70$$

$$\sum_{j=1}^{3} \sum_{i=1}^{n_j} (X_{ij} - 4)^2 = \sum_{i=1}^{4} (X_{i1} - 4)^2 + \sum_{i=1}^{3} (X_{i2} - 4)^2 + \sum_{i=1}^{5} (X_{i3} - 4)^2 =$$

$$= [(4 - 4)^2 + (2 - 4)^2 + (8 - 4)^2 + (5 - 4)^2] + [(5 - 4)^2 + (4 - 4)^2 + (2 - 4)^2] + [(10 - 4)^2 + (8 - 4)^2 + (6 - 4)^2 + (7 - 4)^2 + (9 - 4)^2] = (0 + 4 + 16 + 1) + (1 + 0 + 4) + (36 + 16 + 4 + 9 + 25) = 21 + 5 + 90 = 116$$

2.2. Propiedades

a) $\sum_{j=1}^{m} \sum_{i=1}^{n} (c\,X_{ij} + k) = c \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij} + mn\,k$, donde c y k son dos constantes arbitrarias.

En efecto:

$$\sum_{j=1}^{m} \sum_{i=1}^{n} (cX_{ij} + k) = \sum_{i=1}^{n} (cX_{i1} + k) + \sum_{i=1}^{n} (cX_{i2} + k) + \cdots + \sum_{i=1}^{n} (cX_{im} + k) =$$

$$= \left(c \sum_{i=1}^{n} X_{i1} + nk \right) + \left(c \sum_{i=1}^{n} X_{i2} + nk \right) + \cdots +$$

$$+ \left(c \sum_{i=1}^{n} X_{im} + nk \right) = c \left(\sum_{i=1}^{n} X_{i1} + \sum_{i=1}^{n} X_{i2} + \cdots + \sum_{i=1}^{n} X_{im} \right) + mnk = \sum_{j=1}^{m} c \sum_{i=1}^{n} X_{ij} + mnk = c \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij} + mnk$$

b) $\sum_{j=1}^{m} \sum_{i=1}^{n} (X_{ij} + k) = \sum_{j=1}^{m} \sum_{i=1}^{n} X_{ij} + mnk$. Es un caso particular de *a*), con $c = 1, k \neq 0$.

c) $\sum_{j=1}^{m}\sum_{i=1}^{n} cX_{ij} = c\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}$. Es un caso particular de *a*), con $c \neq 0, k = 0$.

d) $\sum_{j=1}^{m}\sum_{i=1}^{n} k = mnk$. Es un caso particular de *a*), con $c = 0, k \neq 0$.

e) $\sum_{j=1}^{m}\sum_{i=1}^{n} (X_{ij} + Y_{ij}) = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij} + \sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}$.

En efecto:

$$\sum_{j=1}^{m}\sum_{i=1}^{n} (X_{ij} + Y_{ij}) = \sum_{i=1}^{n} (X_{i1} + Y_{i1}) + \sum_{i=1}^{n} (X_{i2} + Y_{i2}) + \cdots +$$

$$+ \sum_{i=1}^{n} (X_{im} + Y_{im}) = \left[\sum_{i=1}^{n} X_{i1} + \sum_{i=1}^{n} X_{i2} + \cdots + \sum_{i=1}^{n} X_{im}\right] +$$

$$+ \left[\sum_{i=1}^{n} Y_{i1} + \sum_{i=1}^{n} Y_{i2} + \cdots + \sum_{i=1}^{n} Y_{im}\right] =$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij} + \sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}$$

f) $\sum_{j=1}^{m}\sum_{i=1}^{n} (X_{ij} - k)^2 = \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 - 2k\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij} + mnk^2$. Su legitimación queda como ejercicio.

g) En general, $\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 \neq \left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2$. Su legitimación queda como ejercicio.

h) En general, $\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}Y_{ij} \neq \sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}$. Su legitimación queda como ejercicio.

Vamos a proponer ahora dos ejemplos comprobatorios de *g*) y *h*).

Para comprobar *g*), supongamos el cuadro siguiente con $m = 3$ y $n = 2$.

EJEMPLO A.2

4	5	2
3	6	4

$$\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}^2 = 4^2 + 3^2 + 5^2 + 6^2 + 2^2 + 4^2 = 16 + 9 + 25 + 36 + 4 + 16 = 106$$

$$\left(\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\right)^2 = (4 + 3 + 5 + 6 + 2 + 4)^2 = 24^2 = 576$$

Para comprobar *h*), supongamos los dos cuadros siguientes, ambos con $m = 2$ y $n = 3$.

EJEMPLO A.3

X_{ij}

2	6
3	4
5	2

Y_{ij}

1	4
5	1
2	3

$$\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}Y_{ij} = (2)(1) + (3)(5) + (5)(2) + (6)(4) + (4)(1) + (2)(3) = 2 + 15 + 10 + 24 + 4 + 6 = 61$$

$$\sum_{j=1}^{m}\sum_{i=1}^{n} X_{ij}\sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij} = (2 + 3 + 5 + 6 + 4 + 2)(1 + 5 + 2 + 4 + 1 + 3) = (22)(16) = 352$$

EJERCICIOS

1. Siendo $X_1 = 3$, $X_2 = -1$, $X_3 = 4$, calcular el valor numérico de las siguientes expresiones, donde suponemos que el subíndice i va de 1 a 3.

a) ΣX_i

b) ΣX_i^2

c) $(\Sigma X_i)^2$

d) $\Sigma 4X_i$

e) $\Sigma (X_i/3)$

f) $\Sigma (X_i - 5)$

g) $\Sigma [(X_i - 7)/6]$

h) $\Sigma (3X_i + 2X_i^2)$

i) $\Sigma (2X_i^2 - X_i/3 + 4)$

j) $\Sigma (X_i - 3)^2$

k) $[\Sigma (X_i - 3)]^2$

2. Supongamos:

a) $X_1 = 1, X_2 = 1, X_3 = -2$; $Y_1 = -1, Y_2 = 3, Y_3 = 2$

b) $X_1 = 2, X_2 = -2$; $Y_1 = 1, Y_2 = 4$

c) $X_1 = 2, X_2 = 0, X_3 = 1$; $Y_1 = 3, Y_2 = -1, Y_3 = -2$

d) $X_1 = 2, X_2 = 1/2$; $Y_1 = -1/2, Y_2 = 1/3$

En este supuesto, calcular el valor numérico de las siguientes expresiones:

2.1. $\Sigma (2X_i + Y_i)$

2.2. $\Sigma (X_iY_i)$

2.3. $\Sigma (X_iY_i)^2$

2.4. $(\Sigma X_iY_i)^2$

2.5. $\Sigma X_i \Sigma Y_i$

2.6. $\Sigma (6X_i)(Y_i/2)$

2.7. $\Sigma (2X_i + Y_i)^2$

2.8. $\Sigma [(2X_i)^2 + Y_i^2]$

2.9. $[\Sigma (2X_i + Y_i)]^2$

3. Demostrar que, en general, $\Sigma (X_i + Y_i)^2 \neq \Sigma (X_i^2 + Y_i^2)$.

4. Demostrar que, en general, $[\Sigma (X_i + Y_i)]^2 \neq \Sigma (X_i + Y_i)^2$.

5. Demostrar que, en general, $[\Sigma (X_i + Y_i)]^2 \neq \Sigma (X_i^2 + Y_i^2)$.

6. Supongamos que a tres personas les han sido aplicadas dos pruebas. Llamemos X_{ij} a la puntuación obtenida por la persona i en la prueba j. Es decir, X_{11} es la puntuación obtenida por la persona primera en la prueba primera, X_{12} es la puntuación obtenida por la persona primera en la prueba segunda, etc. Supongamos, además, que las puntuaciones obtenidas por estas tres personas en las dos pruebas son las siguientes:

	Pruebas (j)	
Personas (i)	$X_{11} = 2$	$X_{12} = 1$
	$X_{21} = -1$	$X_{22} = 2$
	$X_{31} = 3$	$X_{32} = -2$

En este supuesto, desarrollar las siguientes expresiones y calcular su valor numérico:

a) $\sum_i \sum_j X_{ij}$ *b)* $\sum_i \sum_j X_{ij}^2$ *c)* $\left(\sum_i \sum_j X_{ij}\right)^2$

d) $\sum_i \sum_j 7$ *e)* $\sum_i \sum_j 5X_{ij}$ *f)* $\sum_i \left(\sum_j X_{ij}\right)^2$

g) $\sum_j \left(\sum_i X_{ij}\right)^2$ *h)* $\sum_i \sum_j (X_{ij} + 9)$ *i)* $\sum_i \left(\sum_j X_{ij} + 9\right)$

j) $\sum_j \left(\sum_i X_{ij} + 9\right)$.

