Principios de matemática moderna

ESTADISTICA Y PROBABILIDADES

A.J. SHERLOCK

segunda serie

Colección

«Principios de matemática moderna para la enseñanza»

A. J. Sherlock

ESTADISTICA Y PROBABILIDADES

Títulos que constituyen esta colección:

Primera serie

G. MATTHEWS: *Matrices*, 1.
C. A. R. BAILEY: *Conjuntos y lógica*, 1.
F. B. LOVIS: *Computadores*, 1.
J. A. C. REYNOLDS: *Forma, tamaño y lugar.*

Segunda serie

G. MATTHEWS: *Matrices*, 2.
C. A. R. BAILEY: *Conjuntos y lógica*, 2.
F. B. LOVIS: *Computadores*, 2.
A. J. SHERLOCK: *Estadística y probabilidades.*

La versión española de estas series ha sido dirigida por el Dr. R. RODRÍGUEZ VIDAL, catedrático de Matemáticas.

ESTADISTICA Y PROBABILIDADES

A. J. Sherlock

Traducción, prólogo y notas

R. Rodriguez Vidal

Catedrático de la
Universidad de Zaragoza
y de la
Escuela de Formación del
Profesorado de Enseñanza Media

editorial vicens—vives • barcelona

Título original
PROBABILITY AND STATISTICS

Primera edición, 1968
Primera reedición, 1974

Depósito Legal: B. 33.659-1968
ISBN: 84 - 316 - 1238 - X
N.º de Orden V. V.: A 685

IMPRESO EN ESPAÑA
PRINTED IN SPAIN

Editado por Editorial VICENS–VIVES. Avda. de Sarriá, 136, Barcelona-17
Impreso por Juvenil S. A. Maracaibo, 9. Barcelona. Octubre 1974

ISBN: 84 - 316 - 1238 - X

PRESENTACION DE LA VERSION ESPAÑOLA

Hemos tenido una verdadera satisfacción al aprovechar la posibilidad que nos ha brindado Editorial Vicens—Vives de facilitar el acceso del público hispanoparlante a esta colección de manuales, concebidos por muy distinguidos didactas ingleses de la Matemática actual.

La traducción de los diversos volúmenes ha sido confiada a personas de gran experiencia y prestigio docentes, a quienes agradezco su colaboración, y realizarla ha sido un trabajo agradable para todos nosotros por la novedad, ingenio y fantasía de las cuestiones propuestas a los alumnos.

En todo el mundo culto se vienen haciendo ensayos para conseguir las obras del tono que requiere la adaptación de la enseñanza tradicional al nuevo modo de estructurar la matemática. En España se han hecho para esto trabajos verdaderamente excelentes. Pero es necesario un conocimiento lo más general posible de cuanto se va intentando, y esta razón sería bastante para hacer conveniente la publicación de esta serie.

Pero otra razón es, obviamente, su propia excelencia. Me limito a señalar a la atención del profesor que decida ensayar este método, que ni en un párrafo se mantiene al alumno como sujeto pasivo de la clase; en cada momento tiene que estar «haciendo algo». Que con una enseñanza tan fuertemente activa se busque el acceso a una matemática cada vez más abstracta, puede parecer paradójico a muchos, pero no sorprenderá demasiado a quienes hayan meditado largamente sobre los fines y la metodología de la enseñanza necesaria al estudiante de hoy.

R. RODRÍGUEZ VIDAL

INTRODUCCION GENERAL A LA SERIE

Matemática moderna para la enseñanza elemental

«Multiplica 7 semanas, 5 horas, 23 minutos y medio por 37.»

«Simplifica la expresión $\dfrac{3x^2+5}{x^2-5x+6} - \dfrac{7x^2+4x+1}{x^2-9x+14}$.»

«Demuestra que los ángulos adyacentes a la base de un triángulo isósceles son iguales.»

Ocurre, lectores, que preguntas de este tipo ya no representan bien el significado de las matemáticas actuales. De hecho, gran parte de la «matemática» tradicionalmente enseñada en las escuelas está moribunda y necesita de reforma. Esto no quiere decir, sin embargo, que toda idea familiar deba ser indiscriminadamente barrida. Los libritos de esta serie sugieren posibles cambios en los libros elementales de matemáticas, pero al mismo tiempo muestran cómo lo ya existente puede ser vivificado con ideas contemporáneas. Son para completar textos usuales, mejor que para reemplazarlos.

Los niños a los que «no les gustan las matemáticas» pueden cambiar de idea cuando vean su trascendencia actual. Algún tiempo empleado en temas tales como computadores y estadísticas no será, pues, tiempo perdido, incluso tratándose de aprobar los exámenes oficiales. Asimismo, el lenguaje de grupos y de matrices ayudará a mostrar las matemáticas como lo que son: un tema coherente con muchas aplicaciones, y *no* una colección de técnicas aisladas.

Estos libros han sido escritos por miembros del Claustro del *St. Dunstan's College*. El primer grupo ha sido utilizado con alumnos de 11 a 13 años, y el segundo con clases de nivel algo superior. Los alumnos han progresado de una manera natural más allá de las prácticas de clase, hasta el punto de que se juzgó conveniente un nuevo plan de exámenes. Este plan ha sido aprobado por los comités de examen de las escuelas de Oxford y Cambridge y está empezando a entrar en servicio con temas especiales a nivel «O»* basados en estos textos.

Escuelas de toda clase están comenzando a experimentar esta reforma de los libros elementales de matemáticas. Los textos de esta serie se han elaborado para ayudar a esta tarea.

G. M.

* El nivel «O» corresponde a nuestro Bachillerato Elemental.

Indice

PREFACIO A:

Introducción al cálculo de probabilidades y estadística

Este librito está concebido para introducir en forma moderna las ideas básicas de estadística y probabilidades. Se ha insistido en los conceptos y el trabajo numérico ha sido reducido al mínimo; sería una buena ayuda en el último capítulo poder manejar una sencilla regla de cálculo. Los ejercicios propuestos implican solamente ideas simples y esperamos que todos ellos puedan contestarse: al final del libro hay una sección con las respuestas y sugerencias oportunas.

La característica esencial de la estadística es tomar contacto con una verdadera necesidad de las teorías experimentales, tanto en el campo sociológico como en el físico; esta necesidad es encontrar el modo de que pueda hacerse el mejor uso posible de los datos numéricos sobre un tema. Quizá prefieran algunos profesores empezar por el cálculo de probabilidades (cap. IV), antes de la estadística: el curso ha sido planteado de modo que ello sea posible.

En este texto no se emplea el lenguaje de los conjuntos para las probabilidades, pero para una indicación de cómo puede introducirse remitimos al lector a *Conjuntos y Lógica, 2,* de esta misma colección. Las matrices de probabilidad son asimismo consideradas en *Matrices, 2.*

A. J. S.

ADVERTENCIA DEL TRADUCTOR

De todos los textos de la colección, éste (que cierra la serie) es el que trata de temas más concretos y habituales en la vida corriente. Sin embargo, creemos que es el que va a ofrecer más dificultad a los jóvenes estudiantes. Esto (que podría darse como ejemplo para sostener la tesis de que la Matemática no es más difícil por ser más abstracta) se dice para advertir al profesor de que su colaboración se requiere ahora de modo esencial, para ayudar al alumno lector y para decidir qué partes pueden aplazarse hasta que el bagaje matemático de los alumnos sea adecuado.

Por supuesto que partes muy interesantes del libro pueden ser comprendidas, y realizadas, por niños de 14 años (y ello les será altamente formativo), pero no todo el libro.

De otra parte, la consideración de los temas socio-económicos se hace en la educación sajona mucho más precozmente que en la latina, donde los más sencillos temas de mercados y finanzas se reservan para la enseñanza casi superior. Tal vez esto dé un aire de anticipación a este libro entre los otros de la serie.

En algunos ejemplos, tomados de la prensa, se emplean las siguientes siglas:

E.U.A. = Estados Unidos de América.
O.E.C.E. = Organización Europea de Cooperación Económica.
E.F.T.A. = Organización Europea de Libre Comercio (= European Free Trade Association).

Con gusto manifestamos nuestra gratitud a la licenciada María Lucía Yagüe, por su valiosa colaboración en esta traducción.

R. R. V.

INTRODUCCION

Estadística es el nombre dado a la ciencia que se ocupa de la contabilidad y análisis de hechos en gran número. No es una materia reciente, pues su origen se pierde en la noche de los tiempos. Tiene carácter de gran generalidad, abarcando desde los censos de población a la física nuclear; desde las tendencias financieras e industriales a cuestiones de medicina y sanidad.

Las primeras estadísticas se referían, al parecer, a los problemas de censo (¿cuánta gente?), impuesto (¿cuánto dinero?) y guerra (¿cuántos soldados y cuánto material?).

Hoy en día hemos superado esta forma primaria de usar las estadísticas. Las estadísticas populares nos rodean por todas partes, y así se dice: 9 de cada 10 mujeres no advierten la diferencia entre margarina y mantequilla (incidentalmente diremos que es sorprendente lo distinto que es el impacto al decirse 9 de 10, 8 de 10 o 7 de 10). Nuestros periódicos nos exhiben cifras: cifras del paro obrero, índices del coste de vida, el promedio de lo que un hombre gasta en cerveza y tabaco, el promedio de salarios de ciertos sectores de la comunidad, y así sucesivamente. A menudo, estas cifras van acompañadas por alguna clase de diagrama, gráfico o esquema.

Una cosa afirma cada periódico: ¡sus cifras están por encima de toda discusión! Desgraciadamente, las cifras pueden ser de lo más engañoso, a menos que se relacionen correctamente con el problema que tenemos entre manos.

En este libro estudiaremos la forma en que pueden reunirse los datos (así llamamos a los hechos) para dar una imagen de la realidad tan exacta como sea posible, y cómo se pueden interpretar los resultados: justa o injustamente; y veremos los tipos más comunes de engaño.

Es aconsejable examinar toda gráfica o serie de resultados numéricos con ojos de detective. Puede haberse cometido un crimen, y muy sutilmente por cierto.

CAPITULO I

Reunión y representación de datos

Reunión de datos

La primera cuestión que hemos de plantearnos al ver cifras que, supuestamente, representan algún dato es: ¿cómo llegaron aquí? A primera vista puede parecer fácil reunir cifras exactas. Por desgracia esto no es así: en la mayor parte de los casos es muy difícil reunir datos ciertos. Naturalmente, esto depende en cierta medida del tipo de datos en que estemos interesados. Por ejemplo, si enviamos un cuestionario preguntando por los salarios de un cierto sector de la comunidad, podemos estar seguros de recibir pocas y cautelosas respuestas. Podremos tener más confianza en nuestras cifras si lo que hacemos es medir cosas: por ejemplo, contar el número de vehículos que atraviesan el puente de Toledo cada hora.

Abordemos, pues, en primer lugar, el problema del *cuestionario.* La idea es simple. La información requerida está escrita en él en forma de preguntas y respuestas. Entonces se distribuyen los cuestionarios a todas las personas en las que estamos interesados. Por ejemplo, consideremos un escrutinio sobre los ingresos de los antiguos alumnos marrovianos (Marrovia es una antigua escuela pública, imaginaria). En primer lugar tenemos que encontrar los nombres y direcciones de todos los antiguos alumnos. Parece fácil, pero de hecho encontraremos dificultades para ello. Con el antiguo alumno que ha triunfado será fácil establecer contacto, mientras lo más probable es que un gran número de los que no han tenido éxito hayan desaparecido sin dejar sus señas. Así cambia ligeramente el problema, pues sólo podemos mandar cuestionarios a aquellos cuyos nombres y direcciones podamos conseguir (admitamos que sean la mayor parte, pero ya no serán todos ellos). ¿Qué ocurre ahora cuando nuestro antiguo alumno marroviano recibe el cuestionario? Hay varios caminos abiertos para él:

1. Lo tira a la papelera; lo pierde o se olvida de echarlo al correo. En cualquier caso, el cuestionario no vuelve.
2. Responde con sinceridad.
3. Falsea su respuesta (por diversidad de motivos).

Pasado cierto tiempo, llegan las respuestas; posiblemente un 70 o un 80 % de los cuestionarios enviados, pero es mucho más probable que sean del orden del 20 al 30 %. También hay que tener en cuenta que estas respuestas serán de un tipo definido de personas: el antiguo alumno que ha triunfado, o por lo menos ha tenido cierto éxito, que quizá haya contestado imparcialmente, pero que también puede tener toda clase de intereses que darán a sus respuestas diversos grados de parcialidad.

Esta clase de resultados no es buena para nosotros en absoluto. Si queremos conocer los ingresos de estas personas nos veremos obligados a ser mucho más sutiles en nuestro acercamiento. Una gran dificultad en los cuestionarios es la interpretación de las preguntas por parte del público. Por ejemplo, si planteamos la cuestión «¿Cuántas veces se baña usted cada semana?», ¿se deben incluir los baños en el mar y las duchas? Muchas preguntas dejan un elemento de duda en la mente del que las recibe. Los formularios de impuestos son un buen ejemplo de este tipo de ambigüedad de pregunta. A causa de estos problemas, proyectar cuestionarios es un trabajo difícil y un problema complicado. Un problema, de hecho, para un estadístico de primera categoría.

Se ha comprobado que si investigamos los periódicos que lee la gente nos encontramos con un tipo similar de prejuicios. Probablemente, la mejor forma de recoger resultados exactos sería hacerse basurero y atender a los diversos tipos de artículos arrojados a la basura: los periódicos leídos por cada familia serían entonces evidentes. Si no podemos recoger todos los datos por nosotros mismos, el camino mejor y más exacto es investigar *una muestra.* Ésta debe ser cuidadosamente elegida de forma que no pueda falsearse en absoluto. Si cada elemento del conjunto con que tratamos tiene igual posibilidad de selección, se la llamará entonces una muestra *al azar,* o muestra *aleatoria.*

Estas muestras pueden elegirse de muchas formas. Podemos tomar una guía telefónica e investigar una persona de cada 50, pero esto nos hace parciales contra los que no poseen teléfono. Podemos parar una de cada 20 personas en una calle un día laborable, pero esto nos desvía de los que están trabajando y de los que van conduciendo un coche. Podemos parar a la gente en una estación del metro, pero tampoco esta muestra será imparcial. La gente que tiene prisa no se para, habría pocas madres con hijos pequeños (excepto en las primeras horas de la tarde), y también pocos conductores de coche.

En general, encontraremos que nuestras muestras se falsean sistemáticamente hacia la gente con más dinero, mejor educación, más información, mejor apariencia, conducta más convencional y costumbres más arregladas que el promedio de la población que se supone representan. Es fácil ver cómo ocurre esto. Supongamos que estás en la calle recogiendo datos de la gente. Ves a las personas que llegan hacia ti. El primero es evidente que tiene prisa y no puedes pararlo. El segundo ha salido a dar un paseo con su perro y el último es un sujeto con aspecto alarmante. Naturalmente, paras al hombre del perro, la persona «normal». Esta clase de lance puede ocurrir con frecuencia mientras recogemos datos. Tendemos a preguntar a la gente más amigable, a la que parece más decente, y no a los que vemos más sucios y desagradables. De esta forma es muy fácil obtener una muestra falseada.

Puedes ver ahora lo difícil que es encontrar una muestra auténtica. La forma más segura es recoger por uno mismo todos los datos, porque entonces tendrás al menos alguna idea de las dificultades con que te enfrentas.

Los resultados experimentales están también sujetos a otra deformación, que es la que los muchachos introducen para lograr lo que ellos piensan que es la respuesta «correcta». Pero en general son mucho más dignos de confianza.

RESUMIENDO: Antes de que podamos decidir sobre la exactitud de las cifras que se nos presentan debemos saber:

1. Cómo se recogieron los datos.
2. Los diversos tipos de falsedad que se han introducido.

Es *muy difícil* llegar a cifras exactas en cualquier clase de encuesta en que se presenten prejuicios emocionales, políticos o de cualquier otra clase (por ejemplo, cuestiones de religión, cantantes de moda o emigración a Hispanoamérica). Los datos recogidos en experimentos o medidas de cosas serán mucho más exactos que los que dependan de las opiniones diversas.

Métodos de recoger datos:

1. Observación personal.
2. Cuestionarios.
3. Investigación en una muestra aleatoria.

De estos métodos, el primero es el más conveniente.

Representación

Supongamos ahora que hemos recogido nuestros datos y ajustado los resultados para eliminar errores en la medida de lo posible. ¿Qué haremos a continuación? Muy poca gente puede mirar una masa de números y analizar su significado de una ojeada. Para hacer nuestros resultados fáciles de comprender y dar una impresión exacta de su importancia, podemos recurrir a dibujar gráficas o diagramas. Hay gran variedad de modos para hacer esto: unos correctos y otros engañosos o incorrectos.

En este capítulo consideraremos los resultados de investigaciones en las que sólo se haya hecho un tipo de medida u observación. Por ejemplo, la altura de los muchachos de edades comprendidas entre los 15 años y los 15 años y medio; lluvia recogida; número de parados; o la temperatura a cada hora de un día.

El primer tipo que vamos a considerar es la gráfica *lineal*.

Espero que estarás completamente familiarizado con este tipo de gráfica. Aquí hay algunos ejemplos; unos buenos y otros malos.

Cómo ha crecido el número de radios, televisores y coches en uso

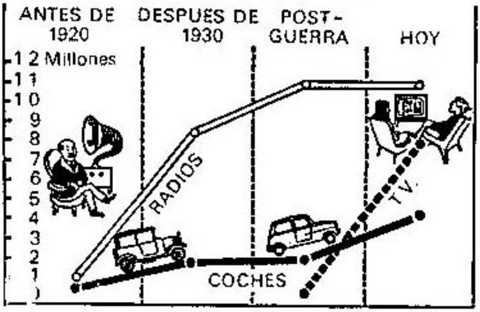


Fig. 1.1

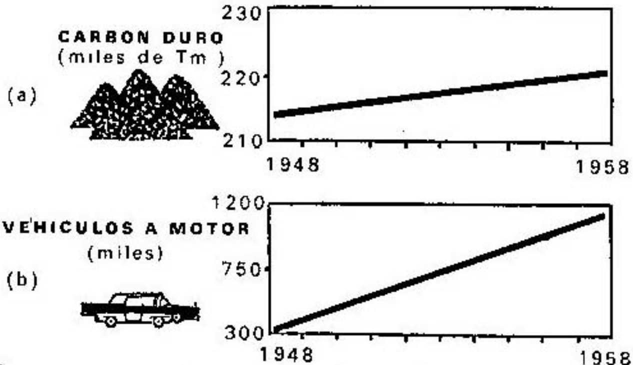


Fig. 1.2

Alguna producción crece en los siete países de la EFTA

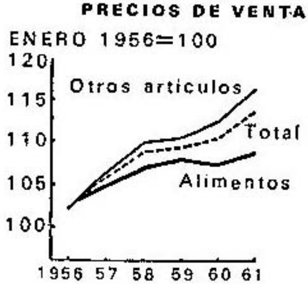


Fig. 1.3

Fig. 1.4

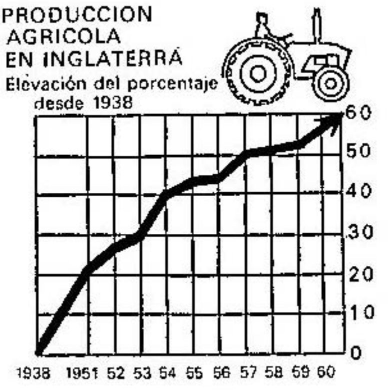


Fig. 1.5

Ésta es una selección típica de gráficas lineales publicadas en periódicos y folletos. Examinando más de cerca estas gráficas, sacaremos a la luz varios hechos interesantes. En primer lugar, tienen dos cosas en común:

1. (Un punto que *debe* ser recordado.) La *escala de tiempos* está sobre la horizontal o eje x, mientras la variable (aparatos de radio, carbón, precios, etc.) se mide sobre la vertical o eje y.
2. No se dan detalles acerca de cómo se recogieron las cifras.

Consideremos las gráficas una por una.

1. Éste es un tipo popular de gráfica que compara el crecimiento con los años del número de aparatos de radio, televisores y coches. ¿Qué ocurre con las escalas? La vertical es bastante fácil de comprender, pero ¿qué pasa con la horizontal? De qué datos están hablando? No tenemos ni idea. Cada gráfica consta de muy pocos puntos, que, aunque estén bien elegidos, pueden ser muy engañosos. Es una gráfica de bonita apariencia y es probable que sea razonablemente exacta a grandes rasgos, pero tiene insuficiente detalle y una escala pobre.

2. *a* y *b*. Primero, como siempre, miremos las escalas. Las horizontales son claras, pero las verticales empiezan en un caso en

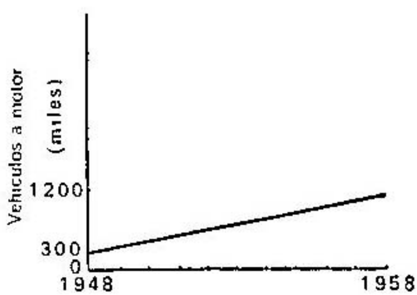
210 000 Tm y en el otro en 300 000 vehículos. Ésta es una forma común de lograr una gráfica exagerada: arreglando las escalas. Puede hacerse de tres maneras:

1. Arreglando la escala vertical.
2. Arreglando la escala horizontal.
3. Arreglando ambas escalas.

Consideremos, por ejemplo, 2 (*b*). Veamos primero cómo sería (con la misma escala).

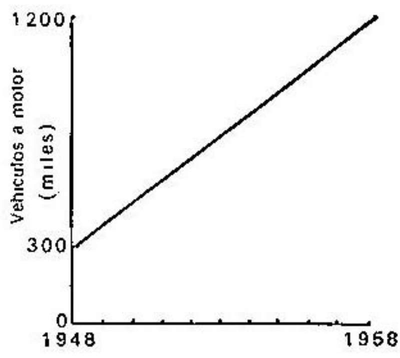
Fig. 1.6

No difiere apenas de la anterior. Ahora alteramos la escala vertical de forma que en la figura (*a*) hemos duplicado la escala y en la (*b*) la hemos dividido por dos.



¡ Una gráfica que anima mucho menos

Fig. 1.7



Bien, bien: ¡ desde luego la producción se ha disparado hacia arriba !...

Fig. 1.8

Efectos similares pueden conseguirse si alargamos o acortamos la escala horizontal.

Sin duda, la gráfica más impresionante se consigue si consideramos la figura (*b*) de antes y acortamos su escala horizontal. Si la dividimos por la mitad:

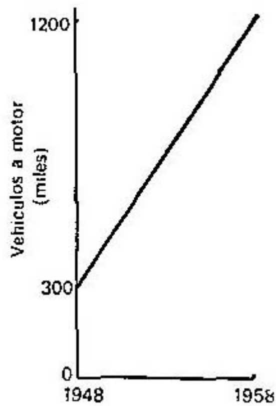


Fig. 1.9

¡conseguimos el resultado más espectacular de todos! Demasiado para 2.

¡Cuidado con las escalas! Desconfiemos también de la bella línea recta. Puede señalar una tendencia, pero nada más.

3. Esta gráfica es correcta y completamente buena, excepto que no se señala bien la indicación del punto inicial (enero 1956, 100).

4. Ésta es una gráfica horrible. ¡Mira *las escalas*! La vertical empieza por 5, habiéndose excluido en ella del 0 al 5 para exagerarla. La escala horizontal de tiempos es peculiar. Se modifica en su mitad. Un año después del período 58-59-60, la escala se pone en el primer trimestre de 1961.

Una gráfica de uso muy dudoso.

5. Una gráfica muy buena que, una vez más, muestra una tendencia. La inclinación inicial, sin embargo, da una impresión falsa

a causa de la escala de tiempos y desgraciadamente la punta de la flecha tiende a estropear esto cuando no hay certeza de que la gráfica continúe en esta dirección.

Con esto, ya te habrás dado cuenta de la gran importancia de las escalas para provocar la forma de tu gráfica.

La escala es de máxima importancia, y es lo primero a que hay que atender. Recientemente vimos en cierta televisión el anuncio de una bebida para la hora de ir a dormir, mostrando que el producto contenía menos calorías (un dudoso mérito para inducir al sueño) que otras bebidas similares. Estaba ilustrado gráficamente, pero carecía en absoluto de escala y por tanto *no decía nada.*

Antes de sacar ninguna conclusión de este tipo de gráfica no estaría de más investigar la posibilidad de variaciones estacionales o de variaciones periódicas típicas. Un ejemplo clásico de esto es la gráfica de las ventas de helados.

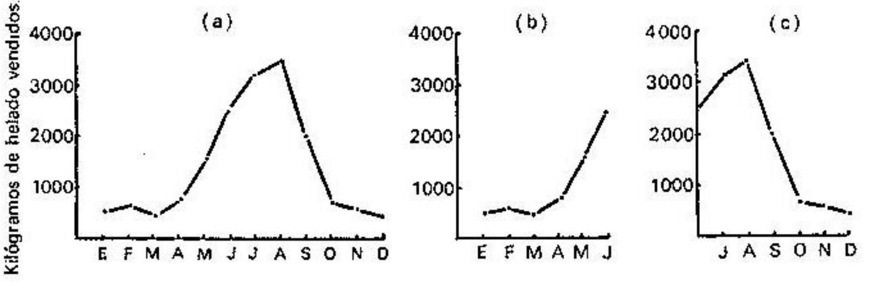


Fig. 1.10

Otro tema sujeto a variaciones estacionales es el desempleo. La enfermedad es otro más.

