Col·lecció
«Ciències experimentals»
Núm. 4

INTRODUCCIÓN AL ANÁLISIS DE DATOS EXPERIMENTALES
Tratamiento de datos en bioensayos

Roque Serrano Gallego

UNIVERSITAT JAUME·I

BIBLIOTECA DE LA UNIVERSITAT JAUME I. Dades catalogràfiques

SERRANO GALLEGO, Roque

Introducción al análisis de datos experimentales : tratamiento de datos en bioensayos / Roque Serrano Gallego. — Castelló de la Plana : Publicacions de la Universitat Jaume I, D.L. 2003

p. : gràf. ; cm. — (Ciencies experimentals ; 4)

ISBN 84-8021-429-5

1. Anàlisi de dades. 2. Estadística matemàtica. 3. Bioassaigs-Mètodes estadístics. I. Universitat Jaume I (Castelló). Publicacions de la Universitat Jaume I, ed. II. Títol. III. Sèrie.

519.6

519.22

57:519.2

Edita: Publicacions de la Universitat Jaume I. Servei de Comunicació i Publicacions
Campus del Riu Sec. Edifici Rectorat i Serveis Centrals. 12071 Castelló de la Plana
Tel. 964 72 88 19. Fax 964 72 88 32
http://sic.uji.es/publ e-mail: publicacions@uji.es

ISBN 84-8021-429-5

Dipòsit legal:CS-391-2003

Imprimeix: Innovació Digital Castelló, S.L.

Índice

Prólogo

A través de mi labor docente en la Universitat Jaume I de Castellón, he podido darme cuenta del enorme interés que el curso de doctorado que vengo impartiendo desde 1993 «Análisis de datos en Química Analítica Ambiental» ha despertado no sólo entre los químicos, sino también en geólogos y biólogos. Este interés se debe, desde mi punto de vista, al conocimiento limitado que estos licenciados en ciencias tienen de las herramientas estadísticas a su disposición, especialmente en el campo del medio ambiente.

De hecho, existe un interés creciente de los científicos experimentales por la estadística. Hasta hace relativamente poco tiempo, los científicos del ámbito de la química, la geología o la biología no han obtenido todos los beneficios que la estadística les podía brindar.

Todo lo expuesto me ha llevado a madurar la idea de escribir un libro sobre análisis de datos experimentales, dirigido a todos aquellos que echan en falta un texto introductorio a esta disciplina que explique cómo seleccionar adecuadamente las técnicas estadísticas que permitan obtener la máxima información relevante a partir de los datos obtenidos en sus investigaciones. Si bien existen magníficas obras sobre estadística, la mayoría de ellas profundizan en aspectos teóricos y presentan ejemplos de aplicaciones típicas, sin concretar la manera de elegir la prueba más adecuada, de preparar los datos, de realizar los cálculos y de interpretar los resultados.

A través de este texto se enumeran, clasifican y describen las técnicas estadísticas que pueden ser de mayor utilidad en este campo, aplicándolas a problemas concretos. A menudo, se utilizan ejemplos que hacen más fácil la comprensión de cada una de ellas. Cuando es necesario, se incluye un apartado dedicado a explicar la manera más sencilla de realizar los cálculos utilizando programas estadísticos para ordenadores personales o sin ellos cuando es posible.

El tratamiento del aparato matemático de cada técnica se reduce al mínimo indispensable para conseguir un conocimiento suficiente que nos permita su correcta aplicación, así como para obtener conclusiones válidas en la interpretación de los resultados generados.

La obra comienza con una breve Introducción, donde se trata la importancia y la necesidad de la estadística en las Ciencias Experimentales, y lo que ha supuesto el desarrollo de la Informática en la aplicación de técnicas estadísticas que suponían en el pasado un gran esfuerzo de cálculo.

En la segunda parte de la Introducción se abordan sucintamente los fundamentos de la Teoría de la Probabilidad, base sobre la que se sustenta la estadística, lo que nos permitirá comprender e interpretar mejor las diferentes técnicas estadísticas en su aplicación a las diferentes ramas de las Ciencias Experimentales.

A continuación, se dedica un capítulo a la estadística básica que puede servir de repaso o recordatorio de los conceptos necesarios para la correcta comprensión de las técnicas estadísticas más complejas, las cuales se estudian en capítulos posteriores. La distribución Normal y otras distribuciones que pueden sernos de utilidad en el tratamiento de datos experimentales se estudian en este capítulo.

Seguidamente, se estudian las pruebas de significación para la comparación de dos muestras, tanto las paramétricas como las de distribución libre. Se hace especial énfasis en la utilización correcta de las primeras, que requieren para su aplicación datos que se distribuyan según una Normal. En la segunda parte del capítulo se ofrece la alternativa de las pruebas no paramétricas que no tienen este condicionante. Debido a la importancia de las pruebas de significación en el tratamiento de datos, este aspecto se trata con la profundidad necesaria, incluyendo el estudio de los tipos de errores que se pueden cometer en la toma de decisiones basadas en éstas.

En el cuarto capítulo se aborda el estudio del Análisis de la Varianza, profundizándose en los requerimientos de los supuestos paramétricos en datos compuestos por más de dos variables aleatorias. Se tratan el Análisis de la Varianza de una vía, de dos vías y múltiple, así como los conceptos de modelo I y II.