7. Demostrar que $\sum_{i=1}^{r}\left(\sum_{j=1}^{s} X_{ij} + k\right) = \sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij} + rk$.

8. Demostrar que $\sum_{j=1}^{s}\left(\sum_{i=1}^{r} X_{ij} + k\right) = \sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij} + sk$.

9. Demostrar que $\sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}^2 = \sum_{i=1}^{r}\left(\sum_{j=1}^{s} X_{ij}^2\right) = \sum_{j=1}^{s}\left(\sum_{i=1}^{r} X_{ij}^2\right)$.

10. Demostrar que, en general, $\left(\sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}\right)^2 \neq \sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}^2$.

11. Demostrar que, en general, $\sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}Y_{ij} \neq \sum_{i=1}^{r}\sum_{j=1}^{s} X_{ij}\sum_{i=1}^{r}\sum_{j=1}^{s} Y_{ij}$.

12. Demostrar que, en general, $\sum_{i=1}^{r}\left(\sum_{j=1}^{s} X_{ij}\right)^2 \neq \sum_{j=1}^{s}\left(\sum_{i=1}^{r} X_{ij}\right)^2$.

Apéndice II. Soluciones a los ejercicios propuestos

CAPÍTULO 4

4.1. *a)* 14,5 – 24,5 ; *b)* 62,45 – 68,55 ; *c)* 19,95 – 20,85 ; *d)* 44,345 – 54,355

4.2. *a)* 10 ; *b)* 6,10 ; *c)* 0,90 ; *d)* 10,010

4.3. *a)* 19,5 ; *b)* 65,5 ; *c)* 20,4 ; *d)* 49,35

4.4.

X	n_j	n_a	p	p_a	%	$\%_a$
26-28	5	40	0,125	1,000	12,5	100,0
23-25	6	35	0,150	0,875	15,0	87,5
20-22	12	29	0,300	0,725	30,0	72,5
17-19	10	17	0,250	0,425	25,0	42,5
14-16	7	7	0 175	0,175	17,5	17,5
	40		1,000		100,0	

También habrían sido posible los intervalos 13-15, 16-18, 19-21, 22-24 y 25-27.

4.6.

X	n_j	n_a	p	p_a	%	$\%_a$
17-19	8	50	0,16	1,00	16	100
14-16	9	42	0,18	0,84	18	84
11-13	12	33	0,24	0,66	24	66
8-10	10	21	0,20	0,42	20	42
5-7	7	11	0,14	0,22	14	22
2-4	4	4	0,08	0,08	8	8
	50		1,00		100	

4.7.

n_j	p	%
42.572	0,2267	22,67
25.683	0,1368	13,68
22.665	0,1207	12,07
8.083	0,0430	4,30
49.049	0,2612	26,12
37.578	0,2001	20,01
2.166	0,0115	1,15
187.796	1,0000	100,00

4.8.

X	n_j	n_a	p	p_a	%	$\%_a$
20-24	12	80	0,150	1,000	15,00	100,00
15-19	18	68	0,225	0,850	22,50	85,00
10-14	24	50	0,300	0,625	30,00	62,50
5-9	16	26	0,200	0,325	20,00	32,50
0-4	10	10	0,125	0,125	12,50	12,50
	80		1,000		100,00	

4.9.

X		n_j	n_a	p	p_a	%	$\%_a$
92-95	////	4	20	0,20	1,00	20	100
88-91	//// ///	8	16	0,40	0,80	40	80
84-87	//// /	6	8	0,30	0,40	30	40
80-83	//	2	2	0,10	0,10	10	10
		20		1,00		100	

CAPÍTULO 5

5.1. *a*) 5,25 ; *b*) 20 ; *c*) 2,8 ; *d*) −0,5 ; *e*) 0,633 ; *f*) −0,5875 ; *g*) −0,051

5.2. *a*) 7,2 ; *b*) 91,2 ; *c*) 75,7 ; *d*) 34,237 ; *e*) 31,875 ; *f*) 0,1379 ; *g*) 10,417

5.3. Al intervalo 11-13 le corresponde 5; al intervalo 2-4 le corresponde 2

5.4. 4 y 12

5.6. 15

5.7. $\bar{X}_g = 2,20$ $(\bar{X} = 2,25)$

5.8. 3 y 12

5.9. También, k

5.10. 30

5.11. 65,727

5.12. 6,5

5.13. Médicos: 0,308; Abogados: 0,467; Ingenieros: 0,225

5.14. Solteros: 0,27; Casados: 0,51; Viudos: 0,22

5.15. *a*) 4,9 ; *b*) 22 ; *c*) 29,75 ; *d*) 42,278 *e*) 28,324 ; *f*) 123,773 ; *g*) 88,5 ; *h*) 100,2

5.16. *a*) 5 ; *b*) 42 ; *c*) 17 ; *d*) 18 ; *e*) 11,75 ; *f*) 4,9 ; *g*) 7,625 ; *h*) 22,25 ; *i*) 3,6

5.17. *a*) Sí ; *b*) No

5.18. Mediana

5.19. *a*) Mediana ; *b*) Mediana ; *c*) Media

5.20. 4,375

5.21. *a*) $P_{15} = 70,2353$; $P_{25} = 73,1765$; $P_{36} = 76,2105$; $P_{75} = 86,6667$; $P_{82} = 89,0000$

5.22. *a*) $P_{25} = 89,000$; $P_{36} = 91,625$; $P_{75} = 100,045$

5.23. 24,7 (percentil 12), 38,5 (percentil 88)

5.24. No tiene que serlo (aunque puede serlo)

5.25. No. Un percentil es una puntuación que puede ser positiva, negativa o nula

CAPÍTULO 6

6.1. *a*) 9,20; 3,033 ; *b*) 1,50; 1,225 ; *c*) 4,40; 2,098 ; *d*) 2; 1,414 ; *e*) 5,33; 2,309 ; *f*) 8,67; 2,944

6.2. *a*) 2,4 ; *b*) 1 ; *c*) 2 ; *d*) 1,333 ; *e*) 2 ; *f*) 2,667

6.3. *a*) 7,5; 2,739 ; *b*) 7,56; 2,75 ; *c*) 6,56; 2,561 ; *d*) 24,01; 4,9 ; *e*) 8,64; 2,939 ; *f*) 26,188; 5,117 ; *g*) 7,29; 2,7 ; *h*) 24,422; 4,942

6.4. *a*) 2 ; *b*) 2,4 ; *c*) 2,24 ; *d*) 3,78 ; *e*) 2,4 ; *f*) 3,85 ; *g*) 2,24 ; *h*) 4,07

6.5. *a*) 6,7451 ; *b*) 5,5225

6.7. $(1-3)^2 + (2-3)^2 + (6-3)^2 = 14$; $(1-5)^2 + (2-5)^2 + (6-5)^2 = 26$; $(3)(3-5)^2 = 12$; $14 = 26 - 12$

6.8. $\Sigma (X_i - 5)^2 = 12$, $n(\bar{X} - 5)^2 = (8)(4,75 - 5)^2 = 0,5$
$\Sigma (X_i - \bar{X})^2 = 12 - 0,5 = 11,5$