Otro abuso común de estas gráficas es emplearlas para repre-

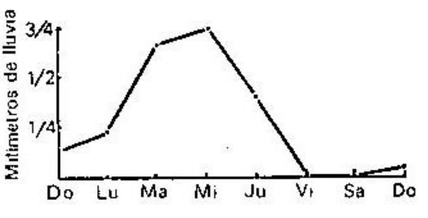


Fig. 1.11

sentar un tipo inapropiado de datos. Una muestra de ello sería dibujar una gráfica de línea para representar, por ejemplo, la lluvia caída durante una semana.

Esto significaría (por ejemplo) que el miércoles por la tarde parece haberse recogido medio milímetro de lluvia, mientras que de hecho posiblemente hizo una tarde soleada después de una tormenta. Este tipo de datos debería representarse por otra clase de gráfica: la gráfica puntual. Ésta es una gráfica discontinua, o sea que salta de punto a punto, pero no tiene significado en ningún punto entre los saltos. Un ejemplo sencillo de esto sería una gráfica expresando si llovió o no en cada hora de un cierto día:

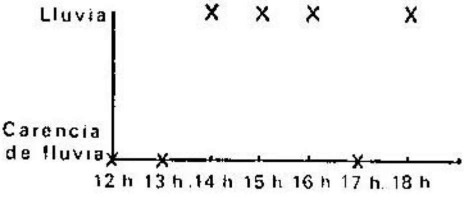


Fig. 1.12

Esta gráfica consta simplemente de un conjunto de puntos: no es una recta ni una curva. El hecho de que lloviera a una hora del día no guarda relación con si llovió o no a la siguiente. Los hechos no están relacionados.

En conclusión, pues, hay tres cosas principales que debemos observar al contemplar una gráfica:

1. ¿Es el tipo apropiado de gráfica para representar los datos?
2. ¿Qué ocurre con las escalas?
3. ¿Hay variaciones estacionales o periódicas que influencien el dibujo?

EJERCICIO 1

1. Explica por qué es fácil introducir errores con los siguientes métodos de elegir una muestra al azar:

(*a*) preguntando a 1 de cada 50 personas, siguiendo la lista de votantes;

(*b*) llamando a 1 de cada 5 casas en una calle durante las horas centrales del día;

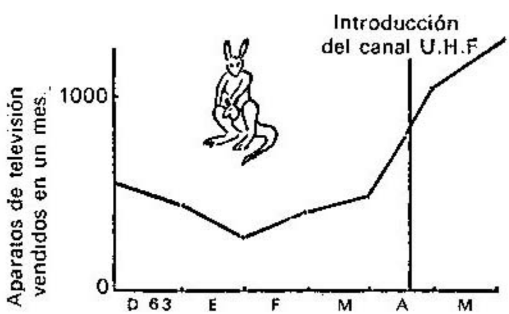
(*c*) preguntando a los espectadores de un partido de fútbol.

2. ¿Cómo escogerías tu muestra al azar para sondear las preferencias del público por:

(*a*) un cantante de moda;
(*b*) una marca de sopas en polvo;
(*c*) una marca de cerveza;
(*d*) una marca de cigarrillos?

¿Es aconsejable elegir una muestra? Razonarlo.

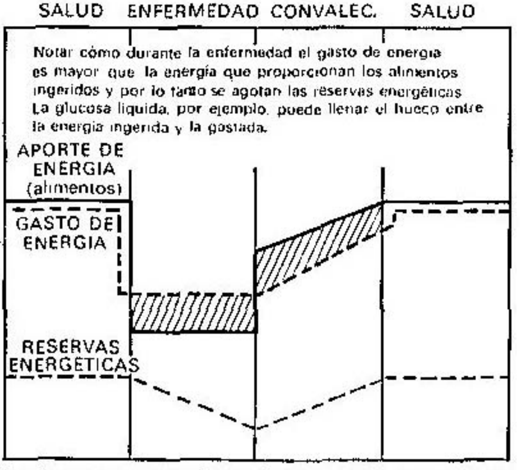
3. Comenta esta gráfica imaginaria



El canal U.H.F. y el incremento de la venta de televisores

Fig. 1.13

4. Comenta esta gráfica



Este diagrama representa el curso relativo de aporte, gasto y reservas de energía en una típica enfermedad febril (Lucozade).

Fig. 1.14

5. ¿Qué es lo que está mal en esta gráfica?

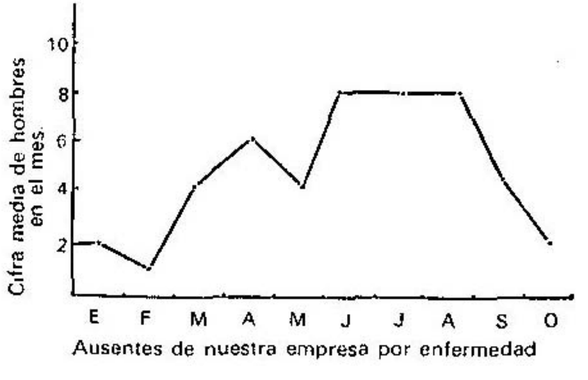


Fig. 1.15

CAPITULO II

Otras representaciones

Histogramas

El segundo tipo de gráfica es el *histograma*, un diagrama de bloques en el cual las cantidades se representan por áreas. Se comprenderá esto mejor examinando algunos ejemplos.

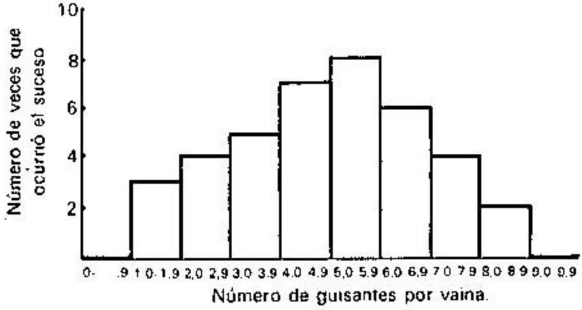


Fig. 2.1

Esta gráfica expone el número de veces que ocurre un suceso (su *frecuencia*) de cierto tipo. Aquí, un hombre estuvo cogiendo vainas de guisantes en su jardín y luego contó el número de guisantes de cada vaina.

Sus resultados fueron:

Núm. de guisantes por vaina:	0	1	2	3	4	5	6	7	8	9
Núm. de veces que esto ocurrió:	0	3	4	5	7	8	6	4	2	0

Este histograma parece ser una figura bastante regular: de hecho, si nos fuera posible continuar desgranando guisantes durante mucho tiempo, encontraríamos resultados ciertamente simétricos.

Naturalmente, *no* debe esperarse que la gráfica sea siempre simétrica; antes bien, usualmente no lo será. Un buen ejemplo de esto es un histograma de los resultados de fútbol en el campeonato de liga: aquí tenemos un histograma de los resultados de todos los partidos jugados en un domingo determinado.

Núm. de goles marcados por un equipo	0	1	2	3	4	5	6	7
Núm. de equipos que marcaron ese número de goles, es decir, número de veces que el suceso ocurrió . . .	14	28	22	9	5	3	0	1

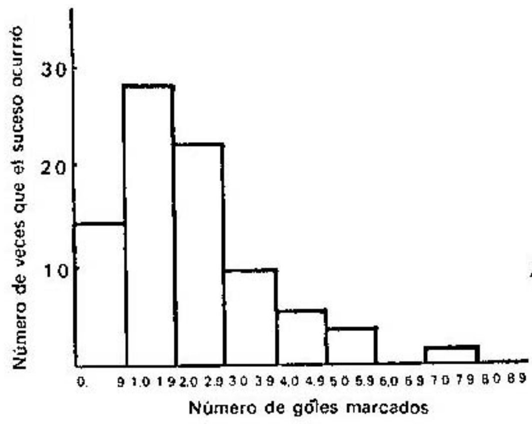


Fig. 2.2

Como podemos ver, la figura no es simétrica. Un histograma (o cualquier otra representación estadística para esta materia) que no sea *simétrico* se llama *asimétrico* u *oblicuo.* Prueba por ti mismo con los resultados de los partidos de liga que se hayan jugado esta semana. O bien, si es la temporada de toros, dibuja un histograma de los resultados de las corridas del último domingo que reseñe tu periódico. Indica el número de corridas en que se corten 0, 1, ..., o 12 orejas; o se registren de 0 a 12 broncas. ¿Observas alguna diferencia entre estas dos gráficas?

Encontramos con frecuencia histogramas en nuestros libros de geografía, mostrando, por ejemplo, importaciones o exportaciones.

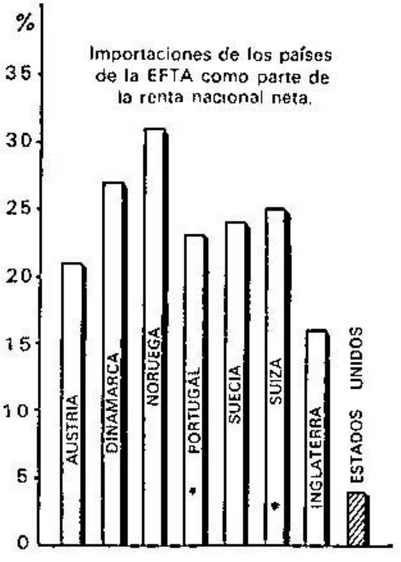


Fig. 2.3

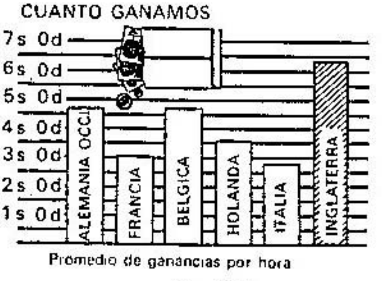


Fig. 2.4

Ambas gráficas son muy buenas en el sentido de que se comprenden fácilmente y tienen escalas precisas, pero la primera presenta síntomas de un engaño muy sutil, frecuente en Estadística. ¿Es el *área* o es el *volumen* del bloque lo que debe representar el porcentaje? (¡No nos dan las cantidades, por lo cual debemos adivinarlo!) Nos encontramos de nuevo con esta observación al tratar de otro tipo de gráfica, el ideograma.

Las barras verticales en un histograma se dibujan con espacios entre ellas, para mostrar que los datos no están relacionados; por ejemplo, los salarios de varios países son figuras aisladas que sólo se muestran sobre un histograma para comparar, mientras que los salarios de la población por edades dentro de cada país están relacionados y pueden mostrarse sobre un histograma sin espacios.

He aquí una pregunta que quizá te hayas hecho. ¿Se altera la forma de un histograma si cambiamos las escalas, como ocurría en la gráfica lineal? Naturalmente, la respuesta es que sí se altera, pero no exactamente de la misma forma que antes. Por ejemplo, volvamos a la primera gráfica de este capítulo, pero esta vez tomaremos como escala horizontal 0 a 1,9; 2 a 3,9; etc., en lugar de 0 a 0,9; 1 a 1,9; etc.

Núm. de guisantes por vaina	*Núm. de veces que esto ocurrió*
de 0 a 1,9	3
de 2 a 3,9	9
de 4 a 5,9	15
de 6 a 7,9	10
de 8 a 8,9	2

Da una gráfica de la siguiente forma

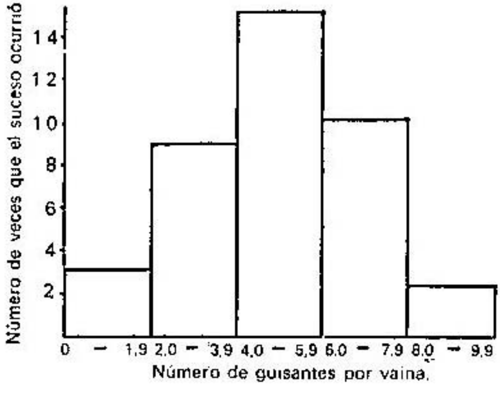


Fig. 2.5

Este histograma da la impresión de ser mucho más empinado y puede hacerse que lo parezca todavía más acortando la escala horizontal. De la misma forma, tomando más intervalos a lo largo de este eje, haríamos un histograma que parecería mucho menos empinado. En este caso sería imposible tener intervalos menores que un guisante, pero frecuentemente es posible hacer los intervalos tan pequeños como se desee; por ejemplo, cuando comparamos las alturas de los muchachos de 15 años de una población.

Repetimos, pues: *cuidado con la escala.* Una elección inteligente de ésta conduce a gráficas seductoras a la vista, sean o no sorprendentes los hechos por sí mismos. Por otra parte, la elección de escala puede muy fácilmente ocultar los hechos al lector poco avisado.

Como verás más tarde, los histogramas tienen usualmente una forma definida. Si tomamos intervalos muy pequeños sobre la escala horizontal y tenemos suficientes datos, obtendremos aproximadamente una curva. Este tipo de figura se llama curva de *distribución de frecuencias.* Habitualmente tiene forma de campana (ver capítulo V).

El diagrama de frecuencia acumulada

Nuestro siguiente método para representar datos es el diagrama de *frecuencia acumulada,* llamado así por razones que van a aparecer como evidentes. Volvamos de nuevo a nuestro primer grupo de cifras sobre los guisantes. El diagrama de frecuencia acumulada muestra simplemente cuántas veces ocurrió un suceso de magnitud *mayor que* o *menor que* la de un suceso dado; por ejemplo, cuántos equipos marcaron menos de tres goles, etc. Volvamos a nuestro primer ejemplo, cuyo resultado puede presentarse así:

Núm. de guisantes por vaina	*Núm. de veces que esto ocurrió*	*Núm. de veces en que hubo MÁS DE*	*Núm. de veces en que hubo MENOS DE*
0	0	39	0
1	3	36	0
2	4	32	3
3	5	27	7
4	7	20	12
5	8	12	19
6	6	6	27
7	4	2	33
8	2	0	37
9	0	0	39
	—		
	39 número total de vainas desgranadas		

Por otra parte, el tipo consiguiente de gráfica nos dice muy poco más que el histograma anterior sobre frecuencia de los datos, y es rara vez utilizado en publicidad. No lo estudiaremos, pues, limitándonos a señalar que su forma depende de la del histograma de frecuencias que representa y, naturalmente, de la escala elegida.

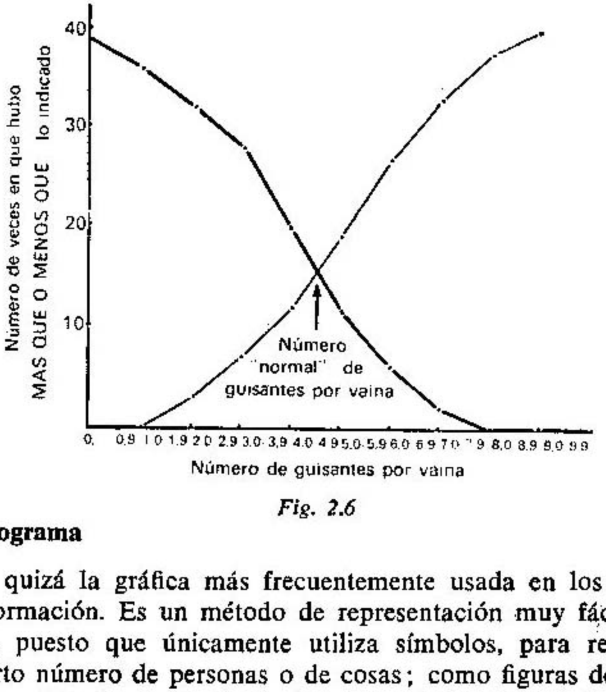


Fig. 2.6

El ideograma

Es quizá la gráfica más frecuentemente usada en los servicios de información. Es un método de representación muy fácil de entender, puesto que únicamente utiliza símbolos, para representar un cierto número de personas o de cosas; como figuras de hombre en una gráfica de empleo (por ejemplo, cada hombre símbolo puede representar 1000 hombres); o esquemas de casas en una gráfica de construcción de viviendas. De hecho, no hay límite para la imaginación en los diferentes tipos de símbolos que pueden utilizarse. La única dificultad surge cuando necesitamos usar fracciones de la unidad. No es fácil indicar la diferencia entre medio hombre símbolo y un tercio de hombre símbolo en una gráfica de empleo, lo cual, según la escala, puede representar una importante diferencia en números efectivos. Aparte de este problema, los ideogramas pueden ser razonablemente exactos y ofrecen interesantes y claras ilustraciones de los datos que representan.

Consideremos ahora unos cuantos ejemplos de ideogramas.

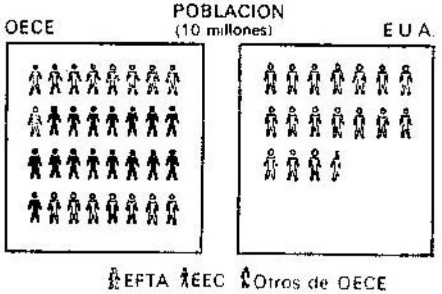


Fig. 2.7

Éste es el tipo más honrado de ideograma. Cada hombre símbolo representa 10 millones de hombres reales. Puedes ver cómo el último símbolo de Estados Unidos es difícil de interpretar. Puede representar 6, 7 u 8 millones de habitantes, pero sin las cifras es imposible indicarlo. Este ideograma es muy engañoso, porque hay 8 símbolos OECE por fila y sólo 7 de E.U.A.

Fig. 2.8

En nuestro segundo ideograma, relativo a los accidentes de «moto» por choque, los símbolos se modifican de acuerdo con el suceso que representan.

Podemos incluir también entre los ideogramas representaciones como las que siguen, en las que los datos están representados por el tamaño variable del símbolo. Es a menudo difícil saber si los datos están representados por la *altura*, el *área* o el *volumen*. En este ideograma es el volumen el que representa los datos, pero también podría creerse que es el área del saco lo que representa el dinero.

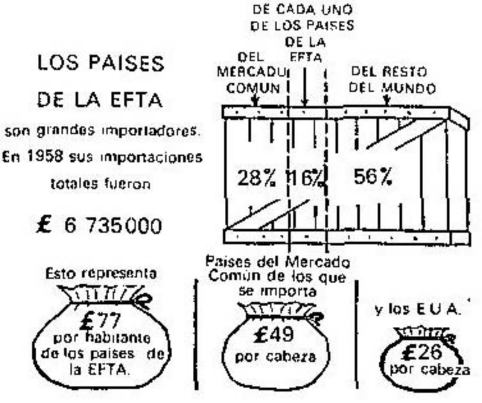


Fig. 2.9

El presentar mal una gráfica de área como una gráfica de volumen puede hacerse como engaño deliberado. La impresión que da es muy diferente, como puedes ver en los siguientes ejemplos.

El grupo de diagramas (*a*) representa los datos correctamente. El (*b*) está dibujado al mismo tamaño, pero se ha convertido en una gráfica de volumen añadiendo curvas a cada uno de los rectángulos. La impresión producida es muy distinta: B_1 parece mayor que A_1, B_2 mayor que A_2, etc., y lo que es más importante: la relación de $A_1:A_2$ parece diferente que la $B_1:B_2$ *.

* En este ejemplo, el error de interpretación es sólo psicológico, ya que los cilindros tienen igual base y por ende sus volúmenes son proporcionales a sus alturas, luego $A_1:A_2 = B_1:B_2$, como debía ser. Pero el error es verdadera falsedad en un caso como el siguiente. Las relaciones representadas por los segmentos (o los rectángulos) a_1, a_2, a_3 son distintas de las representadas por los círculos de radios a_1, a_2, a_3, y todavía más si los círculos se convierten en esferas con sólo añadirles un «ecuador». *(N. del T.)*

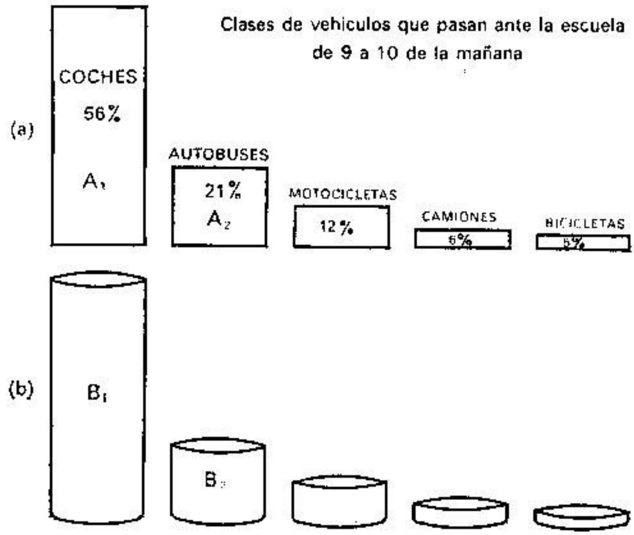


Fig. 2.10

Gráfica de sectores

Es un método muy común de representar datos. La idea es representar el total de los datos como un círculo (o pastel) y sus componentes como sectores del círculo (o porción del pastel). El ángulo central (360°) se divide proporcionalmente a los datos para determinar el tamaño de cada porción.

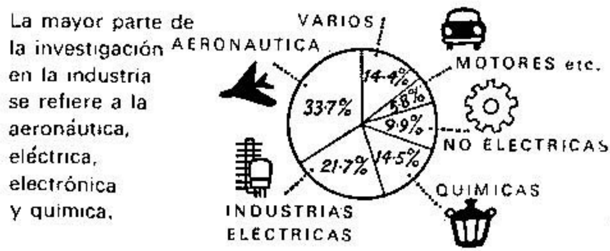


Fig. 2.11

Es ésta una buena forma de representar los datos: especialmente cuando se incluyen los porcentajes. La gráfica de sectores hecha de esta manera no es verosímil que sea falseada, ya que los datos se incluyen en ella.

Sin embargo, una cosa no queda clara aquí, y es lo que representan exactamente los datos. ¿Son los porcentajes de todo el dinero gastado en investigación lo que se representa? La investigación es fácil que sea más cara en unas industrias que en otras. Mientras una firma puede gastar millones de libras en un avión para fines de investigación empleando 50 hombres en el proyecto, otra firma en la industria de coches, donde los materiales son más baratos, puede emplear 5000 hombres por la misma cantidad de dinero. ¿Quién habrá investigado más? Este punto no se aclara en la información que acompaña a la gráfica, aunque por el tono del informe parecería que la base de los datos fue el dinero gastado en investigación, lo que realmente no da una buena medida del total de lo investigado. Cuando veas un gráfico de esta clase, siempre es prudente, si puedes hacerlo, descubrir exactamente lo que representan los datos. Puede ocurrir que no tengan mucha relación con el título que lleva la gráfica.

La gráfica de barras horizontales

Puede tener al menos dos formas distintas. El primer tipo es un histograma girado lateralmente y llena exactamente la misma fun-

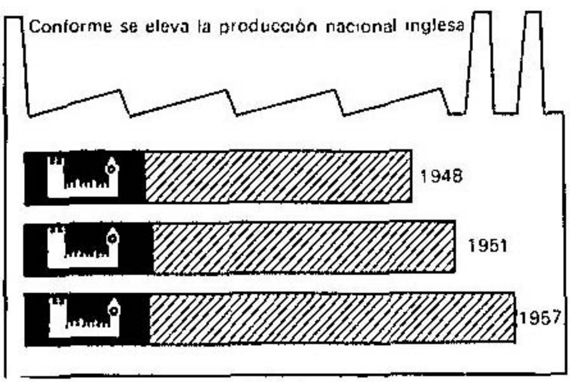


Fig. 2.12

ción que aquél. Está sujeto también a los mismos cambios de escala y a los mismos abusos.

Desgraciadamente, aquí no se muestran escalas ni se dan los datos. La impresión es clara, pero sin mayor información el gráfico es de muy dudoso valor.

El segundo tipo consta de barras horizontales de la misma longitud: representando en cada caso el dato total dividido proporcionalmente para mostrar sus diversos componentes. Esta gráfica indica sólo las proporciones implicadas. Debería aclararse que las barras horizontales representan en cada caso diferentes valores totales de importaciones. A primera vista parece que Gran Bretaña y Suiza importan aproximadamente un mismo total del resto de los países de la EFTA. En realidad esto no es cierto. Ambos países importan aproximadamente el mismo *porcentaje* de los otros de la EFTA. Para Suiza esto sería solamente unos 171 000 dólares, mientras que para Inglaterra sería unos 1 058 000 dólares.

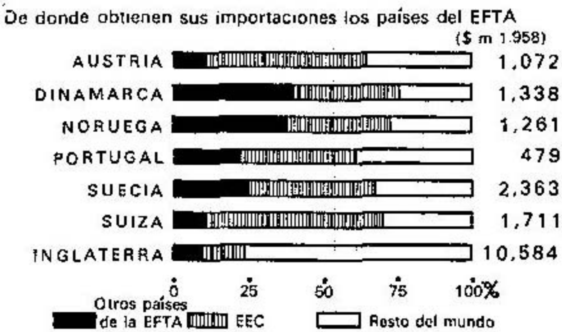


Fig. 2.13

Gráfica de barra compuesta (De la que la anterior es en realidad un ejemplo)

Este tipo de gráfica representa el dato total por una barra vertical u horizontal (o una pila de dinero) que después se divide en las proporciones adecuadas. Este método es el que habitualmente elige el Gobierno inglés en su gráfica del presupuesto. Muestra cómo se recaudan los ingresos y cómo se gastan. Los datos se reúnen con mucho cuidado y se representan con exactitud: al fin y al cabo el Ministro de Hacienda debe poder dar detalles exactos de cómo se ha llegado a cada una de las cifras.