6.9. Elegir $k = 1,75$, $s_x^2 = \dfrac{127,5}{8} = 15,9375$

6.10. 69

6.11. 65
6.12. *a*) 60,66 ; *b*) 40,833 ; *c*) 69,933 ; *d*) 70,70 ; *e*) 57,725 ; *f*) 49,067
6.13. 25; 24
6.14. 100
6.15. $X = 5$, Med = 2

CAPÍTULO 7

7.1. 0,6
7.2. 2.2
7.4. a) $a_3 = 0{,}5145$, $As = 0{,}059$
b) $a_3 = -0{,}328$, $As = -0{,}355$
7.5. a) $a_4 = -0{,}6345$, b) $a_4 = -0{,}336$
7.6. $P_{75} = 12{,}5$, $CV = 40{,}32$
7.7. 50

CAPÍTULO 8

8.1. *a*) dif.: -3; 1; 3; 0; -1 ; típ.: $-1{,}5$; 0,5; 1,5; 0; $-0{,}5$
b) dif. 3; 0; -3; 1; -1 ; típ.: 1,5; 0; $-1{,}5$; 0,5; $-0{,}5$
c) dif.: -5; 5; 5; -5 ; típ.: -1; 1; 1: -1
d) dif.: -1; 0; 2; -1 ; típ.: $-\frac{1}{\sqrt{1{,}5}}$, ó, $\frac{2}{\sqrt{1{,}5}}, \frac{-1}{\sqrt{1{,}5}}$
e) dif.: -3; 0; 3 ; típ.: $-\frac{3}{\sqrt{6}}$, ó, $\frac{3}{\sqrt{6}}$

8.2. *a*) dif.: 8 , directa: 28
b) dif.: 6 , directa: 26
c) dif.: -4, , directa: 16
d) dif.: -1 , directa: 19
e) dif.: 3,6 , directa: 23,6
8.4. 40
8.5. No (su varianza vale 2)
8.6. $Y_i = \frac{X_i - 7}{4}\,12 + 50$; 56, 44, 50, 32, 68
8.7. 120
8.8. 9, 5, 1. 7, 13
8.9. X_i: 13, 9, 7, 10, 11 ; Y_i: 28, 12, 4, 16, 20
8.10. 10
8.11. $\bar{X} = 20$, $\bar{Y} = 12$, $s_x = 8$, $s_y = 6$
8.12. *a*) aplicar la misma transformación $3X + 45$
b) media por debajo, media (aproximadamente) por encima

8.13. *a*) 6,68 %; 10 ; *b*) 84,13 %; 126 ; *c*) 69,15 %; 104 ; *d*) 22,66 %; 34 ; *e*) 28,57 %; 43 ; *f*) 20,29 %; 30 ; *g*) 53,28 %; 80
8.14. *a*) 40,16 ; *b*) 48,80 ; *c*) 50,40 ; *d*) 47,12 ; *e*) 44,08; 55,92 ; *f*) 48; 52
8.15. 30,50 **8.16.** 16,48 **8.17.** 6 **8.18.** 4,50
8.19. 60; 5 **8.20.** 34; 5 **8.21.** 37.900 **8.22.** 110
8.23. 410 **8.24.** 58; 8 **8.25.** 75; 10
8.26. 23,52; 192,48 $\simeq$ 192 **8.27.** 6,68 % **8.28.** 6,67
8.29. 20; 5
8.30. Sí **8.31** No **8.32** No **8.33** No
8.34.

a) T	*b*) T
66,4	67,5
59,9	59,9
55,5	54,1
50,5	48,0
42,9	40,1
32,5	29,5

CAPÍTULO 9

9.1.

		X		
		2-5	6-9	10-13	
Y	11-15	0	4	6	10
	6-10	8	2	0	10
		8	6	6	20

9.2. 8,0375
9.3. *a*) 7,1; 10,5 ; *b*) $\bar{X}_{Y=8} = 4{,}3$, $\bar{X}_{Y=13} = 9{,}9$, $\bar{Y}_{X=3,5} = 8$, $\bar{Y}_{X=7,5} = 11{,}33$; $\bar{Y}_{X=11,5} = 13$; *c*) 11,04 ; 6,25 ; *d*) $s^2_{xy=8} = 2{,}56$, $s^2_{xy=13} = 3{,}84$, $s^2_{yx=3,5} = 0$, $s^2_{yx=7,5} = 5{,}556$, $s^2_{yx=11,5} = 0$; *e*) 7
9.4. *a*) 3,2 , 7,4 ; *b*) $\bar{X}_{Y=5} = 3{,}6$, $\bar{X}_{Y=8} = 3{,}1$, $\bar{Y}_{X=2} = 7{,}7$, $\bar{Y}_{X=4} = 7{,}2$; *c*) 0,96 , 1,44 ; *d*) $s^2_{xy=5} = 0{,}64$, $s^2_{xy=8} = 0{,}99$, $s^2_{yx=2} = 0{,}81$, $s^2_{yx=4} = 1{,}76$; *e*) $-0{,}24$
9.5. *a*) 5,225 , 5,8 ; *b*) 5 ; *c*) 8,154 ; *d*) 5,125 , 10,26 ; *e*) 4,5 ; *f*) 5,589 ; *g*) 4,2075

9.10.

X	Y	XY
2	4	8
4	4	16
8	10	80
10	14	140

CAPÍTULO 10

10.1. *a*) 0,9 ; *b*) −0,6 ; *c*) 0,189 ; *d*) −0,913 ; *e*) 0,575 ; *f*) −0,789

10.2. *a*) 0,356 ; *b*) −0,853 ; *c*) 0,519 ; *d*) −0,556 ; *e*) −0,657

10.5. Recuerde que el numerador de r_{xy} es $n \Sigma XY - \Sigma X \Sigma Y$, que será necesariamente 0 si $r_{xy} = 0$

10.6. No **10.7.** No **10.9.** *a*) Tiene perfecto sentido ; *b*) El mismo (recuerde las propiedades de r_{xy})

10.10. *a*) 244 , *b*) 484 , *c*) 4

10.12. 6 ; **10.13.** $r_{xv} = \dfrac{s_x - rs_y}{\sqrt{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}}$, $r_{yv} = \dfrac{rs_x - s_y}{\sqrt{s_x^2 + s_y^2 - 2r_{xy}s_x s_y}}$

10.14. 0,65

10.15. Para $r_{xy} = 1$, 37; para $r_{xy} = 0$, 41; para $r_{xy} = -1$, 45

10.16. 0,875

10.17. Cuando $r_{xv} = 0$, $s_w = 13$; cuando $r_{xv} = 1$, $s_w = 1$

10.18. $r_{DH} = 0$

CAPÍTULO 11

11.1. *a*) $Y' = 0{,}3 + (0{,}9)X$, $y' = (0{,}9)x$, $z'_y = (0{,}9)z_x$
b) $Y' = 7 + (-0{,}6)X$, $y' = (-0{,}6)x$, $z'_y = (-0{,}6)z_x$
c) $Y' = 1{,}786 + (0{,}071)X$, $y' = (0{,}071)x$, $z'_y = (0{,}189)z_x$
d) $Y' = 3{,}5 + (-0{,}5)X$, $y' = (-0{,}5)x$, $z'_y = (-0{,}913)z_x$
e) $Y' = 3{,}4 + (1{,}15)X$, $y' = (1{,}15)x$, $z'_y = (0{,}575)z_x$
f) $Y' = 5{,}422 + (-0{,}692)X$, $y' = (-0{,}692)x$, $z'_y = (-0{,}879)z_x$

11.2. $Y'(0{,}4;\ 10{,}0;\ 19{,}6;\ 6{,}8;\ 13{,}2)$, $y'(-9{,}6;\ 0;\ 9{,}6;\ -3{,}2;\ 3{,}2)$
$z'_y(-1{,}2;\ 0;\ 1{,}2;\ -0{,}4;\ 0{,}4)$

11.3. *a*) 20 ; *b*) 30 ; *c*) 40

11.4. $Y' = 5 + (2)X$, $y' = (2)x$

11.5. 5 , 7 **11.6.** $y' = (0{,}50)x$

11.7. $Y' = 6 + (0{,}8)X$

CAPÍTULO 12

12.2. Basta con recordar 10.2.3.*b* y tener en cuenta que $y' = r_{xy}(s_y/s_x)x$, donde $r_{xy}(s_y/s_x)$ es una constante

12.3. Basta con demostrar que $\Sigma x(y - y') = 0$, ó $\Sigma xy = \Sigma xy'$

12.4. Basta con demostrar que $\Sigma y'(y - y') = 0$, ó, $\Sigma yy' = \Sigma y'^2$

12.5. Basta con recordar que $\bar{Y} = \bar{Y}'$

12.6. La igualdad propuesta equivale a $\Sigma (y - y')y' = 0$, que ha quedado demostrada en 12.4

12.7. Sí **12.8.** No necesariamente

12.9. 4

12.13

x	y'	y'^2
−6	−9	81
−2	−3	9
6	9	81
0	0	0
2	3	9

12.14.

X	Y
1	6
13	10
4	7
−5	4
7	8

12.15. 0,80 **12.18.** 1,2 **12.21.** 1,44

CAPÍTULO 13

13.1. *a*) $\eta^2_{yx} = 0{,}787$; *b*) $\eta^2_{yx} = 0{,}423$; *c*) $\eta^2_{yx} = 0{,}808$

13.2. *a*) $\eta^2_{yx} = 0{,}592$; *b*) $\eta^2_{yx} = 0{,}625$; *c*) $\eta^2_{yx} = 0{,}200$; *d*) $\eta^2_{yx} = 0{,}663$

13.4.

X	x	Y'
1	−3	5
6	2	5
4	0	5
5	1	5

CAPÍTULO 14

14.1. 0,57 **14.2.** 0,43 **14.3.** 0,69 **14.4.** 0,43

14.5. *a*) 0,7 ; *b*) −0,8 ; *c*) 0,714 ; *d*) 0411

14.6. *a*) 0,667 ; *b*) −0,2 ; *c*) 0,067 ; *d*) −0,524

14.7. $I = 38$; $NI = 152$
14.8. *a*) 0,739 ; *b*) −0,806 ; *c*) −0,606 ; *d*) 0,601 ; *e*) 0,460 ; *f*) −0,282

CAPÍTULO 15

15.1. *a*) 0,667 ; *b*) 0,6 ; *c*) −0,556 ; *d*) 0,280
15.2. *a*) 24,242 ; *b*) 0,284 ; *c*) 15,139 ; *d*) 7,778 ; *e*) 19,833 ; *f*) 59,686
15.3. *a*) 0,442 ; *b*) 0,06 ; *c*) 0,524 ; *d*) 0,529 ; *e*) 0,533 ; *f*) 0,611
15.4. 24
15.5. 0,8485

CAPÍTULO 16

16.1. *a*) −0,41 ; *b*) 0,30 ; *c*) 0,75 ; *d*) 0,46
16.2. *a*) 0,32 ; *b*) 0,37 ; *c*) 0,30 ; *d*) −0,38
16.4. *a*) 0,45 ; *b*) −0,10 ; *c*) −0,20 ; *d*) 0,15 ; *e*) 0,21
16.6. *a*) −0,51 ; *b*) 0,38 ; *c*) 0,95 ; *d*) 0,58
16.7. *a*) 0,40 ; *b*) 0,47 ; *c*) 0,38 ; *d*) −0,49
16.8. *a*) 0,72 ; *b*) −0,16 ; *c*) −0,31 ; *d*) 0,24 ; *e*) 0,33

CAPÍTULO 17

17.1. *a*) 0,722 ; *b*) 0,082 ; *c*) 0,665
17.2. 0,768
17.3. $r_{12.3} = r_{12}$
17.10. $z'_1 = (0{,}438)\, z_2 + (0{,}312)\, z_3$, $x'_1 = (0{,}875)\, x_2 + (0{,}625)\, x_3$
17.11. $z'_1 = (0{,}750)\, z_2 + (0{,}250)\, z_3$, $x'_1 = (0{,}6)\, x_2 + (0{,}5)\, x_3$
17.12. *a*) $z'_1 = (0{,}250)\, z_2 + (0{,}250)\, z_3$, $x'_1 = (0{,}5)\, x_2 + (0{,}25)\, x_3$
$X'_1 = 2{,}75 + (0{,}5)\, X_2 + (0{,}25)\, X_3$
b) $z'_1 = (0{,}92)\, z_2 + (0{,}12)\, z_3$, $x'_1 = (1{,}84)\, x_2 + (0{,}24)\, x_3$
$X'_1 = -0{,}32 + (1{,}84)\, X_2 + (0{,}24)\, X_3$
c) $z'_1 = (0{,}60114)\, z_2 + (0{,}13960)\, z_3$ $x'_1 = (0{,}20038)\, x_2 + (0{,}0698)\, x_3$
$X'_1 = 1{,}77778 + (0{,}20038)\, X_2 + (0{,}0698)\, X_3$
d) $z'_1 = (1/3)\, z_2 + (1/3)\, z_3$; $x'_1 = (1/3)\, x_2 + (1/3)\, x_3$;
$X'_1 = (1/3)\, X_2 + (1/3)\, X_3$
17.13. *a*) típ.: −0,50 ; −0,25 ; 0,00 ; 0,75 ; 0,00
dif.: −2,00 ; −1,00 ; 0,00 ; 3,00 ; 0,00
dir.: 4,00 ; 5,00 ; 6,00 ; 9,00 ; 6,00

b) típ.: −1,56 ; −0,28 ; 0,06 ; 0,40 ; 1,38
dif.: −6,24 ; −1,12 ; 0,24 ; 1,60 ; 5,52
dir.: 1,76 ; 6,88 ; 8,24 ; 9,60 ; 13,52

c) típ.: 0,20940 ; 0,37037 ; −0,97151 ; 0,90171 ; −0,50997
dif.: 0,41880 ; 0,74074 ; −1,94302 ; 1,80342 ; −1,01994
dir.: 4,41880 ; 4,74074 ; 2,05698 ; 5,80342 ; 2,98006

d) típ.: $-\sqrt{2/3}$; $-\sqrt{2/3}$; 0,00 ; $2\sqrt{2/3}$
dif.: −1/3 ; −1/3 ; 0,00 ; 2/3
dir.: 2/3 ; 2/3 ; 1,00 ; 5/3

17.15. *a*) 0,418 , *b*) 0,957 , *c*) 0,663 , *d*) 0,577

CAPÍTULO 18

18.1. *a*) 2,8 ; 13,2 ; 0,175 ; 0,825
b) 14,656 ; 1,344 ; 0,916 ; 0,084
c) 8,79 ; 11,21 ; 2,802 ; 2,198
d) 1/6 ; 1/3 ; 1/3 ; 2/3
18.11. 9 ; 0,80

SOLUCIONES AL APÉNDICE I

1 *a*) 6, *b*) 26, *c*) 36, *d*) 24, *e*) 2, *f*) −9
g) −15/6, *h*) 70, *i*) 62, *j*) 17, *k*) 9

2.1 *a*) 4, *b*) 5, *c*) 6, *d*) 29/6

2.2 *a*) −2, *b*) −6, *c*) 4, *d*) −5/6

2.3 *a*) 26, *b*) 68, *c*) 40, *d*) 37/36

2.4 *a*) 4, *b*) 36, *c*) 16, *d*) 25/36

2.5 *a*) 0, *b*) 0, *c*) 0, *d*) −5/12

2.6 *a*) −6, *b*) −18, *c*) 12, *d*) −5/2

2.7 *a*) 30, *b*) 25, *c*) 50, *d*) 505/36

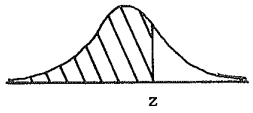
2.8 *a*) 38, *b*) 49, *c*) 34, *d*) 625/36