Notaremos que cada libra está correctamente constituida por 240 peniques, y éste es un buen ejemplo de una correcta representación de volumen. El Gobierno también publica los detalles completos de los datos empleados, aunque sería imposible comprobarlos.

La última figura (fig. 2.15, pág. 40) es otra forma de ideograma. No se dan escalas ni es posible decir con ningún viso de certidumbre la proporción de préstamos-concesiones en ningún caso particular. Tampoco está claro si el saco de dinero es una representación de área o de volumen. Probablemente es de volumen, pero a primera vista esto no es obvio, y así, ¿qué representa la rebanada que le falta? Da la impresión de que el saco de dinero (volumen) ha sido tratado como una gráfica de sectores (área). Todo lo que podemos decir sobre esta gráfica es que muestra a qué lugares del mundo va la ayuda del Gobierno inglés, y qué cantidad va a cada lugar. Pero como representación estadística es pobre.

Hemos visto ya un cierto número de formas en las cuales se pueden representar los datos por medio de diagramas, gráficas o dibujos. Hemos visto también cuán engañosa y descuidadamente pueden ser dibujados. Recuerda esto cuando en lo sucesivo leas los periódicos y, lo que es más importante todavía, cuando dibujes tus propias gráficas.

Cuando encuentres una gráfica en un periódico, saca tu equipo de detective y examínala cuidadosamente:

1. ¿Puedes descubrir cómo se reunieron los datos y si se eligió una buena muestra? La mayor parte de los periódicos no menciona nunca este detalle tan importante.
2. ¿Es el tipo apropiado de gráfica para representar los datos?
3. ¿Qué dato está siendo representado ACTUALMENTE? Comprueba que el título no es engañoso.
4. ¿Qué ocurre con la ESCALA? ¿Cómo alteraría la figura un cambio en la escala?
5. ¿Hay allí variaciones estacionales u otras típicas que deberían haberse tenido en cuenta? Si es así, prueba que las hay.

6. ¿Los datos están representados por el volumen o por el área?

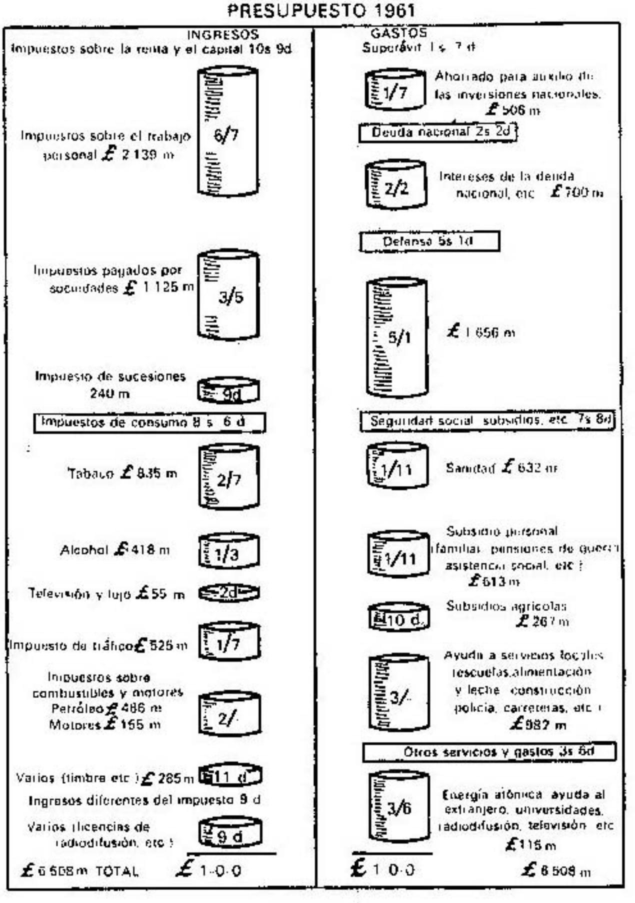
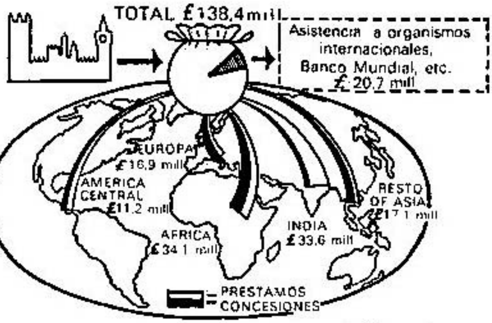


Fig. 2.14

7. ¿Es probable que haya prejuicios políticos o de alguna otra clase tras la presentación de los datos?

Éstos son la mayor parte de los defectos que puede tener una gráfica. Sin embargo, el anunciante mañoso todavía tiene un arma más en su manga, y quizá la más poderosa: las cifras mismas.



Fondos de desarrollo y beneficencia colonial £3.9 mill
Además se entregó un millón de libras a los territorios británicos del Pacífico Occidental.

Fig. 2.15

Un ejemplo muy bueno de esto ocurrió el 27 de febrero de 1962, cuando el canciller inglés Mr. Selwyn Lloyd anunció en los Comunes que los gastos del Gobierno excederían del nivel previsto. Dos periódicos publicaron esto, ambos correctamente:

(*a*) Forma sensacionalista
«Nos pasamos del presupuesto en 111 millones de libras.» *(Evening Standard.)*

(*b*) Forma convencional
«El Canciller se sale un 2 % del presupuesto.» *(The Guardian.)*

Ambas cifras son correctas, pero el tono de los titulares es completamente distinto.

Otro ejemplo es el uso típico de cifras en los anuncios, «9 de cada 10 mujeres no advierten la diferencia entre la margarina y la mantequilla».

Esto implica que 9 de *cada* 10 mujeres no notan la diferencia. De hecho las cifras pueden haberse recogido dejando a las mujeres probar el producto hasta encontrar 9 mujeres que no lo diferenciaron y una que lo hizo. Aunque entonces sería verdad que 9 de estas 10 «particulares» mujeres no habían notado la diferencia, se trataría de una muestra tan falseada que sería completamente inválida. Esta clase de anuncios aparece con frecuencia para los más variados productos: de modo que cuentan los testimonios afirmando lo buenos que son, mientras todas las quejas son indudablemente olvidadas.

Un tercer ejemplo lo constituye el uso de porcentajes. A veces no queda claro de qué es porcentaje cada figura.

Esto ocurre con frecuencia cuando hablamos de índices de precios, que son una clase de porcentajes que se usa comúnmente. El precio de un artículo en un momento determinado se toma como 100 y en fechas posteriores se compara el precio con éste y se establece, por ejemplo, en 110, 112 o 118. La desventaja de estas cifras es que la base del índice se altera frecuentemente, o se adapta el precio original para dar resultados completamente distintos según la técnica usada. (Ver cap. III).

EJERCICIO 2

1. Cierto lunes en una escuela imaginaria se dieron las siguientes cifras para justificar los casos de alumnos ausentes o retrasados.

Razón	*% de justificaciones*
Gripe	25,1
Vacunación	21,5
Catarro, tos, etc.	18,9
Retraso causado por huelga de ferrocarril	14,7
Retraso causado por caos del tráfico	11,0
Heridas serias (brazos rotos, etc.)	4,2
Visitas	3,1
Indispuestos (debido al examen de francés)	1,5

Representa los datos anteriores con tantas clases de gráficas como puedas. (Puedes encontrar por lo menos cinco.)

2. Criticar la siguiente gráfica:

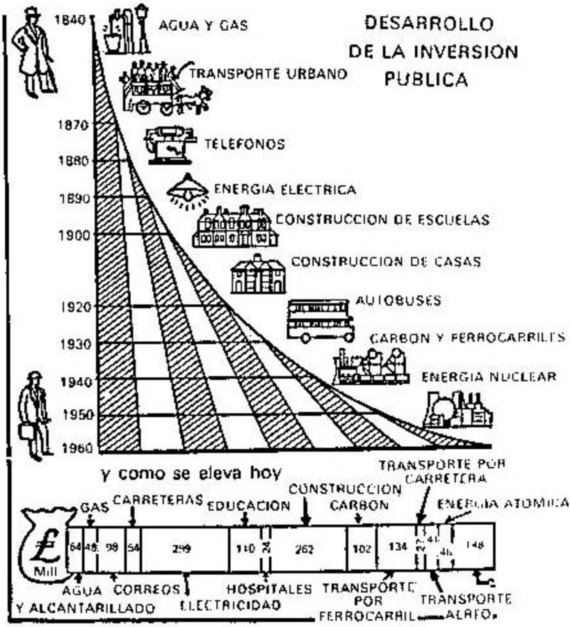


Fig. 2.16

3. Recientemente, después de que un boxeador resultó mortalmente herido durante un combate en América, se levantó un clamor contra el boxeo. En defensa del boxeo se indicó que moría mucha más gente en el fútbol y ciclismo que en el boxeo, deduciendo que éste era más seguro. ¿Puedes ver algo falso en este argumento? Prepara una carta a un periódico explicando este engaño.

4. Reúne tantas gráficas como te sea posible de los periódicos de esta semana. Averigua todo lo que puedas sobre las mismas y escribe breves críticas de cuatro de ellas.

Proyecto de formulario

Prepara un cuestionario relativo al dinero que lleven en su bolsillo tus compañeros de colegio de todos los cursos. Los resultados serán analizados y presentados en forma gráfica.

CAPITULO III

Promedios e índices de precios de venta

Una de las palabras que con más frecuencia se utilizan mal en temas de estadística es «medio». La encontramos en la conversación y en la prensa por todas partes: la familia media, los ingresos medios, el bebedor medio, el fumador medio. ¿Por qué puede engañarnos esta palabra? Al fin y al cabo, parece, todos sabemos lo que «medio» significa.

La dificultad estriba en que la estadística maneja una gran variedad de valores medios distintos y en publicidad se emplea el que más conviene en cada caso. En realidad, cualquier valor comprendido entre el máximo y el mínimo de una serie puede ser un valor medio de algún tipo, por lo que debemos dar un sentido más preciso a esa palabra.

Cuando decimos que el promedio semanal de ingresos de un cierto grupo es 1200 pesetas por persona, ¿qué queremos decir? ¿Queremos decir que una persona normal de ese grupo está ganando 1200 pesetas cada semana? Esto es lo que indican las cifras, pero lo que puede ocurrir es que nuestro grupo conste de cinco personas cuyos ingresos semanales son 2700 ptas., 1500 ptas., 700 ptas., 700 ptas. y 400 ptas. Entonces el resultado medio sería, efectivamente,

$$\frac{2700 + 1500 + 700 + 700 + 400}{5} = 1200 \text{ ptas./semana}$$

y sin embargo *ninguna* persona del grupo gana actualmente 1200 pesetas semanales, mientras que el sueldo más frecuente es de 700 pesetas por semana.

Consideremos otro ejemplo: el de una empresa con varios empleados y su sueldo anual distribuidos como sigue:

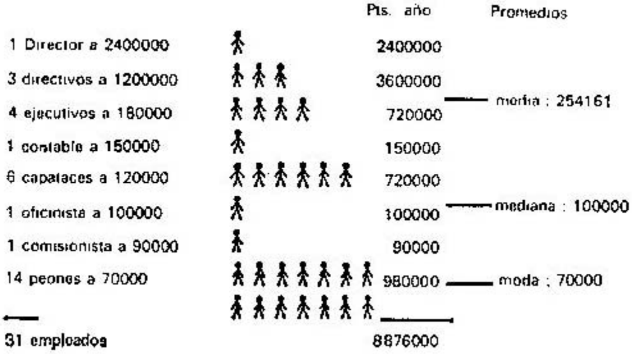


Fig. 3.1

Esto muestra lo distintos que pueden ser algunos de los «promedios». (Media, mediana y moda se definen en el párrafo siguiente.) En el mundo de los negocios, una gran parte de engaños se deben al uso de estos diversos promedios en la distribución de beneficios (con lo que se puede aparentar que los sueldos promedios son más altos de lo que son).

Definiciones

La *media aritmética* (que también se llama, simplemente, *media* o *valor medio*), se halla sumando todos los valores de la serie y dividiendo por el número de sumandos.

Si llamo $x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$ a los valores, la media aritmética viene dada por

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_i + \ldots + x_n}{n} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Σ (sigma) es la letra griega *S*, que representa «la suma de todas las cosas como ...», en este caso «como x_i».

La *mediana* es el valor del centro, es decir, que tiene tantos por debajo como por encima de él; en el ejemplo anterior hay 15 sueldos superiores y 15 inferiores. Si n es par, se toma por mediana la media aritmética de los valores que corresponden a los términos $n/2$ y $n/2 + 1$.

La *moda* es el valor que se repite más veces: en este caso 70 000 ptas./año.

Todavía hay otros promedios de interés:

La *media armónica* es un promedio que debe usarse en muchas cuestiones sobre velocidad.

$$\text{Media armónica} = \frac{n}{\sum_{i=1}^{i=n} \left(\frac{1}{x_i}\right)}$$

Por ejemplo, supongamos que conducimos un coche durante 1 km a la velocidad de 20 km/hora, 1 km a 30 km/h y 1 km a 40 km/h. ¿Cuál es la velocidad media en los tres kilómetros recorridos?

$$\text{Media armónica} = \frac{3}{1/20 + 1/30 + 1/40} \text{ km/h} = \frac{360}{13} \simeq 28 \text{ km/h}$$

También pudimos calcular esa velocidad poniendo

$$\text{Velocidad media} = \frac{\text{distancia total recorrida}}{\text{tiempo total empleado}}$$

Tiempo empleado en recorrer el primer km $= 60/20 = 3$ min

Tiempo empleado en recorrer el segundo km $= 60/30 = 2$ min

Tiempo empleado en recorrer el tercer km $= 60/40 = 1\frac{1}{2}$ min

Tiempo total empleado $= 6\frac{1}{2}$ min $= 13/120$ horas.

$$\text{Velocidad media} = \frac{3 \times 120}{13} = \frac{360}{13} \simeq 28 \text{ km/h}$$

La *media geométrica* es el promedio que debe usarse cuando tratemos, por ejemplo, de la población, cuyo incremento depende del número de individuos (si tenemos un índice estable de natalidad y mortalidad).

$$\text{Media geométrica} = \sqrt{x_1 \times x_2 \times \ldots \times x_i \times x_n}$$

Ejemplo 1

Si la población de una ciudad era de 90 000 habitantes en 1950 y de 160 000 en 1960, ¿cuál era la población aproximada en 1955?

$$\text{Media geométrica} = \sqrt{90\,000 \times 160\,000} = 120\,000$$

Ejemplo 2

Si la población actual es de 60 000 habitantes y hace cinco años era de 40 000, ¿cuál es la población probable para dentro de cinco años? (Problema muy importante para hacer planes de escuelas, puestos en las universidades, etc.) Supongamos que la población futura es P. Entonces

$$60\,000 = \sqrt{40\,000 \times P}$$

Elevando al cuadrado ambos miembros

$$60\,000 \times 60\,000 = 40\,000 \times P \Rightarrow P = 90\,000$$

Ejemplo 3

Es también el promedio que debe usarse cuando consideramos el desarrollo físico de un bebé: *Si un recién nacido pesó 3 kg el 1 de enero y 4,5 kg el 1 de marzo, ¿cuál era su peso aproximado el 1 de febrero?*

$$\text{Media geométrica} = \sqrt{3 \times 4{,}5} \simeq 3{,}7 \text{ kg}$$

Esta fórmula solamente es utilizable durante los tres primeros meses del recién nacido.

Estos diversos promedios han sido considerados por muchas generaciones de matemáticos, y su historia merece conocerse: está también relacionada con temas de música y de arquitectura. Otro tipo de media, la parte áurea, tiene asimismo gran interés. Procura averiguar algo acerca de ella.

Nota sobre el cálculo de la media aritmética

Si queremos hallar la media aritmética de gran número de datos (todos razonablemente próximos entre sí) podemos hacerlo más fácilmente proponiendo una media aproximada y realizando un breve cálculo. Por ejemplo, supongamos que vamos a considerar la altura de los muchachos de una clase de 30, cuyos resultados, expresados en centímetros, pueden ser:

162	161	164	163	161
165	167	170	165	167
163	165	165	168	171
164	164	164	162	160
160	164	166	164	163
158	166	166	165	164

Debes calcular la media aritmética. Parece ser algo así como 1,64 m. Ahora escribe en la tabla anterior la diferencia de cada valor al promedio que tú has previsto, como mostramos a continuación:

—2	—3	0	—1	—3
+1	+3	+6	+1	+3
—1	+1	+1	+4	+7
0	0	0	—2	—4
—4	0	+2	0	—1
—6	+2	+2	+1	0

Suma las diferencias. El total es +8.

Divide por el número de lecturas, 8/30.

La verdadera media aritmética es 164 + 8/30 (lo que indica que nuestra suposición «a ojo» fue muy buena). Este método tiene la ventaja de que nos evita sumar números grandes, y es además mucho más fácil y rápido de realizar que sumar rutinariamente todas las alturas y dividir por 30.

EJERCICIO 3

En una empresa se paga a los 10 obreros 12 000 pesetas mensuales, mientras que el gerente recibe 200 000.

(*a*) Si fueras obrero y desearas demostrar lo pobre que es el sueldo promedio, ¿qué promedio usarías?

(*b*) Si fueras gerente, ¿qué promedio emplearías para demostrar lo altos que son los sueldos?

¿Cuál piensas que es el promedio más justo para utilizar en este caso?

«Índice del coste de vida»

En primer lugar, aclaremos la cuestión del nombre. No existe nada que sea el «índice del coste de vida». Es una expresión común mal aplicada en vez de la correcta, el «índice de precios de venta», que tiene un significado completamente distinto. El índice de precios de venta incluye todos los artículos que pueden ser comprados por una familia «media», mientras que el índice del «coste de vida» supone atención exclusiva a las necesidades básicas e imprescindibles.

¿Qué es un «índice»?

Un índice es un número que se utiliza como recurso adecuado para la comparación de precios. Para establecer el índice se toma un patrón tipo, por ejemplo, el precio del pan el 1 de enero de 1962; sea éste de 10,50 ptas./kg. Si en el día de hoy el precio es de 12 ptas./kg quiere decir que, atribuyéndole el índice 100 al patrón tipo, nuestro índice actual es $\frac{120}{105} \times 100 = 114$ (se da el entero más próximo).

El índice de precios de venta

Antes de que pueda hacerse ningún cálculo se debe investigar el presupuesto de una muestra representativa de familias. Por ejemplo, en Inglaterra se hizo esto en 1953, con una muestra de 13 000 presupuestos familiares. La razón de esto es averiguar la importancia relativa para el consumidor de las varias mercancías para tenerla en cuenta en los cálculos; esto se llama *ponderación* de las mercancías. Cierto tipo de gastos no se incluyó en los presupuestos (seguiremos refiriéndonos a la citada muestra de 1953) por ser demasiado variables. (Por ejemplo, limosnas, documentos personales, etcétera.)

El índice, pues, debe tomar en cuenta la mayor parte de las mercancías usuales, que se incluyeron en los diez grupos siguientes:

1. Alimentos
2. Bebidas alcohólicas
3. Tabaco
4. Vivienda
5. Electricidad y combustible
6. Artículos domésticos durables
7. Ropas y calzado
8. Transportes y vehículos
9. Diversos (prensa, juguetes...)
10. Servicios

Cada grupo se subdivide en varias secciones, por ejemplo, en el grupo 1 se distinguen diversos tipos de alimentos. El índice así proyectado debe darnos exclusivamente una idea de la variación global de los precios, y se calcula cada mes, ahora diremos cómo.

Ponderación

El índice va a indicarnos sólo variaciones globales de los precios, de modo que es esencial conseguir cada mes los precios de los mismos artículos y servicios. Hay dos modos sencillos de hacerlo.

En primer lugar, para cierta fecha inicial podemos calcular el coste de una selección muy amplia y representativa de artículos y servicios. Éste es el que llamaremos *coste inicial*. Luego, cada mes, averiguaremos el coste de esa misma selección, y lo compararemos con el coste inicial, al que tomamos como 100.

Alternativamente, y de hecho este segundo método es más cómodo, puede calcularse en tantos por ciento la alteración del precio (por litro, por kg, etc.) de cada artículo, y reunirlos en una alteración global mediante el uso de pesos como sigue: Evidentemente, algunos artículos son más importantes para las familias que otros. Un aumento del 15 % en el precio del café es posible que altere más a la economía familiar que un aumento del 20 % en el precio del azafrán (pues probablemente se gasta muchísimo más en café que en azafrán). Con el fin de tener en cuenta la importancia que tiene cada artículo en la economía familiar, se atribuye a éste un «peso». Luego, los porcentajes de peso de cada artículo se multiplican por su peso antes de ser promediado, y el resultado es el mismo que antes. Los pesos se atribuyen en proporción a las cantidades gastadas en cada concepto en la fecha inicial.

El peso total de todos los artículos se toma como 1000. El resultado de esta ponderación en el citado cuestionario de 1953 fue como sigue:

1.	Alimentos	350
2.	Bebidas alcohólicas	71
3.	Tabaco	80
4.	Vivienda	87
5.	Luz y combustible	55
6.	Artículos domésticos	66
7.	Ropas y calzados	106
8.	Transportes y vehículos	68
9.	Bienes diversos	59
10.	Servicios	58
	TOTAL	1000

Los precios se recogen en todo el territorio nacional tomando en cuenta una muestra representativa de cada artículo, aunque deben hacerse bastantes ajustes, como en el caso de las patatas (nuevas o viejas), frutas (en temporada o fuera de tiempo) y otros.

Cálculo del índice

Vamos a imaginar, para concretar y simplificar, que nuestra selección inicial se compone de 5 artículos, cuyo valor es el que sigue:

8 barras de pan	a 12 ptas. la barra	96
1 litro de aceite	a 40 ptas. litro	40
500 g de café	a 64 ptas. kg	32
1 kg de fruta	a 20 ptas. kg	20
6 pasteles	a 2 ptas. cada uno	12
	Coste total:	200

Algún tiempo después compramos una selección de mercancías exactamente igual, y los precios han variado, de modo que es

8 barras de pan	a 12,50 ptas. la barra	100
1 litro de aceite	a 40 ptas. litro	40
500 g de café	a 66 ptas. kg	33
1 kg de fruta	a 21 ptas. kg	21
6 pasteles	a 3 ptas. cada uno	18
	Coste total:	212

Por lo tanto, el correspondiente índice es $(212/200) \times 100 = 106$.

Alternativamente, para el cálculo por el segundo método las cosas se disponen así:

Artículo	*% de aumento del precio inicial*	*Peso (dado por el gasto inicial)*	*Producto de ambos (100 veces el aumento de coste)*
Pan	$4\frac{1}{6}$	96	400
Aceite	0	40	0
Café	$3\frac{1}{8}$	32	100
Fruta	5	20	100
Pasteles	50	12	600
		200	1200

De este modo el aumento ponderado es $\frac{12}{200} \times 100 = 6\ \%$, y el índice resulta 106, como antes.

EJERCICIO 4

En ciertas fechas de 1964 y 1965, el gasto en pesetas de una familia en la tienda de comestibles varió como sigue, poniendo la misma cantidad de cada artículo:

	Fecha					
	1964		1965			
Artículo	1-IX	1-XII	1-III	1-VI	1-IX	1-XII
Mantequilla	17	20	19	19,50	19,50	20,50
Aceite	8	9	8,50	8,50	7	9,50
Queso	6,50	5	6,50	6	6	5,50
Huevos	25,50	21	18	19,50	21	27
Café	8	9	8	8	9	8
Pan	21	21,75	21	21	21	21,75
Azúcar	8	8	8	8,25	8,25	8,50
Patatas	10	12,50	17,50	15	10	12,50
Carne	37,50	37,25	30	39	39	39,50
Verdura	5	6	4	2,50	3	6
Legumbres	3,50	5	6	6	4	6
Totales	150,00	154,50	155,50	153,25	147,75	164,75

Tomando la primera fecha dada como inicial y típica, calcula el índice que corresponde a las otras fechas dadas. ¿Cómo se han alterado los pesos? ¿Por qué hay variaciones en los precios?

Haz para ti una selección representativa de los artículos que juzgues de interés y revisa su coste cada mes. Compara este índice particular con los que se publican. ¿Por qué es razonable que haya diferencias?

Las normas adoptadas en España para el cálculo de índices se explican en la monografía del *Instituto Nacional de Estadística*: «Coste de vida. Nuevo sistema de números índices, base 1958 = 100». Madrid, 1962. En la edición del Anuario Estadístico de diciembre de 1965, se dan para el conjunto nacional los siguientes resultados (con 1958 = 100):

	1961	1962	1963	1964
Alimentación	110,2	118,7	130,6	137,2
Vestido	109,0	113,6	124,8	141,7
Vivienda	104,6	111,2	116,1	122,6
Gastos de casa	109,9	112,1	120,9	130,4
Diversos	118,6	121,6	128,1	138,8
Índice global	111,3	117,6	127,9	136,8

N. del T.

CAPITULO IV

Probabilidades

«Hay 100 probabilidades contra 1 de que su coche no sea robado este año. Como conductor, resulta tres veces más probable que tenga un accidente que cause lesiones que el que le roben el coche. Por otra parte, es 10 millones de veces más probable que le roben a usted su coche que acertar una triple quiniela para 54 juegos de 8 premios.»

(De acuerdo con algunas estadísticas publicadas en la prensa en 1962.)

Nunca hubo una cosa más impropiamente considerada que la *probabilidad*. En verdad es una medida tan precisa como cualquier otra de las matemáticas, y descansa en cimientos igual de firmes. El cálculo de probabilidades de que tratamos no tiene nada en común con nociones indeterminadas tales como «¿lloverá mañana?», pero da una respuesta clara a cuestiones precisas, como «¿cuál es la probabilidad de obtener un total de 7 puntos con dos dados?».