2.9 *a*) 16, *b*) 25, *c*) 36, *d*) 841/36

6. *a*) 5, *b*) 23, *c*) 25, *d*) 42, *e*) 25, *f*) 11
g) 17, *h*) 59, *i*) 32, *j*) 23

Apéndice III

TABLA A. Distribución normal. $P(Z \leq z)$.



z	0	1	2	3	4	5	6	7	8	9
-3'5	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002	'0002
-3'4	'0003	'0003	'0003	'0003	'0003	'0003	'0003	'0003	'0002	'0002
-3'3	'0005	'0005	'0004	'0004	'0004	'0004	'0004	'0004	'0004	'0003
-3'2	'0007	'0007	'0006	'0006	'0006	'0006	'0006	'0005	'0005	'0005
-3'1	'0010	'0009	'0009	'0009	'0008	'0008	'0008	'0008	'0007	'0007
-3'0	'0014	'0013	'0013	'0012	'0012	'0011	'0011	'0011	'0010	'0010
-2'9	'0019	'0018	'0017	'0017	'0016	'0016	'0015	'0015	'0014	'0014
-2'8	'0026	'0025	'0024	'0023	'0023	'0022	'0021	'0021	'0020	'0019
-2'7	'0035	'0034	'0033	'0032	'0031	'0030	'0029	'0028	'0027	'0026
-2'6	'0047	'0045	'0044	'0043	'0041	'0040	'0039	'0038	'0037	'0036
-2'5	'0062	'0060	'0059	'0057	'0055	'0054	'0052	'0051	'0049	'0048
-2'4	'0082	'0080	'0078	'0075	'0073	'0071	'0069	'0068	'0066	'0064
-2'3	'0107	'0104	'0102	'0099	'0096	'0094	'0091	'0089	'0087	'0084
-2'2	'0139	'0136	'0132	'0129	'0126	'0122	'0119	'0116	'0113	'0110
-2'1	'0179	'0174	'0170	'0166	'0162	'0158	'0154	'0150	'0146	'0143
-2'0	'0228	'0222	'0217	'0212	'0207	'0202	'0197	'0192	'0188	'0183
-1'9	'0287	'0281	'0274	'0268	'0262	'0256	'0250	'0244	'0238	'0233
-1'8	'0359	'0352	'0344	'0336	'0329	'0322	'0314	'0307	'0300	'0294
-1'7	'0446	'0436	'0427	'0418	'0409	'0401	'0392	'0384	'0375	'0367
-1'6	'0548	'0537	'0526	'0516	'0505	'0495	'0485	'0475	'0465	'0455
-1'5	'0668	'0655	'0643	'0630	'0618	'0606	'0594	'0582	'0570	'0559
-1'4	'0808	'0793	'0778	'0764	'0749	'0735	'0722	'0708	'0694	'0681
-1'3	'0968	'0951	'0934	'0918	'0901	'0885	'0869	'0853	'0838	'0823
-1'2	'1151	'1131	'1112	'1093	'1075	'1056	'1038	'1020	'1003	'0985
-1'1	'1357	'1335	'1314	'1292	'1271	'1251	'1230	'1210	'1190	'1170
-1'0	'1587	'1562	'1539	'1515	'1492	'1469	'1446	'1423	'1401	'1379
-0'9	'1841	'1814	'1788	'1762	'1736	'1711	'1685	'1660	'1635	'1611
-0'8	'2119	'2090	'2061	'2033	'2005	'1977	'1949	'1922	'1894	'1867
-0'7	'2420	'2389	'2358	'2327	'2297	'2266	'2236	'2206	'2177	'2148
-0'6	'2743	'2709	'2676	'2643	'2611	'2578	'2546	'2514	'2483	'2451
-0'5	'3085	'3050	'3015	'2981	'2946	'2912	'2877	'2843	'2810	'2776
-0'4	'3446	'3409	'3372	'3336	'3300	'3264	'3228	'3192	'3156	'3121
-0'3	'3821	'3783	'3745	'3707	'3669	'3632	'3594	'3557	'3520	'3483
-0'2	'4207	'4168	'4129	'4090	'4052	'4013	'3974	'3936	'3897	'3859
-0'1	'4620	'4562	'4522	'4483	'4443	'4404	'4364	'4325	'4268	'4247
-0'0	'5000	'4960	'4920	'4880	'4840	'4801	'4761	4721	'4681	'4641

(continúa)

Los valores interiores indican probabilidades. Delante de la coma decimal, ('), se entiende que va un cero. Así, por ejemplo, '1292 equivale a 0'1292 e indica que $P(Z \leq -1'13) = 0'1292$.

TABLA A. Distribución normal. $P(Z \leq z)$. *(continuación)*

z	0	1	2	3	4	5	6	7	8	9
0'0	'5000	'5040	'5080	'5120	'5160	'5199	'5239	'5279	'5319	'5359
0'1	'5398	'5438	'5478	'5517	'5557	'5596	'5636	'5675	'5714	'5753
0'2	'5793	'5832	'5871	'5910	'5948	'5987	'6026	'6064	'6113	'6141
0'3	'6179	'6217	'6255	'6293	'6331	'6368	'6406	'6443	'6480	'6517
0'4	'6554	'6591	'6628	'6664	'6700	'6736	'6772	'6808	'6844	'6879
0'5	'6915	'6950	'6985	'7019	'7054	'7088	'7123	'7157	'7190	'7224
0'6	'7257	'7291	'7324	'7357	'7389	'7422	'7454	'7486	'7517	'7549
0'7	'7580	'7611	'7642	'7673	'7703	'7734	'7764	'7794	'7823	'7852
0'8	'7881	'7910	'7939	'7967	'7995	'8023	'8051	'8078	'8106	'8133
0'9	'8159	'8186	'8212	'8238	'8264	'8289	'8315	'8340	'8365	'8389
1'0	'8413	'8438	'8461	'8485	'8508	'8531	'8554	'8577	'8599	'8621
1'1	'8643	'8665	'8686	'8708	'8729	'8749	'8770	'8790	'8810	'8830
1'2	'8849	'8869	'8888	'8907	'8925	'8944	'8962	'8980	'8997	'9015
1'3	'9032	'9049	'9066	'9082	'9099	'9115	'9131	'9147	'9162	'9177
1'4	'9192	'9207	'9222	'9236	'9251	'9265	'9278	'9292	'9306	'9319
1'5	'9332	'9345	'9357	'9370	'9382	'9394	'9406	'9418	'9430	'9441
1'6	'9452	'9463	'9474	'9484	'9495	'9505	'9515	'9525	'9535	'9545
1'7	'9554	'9564	'9573	'9582	'9591	'9599	'9608	'9616	'9625	'9633
1'8	'9641	'9648	'9656	'9664	'9671	'9678	'9686	'9693	'9700	'9706
1'9	'9713	'9719	'9726	'9732	'9738	'9744	'9750	'9756	'9762	'9767
2'0	'9772	'9778	'9783	'9788	'9793	'9798	'9803	'9808	'9812	'9817
2'1	'9821	'9826	'9830	'9834	'9838	'9842	'9846	'9850	'9854	'9857
2'2	'9861	'9864	'9868	'9871	'9874	'9878	'9881	'9884	'9887	'9890
2'3	'9893	'9896	'9898	'9901	'9904	'9906	'9909	'9911	'9913	'9916
2'4	'9918	'9920	'9922	'9925	'9927	'9929	'9931	'9932	'9934	'9936
2'5	'9938	'9940	'9941	'9943	'9945	'9946	'9948	'9949	'9951	'9952
2'6	'9953	'9955	'9956	'9957	'9959	'9960	'9961	'9962	'9963	'9964
2'7	'9965	'9966	'9967	'9968	'9969	'9970	'9971	'9972	'9973	'9974
2'8	'9974	'9975	'9976	'9977	'9977	'9978	'9979	'9979	'9980	'9981
2'9	'9981	'9982	'9982	'9983	'9984	'9984	'9985	'9985	'9986	'9986
3'0	'9986	'9987	'9987	'9988	'9988	'9989	'9989	'9989	'9990	'9990
3'1	'9990	'9991	'9991	'9991	'9992	'9992	'9992	'9992	'9993	'9993
3'2	'9993	'9993	'9994	'9994	'9994	'9994	'9994	'9995	'9995	'9995
3'3	'9995	'9995	'9995	'9996	'9996	'9996	'9996	'9996	'9996	'9996
3'4	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9997	'9998
3'5	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998	'9998

Los valores interiores indican probabilidades. Delante de la coma decimal, ('), se entiende que va un cero. Así, por ejemplo, '8925 equivale a 0'8925 e indica que $P(Z < 1'24) = 0'8925$.

FUENTE: BLUM, J. R. y ROSEMBLATT, J. I., *Probabilities and Statistics*, Filadelfia, Launders, 1972.

TABLA B. Funciònes de p, q e y.

p (o, q)	A $\sqrt{pq}$	B $\sqrt{p/q}$	C pq/y	D $\sqrt{pq}/y$	E p/y	F y	q (o, p)
0'99	0'0995	9'950	0'3715	3'733	37'148	0'02665	0'01
0'98	0'1400	7'000	0'4048	2'892	20'240	0'04842	0'02
0'97	0'1706	5'686	0'4277	2'507	14'256	0'06804	0'03
0'96	0'1960	4'899	0'4456	2'274	11'141	0'08617	0'04
0'95	0'2179	4'359	0'4605	2'113	9'211	0'1031	0'05
0'94	0'2375	3'958	0'4735	1'994	7'891	0'1191	0'06
0'93	0'2551	3'645	0'4848	1'900	6'926	0'1343	0'07
0'92	0'2713	3'391	0'4951	1'825	6'188	0'1487	0'08
0'91	0'2862	3'180	0'5043	1'762	5'604	0'1624	0'09
0'90	0'3000	3'000	0'5128	1'709	5'128	0'1755	0'10
0'89	0'3129	2'844	0'5206	1'664	4'733	0'1880	0'11
0'88	0'3250	2'708	0'5279	1'625	4'399	0'2000	0'12
0'87	0'3363	2'587	0'5346	1'590	4'112	0'2115	0'13
0'86	0'3470	2'478	0'5409	1'559	3'864	0'2226	0'14
0'85	0'3571	2'380	0'5468	1'532	3'646	0'2332	0'15
0'84	0'3666	2'291	0'5524	1'507	3'452	0'2433	0'16
0'83	0'3756	2'210	0'5576	1'484	3'280	0'2531	0'17
0'82	0'3842	2'134	0'5625	1'464	3'125	0'2624	0'18
0'81	0'3923	2'065	0'5671	1'446	2'985	0'2714	0'19
0'80	0'4000	2'000	0'5715	1'429	2'858	0'2800	0'20
0'79	0'4073	1'940	0'5756	1'413	2'741	0'2882	0'21
0'78	0'4142	1'883	0'5796	1'399	2'634	0'2961	0'22
0'77	0'4208	1'830	0'5832	1'386	2'536	0'3036	0'23
0'76	0'4271	1'780	0'5867	1'374	2'445	0'3109	0'24
0'75	0'4330	1'732	0'5900	1'363	2'360	0'3178	0'25
0'74	0'4386	1'687	0'5931	1'352	2'281	0'3244	0'26
0'73	0'4440	1'644	0'5961	1'343	2'208	0'3306	0'27
0'72	0'4490	1'604	0'5989	1'334	2'139	0'3366	0'28
0'71	0'4538	1'565	0'6015	1'326	2'074	0'3423	0'29
0'70	0'4583	1'528	0'6040	1'318	2'013	0'3477	0'30
0'69	0'4625	1'492	0'6063	1'311	1'956	0'3528	0'31
0'68	0'4665	1'458	0'6085	1'304	1'902	0'3576	0'32
0'67	0'4702	1'425	0'6106	1'298	1'850	0'3621	0'33
0'66	0'4737	1'393	0'6124	1'293	1'801	0'3664	0'34
0'65	0'4770	1'363	0'6142	1'288	1'755	0'3704	0'35
0'64	0'4800	1'333	0'6158	1'283	1'711	0'3741	0'36
0'63	0'4828	1'305	0'6174	1'279	1'669	0'3776	0'37
0'62	0'4854	1'277	0'6188	1'275	1'628	0'3808	0'38
0'61	0'4877	1'251	0'6200	1'271	1'590	0'3837	0'39
0'60	0'4899	1'225	0'6212	1'268	1'553	0'3863	0'40
0'59	0'4918	1'200	0'6223	1'265	1'518	0'3888	0'41
0'58	0'4936	1'175	0'6232	1'263	1'484	0'3909	0'42
0'57	0'4951	1'151	0'6240	1'260	1'451	0'3928	0'43
0'56	0'4964	1'128	0'6247	1'259	1'420	0'3944	0'44
0'55	0'4975	1'106	0'6253	1'257	1'390	0'3958	0'45
0'54	0'4984	1'083	0'6258	1'256	1'360	0'3969	0'46
0'53	0'4991	1'062	0'6262	1'255	1'332	0'3978	0'47
0'52	0'4996	1'041	0'6264	1'254	1'305	0'3984	0'48
0'51	0'4999	1'020	0'6266	1'253	1'279	0'3988	0'49
0'50	0'5000	1'000	0'6267	1'253	1'253	0'3989	0'50

(continúa)

TABLA B. Funciones de p, q e y. *(continuación)*

p	B $\sqrt{p/q}$	E p/y
0'49	0'980	1'229
0'48	0'961	1'205
0'47	0'942	1'181
0'46	0'923	1'159
0'45	0'904	1'137
0'44	0'886	1'116
0'43	0'869	1'095
0'42	0'851	1'074
0'41	0'834	1'054
0'40	0'816	1'035
0'39	0'800	1'016
0'38	0'783	0'998
0'37	0'766	0'980
0'36	0'750	0'962
0'35	0'734	0'945
0'34	0'718	0'928
0'33	0'702	0'911
0'32	0'686	0'895
0'31	0'670	0'879
0'30	0'655	0'863
0'29	0'639	0'847
0'28	0'624	0'832
0'27	0'608	0'817
0'26	0'593	0'801
0'25	0'577	0'787
0'24	0'562	0'772
0'23	0'546	0'758
0'22	0'531	0'743
0'21	0'516	0'729
0'20	0'500	0'714
0'19	0'484	0'700
0'18	0'468	0'686
0'17	0'453	0'672
0'16	0'436	0'658
0'15	0'420	0'643
0'14	0'403	0'629
0'13	0'387	0'615
0'12	0'369	0'600
0'11	0'352	0'585
0'10	0'333	0'570
0'09	0'314	0'554
0'08	0'295	0'538
0'07	0'274	0'521
0'06	0'253	0'504
0'05	0'229	0'485
0'04	0'204	0'464
0'03	0'176	0'441
0'02	0'143	0'413
0'01	0'100	0'375

FUENTE: GUILFORD, J. P., *Fundamental Statistics in Psychology and Education*, 4.ª edición, 1965, apéndice B, tabla G.