Hoy, cuando tantos adultos emplean gran parte de su tiempo «trabajando» en sus quinielas de fútbol, podría ser interesante calcular la probabilidad que tienen de acertar. Como es natural, básicamente hay tres posibilidades iguales para cada partido: ganar, perder o empatar. Por lo tanto, hay una probabilidad de 1 entre 3 de adivinar el resultado correcto de un partido. Esto significa que la probabilidad de predecir un segundo resultado correcto es también de 1 entre 3; la de acertar ambos es de 1 entre 9, y así sucesivamente. (Si no ves esto claro, no te preocupes, sigue leyendo.) Como verás, la probabilidad de acertar cualquiera de los 14 resultados es más bien pequeña, y la de acertar los 14 partidos exigidos por la quiniela es tan pequeña que resulta casi inexistente. La teoría de probabilidades, decíamos, se refiere a la cuestión «¿cuál es la probabilidad de ...?», que puede traer otras varias cuestiones implicadas.

La probabilidad, como ya hemos dicho, es una medida precisa; de forma que, en primer lugar, debemos asignar probabilidades iguales a acontecimientos igualmente probables. Para hacer esto adecuadamente, hemos de determinar una conveniente escala de probabilidades. Podemos emplear cualquier escala, pero la más útil y casi la más evidente es aquella en la que

0 significa el suceso que no puede ocurrir, o sea que es *imposible*;
1 significa el suceso que necesariamente ha de ocurrir;
y la probabilidad de un resultado entre 10 se expresa como 0,1; etc.

Por ejemplo, estudiemos algunas probabilidades sencillas:

1. Al tirar una moneda, quedará en el suelo mostrando la cara o la cruz (despreciamos cualquier posibilidad de que caiga a través de una rendija del pavimento o de que quede de canto). Así podemos decir que hay una posibilidad entre 2, o probabilidad $\frac{1}{2}$, de que la moneda muestre cara (o cruz).
2. La probabilidad de que al tirar un dado consigamos un 1 (lo mismo con 2, 3, 4, 5, 6) es $\frac{1}{6}$.
3. Hay probabilidades $\frac{1}{4}$ de que al sacar una carta de una baraja española sea de copas (lo mismo con espadas, bastos, oros).
4. La probabilidad de sacar la sota de copas es $\frac{1}{40}$.

A continuación ofrecemos un desarrollo de esta escala:

ESCALA DE PROBABILIDAD

1,0	Certeza de que hemos de morir	$p = 1$
	Mañana amanecerá	$p = 1$
	Este mes no es mayo (lo mismo con en., feb., mar., abr., jun., jul., ag., sep., oct., nov., dic.)	$p = \frac{11}{12}$
	¿Puedes comprar «El Noticiero» de hoy? (el martes, mi., ju., vi., sab. y dom. sí; el lunes no se publica)	$p = \frac{6}{7}$
0,75	La carta que vas a sacar no será de copas (lo mismo con oros, espadas, bastos)	$p = \frac{3}{4}$
	Tendrás clase hoy (sólo el domingo no hay clases)	$p = \frac{6}{7}$
	Tu equipo de fútbol no perderá (esto es suponiendo que las probabilidades de ganar, perder y empatar sean iguales)	$p = \frac{2}{3}$
0,50	Moneda: cara (Cruz) Recién nacido: varón (Hembra)	$p = \frac{1}{2}$
	Perder un partido de fútbol (si son iguales las probabilidades de ganar, perder y empatar)	$p = \frac{1}{3}$
0.25	La carta que vas a sacar es de espadas (oros, bastos, copas)	$p = \frac{1}{4}$
	Es primavera (verano, otoño, invierno)	$p = \frac{1}{4}$

	Sacar 6 con un dado (1, 2, 3, 4, 5)	$p = \frac{1}{6}$
	Hoy es lunes (mar., mi., ju., vi., sáb., dom.)	$p = \frac{1}{7}$
	Sufrir el robo del coche ($\simeq$ se lee: aproximadamente)	$p \simeq \frac{1}{100}$
	Hoy es Navidad	$p = \frac{1}{365}$ (año normal)
0	Acertar las quinielas	$p \simeq 0$
	Atravesar a nado el Océano Atlántico	$p = 0$
	Ver un asno que vuela	$p = 0$

Consideremos ahora unos cuantos ejemplos menos simples:

1. *¿Cuál es la probabilidad de sacar:* (a) *dos caras,* (b) *una cara y una cruz, tirando dos veces una moneda?*

Indicamos con K los casos de cara y con T los de cruz. Las posibilidades son:

Primera tirada	*Segunda tirada*
K	K
K	T
T	K
T	T

Así, en una de las 4 posibilidades tenemos dos caras. Por tanto, la probabilidad preguntada en (*a*) es $\frac{1}{4}$. En 2 de las 4 posibilidades aparece una cara y una cruz. Por tanto, la probabilidad pedida en (*b*) es $\frac{2}{4} = \frac{1}{2}$.

2. *¿Cuál es la probabilidad de conseguir:* (a) *tres caras,* (b) *dos caras y una cruz,* (c) *tres marcas iguales, en una prueba en la cual tiramos tres monedas?*

Las posibilidades son:

Primera moneda	*Segunda moneda*	*Tercera moneda*
K	K	K
K	K	T
K	T	K
K	T	T
T	K	K
T	K	T
T	T	K
T	T	T

Por tanto, la probabilidad de tres caras es $\frac{1}{8}$.

La probabilidad de dos caras y una cruz es $\frac{3}{8}$.

La probabilidad de tres marcas iguales es $\frac{1}{8}$ (caras) + $\frac{1}{8}$ (cruces), total $\frac{2}{8} = \frac{1}{4}$.

Esta idea puede extenderse a todos los casos en los que haya igual probabilidad de que ocurra uno u otro de los sucesos; por ejemplo, para una familia de tres hijos (si hay igual probabilidad de que haya nacido un niño o una niña) hay probabilidad de $\frac{1}{4}$ de que los tres sean del mismo sexo, probabilidad de $\frac{3}{8}$ de que sean dos chicos y una chica (o dos chicas y un chico) y probabilidad de $\frac{1}{8}$ de que sean los tres chicos (o chicas).

El mismo tipo de análisis puede aplicarse en cualquier examen de muestras en el que haya iguales posibilidades de que la muestra sea aceptada o rechazada. Para facilitar el cálculo de probabilidades en este tipo de cuestión, introducimos la idea de «árbol de probabilidades». Las probabilidades en cada etapa se escriben sobre las ramas, y para hallar la probabilidad final de que ocurra un suceso, multiplicamos las probabilidades de cada una de las ramas. (¿Puedes ver por qué? Si no, vuelve sobre este punto después de leer unas cuantas páginas más.) Es innecesario decir que sólo se refiere a experimentos en los cuales únicamente puede ocurrir un suceso cada vez (podemos tener cara o cruz, pero no ambas cosas a la vez). Tales sucesos se llaman *incompatibles.* Ahora refiramos a nuestro árbol de probabilidades el anterior supuesto de lanzar una moneda:

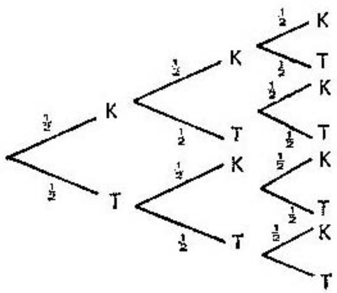


Fig. 4.1

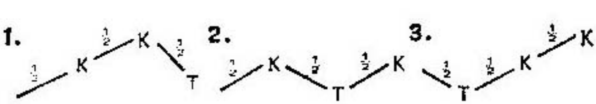
de forma que la probabilidad de obtener tres caras es

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$$

La probabilidad de obtener dos caras y una cruz es

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \quad + \quad \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \quad + \quad \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$$

Esto corresponde a los tres caminos:



Hemos estado utilizando intuitivamente las dos leyes de la probabilidad: la *ley de multiplicación* y la *ley de adición*. Antes de enunciar formalmente dichas leyes, estudiaremos unos cuantos problemas más para ver cómo se emplean.

Ley de adición

1. ¿Cuál es la probabilidad de sacar el punto 2 o el punto 3 con un dado?

La probabilidad de sacar un 2 es $\frac{1}{6}$.

La probabilidad de sacar un 3 es $\frac{1}{6}$.

Por tanto, la probabilidad de sacar un 2 o un 3 es $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

2. ¿Cuál es la probabilidad de sacar 5 o menos de 5 lanzando un dado?

La probabilidad de sacar un 1 es $\frac{1}{6}$.

La probabilidad de sacar un 2 es $\frac{1}{6}$.

La probabilidad de sacar un 3 es $\frac{1}{6}$.

La probabilidad de sacar un 4 es $\frac{1}{6}$.

La probabilidad de sacar un 5 es $\frac{1}{6}$.

Por lo tanto, la probabilidad de sacar un 1, o 2, o 3, o 4, o 5 es

$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6}$$

Alternativamente, la probabilidad de sacar *más* de 5 (es decir, un 6) es $\frac{1}{6}$. Por lo tanto, la probabilidad de sacar menos de 6 es

$$1 - \frac{1}{6} = \frac{5}{6}$$

La ley de adición dice que si las probabilidades de que ocurran n sucesos incompatibles son p_1, p_2, p_3, ..., p_n, entonces hay una probabilidad $p = p_1 + p_2 + p_3 + \ldots + p_n$ de que ocurra uno de los sucesos dichos.

Ejemplo

Supongamos una urna que contiene siete bolas blancas y tres negras, idénticas salvo en el color, ¿cuál es la probabilidad de extraer a ciegas una bola negra o una blanca?

La probabilidad de extraer una bola blanca es $\frac{7}{10}$.

La probabilidad de extraer una bola negra es $\frac{3}{10}$.

Por tanto, la probabilidad de extraer una bola negra o una blanca es $\frac{7}{10} + \frac{3}{10} = 1$ (la seguridad). Esto podía haberse previsto, ya que tenemos que extraer forzosamente una bola blanca o una bola negra.

Ley de multiplicación

Ejemplos

1. *¿Cuál es la probabilidad de obtener cuatro veces cara tirando cuatro veces una moneda?*

Probabilidad de cara en la 1.ª tirada $= \frac{1}{2}$.

Probabilidad de cara en la 2.ª tirada $= \frac{1}{2}$.

Probabilidad de cara en la 3.ª tirada $= \frac{1}{2}$.

Probabilidad de cara en la 4.ª tirada $= \frac{1}{2}$.

Por tanto, la probabilidad de cuatro caras es

$$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{16}$$

(Podríamos haber resuelto este problema utilizando el árbol de probabilidades.)

2. *¿Cuál es la probabilidad de sacar dos cincos tirando dos veces un dado?*

Probabilidad de un 5 en la 1.ª tirada $= \frac{1}{6}$.

Probabilidad de un 5 en la 2.ª tirada $= \frac{1}{6}$.

Por tanto, la probabilidad de sacar dos cincos es $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$.

La *ley de multiplicación* dice que si las probabilidades de n sucesos independientes * son $p_1, p_2, p_3, \ldots, p_n$, entonces la probabilidad de que ocurran todos los sucesos es $p = p_1 \times p_2 \times \ldots \times p_n$.

Ejemplo

Imagina que tiramos un dado dos veces. ¿Cuál es la probabilidad de que en la primera tirada saquemos 3 o menos de 3 y en la segunda 4 o menos de 4?

* Se dice que los sucesos son independientes si ninguno de ellos afecta a la realización de los otros. *(N. del T.)*

La probabilidad de sacar en la primera tirada 3 o menos de 3 es $\frac{3}{6}$.

La probabilidad de sacar en la segunda tirada 4 o menos de 4 es $\frac{4}{6}$.

Por tanto, la probabilidad de que ocurran ambas cosas es

$$\frac{3}{6} \times \frac{4}{6} = \frac{1}{3}$$

Muchos problemas de probabilidad requieren el conocimiento de estas dos leyes.

Ejemplos

1. *Supongamos que tenemos una urna conteniendo ocho bolas: cinco blancas y tres negras. ¿Cuál es la probabilidad de que si sacamos dos sean una blanca y otra negra?* (Sucesos dependientes.)

Supongamos que primero sacamos una blanca, lo que tiene probabilidad $p_b = \frac{5}{8}$; quedan entonces siete bolas y la probabilidad de sacar una negra es $\frac{3}{7}$. Por lo tanto, la probabilidad de sacar primero una blanca y luego una negra es $\frac{5}{8} \times \frac{3}{7} = \frac{15}{56}$.

Análogamente, si sacamos primero la negra es $p_n = \frac{3}{8}$; quedan entonces siete bolas y la probabilidad de sacar una blanca es $\frac{5}{7}$. De donde se deduce que la probabilidad de sacar una bola negra y después una blanca es $= \frac{3}{8} \times \frac{5}{7} = \frac{15}{56}$.

Por tanto, la probabilidad total de que las bolas sean una blanca y otra negra es $\frac{15}{56} + \frac{15}{56} = \frac{15}{28}$.

2. *¿Cuál es la probabilidad de conseguir un total de 8 puntos tirando dos veces un dado?* (Sucesos independientes.)

En primer lugar hay 36 formas en las que puede caer el dado (6 para la primera tirada y 6 para la segunda).

¿Cómo puede alcanzarse el tanteo de 8?

Como $6 + 2$, $5 + 3$, $4 + 4$, $3 + 5$, $2 + 6$, es decir, de cinco formas distintas, o sea que la probabilidad de obtener un total de 8 es $\frac{5}{36}$.

3. *Problema de los cumpleaños. Imagina que tienes 24 amigos que celebran la fiesta de su cumpleaños. En un mismo día, sólo puedes asistir a una fiesta. ¿Cuál es la probabilidad de que pierdas alguna invitación debido a coincidencia de fechas de nacimiento?*

En un año normal de 365 días, si no han de coincidir las fechas:

(*a*) El cumpleaños de tu primer amigo puede ser cualquier día.

(*b*) El de tu segundo amigo, cualquiera de los 364 restantes.

(*c*) El cumpleaños de tu tercer amigo, cualquiera de los 363 días restantes, etc.

Por tanto, la probabilidad de no perder ninguna fiesta es

$$\frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \ldots \times \frac{342}{365}$$

que podemos calcular fácilmente usando logaritmos, y da la probabilidad

$$p = 0{,}462$$

Así, la probabilidad de que haya coincidencia de fecha es $1 - 0{,}462 = 0{,}538$. Dicho de otro modo, hay muchas probabilidades (más del 50 %) de que tengas que perderte alguna fiesta.

EJERCICIO 5

1. ¿Cuál es la probabilidad de sacar un 5 con un dado?

2. ¿Cuál es la probabilidad de totalizar: (*a*) 9, (*b*) 7, (*c*) 5, tirando un dado dos veces?

4. ¿Cuál es la probabilidad de tirar cinco veces una moneda y obtener: (*a*) 5 caras, (*b*) 3 caras y 2 cruces, (*c*) 4 marcas iguales?

5. En una familia con seis hijos, ¿cuál es la combinación más probable de chicos y chicas?

6. ¿Cuál es la probabilidad de sacar: (*a*) el as de espadas, (*b*) un as, (*c*) el as, rey y caballo de espadas tirando tres cartas, (*d*) cuatro cartas del mismo palo de una baraja española de 40 naipes?

7. Cinco hombres van a un banquete de boda y cuelgan sus americanas en el guardarropa. Desgraciadamente las papeletas del guardarropa se extravían y por esa razón el empleado cuelga sus americanas al azar. ¿Cuál es la probabilidad de que: (*a*) cada uno recoja su propia americana; (*b*) nadie recoja su americana?

9. Durante el mes de marzo, la probabilidad de que haya un día lluvioso es 0,6. El Real Madrid tiene 0,8 probabilidades de ganar en un día claro, pero sólo 0,4 en uno lluvioso. Si sabemos que ganó cierto partido en marzo, ¿cuál es la probabilidad de que lloviera ese día?

10. Diez hombres toman parte en un juego de azar. Cada uno tiene el mismo capital. Los dos primeros (seleccionados a suertes) juegan hasta que uno de ellos gana al otro todo el dinero. Entonces el ganador se enfrenta con otro jugador, y así sucesivamente. ¿Quién tiene más probabilidades de ganar: uno de los primeros jugadores o el último que entra en el juego?

El árbol de probabilidades

Volvemos ahora al juego de cara y cruz y construimos un árbol mayor, para cinco tiradas de una moneda. En cada escalón calcularemos la probabilidad de llegar a ese punto.

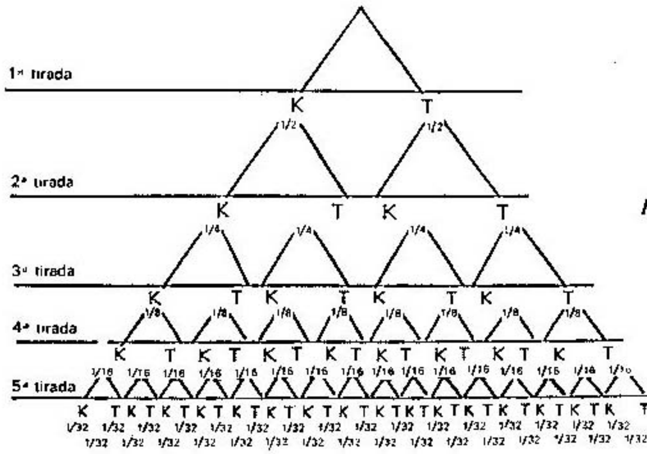


Fig. 4.2

Ahora podemos combinar algunos de los resultados de las ramas, por ejemplo en dos tiradas tenemos una probabilidad de $\frac{1}{2}$ para dos caras y una de $\frac{1}{4} + \frac{1}{4}$ para una cara y una cruz (sin hacer distinción en el orden).

Trazando ahora un nuevo árbol, aplicando esta idea, tenemos:

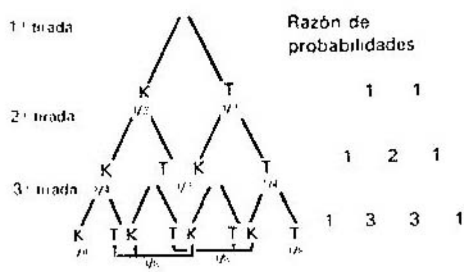


Fig. 4.3

Este triángulo que expresa las razones entre las probabilidades de los resultados de 2, 3, 4, ... jugadas se llama *triángulo de Pascal.* Está formado como sigue:

$$\begin{array}{c}
1 \quad 1 \\
1 \quad 2 \quad 1 \\
1 \quad 3 \quad 3 \quad 1 \\
1 \quad 4 \quad 6 \quad 4 \quad 1 \\
1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\
1 \quad 6 \quad 15 \quad 20 \quad 15 \quad 6 \quad 1 \\
1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1 \\
1 \quad 8 \quad 28 \quad 56 \quad 70 \quad 56 \quad 28 \quad 8 \quad 1 \\
1 \quad 9 \quad 36 \quad 84 \quad 126 \quad 126 \quad 84 \quad 36 \quad 9 \quad 1 \\
1 \quad 10 \quad 45 \quad 120 \quad 210 \quad 252 \quad 210 \quad 120 \quad 45 \quad 10 \quad 1
\end{array}$$

Un número cualquiera de cada fila se obtiene sumando los dos números que están sobre él, como puede verse en el diagrama. El triángulo de Pascal es muy importante, pues nos da los coeficientes del desarrollo binómico, por ejemplo:

$$\begin{aligned}
a + b &= +\mathbf{1}\,a + \mathbf{1}\,b \\
(a + b)^2 &= +\mathbf{1}\,a^2 + \mathbf{2}\,a\,b + \mathbf{1}\,b^2 \\
(a + b)^3 &= +\mathbf{1}\,a^3 + \mathbf{3}\,a^2\,b + \mathbf{3}\,a\,b^2 + \mathbf{1}\,b^3 \\
(a + b)^4 &= +\mathbf{1}\,a^4 + \mathbf{4}\,a^3\,b + \mathbf{6}\,a^2\,b^2 + \mathbf{4}\,a\,b^3 + \mathbf{1}\,b^4 \\
(a + b)^5 &= +\mathbf{1}\,a^5 + \mathbf{5}\,a^4\,b + \mathbf{10}\,a^3\,b^2 + \mathbf{10}\,a^2\,b^3 + \mathbf{5}\,a\,b^4 + \mathbf{1}\,b^5
\end{aligned}$$

. . . .

10.ª fila:

$$\begin{aligned}
(a + b)^{10} = &+ 1\,a^{10} + 10\,a^9\,b + 45\,a^8\,b^2 + 120\,a^7\,b^3 + 210\,a^6\,b^4 \\
&+ 252\,a^5\,b^5 + 210\,a^4\,b^6 + 120\,a^3\,b^7 + 45\,a^2\,b^8 \\
&+ 10\,a\,b^9 + 1\,b^{10}
\end{aligned}$$

En los problemas de muestreo usaremos con frecuencia el *teorema del binomio* (que es una generalización de estas ideas). El teorema del binomio dice que

$$(a + b)^n = a^n + n\,a^{n-1}\,b +$$

$$+ \frac{n\,(n-1)}{2 \times 1}\,a^{n-2}\,b^2 + \frac{n\,(n-1)\,(n-2)}{3 \times 2 \times 1}\,a^{n-3}\,b^3 +$$

$$+ \frac{n\,(n-1)\,(n-2)\,(n-3)}{4 \times 3 \times 2 \times 1}\,a^{n-4}\,b^4 + \ldots + b^n$$

Experimento 1

Billarín de Galton

El anterior teorema puede experimentarse empleando un tablero de alfileres cuidadosamente clavados al tresbolillo, como en el esquema que sigue

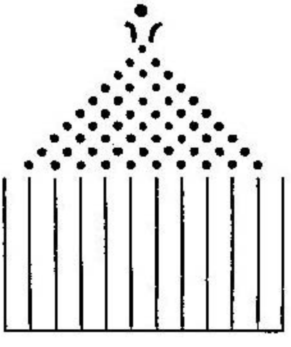


Fig. 4.4

Podemos inclinar el tablero cuidadosamente de forma que las bolitas caigan a través del embudo y hacia abajo, pasando entre los alfileres. Si se hace bien esto, habrá una probabilidad de $\frac{1}{2}$ de que cada bolita se dirija hacia la izquierda o hacia la derecha, después de cada choque con un alfiler. Si empleamos 1024 bolitas, podemos esperar (aplicando el teorema del binomio) que en los conductos del fondo haya el siguiente número de bolas:

1 10 45 120 210 252 210 120 45 10 1

Experimento 2

El cálculo de la aguja de Buffon *

Requiere mucho menos aparato. Todo lo que se necesita es una aguja o una cerilla y una hoja grande de papel sobre la cual hay un conjunto de rectas paralelas igualmente espaciadas. Si colocamos alisado y horizontal el papel y lanzamos sobre él la aguja al azar, podemos contar el número de veces que la aguja queda cortando a una de las rectas. La probabilidad de que la aguja corte a una recta tal viene dada por:

$$P = \frac{2l}{\pi a}$$

donde l es la longitud de la aguja y a la distancia entre las paralelas.

Así, si n es el número de veces que la aguja queda secante a una recta y N el número total de tiradas,

$$\frac{n}{N} = \frac{2l}{\pi a}$$

y podemos obtener un valor aproximado para π. Cuantas más tiradas hagamos, más exactos serán nuestros resultados (a menos que tiremos la aguja de forma premeditada).

Experimento 3

Cada muchacho de la clase debe llevar a la escuela una caja de cerillas de madera. Se cortan todas las cabezas y se marca una cara de cada cerilla. Las cerillas se dejan caer al azar sobre el pupitre y las que caigan con la cara marcada hacia arriba se retiran. Se repite el experimento hasta que no quede ninguna.

* La realización del experimento de la aguja de Buffon requiere una precisión y unas precauciones prácticas que lo sitúan ordinariamente fuera de las posibilidades de una clase. Se puede, pues, prescindir de este párrafo. *(N. del T.)*

La gráfica que representa el número de cerillas que quedan en relación al número de tiradas es como sigue:

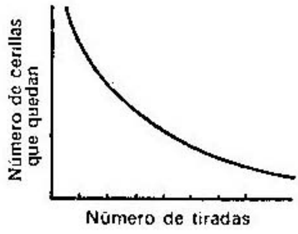


Fig. 4.5

Cada muchacho deberá trazar su propia gráfica y también la de los resultados totales de la clase. Este experimento ilustra la relación entre la desintegración radiactiva y la teoría de la probabilidad. En una sustancia radiactiva, el grado de desintegración depende de la masa (que a su vez depende del número de partículas que no se han desintegrado); análogamente, en nuestro experimento, la probabilidad de que retiremos una cerilla depende del número de éstas que queda en ese momento. Si fuera posible, debería realizarse el experimento complementario en el laboratorio de física; por ejemplo, hallando la vida media del torio, relacionando el número de choques en un electroscopio con el tiempo.

Los resultados de estos experimentos *muestran dos características* muy importantes:

1. Los resultados experimentales difieren algo de los esperados (los que habíamos calculado).
2. Cuanto mayor número de pruebas realicemos, tanto más estrechamente concordarán nuestros resultados teóricos y experimentales.

Recuerda que los resultados teóricos rara vez concuerdan exactamente con los experimentales.