TABLA C. Cálculo del coeficiente de correlación tetracórica, r_t. (*)

r_t	cb/ad o ad/cb	r_t	cb/ad o ad/cb	r_t	cb/ad o ad/cb
0'00	1'000	0'35	2'492-2'563	0'70	8'500-8'910
0'01	1'013-1'039	0'36	2'564-2'638	0'71	8'911-9'351
0'02	1'040-1'066	0'37	2'639-2'716	0'72	9'352-9'828
0'03	1'067-1'093	0'38	2'717-2'797	0'73	9'829-10'344
0'04	1'094-1'122	0'39	2'798-2'881	0'74	10'345-10'903
0'05	1'123-1'151	0'40	2'882-2'968	0'75	10'904-11'512
0'06	1'152-1'180	0'41	2'969-3'059	0'76	11'513-12'177
0'07	1'181-1'211	0'42	3'060-3'153	0'77	12'178-12'905
0'08	1'212-1'242	0'43	3'154-3'251	0'78	12'906-13'707
0'09	1'243-1'275	0'44	3'252-3'353	0'79	13'708-14'592
0'10	1'276-1'308	0'45	3'354-3'460	0'80	14'593-15'574
0'11	1'309-1'342	0'46	3'461-3'571	0'81	14'575-16'670
0'12	1'343-1'377	0'47	3'572-3'687	0'82	16'671-17'899
0'13	1'378-1'413	0'48	3'688-3'808	0'83	17'900-19'287
0'14	1'414-1'450	0'49	3'809-3'935	0'84	19'288-20'865
0'15	1'451-1'488	0'50	3'936-4'067	0'85	20'866-22'674
0'16	1'489-1'528	0'51	4'068-4'205	0'86	22'675-24'766
0'17	1'529-1'568	0'52	4'206-4'351	0'87	24'767-27'212
0'18	1'569-1'610	0'53	4'352-4'503	0'88	27'213-30'105
0'19	1'611-1'653	0'54	4'504-4'662	0'89	30'106-33'577
0'20	1'654-1'697	0'55	4'663-4'830	0'90	33'578-37'815
0'21	1'698-1'743	0'56	4'831-5'007	0'91	37'816-43'096
0'22	1'744-1'790	0'57	5'008-5'192	0'92	43'097-49'846
0'23	1'791-1'838	0'58	5'193-5'388	0'93	49'847-58'758
0'24	1'839-1'888	0'59	5'389-5'595	0'94	58'759-71'035
0'25	1'889-1'940	0'60	5'596-5'813	0'95	71'036-88'964
0'26	1'941-1'993	0'61	5'814-6'043	0'96	88'965-117'479
0'27	1'994-2'048	0'62	6'044-6'288	0'97	117'480-169'503
0'28	2'049-2'105	0'63	6'289-6'547	0'98	169'504-292'864
0'29	2'106-2'164	0'64	6'548-6'822	0'99	292'865-923'687
0'30	2'165-2'225	0'65	6'823-7'115	1'00	923'688- ∞
0'31	2'226-2'288	0'66	7'116-7'428		
0'32	2'289-2'353	0'67	7'429-7'761		
0'33	2'354-2'421	0'68	7'762-8'117		
0'34	2'422-2'491	0'69	8'118-8'499		

(*) Si $cb>ad$ calcúlese cb/ad y acéptese como positivo el valor obtenido; indica relación positiva entre la categoría 0 (1) de X y la categoría 0 (1) de Y. Si $cb<ad$, calcúlese ad/cb y acéptese como negativo el valor obtenido; indica relación negativa entre la categoría 0 (1) de X y la categoría 0 (1) de Y.

FUENTE: GLASS, G. V. y STANLEY, J. C., *Statistical Methods in Education and Psychology*, 1970, apéndice A, tabla H.

BIBLIOGRAFÍA

Alcaide Inchausti, A.: *Estadística aplicada a las Ciencias Sociales*, Ediciones Pirámide, Madrid, 1976.

Amón, J.: *Multidimensionalidad de la religiosidad utilitaria a través de la afinidad entre las cuestiones intrínsecas y las extrínsecas*, «Rev. de Psicol. Gen. y Aplic.», 1968, vol. XXIII, núm. 95, págs. 983-988.

Amón, J.: *Prejuicio antiprotestante y religiosidad utilitaria*, Aguilar, Madrid, 1969.

Amón, J.: *La construcción de escalas psicológicas en función del método elegido*, «Rev. de Psicol. Gen. y Aplic.», 1972, vol. XXVII, núms. 116-117, págs. 423-431.

Atkinson, R. C., Bower, G. H. y Crothers, E. J.: *An Introduction to Mathematical Learning Theory*, John Wiley, Nueva York, 1965.

Bailey, N. T. J.: *The Mathematical Approach to Biology and Medicine*, John Wiley, Nueva York, 1967.

Balow, B., Fulton, H. y Peploe, E.: *Reading Comprehension Skills Among Hearing Impaired Adolescents*, «Volta Review», 1971, 73, págs. 113-119.

Berman, P. W., Waisman, H. A. y Graham, F. K.: *Intelligence in Treated Phenylketonuric Children: A Developmental Study*, «Child Development», 1966, 37, páginas 731-747.

Bingham, W. Wd., Moore, B. V. y Gustad, J. W.: *How to Interview*, Harper & Row, Nueva York, 1959.

Bisset, B. M. y Rieber, M.: *The Effects of Age and Incentive Value on Discrimination Learning*, «J. Exp. Child Psych.», 1966, 3, págs. 199-206.

Blommers, P. y Lindquist, E. F.: *Elementary Statistical Methods in Psychology and Education*, University of London Press, Londres, 1965.

Burt, C.: *Is Intelligence Distributed Normally?*, «Brit. J. Statist. Psychol.», 1963, 16, págs. 175-90.

Calot, G.: *Cours de Statistique descriptive*, Dunod, París, 1969. (Traducida al castellano con el título de *Curso de estadística descriptiva*, editada por Paraninfo, 1970.)

Casa Aruta, E.: *200 problemas de Estadística descriptiva*, Vicens Vives, Barcelona, 1969.

Collman, R. D. y Stoller, A.: *A Survey of Mongoloid Births in Victoria, Australia*, 1942-1957, «Amer. J. Public Health», 1962, 52, págs. 813-829.

Conde, V. y Domenech, B.: *Considerations on the I. Q. With the WAYS in a Sample of Schizophrenics in Relation With the Sex, Age, Cultural Level, Origin, Residence and Civil State*, «Rev. de Neurol., Neuroch. y Psich.», Oviedo, 1976, 27, págs. 258-291.

Coombs, C. H., Dawes, R. M. y Tverski, A.: *Mathematical Psychology: An Elementary Introduction,* Englewood Cliffs, N. J.: Prentice-Hall, 1970.
Cravioto, J. y Robles, B.: *Evolution of Adaptive and Motor Behavior During Rehabilitation from Kwashiorkor,* «Amer. J. of Orthopsych.», 1965, 35, págs. 449-464.
Croxton, F. E., Cowden, D. J. y Klein, S.: *Applied General Statistics,* Pitman, Londres, 1968. (Traducida al castellano con el título *Estadística general aplicada,* editada por Fondo de Cultura Económica, México).
Chesire, L., Saffir, M. y Thurstone, L. L.: *Computing Diagrams for the Tetrachoric Coefficient,* The University of Chicago, Chicago, 1933.
Downie, N. M. y Heath, R. W.: *Basic Statistical Methods,* Harper & Row, Nueva York, 1970. (Traducida al castellano con el título *Métodos estadísticos aplicados,* editada por Ediciones del Castillo, Madrid, 1971.)
Edwards, A. L.: *Statistical Methods,* Holt, Rinehart & Winston, Nueva York, 1973.
Engelman, S.: *The Effectiveness of Direct Verbal Instruction on I. Q., Performance and Achievement in Reading and Arithmetic,* en G. Hullmuth (Ed.) «Disadvantaged Child», vol. 3. *Compensatory Education: A National Debate,* Brunner/Mazel, Publishers, 1970, págs. 339-361.
Felsinger, J. M., Gladstone, A. I., Yamaguchi, H. G. y Hull, C. L.: *Reaction Latency ($_{s}t_{r}$) as a Function of the Number of Reinforcements,* «J. Exper. Psychol.», 1947, 37, págs. 214-288.
Ferguson, G. A.: *Statistical Analysis in Psychology and Education,* McGraw-Hill, Nueva York, 1976.
Freeman, H.: *Introduction to Statistical Inference,* Addison-Wesley, Reading, Mass., 1963.
Freeman, L. C.: *Elementary Applied Statistics,* John Wiley, Nueva York, 1968.
Freund, J. E.: *Statistics, a First Course,* Englewood Cliffs, N. J.: Prentice-Hall, 1970.
Games, P. A. y Klare, G. R.: *Elementary Statistics: Data Analysis for the Behavioral Sciences,* McGraw-Hill, Nueva York, 1967.
Getzels y Jackson: *Creativity and Intelligence. Explorations with Gifted Students,* John Wiley, Nueva York, 1968.
Glass, G. V. y Stanley, J. C.: *Statistical Methods in Psychology and Education,* Englewood Cliffs, N. J.: Prentice-Hall, 1970. (Traducida al castellano con el título de *Métodos estadísticos aplicados a las Ciencias Sociales,* editada por Prentice-Hall, 1974.)
Goodman, L. A. y Kruskal, W. H.: *Measures of Association for Cross-Classifications,* «Journal J. Amer. Stat. Assoc.», 1954, 49, págs. 732-764.
Guilford, J. P. y Fruchter, B.: *Fundamental Statistics in Psychology and Education,* McGraw-Hill, Nueva York, 1973.
Haber, A. y Runyon, R. P.: *General Statistics,* Addison-Wesley, Reading, Mass., 1973. (Traducida al castellano, con el título de *Estadística General,* editada por Fondo Educativo Interamericano, 1973.)
Harrower, M.: *Psychodiagnostic Testing: An Empirical Approach Based in a Follow-up of 2.000 Cases,* Charles C. Thomas, Springfield, Ill., 1973.
Hays, W. L.: *Statistics for the Social Sciences,* Holt, Rinehart & Winston, Nueva York, 1973.
Horowitz, L. M.: *Elements of Statistics for Psychology and Education,* McGraw-Hill, Nueva York, 1974.
Horst, P.: *Psychological Measurement and Prediction,* Wadsworth, Belmont, Cal., 1966.