EJERCICIO 6

1. En un experimento tiramos nueve veces una moneda y anotamos los resultados. Si repetimos esto 512 veces, ¿en cuántas esperarías:

(*a*) 4 caras y 5 cruces;
(*b*) 7 caras y 2 cruces?

2. En otro experimento tiramos diez veces una moneda. ¿Qué conclusiones (si las hay) podrías sacar en el caso de que los resultados fueran:

(*a*) 3 caras y 7 cruces;
(*b*) 4 caras y 6 cruces;
(*c*) 5 caras y 5 cruces?

Calcula las probabilidades en cada caso.

3. En un puñado de nueces hay una probabilidad de $\frac{1}{3}$ de que esté mala una nuez y de $\frac{2}{3}$ de que sea buena. En un paquete de 15 nueces, ¿cuántas es lo más probable que estén malas? Calcula las probabilidades de:

(*a*) 6 malas y 9 buenas;
(*b*) 6 buenas y 9 malas;
(*c*) 8 buenas y 7 malas;
(*d*) todas malas;
(*e*) todas buenas.

4. Una empresa construye transistores. Hay una probabilidad 0,9 de que cualquiera de ellos se ajuste bien a su esquema. En una caja de una docena, ¿cuál es la probabilidad de:

(*a*) 9 buenos y 3 defectuosos;
(*b*) 5 buenos y 7 defectuosos?

¿Cuál es el número de transistores defectuosos por caja que tiene mayor probabilidad?

5. Si sólo tenemos nueve filas de alfileres en nuestro tablero y 1024 bolitas, ¿cuántas puedes esperar que queden en cada conducto? ¿Puedes dar razones por las cuales (muy probablemente) en realidad no nos resultaría esto?

6. ¿Cuál es la probabilidad de: (*a*) sacar exactamente 5 caras tirando 10 veces una moneda; (*b*) sacar 4 caras y 6 cruces, 5 caras y 5 cruces, o 6 caras y 4 cruces?

7. ¿Cuál es la probabilidad de totalizar entre 5 y 9 puntos tirando dos dados?

8. Un hombre está bebiendo en un bar situado en la encrucijada de cinco calles. Bebe demasiado y es incapaz de recordar qué calle es la que conduce a su casa. Recurre a probar cada calle al azar sin elegir nunca dos veces la misma, hasta que encuentra la que conduce a su casa. Dibuja un árbol de probabilidades que represente cada camino posible hasta su casa, presentando las probabilidades de cada trayecto.

9. Entras en un local en el que hay dos máquinas tragaperras. Sabes que una de ellas paga premio con una probabilidad de $\frac{1}{2}$, mientras que la otra solamente lo hace con una probabilidad de $\frac{1}{3}$. Tú no sabes cuál de estas máquinas es cada una, por lo que eliges una al azar y ganas. ¿Cuál es la probabilidad de que hayas elegido la primera máquina? Si hubieras perdido, ¿qué máquina probarías a continuación?

10. Una caja contiene 3 huevos frescos y 5 pasados. Construye un árbol de probabilidades para mostrar los casos posibles si se toman y emplean 3 huevos consecutivamente (y, naturalmente, no se reemplazan). Construye otro árbol para mostrar qué ocurriría si tuviéramos 4 frescos y 5 pasados.

Probabilidades y áreas

La probabilidad y las áreas de los histogramas o gráficas de distribución están muy estrechamente relacionadas. Volvamos a considerar dos de los ejemplos propuestos en el capítulo II. En primer lugar, la cuestión de los guisantes contenidos en una vaina. Podemos considerar la probabilidad de sacar una vaina al azar y encontrar cuatro guisantes en ella; esto ocurrió siete veces de un total de treinta y nueve, por lo cual la probabilidad es $\frac{7}{39}$. Tomemos ahora una escala conveniente para que el área total encerrada por los «bloques» de nuestro histograma sea 1 (idea que usaremos con frecuencia); entonces la probabilidad de sacar una vaina que contenga 4 guisantes será el área del bloque que corresponde a la aparición del suceso, es decir,

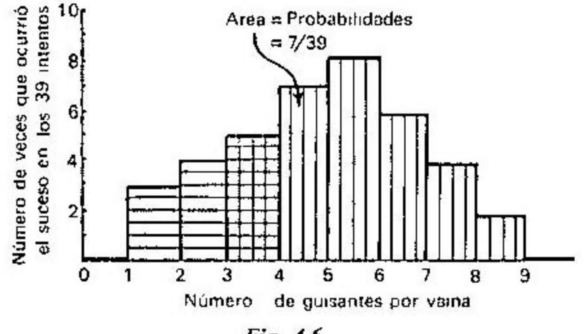


Fig. 4.6

Nota. Normalmente se ajusta la escala horizontal y no la vertical, pues es más fácil fijar las reglas de este procedimiento (en el capítulo siguiente se explicará esto con más detalle).

La probabilidad de tomar una vaina que contenga entre tres y siete guisantes es $\frac{5+7+8+6+4}{39} = \frac{30}{39}$, representada por el área rayada verticalmente.

La probabilidad de extraer una vaina con menos de cuatro guisantes es $\frac{5+4+3}{39} = \frac{12}{39}$, representada por el área rayada horizontalmente en el diagrama.

Problema 1

Vuelve a hacer el histograma de los resultados de fútbol (página 28), ajustando la escala horizontal, de forma que el área total encerrada sea la unidad. Luego utilízalo para calcular la probabilidad de que un equipo marque

(*a*) 2 goles en un partido;
(*b*) menos de 3 goles;
(*c*) más de 5 goles;
(*d*) 1, 2 o 3 goles;
(*e*) 6 goles. (¿Es de esperar que tu respuesta sea correcta para cualquier partido?)

Ejemplo

Supongamos que tienes que atravesar Londres en el metro y dispones de 8 minutos para hacer dos transbordos, cada uno de los cuales puede tomar hasta 7 minutos. ¿Cuál es la probabilidad de que pierdas el tren?

Podemos dibujar un diagrama para representar lo que ocurre.

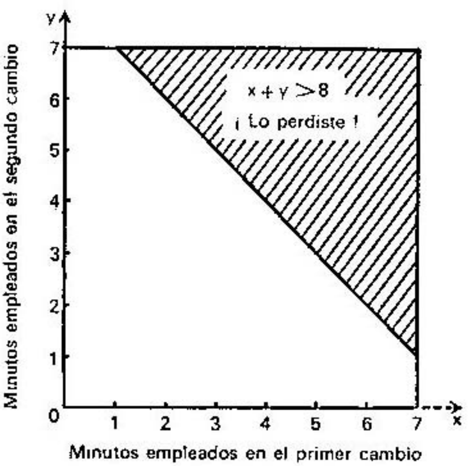


Fig. 4.7

Podemos emplear hasta 7 minutos en un cambio, pero sólo 8 para los dos juntos. El área sin rayar representa las coordenadas que satisfacen esta condición ($x + y < 8$, $x \geqslant 0$, $y \geqslant 0$). La probabilidad de perder el tren es la relación entre el área rayada y la total, es decir, $\frac{18}{49}$.

Problema 2 (mucho más difícil)

Intenta el mismo problema con 8 minutos de tiempo disponible, y teniendo que hacer tres transbordos, cada uno de los cuales puede tomar hasta 3 minutos. (La figura esta vez debe hacerse en tres dimensiones.) Halla la probabilidad de perder el tren. ¿Qué ocurre si realizamos el primer cambio en menos tiempo del promedio?

Probabilidad y física nuclear

La probabilidad y la estadística han desempeñado un gran papel en el desarrollo de la física nuclear. La misma teoría cuántica es de origen estadístico.

Hoy en día se utilizan en Física tres tipos de estadísticas:

1. La estadística de Maxwell-Boltzmann, que se emplea al estudiar la teoría cinética de los gases.

2. La estadística de Bose-Einstein, al tratar de fotones, ciertas moléculas gaseosas y átomos con número *par* de partículas elementales.

3. La estadística de Fermi-Dirac, referente a electrones, neutrones, protones, nucleones y átomos, con número *impar* de partículas elementales (bajo el principio de exclusión de Pauli).

Un modelo muy simple de estas estadísticas es como sigue. Supongamos que tenemos tres muchachos y dos caramelos (uno rojo y otro verde) para distribuir entre ellos (sin romper los dulces).

Hay las siguientes posibilidades:

	Primer muchacho	*Segundo muchacho*	*Tercer muchacho*
1.	rojo, verde		
2.		rojo, verde	
3.			rojo, verde
4.	rojo	verde	
5.	rojo		verde
6.		rojo	verde
7.	verde	rojo	
8.	verde		rojo
9.		verde	rojo

1. Según Maxwell-Boltzmann, todas estas posibilidades tienen la misma probabilidad de ocurrir (9 casos).

2. Se puede relacionar Bose-Einstein con el mismo problema en el caso en que los dulces sean idénticos (indistinguibles) (6 casos).

3. El principio de exclusión de Pauli dice que ningún muchacho puede recibir más de un dulce. Si además los dulces son indistinguibles, tenemos un modelo de la estadística de Fermi-Dirac (3 casos).

Naturalmente, en física nuclear el problema no es tan sencillo. De hecho, una analogía mejor tendría en cuenta las distintas edades de los muchachos, y cómo esto podría influir en las probabilidades de que les correspondiera un dulce.

CAPITULO V

Distribuciones

En el capítulo segundo hemos considerado varios histogramas. Si dispusiéramos de un número muy grande de datos, y los intervalos de la escala horizontal fuesen muy pequeños, nuestro histograma parecería limitado por una curva. Ésta recibe el nombre de *Curva de distribución de frecuencias.*

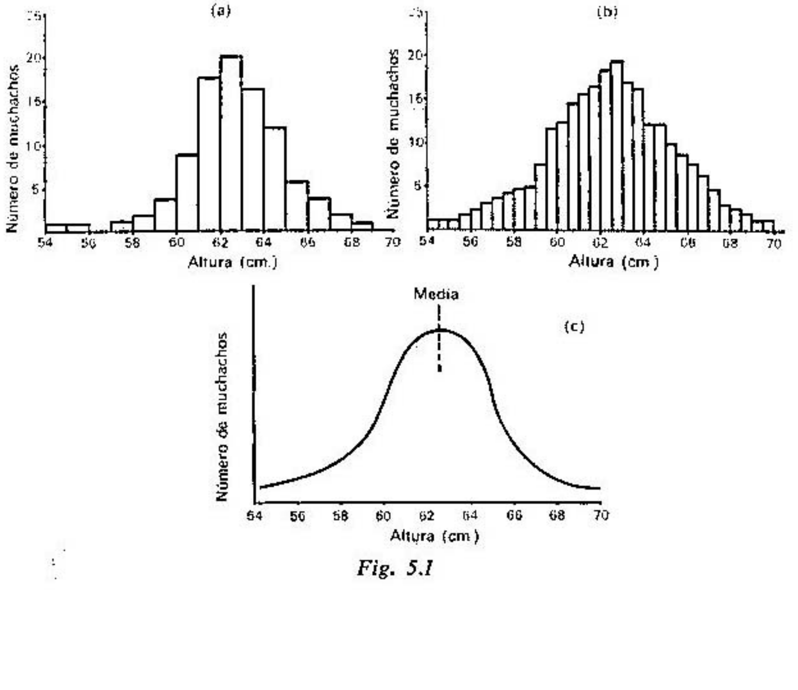


Fig. 5.1

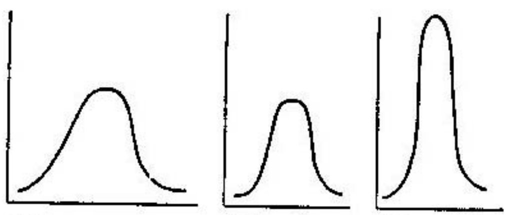
Observación. Sólo es posible tener una curva de distribución de frecuencias cuando la escala horizontal indica una magnitud continua (talla, peso, etc.).

Ejemplo: Talla de muchachos.

(*a*) Histograma para 100 muchachos (medidos con una precisión de 1 cm).
(*b*) Histograma para 200 muchachos (medidos con una precisión de 5 mm).
(*c*) Curva de distribución de frecuencia, basada en datos relativos a un gran número de muchachos.

Distribución normal

Hemos visto que el tipo más común de distribución en la naturaleza (por ejemplo, alturas de un grupo de hombres de una edad determinada, número de guisantes en una vaina, briznas de hierba por dm^2 de césped) es de forma acampanada: el tamaño y forma de la campana vienen determinados por las escalas que se utilizan.



(a) (b) (c) *Fig. 5.2*

En (b) se ha acortado la escala horizontal, y en (*c*) se han permutado las escalas horizontal y vertical.

Esta curva fue estudiada por De Moivre en 1733 y más tarde por Gauss. Aparece fundamentalmente cuando deseamos calcular la probabilidad de que ocurran ciertos sucesos, sabiendo que su distribución es de tipo normal (acampanado). Para hacer este cálculo, debemos partir de la base de que el *área total* (la *probabilidad*) encerrada por la curva y el eje x es la unidad.

Esto se hace utilizando una escala horizontal especial. Es relativamente complicado, ya que sabemos que la forma de la curva depende del número de observaciones y de la concentración (o dispersión) de los resultados.

La escala que vamos a utilizar está basada en la *desviación típica* de la media (llamada internacionalmente desviación *standard*). La desviación standard σ (sigma) se define por

$$\sigma = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - \bar{x})^2}{n}}$$

donde x_i es el valor observado
$\bar{x}$ es el valor medio
n es el número de observaciones.

La desviación standard es una medida de la dispersión de los valores observados. En el diagrama adjunto hay tres curvas de dis-

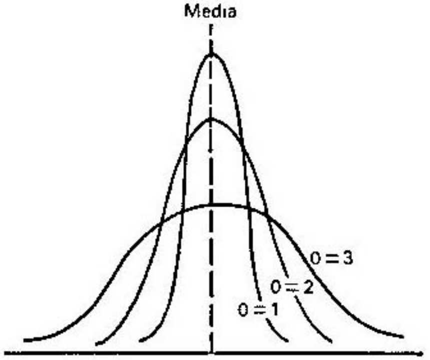


Fig. 5.3

tribución normal, las tres con el mismo valor medio y encerrando áreas iguales, pero con desviación standard diferente. En la práctica se comparan todas las distribuciones normales con aquella en la cual es $\sigma = 1$. Esto se consigue ajustando convenientemente la escala horizontal.

La distribución normal aparece en la naturaleza con mucha frecuencia y por esta causa es de extraordinaria importancia bajo muchos aspectos. Un simple ejemplo puede verse fácilmente en la industria del vestido. La talla y medidas de las personas siguen un modelo de distribución normal, de manera que usando éste se puede calcular el número de prendas de cada talla que es probable sean encargadas.

Debido a la importancia de la curva de distribución normal, todas las áreas y coordenadas de la escala vertical se han tabulado para $\sigma = 1$. En la escala horizontal t es la medida *standard* de la desviación del valor observado respecto a la media.

Escala horizontal

t: 0 0,25 0,50 0,75 1,0 1,25 1,50 1,75 2,0 2,25 2,50 2,75 3,0

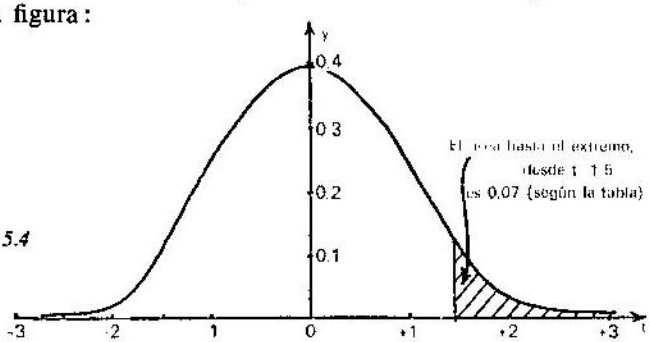
Escala vertical

y: 0,4 0,39 0,35 0,30 0,24 0,19 0,13 0,09 0,05 0,03 0,02 0,01 0,004

La escala vertical se ajusta para $\sigma = 1$ con el fin de que bajo la curva quede comprendida una área unidad. Esta escala no es nada importante por sí misma: todo el interés radica en la escala horizontal y en el área encerrada por la curva.

La gráfica es simétrica respecto al eje y (la recta $t = 0$ es el valor medio de la distribución), de forma que si trazas la gráfica para los valores de t comprendidos entre $t = -3$ y $t = 3$, llegarás a esta figura:

Fig. 5.4



t (medida de la desviación standard desde la media)	0	0,25	0,50	0,75	1,0	1,25

p (probabilidad de estar más desviado de la media que lo dicho) .	0,5	0,4	0,3	0,23	0,16	0,11	
t	1,5	1,75	2,00	2,25	2,50	2,75	3,00
p	0,07	0,04	0,02	0,01	0,006	0,003	0,001

Observa que esa probabilidad corresponde al área situada hacia la derecha del origen de coordenadas si $t > 0$, y hacia la izquierda si $t < 0$.

Por ejemplo, para $t = 1$, $p = 0{,}16$, de forma que la probabilidad de que una observación esté entre las verticales $t = 0$ y $t = +1$ es $0{,}5 - 0{,}16 = 0{,}34$, es decir, el 34 % de los valores se hallan en esta región; la probabilidad de que las observaciones estén dentro de una desviación standard de la media (es decir, de $t = -1$ a $+1$) es de 0,68 o, lo que es lo mismo, el 68 % de los valores se hallan en esta región. Asimismo, aproximadamente el 95 % de los valores están desviados de la media en menos de 2σ; y menos del 1 % está a más de 3 desviaciones standard. De hecho, el 99,73 % se halla entre $t = -3$ y $t = +3$, y el 99,994 % entre $t = -4$ y $t = +4$.

Uno de los trabajos más difíciles para el estadístico es hallar la fórmula que represente los resultados observados. En el caso presente de la distribución normal, esta fórmula es

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2\sigma^2}} \qquad \text{o} \qquad y = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

donde

$$t = \frac{x - \bar{x}}{\sigma} \qquad \text{y} \qquad \sigma = 1$$

(porque ésta es la unidad en nuestra escala).

Hallarás tablas de valores para e^{-x} en tus tablas de logaritmos.

El factor $\dfrac{1}{\sqrt{2\pi}}$ hubo de ponerse para que el área encerrada por la curva y los ejes fuera la unidad.

Ejemplos del uso de la distribución normal

(*a*) Necesitamos hallar en cada caso el valor medio de la distribución.

(*b*) También debemos hallar la desviación standard antes de que podamos utilizar la tabla de valores de la distribución normal. Esto se hace como sigue:

1. Se calcula el valor medio.
2. Se hallan las diferencias entre los valores observados y el valor medio.
3. Se calcula la suma de los cuadrados de esas diferencias.
4. Se divide esto por n (número de observaciones).
5. Se extrae la raíz cuadrada para obtener la desviación standard

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Ejemplo 1

Tomemos los resultados de los capítulos II y IV, cuando comentábamos el número de guisantes hallados en las vainas.

Núm. de guisantes x_i	*Frecuencia del suceso* F_i	$x_i F_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 F_i$ *
0	0	0	—4,46	19,89	0
1	3	3	—3,46	11,97	35,91
2	4	8	—2,46	6,05	24,20
3	5	15	—1,46	2,13	10,65
4	7	28	—0,46	0,21	1,47
5	8	40	+0,54	0,29	2,32
6	6	36	+1,54	2,37	14,22
7	4	28	+2,54	6,45	25,80
8	2	16	+3,54	12,53	25,06
9	0	0	+4,54	20,61	0
	39	174			

$$\sum_{i=1}^{39} (x_i - \bar{x})^2 \times F_i = 139,63$$

* Observa que es más fácil utilizar el producto $(x_i - x)^2 F_i$ que escribir F_i veces el sumando $(x_i - x)^2$, como dice la fórmula. *(N. del T.)*

media:

$$\bar{x} = \frac{174}{39} \simeq 4{,}46$$

y luego sucesivamente,

$$\frac{139{,}6}{39} \simeq 3{,}58 \qquad \sqrt{3{,}58} = 1{,}89$$

Por tanto, la desviación standard de esta distribución es de 1,89 guisantes. Si reunimos nuestros resultados obtenemos:

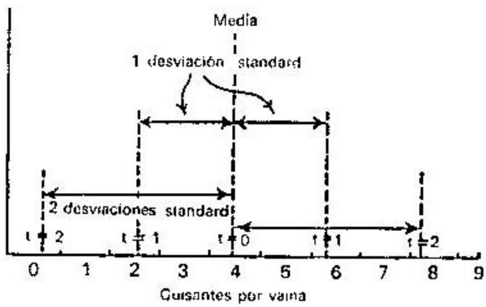


Fig. 5.5

t es la medida de la desviación standard de la distribución, en este caso 1,89.

Ahora debemos esperar que el 68 % de nuestros valores estarán comprendidos entre 2,57 y 6,35, cifras que limitan una desviación standard a cada lado de la media (ver pág. 80). Lo que encontraremos es 15 + 28 + 40 + 36 valores en esta región, o sea $\frac{119}{174}$ de los casos observados, o, lo que es lo mismo, el 68,39 % de las observaciones, lo que está muy de acuerdo con el resultado previsto.

Hallar el valor de la desviación standard es un proceso bastante pesado, pero que puede hacerse más ameno empleando una pequeña máquina de calcular.

Ejemplo 2

Un tren tiene su llegada a la estación más cercana a mi escuela a las 9 h 8 min de la mañana diariamente. Un muchacho que toma dicho tren anotó el número de minutos que llevaba de retraso cada día durante una quincena (12 días de clase):

5	8	10	15	2	1
—3	0	—2	0	5	7

Calcula la hora media de llegada del tren y la desviación standard.

Aceptando que las horas de llegada están normalmente distribuidas alrededor de esta media, calcula la probabilidad de que:

(*a*) El tren llegue con más de 6,5 minutos de retraso.

(*b*) El muchacho llegue puntual a la escuela (hay 8 minutos de camino desde la estación, y la escuela empieza a las 9.15 de la mañana).

(*c*) El tren llegue a la hora prevista en su horario o antes de dicha hora.

Hora media de llegada

$$= 9^h\ 8^m + \frac{\text{suma de minutos de retraso}}{12} =$$

$$= 9^h\ 8^m + \frac{48}{12} = 9^h\ 12^m$$

Para la desviación standard tenemos

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5	1	1	—3	—7	49
8	4	16	0	—4	16
10	6	36	—2	—6	36
15	11	121	0	—4	16
2	—2	4	5	1	1
1	—3	9	7	3	9
					314

Por tanto,

$$\sum_{i=1}^{i=12} (x_i - \bar{x})^2 = 314$$

De donde desviación standard $=\sqrt{\dfrac{314}{12}} \simeq 5$ minutos, con el grado de precisión que alcanzan nuestras observaciones.

(*a*) La hora media de llegada es con 4 minutos de retraso, y si el tren atrasa 6,5 minutos, la medida standard de la desviación de la media es

$$\frac{2{,}5}{5} = 0{,}5$$

Por tanto, según la tabla de áreas, para $t = 0{,}5$, la probabilidad de que lleve más de 6,5 minutos de retraso es 0,3.

(*b*) Si el muchacho ha de entrar a tiempo, debe llegar a la estación a las 9,7 de la mañana, esto es, 5 minutos antes de la hora media de llegada.

La medida standard de tal desviación es, por consiguiente,

$$-\frac{5}{5} = -1$$

Luego la probabilidad de llegar a tiempo vendrá dada por $p = 0{,}16$ (usando la tabla de la pág. 81 el valor a tomar para $t = -1$ es el mismo que para $t = +1$).

(*c*) Para que el tren llegue puntual o antes de tiempo, tenemos que consultar la tabla para $t = -\dfrac{4}{5}$ (el valor más cercano que tenemos es $t = 0{,}75$), de donde la probabilidad es aproximadamente 0,23.

Ejercicio 7

1. Se preguntó a una muestra de 100 personas a qué hora se acostaban. Éstas son sus respuestas:

Hora	*Personas*
8.30 de la noche	0
9	1
9.30	1
10	8
10.30	20
11	38
11.30	23
12	5
12.30	2
1	1
1.30	0
2	1

Halla la hora media de acostarse y la desviación standard. Si la distribución es normal y elegimos la muestra al azar, ¿cuántas personas en una ciudad de 40 000 habitantes será de esperar que

(*a*) estén acostadas a las 10 de la noche;
(*b*) se acostarán entre las 9.30 y las 11?

¿A qué hora estará acostado el 89 % de la población?

2. Una industria está fabricando zapatos destinados a la exportación. Se espera vender 20 000 pares a un país en el cual la longitud media del pie de una persona es 10 pulgadas * y la desviación standard 1 pulgada. Los zapatos se hacen en 11 tamaños, a saber: $7\frac{1}{2}$, 8, etc., hasta $12\frac{1}{2}$ pulgadas. ¿Cuántos pares de zapatos de cada tamaño deberán fabricarse y cuántas personas necesitarán zapatos especialmente hechos para ellas? (Aceptamos que si un zapato excede en más de $\frac{1}{2}$ pulgada del tamaño pedido, no puede utilizarse. Se admite que en toda persona ambos pies tienen la misma longitud.)

3. Un fabricante de cajas de cerillas sabe que el número de cerillas en sus cajas se ajusta a una distribución normal, con una

* 1 pulgada = 2,54 cm.

media de 43 cerillas por caja y una desviación standard de 2 cerillas. Si se venden diariamente 30 000 cajas,

(*a*) ¿Cuántas contendrán menos de 40 cerillas?
(*b*) ¿Cuántas contendrán entre 38 y 47 cerillas?
(*c*) ¿Cuántas contendrán menos de 46 cerillas?

4. Las pilas de linterna tienen una vida media de 80 horas, con una desviación standard de 2 horas. De una partida de 100, ¿cuántas puede esperarse que se consuman después de funcionar:

(*a*) entre 75 y 81 horas;
(*b*) menos de 77 horas?
(*c*) ¿Después de cuánto tiempo debemos esperar que se hayan agotado 70? (Emplea la distribución normal.)

5. Se sabe que la fruta de cierta granja madura usualmente el 13 de agosto, con una desviación standard de una semana (la maduración se supone que obedece a la ley de distribución normal). Se tiene que recoger el día que madura, porque de lo contrario se estropea. El rendimiento anual se estima en 10 000 kg. Calcula:

(*a*) Cuántos kg de fruta perdería el granjero si sólo pudiera conseguir operarios que la recogieran durante tres semanas.
(*b*) Cuántos kg perdería si detiene la recogida y se va de vacaciones del 20 de agosto al 30 de septiembre.
(*c*) Si cada operario puede recoger 160 kg diarios, el número máximo de obreros que se necesitan durante las semanas

(1) del 30 de julio al 6 de agosto
(2) del 27 de agosto al 3 de septiembre.

Sugerencia: Dibuja un diagrama a escala de la distribución normal y halla las áreas aproximadas comprendidas entre $t = -1{,}0$ y $-1\frac{1}{7}$ y entre $t = 2$ y $2\frac{1}{7}$.

Otras distribuciones

La segunda distribución en orden de importancia es la *distribución binomial*, a la cual dedicamos el siguiente capítulo.

Existen otros varios tipos de distribución, uno de los cuales ya encontramos al estudiar los resultados de fútbol. Recordarás que esto nos dio una distribución oblicua; el histograma era así:

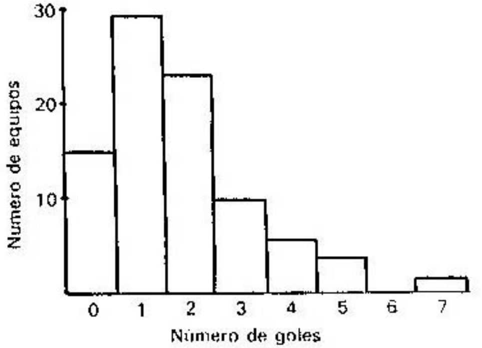


Fig. 5.6

Esta distribución oblicua se aproxima a la distribución de frecuencias de Poisson. De acuerdo con la teoría de Poisson, el número probable de equipos que marcan

0 goles es $n\, e^{-\bar{x}}$

1 gol es $n\, e^{-\bar{x}} \cdot \bar{x}$

2 goles es $n\, e^{-\bar{x}} \cdot \dfrac{\bar{x}^2}{2!}$

3 goles es $n\, e^{-\bar{x}} \cdot \dfrac{\bar{x}^3}{3!}$

4 goles es $n\, e^{-\bar{x}} \cdot \dfrac{\bar{x}^4}{4!}$

y así sucesivamente, donde n es el número de equipos, $\bar{x}$ es el número medio de goles marcados por equipo por partido y, por ejemplo, 4! (léase factorial de 4) significa $4 \times 3 \times 2 \times 1$ y 3! es $3 \times 2 \times 1$.

Puedes encontrar tablas de los valores de $e^{-\bar{x}}$ en muchas tablas de logaritmos.

Si comparamos los resultados predichos con los obtenidos, tenemos:

Número de goles	*Número de equipos*	*Núm. predicho de equipos (aprox.)*
0	14	14
1	28	25
2	22	22
3	9	13
4	5	6
5	3	2
6	0	$\frac{1}{2}$
7	1	$\frac{1}{6}$

lo que demuestra que los dos concuerdan muy bien. Hay algunos puntos interesantes a propósito de los resultados de fútbol.

1. El tanteo medio parece variar muy poco de unos 1,8 goles por partido, a pesar de factores tales como el tiempo, los trofeos en juego, etc. (Esto es lo más importante, porque sólo podemos hacer predicciones con la distribución de frecuencia de Poisson si $\bar{x}$ es constante.)
2. La probabilidad de que un equipo marque o encaje 7 goles es muy pequeña. Si a un equipo le marcan 7 goles, inmediatamente adquirirá un nuevo jugador para rectificar su mal juego.

Proyecto

Si estás en la temporada de Liga, haz una tabla de los resultados reales y de los predichos por la fórmula de Poisson, como acabamos de ver. Mira si ajustan los resultados. En caso contrario, ¿por qué no ajustan? ¿Hay equipos que tengan alto tanteo en contra? Si es así, ¿se ha hecho algo para rectificar esto?

Si no es la temporada de fútbol, usa la distribución de Poisson para predecir el número de equipos que marquen 0, 1, 2, ..., 6 goles si un día determinado hubieran jugado 84 equipos y se hubieran marcado 154 goles.

Compara tus respuestas con la siguiente relación de resultados con esas condiciones (en un día de la temporada 1962-63):

Núm. de goles . . .	0	1	2	3	4	5	6
Núm. de equipos . .	18	25	12	17	8	2	2

Otro grupo de valores que se aproxima a la distribución de Poisson es el de los ingresos personales de una nación. Por ejemplo (imaginario), contando en pesetas por semana, resultará algo de este tipo:

Pesetas por semana	*Miles de personas*
0 - 500	16 385
500 - 1 000	2 830
1 000 - 1 500	440
1 500 - 2 000	160
2 000 - 3 000	122
3 000 - 5 000	70
5 000 - 10 000	32
10 000 - 20 000	9
Más de 20 000	2

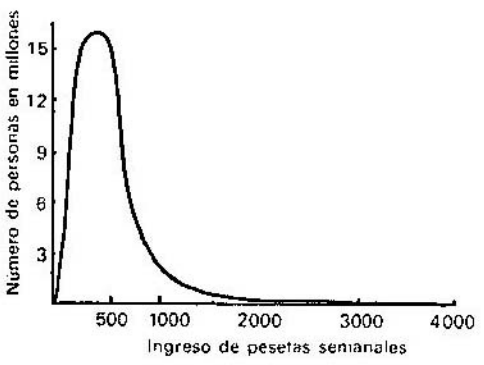


Fig. 5.7

Otra distribución que es oblicua en sentido opuesto a la distribución de Poisson es la del número de fallecimientos en relación a la edad.

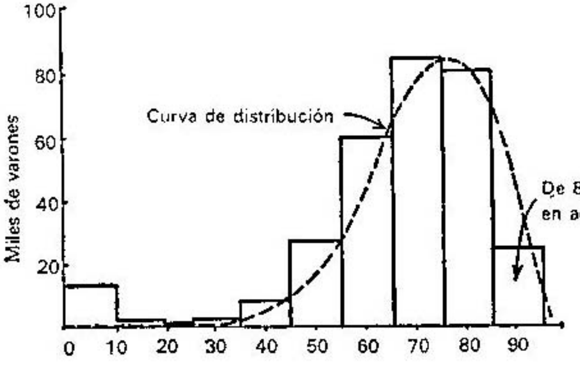


Fig. 5.8

(Tomada de *Resumen Anual de Estadísticas* (1960) para varones en Inglaterra.)

Aquí tenemos un tipo de distribución todavía más raro en la naturaleza, el tipo U. Ésta fue la media del grado de nubosidad en Greenwich en el mes de julio, a lo largo de 14 años.

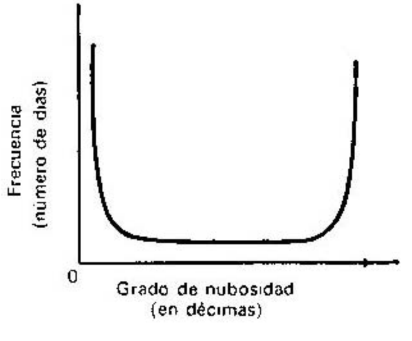


Fig. 5.9

Ya hemos estudiado las distribuciones que aparecen en la naturaleza. Hay otras, como la *distribución triangular*, que se dan en el juego.

Si tiramos dos dados, el resultado final da una distribución así:

Resultado con el primer dado	*Totales posibles*										
1	2	3	4	5	6	7					
2		3	4	5	6	7	8				
3			4	5	6	7	8	9			
4				5	6	7	8	9	10		
5					6	7	8	9	10	11	
6						7	8	9	10	11	12

y la distribución puede ordenarse como sigue

					7					
				6	7	8				
			5	6	7	8	9			
		4	5	6	7	8	9	10		
	3	4	5	6	7	8	9	10	11	
2	3	4	5	6	7	8	9	10	11	12

Con una distribución así es fácil trabajar. El valor medio se ve que es 7, y la desviación standard puede calcularse con prontitud. La probabilidad de obtener determinado total se puede leer también. Por ejemplo, el número total de sumas es 36; hay 4 nueves, y por tanto una probabilidad de $\frac{4}{36}$, o sea $\frac{1}{9}$, de conseguir ese total.

Supongamos que tenemos tres dados, dos de cuyas caras dan el valor 1, dos el valor 2 y dos el valor 3.

Entonces el tanteo total viene dado por:

Tanteo con el primer dado	*Totales posibles*						
1	3	4	5	6	7		
		4	5	6			
			5				
2		4	5	6	7	8	
			5	6	7		
				6			
3			5	6	7	8	9
				6	7	8	
					7		

y de la misma forma para los otros 1, 2, 3 en el primer dado.

De nuevo la distribución es de forma casi triangular y, usando la mitad superior de los valores, puede ordenarse en la siguiente forma:

			6			
		5	6	7		
		5	6	7		
		5	6	7		
	4	5	6	7	8	
	4	5	6	7	8	
3	4	5	6	7	8	9

Un tercer ejemplo similar es el de dos dados cuyas caras tienen los valores (como en el caso anterior) 1, 2, 3, 1, 2, 3. Tenemos:

Resultado con el primer dado	*Totales posibles*				
1	2	3	4		
	2	3	4		
2		3	4	5	
		3	4	5	
3			4	5	6
			4	5	6

y análogamente para los otros 1, 2, 3 del primer dado.

Esto nos da (empleando como antes la mitad de los resultados):

		4		
		4		
	3	4	5	
	3	4	5	
2	3	4	5	6

Si tiramos sólo un dado tenemos la distribución rectangular básica para las probabilidades de sacar un 1, 2, 3, 4, 5 o 6, y ésta es de $\frac{1}{6}$ para cada uno (si el dado está bien hecho).

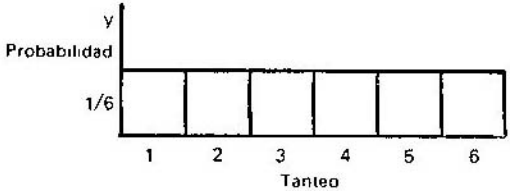


Fig. 5.10

Fácilmente pueden idearse otros tipos similares de distribución.

EJERCICIO 8

1. Calcula la media y la desviación standard de la primera distribución triangular del texto (pág. 92).

2. Repite lo anterior para el segundo ejemplo (pág. 93).

3. Empleando otra vez el segundo ejemplo (pág. 93), ¿cuáles son las probabilidades de sacar un total de: (*a*) 4, (*b*) 6, (*c*) 7, (*d*) más de 5?

4. Encuentra la distribución de los totales de dos dados cuando la numeración de uno es 1, 2, 3, 4, 5, 1 y la del otro 1, 2, 3, 1, 2, 3. ¿Cuál es la probabilidad de sacar: (*a*) 4 o más, (*b*) 3, (*c*) menos de 6?

5. En un juego de cricket de bolsillo se juega con dos pirindolas de metal. Una va marcada con los tantos 1, 2, 3, 4, 4, 6, mientras que la otra va marcada 1, 2, 3, 1, FUERA, FUERA. Halla las distribuciones de los tanteos totales. (FUERA anula el tanteo de la otra pieza, por ejemplo, 2 FUERA quiere decir tanteo 0.) ¿Cuáles son las probabilidades de

(*a*) salirse;
(*b*) sacar 5 o más;
(*c*) sacar exactamente 6?

Dispersión

Hay otras dos formas de medir la *dispersión* (o extensión) de los valores en una distribución, aunque ninguna de las dos es tan importante ni tan útil como la desviación standard. Estas medidas son:

1. La desviación mediana. 2. La cuartila.

Desviación mediana

Para hallar la desviación mediana debemos de nuevo, en primer lugar,

(*a*) calcular la media;
(*b*) hallar la diferencia entre la media y los valores observados;
(*c*) sumar el valor absoluto de todos los resultados de (*b*), es decir, considerar todas esas diferencias como positivas;
(*d*) dividir por *n*, para hallar la desviación mediana

$$\text{Desviación mediana} = \frac{\sum_{i=1}^{i=n} |x_i - \bar{x}|}{n}$$

Ejemplo

Consideremos una vez más los resultados de contar los «guisantes por vaina».

N.º de guisantes x_i	*Frecuencia* F_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert \times F_i$
0	0	—4,46	4,46	0
1	3	—3,46	3,46	10,38
2	4	—2,46	2,46	9,84
3	5	—1,46	1,46	7,30
4	7	—0,46	0,46	3,22
5	8	+0,54	0,54	4,32
6	6	1,54	1,54	9,24
7	4	2,54	2,54	10,16
8	2	3,54	3,54	7,08
9	0	4,54	4,54	0

$$\sum_{i=1}^{39} |x_i - \bar{x}| \times F_i = 61{,}54$$

La desviación mediana es $\dfrac{61{,}54}{39} = 1{,}58$.

Cuartilas

Como vimos en el capítulo III, la mediana es un valor que tiene la mitad de los valores observados sobre él y la otra mitad por debajo de él. Podemos también dividir la fila ordenada de valores en cuatro «cuartilas», conteniendo cada una un 25 % de los valores.

Consideremos el mismo ejemplo.

La mediana es 5 guisantes por vaina.

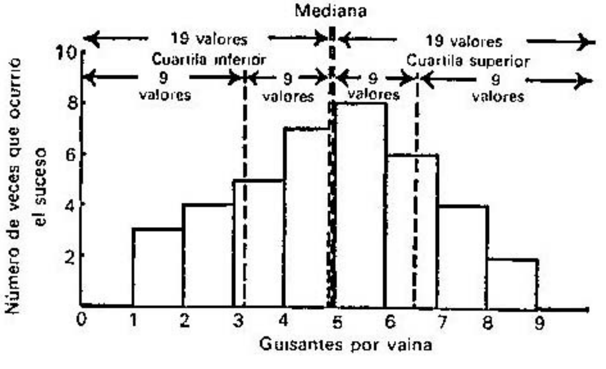


Fig. 5.11

Esta medida de la dispersión se utiliza cuando quedan clases indeterminadas a uno o a ambos extremos de nuestra clasificación (ver tabla de fallecimientos, en uno de cuyos extremos se ha escrito «de 85 en adelante»). Naturalmente, en estos casos es imposible hacer ningún cálculo preciso de la dispersión, pero es posible usar cuartilas, que proporcionan una medida aproximada de ella.

CAPITULO VI

Distribución binomial, muestreo

Distribución binomial

Imagina que tenemos una urna que contiene *un gran numero* de bolas, idénticas excepto en lo referente al color. Sean negras el 20 % y blancas el resto. Entonces es evidente que la probabilidad de que una bola extraída al azar de la urna sea negra será $p = \frac{1}{5}$.

Luego la probabilidad de extraer dos negras seguidas será $\frac{1}{5} \times \frac{1}{5}$, la de tres negras seguidas $= (\frac{1}{5})^3$, y, en general, para n bolas negras la probabilidad será $(\frac{1}{5})^n$.

Análogamente, la probabilidad de extraer n bolas blancas del conjunto es $(4/5)^n$.

Esto es muy fácil, pero en la práctica lo probable es que las bolas extraídas no sean todas del mismo color.

¿Cuáles son, pues, las probabilidades de obtener 0, 1, 2, 3, ... bolas negras en un conjunto de n elegidas al azar (una *muestra* de extensión n).

Este problema es de gran importancia en la industria, donde las mercancías son muestreadas, inspeccionadas y, finalmente, clasificadas bien como aceptables, bien como defectuosas.

Supongamos que de entre un gran número de objetos, el 20 % de los cuales son defectuosos, elegimos una muestra de 2. (Debemos tener *muchos* para elegir; de lo contrario sería posible comprobar con cada uno de ellos.)

Entonces, la probabilidad de que ambos sean defectuosos es $\frac{1}{5} \times \frac{1}{5} = 0{,}04$.

La probabilidad de que ninguno de los dos sea defectuoso es $\frac{4}{5} \times \frac{4}{5} = 0{,}64$.

Por lo tanto, la probabilidad de que uno sea defectuoso y uno aceptable es $1 - 0{,}04 - 0{,}64 = 0{,}32$.

Como has visto (cap. IV), estas probabilidades pueden considerarse como términos del desarrollo $(p + q)^2 = p^2 + 2\,p\,q + q^2$ o, en nuestro caso, poniendo $p = \frac{1}{5}$, $q = \frac{4}{5}$

$$\begin{array}{lcccccccc} \left(\frac{1}{5} + \frac{4}{5}\right)^2 & = & \frac{1}{25} & + & 2\left(\frac{1}{5} \times \frac{4}{5}\right) & + & \frac{16}{25} \\ & = & 0{,}04 & + & 0{,}32 & + & 0{,}64 \\ & & \uparrow & & \uparrow & & \uparrow \\ \text{probabilidad de:} & & \text{2 defect.} & & \text{1 defectuoso} & & \text{2 acept.} \\ & & & & \text{1 aceptable} & & \end{array}$$

Análogamente, si elegimos tres objetos estudiaremos los términos de la suma

$$\begin{array}{lccc} \left(\frac{1}{5} + \frac{4}{5}\right)^3 = & \left(\frac{1}{5}\right)^3 & + 3\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right) + & \\ \text{probabil. de:} & \text{3 defect.} & \text{2 defect. y} & \\ & & \text{1 acept.} & \end{array}$$

$$\begin{array}{cc} + 3\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^2 & + \left(\frac{4}{5}\right)^3 \\ \text{1 defect. y} & \text{3 acept.} \\ \text{2 acept.} & \end{array}$$

Lo mismo ocurre para 4 objetos, 5, etc.

Esto significa que, en general, en una muestra de n objetos en la que cada uno tiene una probabilidad p de ser aceptado y q de ser rechazado, el número de defectuosos puede preverse a partir de los términos del desarrollo de $(p + q)^n$, usando el triángulo de Pascal o el teorema del binomio (cap. IV).

Estudiemos un poco más a fondo el teorema del binomio. Consideremos $(p + q)^5$ (o sea tomando una muestra de 5 artículos):

$$p^5 + 5\,p^4\,q + 10\,p^3\,q^2 + 10\,p^2\,q^3 + 5\,p\,q^4 + q^5$$

Si $p = q = \dfrac{1}{2}$ (es decir, que los artículos tienen la misma probabilidad de ser aceptados o rechazados), podemos dibujar el histograma: es simétrico como aparece aquí:

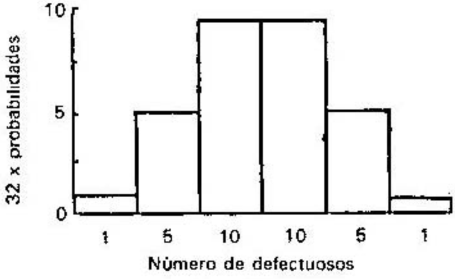


Fig. 6.1

Por ejemplo, podemos leer que la probabilidad de 4 defectuosos es $\dfrac{5}{32}$.

Supongamos ahora que tomamos una muestra de 20 artículos; el histograma será:

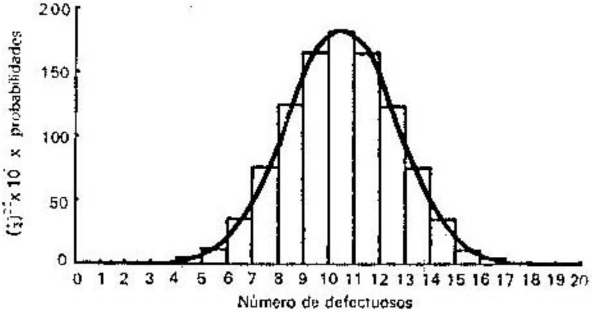


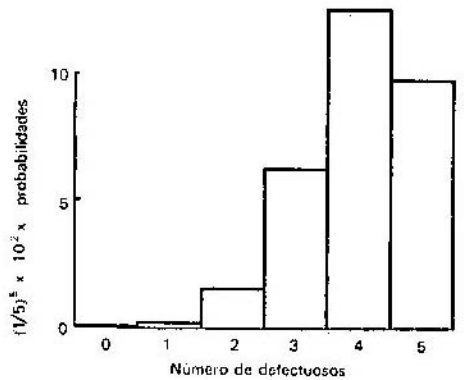
Fig. 6.2

N.º	$(\frac{1}{2})^{20} \times p$
0	1
1	20
2	190
3	1,140
4	4,845
5	15,504
6	38,760
7	77,520
8	125,970
9	167,960
10	184,756
11	167,960
	etc.

Como podemos leer aquí, la probabilidad de que haya precisamente 8 defectuosos, por ejemplo, es: $\left(\frac{1}{2}\right)^{20} \times 10^3 \times 125\,970$.

Como puedes ver, parece ser esto una buena aproximación del caso de la distribución normal. De hecho, si n es mayor que 50, la aproximación entre la distribución normal y la binomial es excelente.

Hasta ahora sólo hemos visto histogramas en los que $p = q = \frac{1}{2}$. Veamos lo que ocurre cuando $p = \frac{1}{5}$, $q = \frac{4}{5}$. Empleando los coeficientes del desarrollo de $(p + q)^5$, la distribución es



N.º	$(\frac{1}{5})^5 \times p$
0	1
1	20
2	160
3	640
4	1280
5	1024

Fig. 6.3

De nuevo podemos leer que la probabilidad de 4 defectuosos es

$$\left(\frac{1}{5}\right)^5 \times 10^2 \times 1280$$

Esta distribución es muy oblicua.

Veamos ahora lo que ocurre si tomamos $n = 20$. Entonces $(p + q)^{20}$ vendrá representado así:

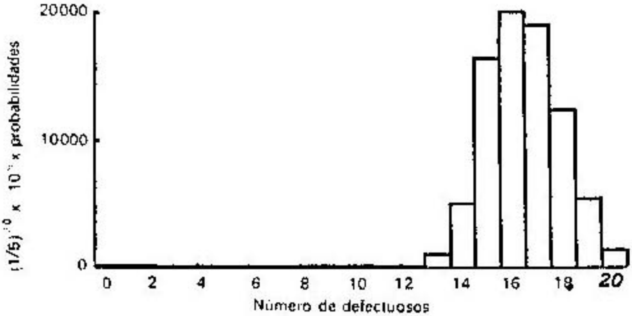


Fig. 6.4

y las probabilidades pueden leerse como antes.

Puedes ver que esta distribución es todavía muy oblicua, pero la «joroba» no es tan asimétrica como en el ejemplo anterior. Afortunadamente, si tomamos n mayor que 50 importa poco lo desiguales que sean p y q (recuerda que $p + q = 1$); de nuevo la curva de distribución se aproxima mucho a una curva de distribución normal.

Ejemplos

1. *Considera un experimento en el cual se tiran 64 veces 6 monedas juntas y los resultados están exactamente de acuerdo con el teorema del binomio, o sea*

Número de caras	0	1	2	3	4	5	6
Frecuencia del suceso anterior . .	1	6	15	20	15	6	1

Calcula: (a) el número medio de caras que salen; (b) la desviación standard del número de caras.

Después de haber calculado estas dos cosas, ¿puedes ver un método más fácil de llegar a esos resultados?

(a) El número promedio de caras es

$$\text{Núm. de monedas} \times \text{probabilidad de cara} = 6 \times \frac{1}{2} = 3.$$

(*b*) La desviación standard es

$\sqrt{\text{Núm. de monedas} \times \text{probabil. de cara} \times \text{probabil. de cruz}} =$

$$= \sqrt{6 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{\frac{3}{2}}$$

Es muy fácil trabajar con la distribución binomial, ya que, si suponemos como antes que tenemos una muestra de n objetos y la probabilidad de que uno de ellos sea defectuoso es p, y q la de que sea aceptable, entonces el valor medio es $\bar{x} = n\,q =$ promedio de artículos aceptables en la muestra; y la desviación standard es $\sigma = \sqrt{n\,p\,q}$.

2. *Supongamos que tiramos 500 veces una moneda. Cada vez que se haga esta experiencia, podemos esperar* $500 \times \frac{1}{2} = 250$ *caras, con una desviación standard de la media igual a*

$$\sqrt{500 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{125} = 11{,}2$$

u 11 *para facilitar el trabajo.*

En otras palabras (usando la curva de distribución normal): debemos esperar que en el 68 % de tales experimentos resulten entre 239 y 261 caras, y en el 95 % de experiencias entre 228 y 272. Si el resultado que obtuviéramos fuese muy diferente de éste, significaría que una de dos: la moneda tiene trampa, o el experimento no se ha realizado correctamente.

3. *Dos restaurantes tratan de atraerse a 100 clientes ofreciendo buenas comidas a la misma hora. Si una persona es tan libre de elegir uno como otro, ¿qué número de plazas debería proporcionar el restaurante si quiere estar seguro de que su capacidad es suficiente el 99 % de las veces?*

$$\text{Número medio de personas} = \frac{1}{2} \times 100 = 50$$

$$\text{La desviación standard} = \sqrt{100 \times \frac{1}{2} \times \frac{1}{2}} = 5$$

Por tanto, utilizando la curva de distribución normal, necesitamos aproximadamente 3×5 asientos sobre la media, o sea 65 asientos.