Hull, C. L., Felsinger, J. M., Gladstone, A. I. y Yamaguchi, H. G.: *A Proposed Quantification of Habit Strength,* «Psychol. Rev.», 1947, 54, págs. 237-254.
Jáñez, L.: *Efectos de la comunidad de asociados en el aprendizaje verbal,* tesis doctoral no publicada, Universidad Complutense de Madrid, 1976.
Kendall, M. G.: *Rank Correlation Methods,* Ch. Griffin, Londres, 1970.
Kitterle, F. L. y Helson, H.: *On the Inhibitory Effect of a Second Stimulus Following the Primary Stimulus to React: A Successfull Replication,* «J. Exp. Psychol.», 1972, 96, págs. 138-141.
Kohuot, F. J.: *Statistics for Social Scientists: A Coordinated Learning System,* John Wiley, Nueva York, 1974.
Korin, B. P.: *Statistical Concepts for the Social Sciences,* Cambridge Mass, Winthrop, 1975.
Kuo, Z. Y.: *The Genesis of the Cat's Response to the Rat,* «J. Comp. Psychol.», 1930, 11, págs. 1-30.
Labrousse, Ch.: *Statistique. Exercices corrigés,* tomo I, Dunod, París, 1970. (Traducida al castellano con el título *Estadística. Ejercicios resueltos,* editada por Paraninfo, 1973.)
Levenstein, P. y Sunley, R.: *Stimulation of Verbal Interaction Between Disadvantaged Mothers and Children,* «Amer. J. of Orthopsych.», 1968, 38, págs. 116-121.
Lewis, D.: *Quantitative Methods in Psychology,* McGraw-Hill, Nueva York, 1960.
Lord, F. M. y Novick, M. R.: *Statistical Theories of Mental Test Scores,* Addison-Wesley, Reading, Mass., 1968.
Lorge, I.: *The Fundamental Nature of Measurement,* en D. N. Jackson y S. Messick (Eds.), *Problems in Human Assessment,* McGraw-Hill, Nueva York, 1967, páginas 43-56.
Magnusson, D.: *Test Theory,* Addison-Wesley, Reading, Mass., 1967.
McCall, R. B.: *Fundamental Statistics for Psychology,* Harcourt, Brace Jovanovich, Nueva York, 1975.
McNemar, Q.: *Psychological Statistics,* John Wiley, Nueva York, 1969.
Miller, G. A.: *Mathematics and Psychology,* John Wiley, Nueva York, 1964.
Nunnally, J. C.: *Psychometric Theory,* McGraw-Hill, Nueva York, 1967.
Nunnally, J. C.: *Introduction to Statistics for Psychology and Education,* McGraw-Hill, Nueva York, 1975.
Osgood, C. E.: *Method and Theory in Experimental Psychology,* Oxford University Press, Londres, 1953.
Pavlov, I. P.: *Conditioned Reflexes (Selected Experiments),* Oxford University Press, Londres.
Pfanzagl, J.: *Theory of Measurement,* Physica-Verlag, Würzburg-Viena, 1971.
Rachman, S. y Hodson, R. J.: *Experimentally Induced «Sexual Fetichism»: Replication and Development,* «The Psychol. Record.», 1968, 18, págs. 25-27.
Runyon, R. P. y Haber, R.: *Fundamentals of Behavioral Statistics,* Addison-Wesley, Cambridge, Mass., 1976.
Savage, I. R.: *Nonparametric Statistics,* «J. Amer. Statist. Ass.», 1957, 52, págs. 331-344.
Sender, V. L.: *Measurement and Statistics,* Oxford University Press, 1958.
Siegel, S.: *Non-parametric Statistics,* McGraw-Hill, Nueva York, 1956.
Sokolov, A. N.: *Inner Speech and Thought,* Plenum Press, Nueva York, 1972.
Stevens, S. S.: *On the Theory of Scales of Measurement,* «Science», 1946, vol. 103, núm. 2.684, págs. 677-680.

Stevens, S. S.: *Mathematics, Measurement and Psychophysics,* en S. S. Stevens (Ed.), *Handbook of Experimental Psychology,* John Wiley, Nueva York, 1951, págs. 1-49.
Stevens, S. S.: *Measurement, Psychophysics and Utility,* en C. W. Churchman y Ph. Ratoosh (Eds.), *Measurement: Definitions and Theories,* John Wiley, Nueva York, 1959, págs. 18-63.
Stevens, S. S.: *A Metric for the Social Consensus,* «Science», 1966, 151, págs. 530-541.
Stevens, S. S.: *Ratio Scales of Opinion,* en D. K. Whitla (Ed.), *Handbook of Measurement and Assessment in Behavioral Sciences,* Addison-Wesley, Reading, Mass., 1968a, págs. 171-199.
Stevens, S. S.: *Measurement Statistics and the Schemapiric View,* «Science», 1968b, 161, págs. 843-856.
Stevens, S. S.: *Psychophysics,* John Wiley, Nueva York, 1975.
Suppes, P. y Zinnes, J. L.: *Basic Measurement Theory,* en R. D. Luce, R. R. Bush y E. Galanter (Eds.), *Handbook of Mathematical Psychology,* vol. I, John Wiley, Nueva York, 1963, págs. 1-76.
Thurstone, L. L.: *The Measurement of Values,* University of Chicago Press, Chicago, 1959.
Torgerson, W. S.: *Theory and Methods of Scaling,* John Wiley, Nueva York, 1958.
Wadsworth, H. G.: *A Motivational Approach Toward the Remediation of Learning of Disabled Boys. Exceptional Children,* 1971, 38, págs. 33-42.
Walker, H. M. y Lev, J.: *Elementary Statistical Methods,* Holt, Rinehart & Winston, Nueva York, 1969.
Wallis, W. A. y Roberts, H. V.: *Statistics: A New Approach,* The Free Press Glencoe, Ill., 1956.
Warren, R. E.: *Norms of Restricted Color Association,* «Bull. of the Psychon. Soc.», 1974, 4, págs. 37-38.
Woodworth, R. S.: *Heredity and Environment,* Nueva York, Social Science Research Council, 1941.
Yule, G. U. y Kendall, M. G.: *An Introduction to the Theory of Statistics,* Griffin, Londres, 1968.
Zelditch, M. Jr.: *A Basic Course in Sociological Statistics,* Holt, Rinehart & Winston, Nueva York, 1966.

ÍNDICES ALFABÉTICOS

ÍNDICE ALFABÉTICO DE AUTORES

ÍNDICE ALFABÉTICO DE MATERIAS