EJERCICIO 9

1. Supongamos que tiramos una moneda 100 veces. ¿Cuál es el número de caras esperado, y entre qué límites podemos esperar que esté el número de caras el 95 % de las veces?

2. Si se lanza 50 veces un dado corriente, ¿cuántas veces esperamos que salga un 3? ¿Cuál es la desviación standard del número de salidas del 3? ¿Cuál es la probabilidad de que el 3 salga exactamente

(*a*) 25 veces
(*b*) 20 veces?

3. Se sabe que un producto de limpieza es efectivo en el 40 % de los casos en que se usa. Se le añade un nuevo ingrediente y, de las 500 veces siguientes en que se utiliza, es efectivo 250. ¿Qué pude decirse sobre la eficacia del nuevo componente?

4. Aceptemos que el 5 % de la población de cierta ciudad padece cáncer. Si se elige al azar una muestra de 200 personas, ¿cuántas debemos esperar que tengan cáncer? ¿Cuál es la desviación standard? Si ninguna de las personas investigadas lo padecía, ¿significaría esto que nuestra suposición era falsa?

5. Durante la guerra, supongamos que cada bomba tenía probabilidad $\frac{1}{10}$ de alcanzar su objetivo. En una noche en que se dejaron caer 5000 bombas, ¿cuántos objetivos debemos esperar que se alcanzaron? En esa noche se alcanzaron realmente 600, ¿fue un raid excepcional? La noche siguiente se lanzaron 10 000 bombas y sólo se alcanzaron 800 objetivos; ¿fue un raid fracasado? ¿Podemos comentar algo sobre nuestra suposición de partida?

Muestreo

¿Por qué se inspeccionan los artículos manufacturados? Las razones son sencillas:

1. Determinar la calidad de los artículos y, si es posible, retirar los defectuosos.
2. Proteger al consumidor y la reputación del fabricante. La verdadera finalidad de la inspección es dejar una proporción tan pequeña como sea posible de artículos defectuosos en las mercancías que se ofrecen al consumidor.

¿Por qué se toman muestras?

1. Como quiera que sería antieconómico comprobar todos los artículos, tenemos que tomar muestras. En una fábrica de cojinetes de bolas supondría malgastar el tiempo y sería muy costoso comprobar tamaño, peso, etc. de cada cojinete.
2. Algunas comprobaciones son destructivas por su naturaleza, por ejemplo hallar la vida de las bombillas eléctricas, resistencias, válvulas, pilas o alimentos. Si comprobáramos cada artículo de este tipo no quedaría nada para vender.

Supongamos que se toma una muestra de 5 artículos de entre un gran número de ellos, de los cuales se sabe que el 10 % son defectuosos. Entonces la probabilidad de obtener 0, 1, 2, 3, 4, 5 artículos defectuosos en esta muestra viene dada por los términos de la suma

$$\left(\frac{1}{10}+\frac{9}{10}\right)^5 = \left(\frac{1}{10}\right)^5 + 5\left(\frac{1}{10}\right)^4\left(\frac{9}{10}\right) + 10\left(\frac{1}{10}\right)^3\left(\frac{9}{10}\right)^2 +$$

$$+ 10\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^3 + 5\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^4 + \left(\frac{9}{10}\right)^5$$

Multiplicando ambos miembros por 10^5,

$$10^5\left(\frac{1}{10}+\frac{9}{10}\right)^5 = 1 + 45 + 810 + 7\,290 + 32\,805 + 59\,049$$

podemos ver que si elegimos 100 000 muestras de cinco artículos podemos esperar obtener

1	con	5	defectuosos
45	con	4	defectuosos
810	con	3	defectuosos
7 290	con	2	defectuosos
32 805	con	1	defectuoso
59 049	con	0	defectuosos

lo cual significaría que

59,049 %	de las muestras sugieren que de los artículos de nuestra manufactura hay	0	defectuosos
32,805 %	ídem	20 %	defectuosos
7,290 %	ídem	40 %	defectuosos
0,810 %	ídem	60 %	defectuosos
0,045 %	ídem	80 %	defectuosos
0,001 %	ídem	100 %	defectuosos

Por otra parte, supongamos que se han tomado muestras de 5 de un lote que contiene el 20 % de defectuosos. Igual que antes, podemos obtener los términos de la serie por

$$5^5\left(\frac{1}{5}+\frac{4}{5}\right)^5 = 1 + 20 + 160 + 640 + 1280 + 1024$$

Ahora (análogamente) podemos calcular que

32,768 %	de las muestras sugieren que en cada cinco productos manufacturados hay	0	defectuosos
40,960 %	ídem	1	defectuoso
20,480 %	ídem	2	defectuosos
5,120 %	ídem	3	defectuosos
0,640 %	ídem	4	defectuosos
0,032 %	ídem	5	defectuosos

Hemos estado estudiando este problema sabiendo el porcentaje global de defectuosos en los artículos manufacturados. De hecho, el problema es usualmente más difícil, puesto que no conocemos el porcentaje de los que se han fabricado deficientemente. Este obstáculo se supera por medio de esquemas de doble muestreo y otros métodos más complicados, dependientes a menudo de la forma de fabricación.

Error standard de la media

Si examinamos una vez más la distribución del número de guisantes por vaina

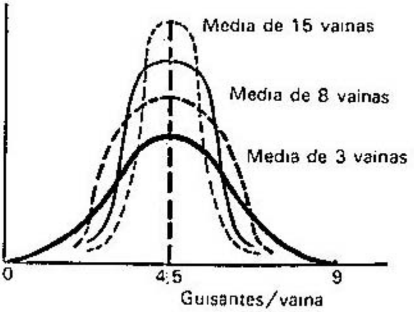


Fig. 6.5

y consideramos las medias de muestras de 3, 8 y 15 vainas, hallamos que la distribución de las medias es también normal alrededor de la media que habíamos encontrado ya (las curvas son sólo aproximadas para dar una idea).

La desviación standard de las medias respecto a la media global, obviamente decrece con la extensión de la muestra. En realidad, el error standard de la media de extensión n (que es la desviación standard de la media de la muestra respecto a la media verdadera o global) es $\frac{\sigma}{\sqrt{n}}$, donde σ es la desviación standard del conjunto total y n el número de artículos de nuestra muestra. Como es imposible medir σ, se suele calcular σ_n (desviación standard de la muestra mayor), y luego el error standard de la media se toma en la práctica $\frac{\sigma_n}{\sqrt{n}}$.

Ejemplo

Supongamos que en las revisiones médicas del ejército se han medido las alturas de 400 hombres, y se encuentra que la distribución tiene un valor medio de 168 cm, con una desviación standard de 6,30 cm. ¿Cuál sería la media verdadera de todos los hombres?

$$\text{Entonces, error standard} = \frac{6{,}30}{\sqrt{400}} = 0{,}315.$$

Se sabe que muy probablemente la media verdadera no distará más de 2 desviaciones standard de 168 cm, es decir, no más de $2 \times 0{,}315$ cm de 168 cm. Por tanto, (con 95 % de seguridad), la media del total está comprendida entre 167,37 cm y 168,63 cm; y es casi seguro (99 % de seguridad) que la desviación real no alcanzará a 3 veces la standard.

EJERCICIO 10

1. ¿Entre qué límites (con certeza del 95 %) debe estar comprendido el verdadero número medio de guisantes por vaina, utilizando los datos de la página 27?

2. Se halla que cierta marca de transistores tiene una vida media de 2000 horas, con una desviación standard de 5 horas. ¿Cuál sería la desviación standard de la vida media de muestras de 25 transistores?

3. Se toman muestras de la longitud de 16 artículos de la cadena de una producción en serie. Se halla que la desviación standard de la longitud de los artículos es 0,10 cm. ¿En cuánto calculas la desviación standard global para la longitud?

Tests de significación

Un matrimonio amigo nuestro tiene cinco hijos, y los cinco son niñas. Desea grandemente tener un niño, y le inquieta la duda de que ello no sea en su caso tan probable como es normal, pues se pregunta si el haber tenido cinco veces niña es una prueba significativa de que para ellos la probabilidad de niño o niña no sea la misma.

Para contestar, utilizaremos el método llamado de *hipótesis nula.* En este caso es lo razonable aceptar inicialmente que es igualmente probable tener un niño o una niña. Así, usando la distribución binomial

$$\left(\frac{1}{2}+\frac{1}{2}\right)^5=\left(\frac{1}{2}\right)^5+5\left(\frac{1}{2}\right)^5+10\left(\frac{1}{2}\right)^5+$$
$$+10\left(\frac{1}{2}\right)^5+5\left(\frac{1}{2}\right)^5+\left(\frac{1}{2}\right)^5$$

la probabilidad de tener 5 niñas es

$$\left(\frac{1}{2}\right)^5=\frac{1}{32}=3{,}1\ \%$$

Por tanto, la probabilidad de tener 5 niñas o 5 niños es

$$3{,}1\ \%+3{,}1\ \%=6{,}2\ \%$$

¿Podemos decir que nuestra hipótesis nula es correcta, o bien que ha ocurrido un suceso demasiado improbable? ¿Es significativo este 6,2 % o no?

Hemos de fijar algún patrón para la significación. Se acepta usualmente que si el suceso tiene una probabilidad de ocurrir de menos del 5 %, entonces es significativo, o sea que nuestra hipótesis nula es incorrecta y debe ser reajustada. El anterior resultado, 6,2 %, no era significativo al nivel 5 %.

La probabilidad $p = 0{,}05$ se llama *límite de aviso* y $p = 0{,}01$ es el llamado *límite de acción.*

Observa que debemos considerar la probabilidad del suceso actual observado junto con otros sucesos igualmente extremos o más extremos todavía que pudieran haber ocurrido. (Por esto es por lo que consideramos los 5 niños al igual que las 5 niñas).

Los niveles de significación son puramente arbitrarios, y si estuviéramos experimentando, por ejemplo, una droga peligrosa, exigiríamos una probabilidad mucho más pequeña, quizá 0,001.

Tests de 1 y de 2 extremos

Como hemos visto, en el ejemplo anterior tuvimos que considerar la probabilidad de 5 niños tanto como el probable suceso de 5 niñas en la familia. Éste sería un test significativo de «dos extremos». En algunos casos, por ejemplo el de la fabricación de una droga peligrosa, sólo estamos interesados en «1 extremo», en el que los ingredientes son más potentes que el nivel de peligro. Realmente no nos interesa lo débil que pueda ser la droga; sólo lo fuerte. En esta ocasión usaríamos un test significativo de «un extremo».

EJERCICIO 11

1. Al tirar un dado 12 veces, hallamos que el 6 salió en 5 ocasiones. ¿Era un dado sin trampa?

2. En relación con los resultados de fútbol de la página 28, halla las probabilidades de que a tu equipo le marquen sucesivamente 6 y 7 goles. ¿Sería esto significativo?

3. Un hombre dice que para él es indiferentes tomar el autobús o el tren para ir a su trabajo. En una quincena (14 días) observas que toma el autobús sólo 5 veces. ¿Es esto significativo?

4. En cierta fábrica las latas deben, de acuerdo con la ley, contener por lo menos 640 g de conserva. La máquina envasadora varía la cantidad que vierte en la lata en distribución normal, con una desviación standard de 10 g. ¿A qué peso de suministro debe colocarse la máquina de manera que haya menos del 1 % de probabilidad de vender latas faltas de peso? En una muestra de 20 latas, a una le faltaba peso. ¿Debe repararse la máquina?

5. En un distrito electoral, en las últimas elecciones, el 60 % votó a los conservadores y el resto a los laboristas. Se preguntó después, al azar, a una muestra de 60 personas y se hallaron 28 conservadores y 32 laboristas. ¿Es significativo este resultado?

CAPITULO VII

Líneas de regresión y correlación

Líneas de regresión

Hasta ahora hemos estudiado problemas en los cuales interviene la medida de un solo carácter variable: por ejemplo, los equipos que marcaron un cierto número de goles, o el número de guisantes en cada vaina. En este capítulo nos interesamos por experimentos en los que se hacen dos medidas, como por ejemplo relacionar la longitud y temperatura de una varilla de cobre. Haremos nuestro experimento con toda la exactitud posible y dibujaremos una gráfica de los resultados.

Dichos resultados aparecerán probablemente así:

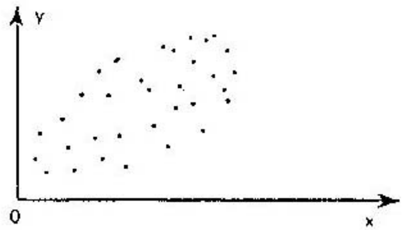


Fig. 7.1

El problema es hallar cómo se relacionan ambas cantidades y, si existe una relación lineal entre ellas, hallarla.

Si dibujamos nuestra gráfica en papel cuadriculado podemos hacer una de estas dos cosas: o bien tomar fajas verticales iguales y hallar la media de los puntos en cada faja, o bien hacer esto mismo en fajas análogas horizontales. El punto medio está situado hacia el centro de cada tira.

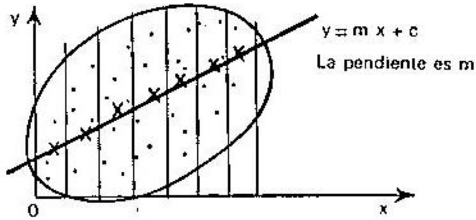


Fig. 7.2

La línea recta que se ajusta mejor a los puntos medios obtenidos se llama aquí recta de regresión de y sobre x. Daremos una definición de este «mejor» más adelante.

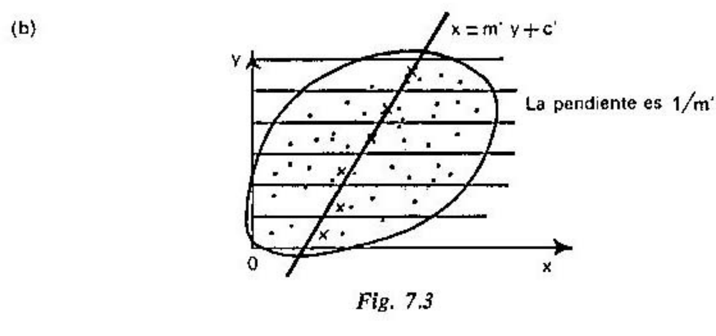


Fig. 7.3

La línea recta que se ajusta mejor a los puntos medios se llama aquí recta de regresión de x sobre y.

Observa que normalmente estas dos rectas son distintas. Éste es sólo un método aproximado de hallar dichas líneas. Ahora consideremos el problema matemáticamente.

Antes imaginemos, para simplificar, que los resultados de un experimento fueron

x:	1	2	3	4	5
y:	1	2,5	2,75	3,0	3,5

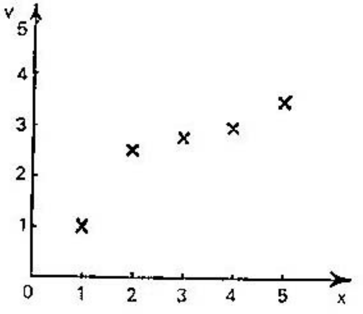


Fig. 7.4

El problema es dibujar una línea recta que represente estos datos con la mayor precisión posible, de forma que en la gráfica lineal podamos hacer posteriores lecturas de y para cualquier valor intermedio de x. Esto se hace por el método de mínimos cuadrados, inventado por el matemático Gauss.

La idea es sencilla. Dibujemos una recta como supuesta solución, y llamemos d_1, d_2, etc. a las distancias verticales de cada punto a esa línea.

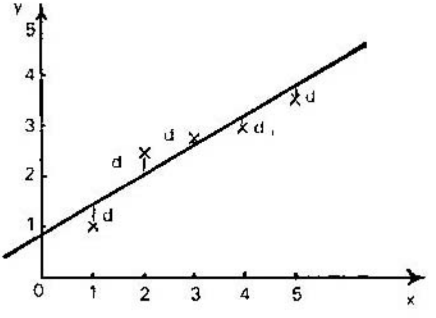


Fig. 7.5

El problema ahora es hacer $(d_1^2 + d_2^2 + \ldots + d_5^2)$ tan pequeño como sea posible o, lo que es lo mismo, hallar el mínimo valor de $\sum_{i=1}^{5} d_i^2$ *.

Hay dos formas de hacer esto. La primera consiste en seguir exactamente la idea expuesta arriba. Supongamos que la ecuación de la recta que deseamos hallar sea $y = mx + c$. Entonces

$$\begin{aligned}
d_1 &= m + c - 1 \\
d_2 &= 2m + c - 2{,}5 \\
d_3 &= 3m + c - 2{,}75 \\
d_4 &= 4m + c - 3 \\
d_5 &= 5m + c - 3{,}5
\end{aligned}$$

y

$$\begin{aligned}
d_1^2 &= m^2 + 2mc + c^2 - 2m - 2c + 1 \\
d_2^2 &= 4m^2 + 4mc + c^2 - 10m - 5c + 6{,}25 \\
d_3^2 &= 9m^2 + 6mc + c^2 - 16{,}5m - 5{,}5c + 7{,}5625 \\
d_4^2 &= 16m^2 + 8mc + c^2 - 24m - 6c + 9 \\
d_5^2 &= 25m^2 + 10mc + c^2 - 35m - 7c + 12{,}25
\end{aligned}$$

$$\sum_{i=1}^{5} d_i^2 = 55m^2 + 30mc + 5c^2 - 87{,}5m - 25{,}5c + 36{,}0625$$

Ahora hemos de hallar el valor mínimo de la función anterior. Si la tratamos primero como una ecuación cuadrática en m y después como una ecuación cuadrática en c, podemos hallar los valores buscados de m y c.

Supongamos que tenemos una ecuación

$$am^2 + bm + c = f(m)$$

$$a\left(m^2 + \frac{b}{a}m + \frac{c}{a}\right) = f(m)$$

$$a\left(\left(m + \frac{b}{2a}\right)^2 + \frac{c}{a} - \frac{b^2}{4a^2}\right) = f(m)$$

* No investigamos el mínimo de los cuadrados de las distancias «perpendiculares» de cada punto a la recta, pues depende de la escala. De aquí que las rectas dibujadas «a ojo» sean siempre incorrectas. *(N. del T.)*

Entonces el valor mínimo sería cuando $m = -\frac{b}{2a}$, ya que $\left(m + \frac{b}{2a}\right)^2 \geqslant 0$. Por tanto, en nuestro caso

$$m = \frac{87{,}5 - 30\,c}{2 \times 55} \Rightarrow 110\,m = 87{,}5 - 30\,c$$

Análogamente,

$$c = \frac{25{,}5 - 30\,m}{2 \times 5} \qquad \text{o sea} \qquad 10\,c = 25{,}5 - 30\,m$$

Encontramos los valores de c y m que satisfacen a la vez ambas ecuaciones, a saber:

$$m = \frac{11}{20} = 0{,}55 \qquad \text{y} \qquad c = \frac{9}{10} = 0{,}9$$

La ecuación que buscamos es, pues,

$$y = 0{,}55\,x + 0{,}9$$

Afortunadamente, hay un método más sencillo de calcular m y c, aunque su demostración es más bien difícil y la hemos omitido aquí. Esto implica:

1. Escribir las ecuaciones como antes y sumarlas miembro a miembro.
2. Multiplicar cada ecuación por el valor correspondiente de x y sumar miembro a miembro estas nuevas ecuaciones.
3. Hallar los valores de c y m que satisfacen al sistema de las dos ecuaciones suma.

En nuestro ejemplo tenemos

x	$y = mx + c$	$xy = x(mx + c)$
1	$1{,}0 = m + c$	$1 = m + c$
2	$2{,}5 = 2m + c$	$5 = 4m + 2c$
3	$2{,}75 = 3m + c$	$8{,}25 = 9m + 3c$
4	$3{,}0 = 4m + c$	$12 = 16m + 4c$
5	$3{,}5 = 5m + c$	$17{,}5 = 25m + 5c$
	$12{,}75 = 15m + 5c$ (1)	$43{,}75 = 55m + 15c$ (2)

Resolvamos ahora estas dos ecuaciones

$$12{,}75 = 15\,m + 5\,c \qquad \text{y} \qquad 43{,}75 = 55\,m + 15\,c$$

para hallar, como antes,

$$m = \frac{11}{20} = 0{,}55 \qquad\qquad c = \frac{9}{10} = 0{,}9$$

con método mucho más fácil y rápido que el primero.

Un punto interesante es que esta recta pasa por el valor medio de nuestras observaciones. Esto puede verse con facilidad, puesto que nuestra ecuación es justamente la

$$\sum_{i=1}^{5} y_i = m \sum_{i=1}^{5} x_i + 5\,c$$

o

$$\sum_{i=1}^{5} \frac{y_i}{5} = m \sum_{i=1}^{5} \frac{x_i}{5} + c$$

o

$$\bar{y} = m\,\bar{x} + c$$

Esta recta que hemos hallado (la que expresa los valores de y a partir de los de x) se llama recta de regresión de y sobre x *. En nuestro ejemplo la ecuación es $y = 0{,}55\,x + 0{,}9$.

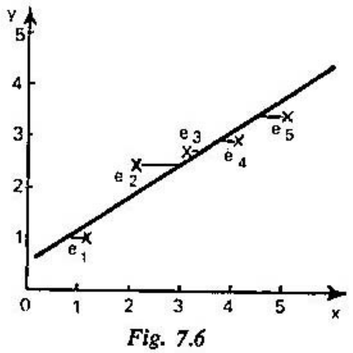


Fig. 7.6

* Adviértase que cuando no hay relación entre las observaciones la recta es horizontal y $m = 0$. *(N. del T.)*

Si queremos medir valores de x a partir de y, el procedimiento es muy parecido: en esta ocasión lo que nos interesa son las distancias horizontales desde los puntos a nuestra línea, y hemos de hallar el valor mínimo de

$$e_1^2 + e_2^2 + \ldots + e_5^2$$

Desarrollaremos esto exactamente igual que antes, sólo que esta vez la ecuación de la recta será $x = m' y + c'$. Entonces (usando el segundo método)

y	$x = m' y + c'$	$yx = y(m' y + c')$
1,0	$1 = m' + c'$	$1 = m' + c'$
2,5	$2 = 2{,}5\, m' + c'$	$5 = 6{,}25\, m' + 2{,}5\, c'$
2,75	$3 = 2{,}75\, m' + c'$	$8{,}25 = 7{,}5625\, m' + 2{,}75\, c'$
3,0	$4 = 3\, m' + c'$	$12 = 9\, m' + 3\, c'$
3,5	$5 = 3{,}5\, m' + c'$	$17{,}5 = 12{,}25\, m' + 3{,}5\, c'$
	$15 = 12{,}75\, m' + 5\, c'$	$43{,}75 = 36{,}0625\, m' + 12{,}75\, c'$

Resolvemos el sistema de dos ecuaciones, cuya solución es ahora

$$m' = 1{,}55 \qquad \text{y} \qquad c' = -0{,}95$$

Por tanto, $x = 1{,}55\, y - 0{,}95$.

Observa que la recta de regresión de x sobre y *no es* la misma que la recta de regresión de y sobre x; que si $m' = 0$, no hay relación entre las dos series de datos, y que las rectas de regresión pasan por el punto medio $(\bar{x}, \bar{y})$.

Si las dos rectas de regresión fueran idénticas, significaría, si los valores estuvieran exactamente sobre una línea recta, que

$$m = \frac{1}{m'} \qquad \text{o} \qquad m\, m' = 1$$

Como cuando no hay interdependencia es $m\, m' = 0$ y ahora hemos visto que para una ligadura estricta $m\, m' = 1$, este producto nos da una medida de lo bien que una línea recta puede ser la representación de nuestro grupo de valores.

Definimos para esto el *coeficiente de correlación* $r = \pm \sqrt{m\, m'}$.

Por convenio, si m y m' son positivas, entonces se toma r como positivo, y si $m, m' < 0$, entonces r es negativo. Esto lo discutimos con más detalle más adelante.

EJERCICIO 12

Las preguntas de este ejercicio deben resolverse con ayuda de una calculadora. Si no puedes disponer de ella, sitúa los valores dados en cada pregunta sobre papel milimetrado, y utiliza el método de aproximación (pág. 111) para obtener en cada caso la recta de regresión. Cuantos más puntos tengamos, mejor opera este método.

1. Para determinar g, aceleración de la gravedad, se realizó un experimento en la forma siguiente. Se puso un péndulo simple y se midió el tiempo de una oscilación completa, T, para longitudes variables, L. Se dibujó la gráfica que relaciona T^2 con L y se calculó g a partir de la fórmula

$$T^2 = \frac{4\pi^2 L}{g}$$

L, en cm	15	50	100	125	150
T^2 (T en seg) . . .	0,75	1,47	2	2,25	2,8

Dibuja la recta de regresión de T^2 sobre L. ¿Opinas que el experimento fue muy exacto? ¿Dónde es probable que se hayan introducido errores? Calcula g.

2. En un experimento para investigar una lente líquida se obtuvieron los siguientes resultados:

y, distancia de la lente a la imagen
x, distancia del objeto a la lente

$\frac{1}{x}$:	0,12	0,07	0,04	0,03
$\frac{1}{y}$:	0,06	0,11	0,12	0,13

Dibuja la recta de regresión de $\frac{1}{y}$ sobre $\frac{1}{x}$.

3. Empleando el tubo de Geiger-Muller se hizo un experimento en el que se midieron los voltajes aplicados y los niveles obtenidos.

V, en voltios . .	280	320	360	400	440	480
Niveles (*N*) . . .	0	2	64	73	80	89

Dibuja la recta de regresión de V sobre N.

4. En un experimento para hallar la resistencia de una bobina, se dibujó la gráfica que relaciona voltaje e intensidad. He aquí los resultados:

V (voltaje en volt)	0,26	0,35	0,4	0,45	0,52	0,72	0,92	1,20
I (int. en amp) .	0,1	0,18	0,2	0,22	0,28	0,37	0,48	0,60

Dibuja la recta de regresión de V sobre I y después halla la resistencia de la bobina (ley de Ohm $V = R\,I$).

5. En un experimento para medir el alcance de partículas radiactivas a través del aire se contó el número de choques por segundo en un electroscopio para diferentes distancias d.

d en cm	0,5	1,4	2,2	3,3	4,1	5,0	5,5	6,25
Choques por seg .	1,73	1,63	1,4	1,13	0,66	0,37	0,20	0,13

¿Existe una buena relación lineal entre estos resultados?

Diagramas de dispersión

Si tenemos cierto número de observaciones, en las cuales hemos realizado dos medidas, podemos hacer una gráfica con ellas, como ya hemos visto. En dicha gráfica es posible ver si existe alguna relación aparente entre nuestras dos medidas. Estas gráficas pueden tomar muchas formas; por ejemplo:

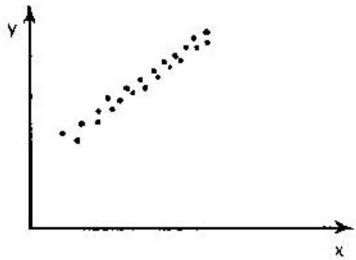


Fig. 7.7

1. Evidentemente, aquí hay una buena relación lineal entre x e y. Esta relación se denomina *correlación* entre x e y.

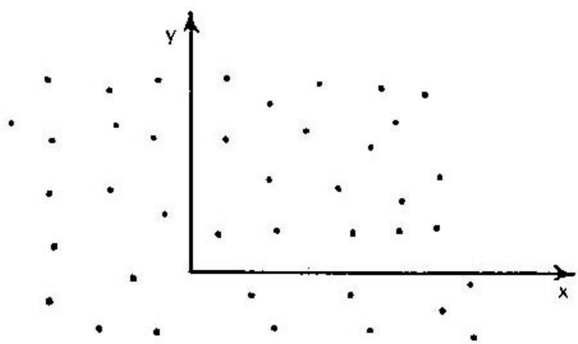


Fig. 7.8

2. Total disociación entre x e y.

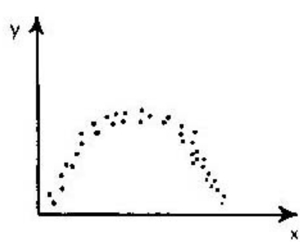


Fig. 7.9

3. En ésta hay una relación más complicada entre x e y: una relación cuadrática.

Estas gráficas se llaman *diagramas de dispersión*. Aquí nos interesamos sólo en las relaciones lineales. La medida de la «precisión de ajuste» de los valores observados con respecto a una línea recta es el *coeficiente de correlación*. Sus valores están comprendidos entre +1 y —1. Cuanto más cercano sea a esos valores el coeficiente, tanto mejor es el ajuste.

Ejemplos

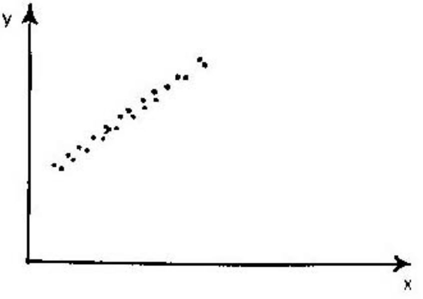


Fig. 7.10

1. Una buena aproximación a una línea recta. En este caso el coeficiente de correlación $r = \sqrt{m\,m'}$ sería aproximadamente igual a +1.

2. Una pobre correlación ($r \simeq 0{,}2$).

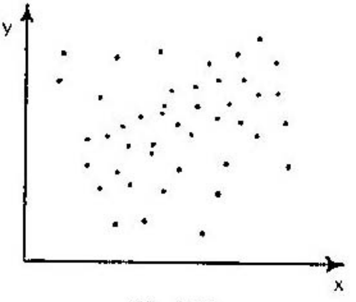


Fig. 7.11

3. Una buena correlación negativa ($r \simeq -0{,}95$).

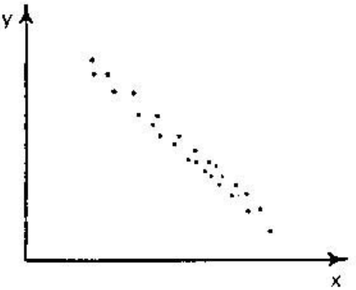


Fig. 7.12

Como ya hemos visto, el coeficiente de correlación viene definido por $r = \pm \sqrt{m\,m'}$. Esto puede escribirse

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_x} \cdot \frac{(y_i - \bar{y})}{\sigma_y}$$

donde σ_x y σ_y son las desviaciones standard de los valores de x e y respectivamente.

Como un simple ejemplo consideramos los valores

x_i	0	1	2	3	4
y_i	1	4	7	10	13

(Sabemos que los puntos están sobre la recta $y = 3\,x + 1$, y así r valdrá 1.)

x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
0	1	4	36	$-2 \times -6 = 12$
1	4	1	9	$-1 \times -3 = 3$
2	7	0	0	$0 \times 0 = 0$
3	10	1	9	$+1 \times +3 = 3$
4	13	4	36	$+2 \times +6 = 12$
10	35	10	90	30. Totales.

$\bar{x} = 2$ $\quad \bar{y} = 7$ $\quad \sigma_x = \sqrt{2}$ $\quad \sigma_y = \sqrt{18}$

Entonces

$$r = \frac{1}{5} \cdot \frac{30}{\sqrt{2}\,\sqrt{18}} = +1$$

4. La siguiente tabla contiene una lista de calificaciones de examen obtenidas por 10 alumnos en Matemáticas, Arte y Física. Calcula el coeficiente de correlación entre: (1) Arte y Matemáticas; (2) Física y Matemáticas.

Matemáticas (x_i)	*Arte* (y_i)	*Física* (z_i)
66	25	43
30	80	40
38	30	5
60	41	55
70	56	52
43	55	69
30	81	10
58	36	25
56	65	26
49	10	75

(1)	x_i	y_i	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	66	25	256	529	—368
	30	80	400	1024	—640
	38	30	144	324	+216
	60	41	100	49	—70
	70	56	400	64	+160
	43	55	49	49	—49
	30	82	400	1156	—680
	58	36	64	144	—96
	56	65	36	289	+102
	49	10	1	1444	+38
Totales	500	480	1850	5072	—1387

$\bar{x} = 50$ $\quad \bar{y} = 48$ $\quad \sigma_x = \sqrt{185}$ $\quad \sigma_y = \sqrt{507{,}2}$

$$r = \frac{1}{10} \cdot \frac{-1387}{13{,}6 \times 22{,}56} = -0{,}45$$

Esto muestra que

(*a*) Como $r = -0{,}45$, no hay una apreciable correlación lineal entre Arte y Matemáticas.

(*b*) Conforme aumentan las notas de Matemáticas, disminuyen las de Arte (o sea es una correlación negativa).

El diagrama de dispersión aparecerá así:

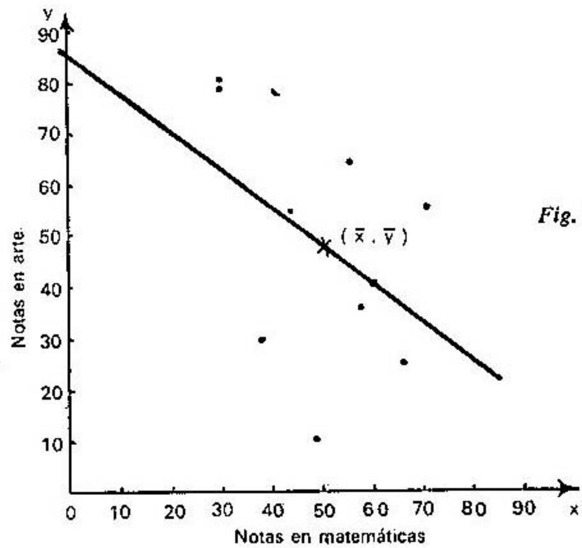


Fig. 7.13

La recta de regresión de y sobre x puede calcularse fácilmente a partir de

$$\left(\frac{y-\bar{y}}{\sigma_y}\right) = r\left(\frac{x-\bar{x}}{\sigma_x}\right)$$

(Observa que hemos omitido la demostración de esta fórmula.)

$$y - 48 = -0{,}45 \cdot \left(\frac{22{,}56}{13{,}6}\right)(x - 50)$$

o

$$y = -0{,}75\,x + 85{,}3$$

(2)

	x_i	z_i	$(x_i - \overline{x})^2$	$(z_i - \overline{z})^2$	$(x_i - \overline{x})(z_i - \overline{z})$
	66	43	256	9	+48
	30	40	400	0	0
	38	5	144	1225	+420
	60	55	100	225	+150
	70	52	400	144	+240
	43	69	49	361	—133
	30	10	400	900	+600
	58	25	64	225	—120
	56	26	36	196	—84
	49	75	1	1225	—35
Totales	500	400	1850	4510	+1086

$\overline{x} = 50$ $\quad\overline{z} = 40\quad$ $\sigma_x = \sqrt{185}$ $\quad\sigma_z = \sqrt{451}$

$$r = \frac{1}{10} \cdot \frac{1086}{13{,}6 \times 21{,}24} = +0{,}376$$

Esta vez el diagrama de dispersión será así:

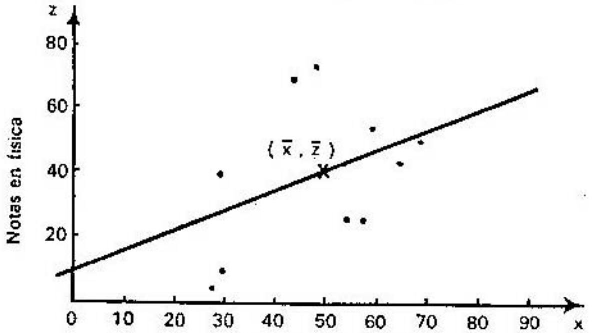


Fig. 7.14

En esta ocasión la recta de regresión es:

$$\left(\frac{z - \overline{z}}{\sigma_z}\right) = r\left(\frac{x - x}{\sigma_x}\right)$$

o bien

$$z - 40 = 0{,}376 \left(\frac{21{,}24}{13{,}6} \right) (x - 50)$$

es decir,

$$z = 0{,}587\, x + 10{,}65$$

Ahora podemos decir que:

(*a*) No hay una correlación fuerte entre las notas de matemáticas y las de física, pero

(*b*) conforme aumentan las notas de matemáticas, aumentan también las de física.

Antes de deducir ninguna conclusión a partir del valor del *coeficiente de correlación* entre dos observaciones, hay que atender cierto número de detalles.

1. Conviene siempre trazar el diagrama de dispersión para ver si es probable que haya una relación lineal entre las observaciones. Algunas veces es posible dibujar la línea de regresión y hallar un coeficiente de correlación bastante alto cuando de hecho las observaciones no están relacionadas linealmente.

2. Es arriesgado suponer que un cambio en uno de los caracteres observados puede ser causa de un cambio en el otro. Es muy difícil establecer causa y efecto, pues con frecuencia hay otras relaciones ocultas. Un ejemplo famoso es el siguiente: Se preguntó a cierto número de personas cuántas tías solteras tenían, y después se investigó el porcentaje de calcio que tenían dichas personas en los huesos. Se halló que había una buena correlación (negativa) entre estas observaciones; pero sería falso decir que el porcentaje de calcio en los huesos era el causante de que una persona tuviera ese número de tías solteras. En este caso la relación oculta era la *edad* de los preguntados.

EJERCICIO 13

Un grupo de niños obtuvo las siguientes notas en sus exámenes de fin de curso.

2.

Aritmética . . .	42	46	50	64	54	53	42	50	49	80
Inteligencia . . .	20	29	51	42	49	50	51	46	40	52
Edad: años . .	8	8	8	8	8	8	8	8	8	8
y meses .	4	8	6	4	2	4	7	3	2	0

Dibuja los diagramas de dispersión de: (*a*) aritmética-inteligencia; (*b*) aritmética-edad, y (*c*) inteligencia-edad.

Calcula en cada caso el coeficiente de correlación y saca tus propias conclusiones. ¿Debería haber un ajuste de las notas de acuerdo con la edad?

Si los exámenes fueran correctos para niños de todas las edades, ¿cuál sería el coeficiente de correlación y por qué?

2. Anota las tallas de las gorras de tus compañeros de clase y comprueba si hay alguna correlación entre los resultados de vuestro último examen de matemáticas y la talla de las gorras.

Si no puedes conseguir las cifras, utiliza las siguientes:

Nota	11	36	26	12	15	33
Talla	$7\frac{1}{4}$	$6\frac{7}{8}$	$7\frac{1}{8}$	$7\frac{1}{4}$	$7\frac{1}{2}$	7

Si no te es posible utilizar calculadora, dibuja los diagramas de dispersión de los resultados anteriores y emplea los métodos de aproximación (pág. 110) para obtener las dos rectas de regresión. Mide m y m' y calcula r de acuerdo con

$$r = \sqrt{m\,m'}$$

Observación

Es muy importante que el alumno intente por sí mismo algunas experiencias, con el fin de que aprecie las dificultades inherentes a las mismas. También debería estar preparado para criticar los métodos y explicar las causas por las cuales se han introducido errores. Es esencial que el estudiante tenga alguna idea de la magnitud de los errores introducidos y sea capaz de valorar la exactitud de sus resultados. Para constatar la necesidad de esto, se pedirá a todos los de la clase que dibujen cada uno un pentágono de 5 cm de lado, tan exactamente como le sea posible, y que midan la diagonal. Los alumnos quedarán convenientemente asombrados de la diversidad de sus respuestas y estarán capacitados para darse una idea del grado de exactitud en sus respuestas. Comprobarlas con el cálculo.

Sugerencias y respuestas

Ejercicio 1

1. (*a*) No están incluidos los no votantes (por ejemplo, todos los menores de 21 años).

(*b*) La mayor parte de los hombres están trabajando y los niños en los colegios.

(*c*) Están excluidos todos aquellos a los que no gusta el fútbol.

2. (*a*) Depende en gran manera de la edad del grupo. Quizás una prueba relámpago en cada casa de una comunidad típica ciudadana en un domingo por la mañana pudiera ser una muestra tipo perfecta.

(*b*) Entrevistar a las amas de casa de una de cada 5 viviendas un lunes, pues las opiniones de hombres y niños no nos interesan en este caso.

(*c*) La única forma correcta sería entrevistar a la gente en un bar, donde todas las bebidas están disponibles y representadas todas las marcas.

(*d*) Preguntar a la gente de un vagón de fumadores de un tren rápido.

Aconsejar la forma de selección de una muestra tipo es muy dudoso, por ejemplo, en (*a*) sería más prudente preguntar al grupo de edad «pop», o sea de 13 a 20 años; en (*b*) sólo es necesario preguntar a las amas de casa; en (*c*) sólo a los hombres, y únicamente a los fumadores en (*d*).

3. La gráfica implica que a causa de la inauguración del canal UHF ha habido un aumento en la venta de televisores. Esto no es

probable que sea cierto, pues la mayor prate de los aparatos, una de dos: (*a*) recogen el segundo canal, o (*b*) se les pone un convertidor.

4. Todo es falso: no hay escalas, ni unidades, etc.

5. No es el tipo apropiado de gráfica, debería ser de barras. Además, la gráfica es altamente improbable, pues mucha gente seguramente estará enferma en invierno.

Ejercicio 2

1. Debería considerarse el número de muertos por «hombre-hora» de deporte. Las características del boxeo y las de los otros deportes son muy distintas.

2. Una bonita curva para causar buen efecto, pero no tiene verdaderas escalas, la de tiempos es vertical en lugar de horizontal, etc.

Ejercicio 3

(*a*) La moda es 12 000 ptas. al mes.
(*b*) Media aritmética = 29 090 ptas. al mes.
Depende por completo de lo que entendamos que es correcto. En este caso la moda puede que sea lo más correcto, pues representa la cantidad que perciben 10 de cada 11 personas.

Ejercicio 4

1. 100, 103, 104, 102, $98\frac{1}{2}$, 110.

Ejercicio 5

1. $\frac{1}{6}$.

2. $\frac{5}{6}$.

3. (*a*) $\frac{4}{36} = \frac{1}{9}$; (*b*) $\frac{6}{36} = \frac{1}{6}$; (*c*) $\frac{4}{36} = \frac{1}{9}$.

4. $(a)\ \frac{1}{32}$; $(b)\ \frac{10}{32}$; $(c)\ \frac{10}{32}$.

5. 3 chicas, 3 chicos; $p = \frac{20}{64}$.

6. $(a)\ \frac{1}{40}$; $(b)\ \frac{4}{40} = \frac{1}{10}$; $(c)\ \frac{1}{40}\cdot\frac{1}{39}\cdot\frac{1}{38}\cdot 6$; $(d)\ 1\cdot\frac{3}{39}\cdot\frac{2}{38}\cdot\frac{1}{37}\cdot 6$.

7. $(a)\ \frac{1}{5}\cdot\frac{1}{4}\cdot\frac{1}{3}\cdot\frac{1}{2}\cdot\frac{1}{1} = \frac{1}{120}$; $(b)\ \frac{119}{120}$.

8. Probabilidad de 4 fallos $= \left(\frac{1}{3}\right)^4$; de 3 fallos y 1 blanco $= 4\left(\frac{1}{3}\right)^4$. Probabilidad total de fallarlo $= \frac{5}{81} \Rightarrow$ probabilidad de echar a pique el barco $= \frac{76}{81}$.

9. Ganar y despejado $p_c = 0{,}8 \times 0{,}4 = 0{,}32$. Ganar y lluvia $p_h = 0{,}6 \times 0{,}4 = 0{,}24$; o sea que $p_{\text{lluvia}} = \frac{0{,}24}{0{,}24 + 0{,}32} = \frac{3}{7}$.

10. Todos tienen igual oportunidad.

Ejercicio 6

1. (a) 126; (b) 36.

2. $(a)\ \frac{15}{2^7}$; $(b)\ \frac{105}{2^9}$; $(c)\ \frac{63}{2^8}$.

3. $(a)\ \frac{15\cdot 14 \ldots 7}{9\cdot 8 \ldots 1}\left(\frac{1}{3}\right)^6\left(\frac{2}{3}\right)^9$; $(b)\ \frac{15\cdot 14 \ldots 10}{6\cdot 5 \ldots 1}\left(\frac{1}{3}\right)^9\left(\frac{2}{3}\right)^6$

$(c)\ \frac{15\cdot 14 \ldots 8}{8\cdot 7 \ldots 1}\left(\frac{1}{3}\right)^7\left(\frac{2}{3}\right)^8$; $(d)\ \left(\frac{1}{3}\right)^{15}$; $(e)\ \left(\frac{2}{3}\right)^{15}$

4. $(a)\ \frac{12\cdot 11\cdot 10}{3\cdot 2\cdot 1}\left(\frac{9}{10}\right)^9\left(\frac{1}{10}\right)^3$; $(b)\ \frac{12\cdot 11 \ldots 6}{7\cdot 6 \ldots 1}\left(\frac{9}{10}\right)^5\left(\frac{1}{10}\right)^7$

(c) 1.

5. 2 18 72 168 252 252 168 72 18 2

6. (*a*) $\frac{252}{1024}$; (*b*) $\frac{672}{1024}$.

7. $\frac{16}{36} = \frac{4}{9}$.

8. $\dfrac{\frac{1}{2}}{\frac{1}{2} + \frac{1}{3}} = \frac{3}{5}$. Cambiaría de máquina.

Problema 1.

(*a*) $\frac{22}{82}$; (*b*) $\frac{64}{82}$; (*c*) $\frac{1}{82}$; (*d*) $\frac{59}{82}$; (*e*) $\frac{0}{82}$. No.

Problema 2.

$\frac{319}{750} = \frac{20\,416}{48\,000}$. Si haces el primer cambio en menos tiempo del promedio, por ejemplo $2\frac{5}{8}$ minutos, aumenta la probabilidad de perder el tren $\left(= \frac{20\,535}{48\,000}\right)$.

Ejercicio 7

1. Hora media $\simeq$ 11 de la noche. Desviac. standard $\simeq$ 40 min.

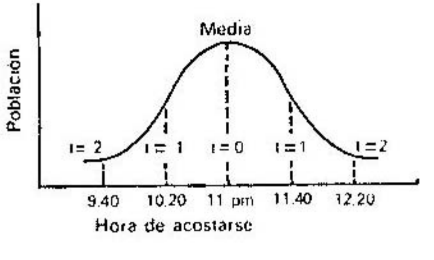


Fig. 7.15

(*a*) Las 10 h es $1\frac{1}{2}$ desviación standard del valor medio, de forma que leemos en nuestra tabla (pág. 81) para $t = -1\frac{1}{2}$ (lo mismo que para $t = 1\frac{1}{2}$) $p = 0{,}07$. Por tanto, en una ciudad de 40 000 habitantes podemos esperar que $40\,000 \times 0{,}07$ estén en la cama a las 10 h.

(*b*) Para hallar el área (probabilidad) comprendida entre las ordenadas que representan las $9\frac{1}{2}$ h y las 11 h leemos el valor de p para $t = -2{,}25$, pues $9\frac{1}{2}$ es 2,25 desviaciones standard, luego $p = 0{,}01$. Por tanto, el área comprendida entre $9\frac{1}{2}$ y 11 de la noche es $0{,}5 - 0{,}01$. O sea que esperamos que $40\,000 \times 0{,}49$ personas se hayan acostado entre las $9\frac{1}{2}$ y las 11 de la noche, o sea 19 600. Para hallar a qué hora se ha acostado ya el 89 % de la población, sabemos que el 50 % lo estará a las 11. El 39 % restante estará en la cama (según la tabla con $p = 0{,}5 - 0{,}39 = 0{,}11$) a la hora dada por $t = 1{,}25$, o sea 50 minutos después del valor medio, es decir, a las 11.50 de la noche.

2. 100, 280, 1000, 1800, 2800, 4000, 4000, 2800, 1800, 1000, 280, 100. Total 19 960 $\Rightarrow$ 40 necesitarán zapatos especiales.

3. (*a*) 2100; (*b*) 29 220; (*c*) 27 900.

4. (*a*) 99; (*b*) 7; (*c*) 81 horas.

5. (*a*) 1400 libras; (*b*) 1590; (*c*) (I) 4, (II) 1.

Ejercicio 8

1. Media 7, desviación standard $= \sqrt{\dfrac{105}{14}}$.

2. Media 6, desviación standard $= \sqrt{\dfrac{54}{27}} = \sqrt{2}$.

3. $\dfrac{3}{27} = \dfrac{1}{9}$; $\dfrac{7}{27}$; $\dfrac{6}{27} = \dfrac{2}{9}$; $\dfrac{17}{27}$.

4. (*a*) $\dfrac{26}{36}$; (*b*) $\dfrac{6}{36}$; (*c*) $\dfrac{24}{36}$.

5. (*a*) $\frac{12}{36} = \frac{1}{3}$; (*b*) $\frac{15}{36} = \frac{5}{36}$; (*c*) $\frac{1}{12}$.

Ejercicio 9

1. (*a*) 50; (*b*) 40 y 60.

2. 8, desviación standard $= \sqrt{\frac{250}{36}}$.

3. Número esperado de casos en que es efectivo = 200. Desviación standard $\simeq 11$, o sea que el agente es efectivo.

4. (*a*) 10; (*b*) $\sqrt{200 \cdot \frac{19}{20} \cdot \frac{1}{20}}$; (*c*) Sí.

5. (*a*) 500. (*b*) Sí (desviación standard $= \sqrt{450}$). (*c*) La suposición original era falsa. (Esperanza 1000, desviación standard = 30.)

Ejercicio 10

1. Entre 3,9 y 5,1.

2. 10 horas.

3. 0,40 cm.

Ejercicio 11

1. Probabilidad de 5 veces un mismo punto determinado = $= \frac{12!}{5!\,7!}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^7$ que multiplica por 6 nos da la probabilidad total de que ocurran sucesos igualmente probables que el producido. El resultado es significativo: el dado está amañado ($p \simeq 0{,}04$).

2. (*a*) $p = 0{,}06$ no significativo; (*b*) $p = 0{,}015$ significativo.

3. No.

4. 670 g. Sí.

5. Número esperado $= 36$. Como $n > 50$ utiliza la distribución binómica, la desviación standard $\simeq 3{,}8$. De la distribución normal $p < 0{,}05$ (2 desviaciones standard del medio); así pues, el resultado es significativo.

Ejercicio 12

1. $T^2 = 0{,}014\,L + 0{,}619$.

2. $\dfrac{1}{y} = -0{,}755\,\dfrac{1}{x} + 0{,}155$.

3. $V = 1{,}733\,R + 291{,}0$.

4. $V = 1{,}903\,I + 0{,}0245$.

5. $B = -0{,}314\,D + 2{,}0136$.

Ejercicio 13

1. Aritmética-edad: $r = 0{,}44$.

2. Aritmética-edad: $r = 0{,}656$.

3. Inteligencia-edad: $r = 0{,}268$.

4. $r \simeq -0{,}9$.