A continuación, se dedica un capítulo a la Regresión y sus diferentes tipos (simple, múltiple y por pasos), haciendo hincapié en los parámetros que indican la validez de los modelos adoptados. Asimismo, en este capítulo se trata de una manera breve el Análisis de la Covarianza y su utilidad en el análisis de la Regresión.

En el capítulo 6 se aborda el tratamiento de datos compuestos por numerosas variables mediante diferentes técnicas de Análisis Multivariante, excepto el Análisis de la Varianza Múltiple, la Regresión Múltiple y el Análisis de la Covarianza Múltiple, tratadas ya en capítulos anteriores. En primer lugar, se estudia el Análisis Discriminante, técnica predictiva. A continuación, se abordan las técnicas reductivas: el Análisis de Componentes Principales, el Análisis del Factor y el Análisis Cluster. Si bien existe una gran variedad de métodos multivariantes, en este capítulo se desarrollan los más ampliamente utilizados en las ciencias experimentales, los cuales sirven de base para la comprensión de otras muchas técnicas pertenecientes a esta misma familia.

Para finalizar, se incluye un capítulo sobre el tratamiento de datos obtenidos a partir de bioensayos. Se estudia la manera correcta de llevar a cabo un bioensayo, desde su diseño hasta el tratamiento de los datos obtenidos. Se incluyen tests de toxicidad, de acumulación y de factores subletales y se trata la modelización toxicocinética.

1. Introducción

En 1969 Robert R. Sokal y F. James Rohlf en su magnífica obra *Biometry* se aventuraron a predecir que el *Análisis de Datos* era la especialidad más prometedora de la estadística moderna. Esta disciplina consiste en la búsqueda sistemática de información y de relaciones a través de conjuntos de datos.

La gran cantidad de información que es posible generar en la actualidad a partir de un sistema experimental o de la naturaleza misma, gracias a la capacidad de los laboratorios analíticos y a los sistemas computerizados de adquisición de información, hace que el científico se encuentre, a menudo, con un número enorme de datos a partir de los cuales debe sacar conclusiones. La Estadística aplicada a las diferentes ramas de las Ciencias Experimentales se ocupa de desarrollar métodos que permitan a los científicos el tratamiento de estos datos. Así, el uso de las herramientas estadísticas a nuestra disposición permite extraer la máxima cantidad de información relevante contenida en un conjunto de datos obtenidos a lo largo de un experimento. Además, la implantación generalizada de los ordenadores en los centros de investigación hace que, en la actualidad, la aplicación de los métodos estadísticos sea extremadamente fácil.

Sin embargo, a menudo se utilizan técnicas estadísticas para la resolución de problemas que surgen en diferentes ramas científicas sin que se tenga la base de conocimientos necesarios para saber elegir y aplicar la técnica adecuada e interpretar los resultados obtenidos. Frecuentemente, el desconocimiento de las bases matemáticas y conceptuales de los métodos estadísticos hace que los utilicemos inapropiadamente, ya sea por elegir un análisis estadístico inadecuado para conseguir la información que deseamos obtener, o bien por aplicarlo a datos que no cumplen las premisas necesarias para que la información que obtengamos sea válida.

El objetivo de este texto es orientar a los potenciales usuarios de las técnicas de análisis de datos en la elección correcta de los tests más adecuados para cada caso, así como en la interpretación de los resultados generados por ellos. Para llegar a este fin, se trata la parte del aparato matemático que existe detrás de cada método estadístico de manera breve pero indispensable para el buen entendimiento de los fundamentos y aplicaciones de cada uno de ellos. En la bibliografía recomendada se inclu-

yen obras donde se encuentra el desarrollo matemático completo de los métodos que se describen, ya que no es el objetivo de este texto profundizar en ese aspecto.

Una de las razones fundamentales para que los científicos experimentales puedan permitirse el dudoso placer de obviar, en una cierta medida, el complicado aparato matemático existente detrás de cada método estadístico, reside en la disponibilidad que actualmente se tiene de una serie de programas para ordenadores personales que nos permiten realizar los cálculos necesarios en cada caso, con sólo pulsar una tecla de nuestro ordenador. El proceso para llevar a cabo el más complicado de los métodos estadísticos a nuestro alcance para un determinado grupo de datos consiste, en primer lugar, en introducir los datos objeto de estudio según los requerimientos del programa a utilizar. Sin embargo, hasta esta labor, quizá la más tediosa con la que nos encontramos, se ve facilitada en la actualidad. Los antiguos programas estadísticos –que sólo permitían el análisis de los datos si se estructuraban siguiendo estrictamente los requerimientos del programa, para lo cual se necesitaban en no pocas ocasiones largas horas de estudio del manual de instrucciones–, han sido sustituidos por los modernos paquetes de *software* estadístico, que permiten la interacción con el usuario hasta el punto de ejecutarse bajo entornos muy sencillos y conocidos por todos, como por ejemplo el MS Windows®.

En el siguiente apartado se describen una serie de programas, que si bien son sólo una pequeña parte de los existentes, pueden servir para que el lector se haga una idea de la oferta disponible.

1.1. Paquetes informáticos más utilizados en el análisis de datos experimentales

Microsoft EXCEL

Esta conocida hoja de cálculo contiene, en sus distintas versiones, un número creciente de técnicas estadísticas muy útiles en la investigación experimental, permitiendo además el diseño de hojas personalizadas para aplicaciones específicas.

En el programa se incluyen diferentes tipos de funciones: matemáticas, trigonométricas, estadísticas básicas, lógicas, financieras, etc. Además, existe un módulo –análisis de datos– que incluye distribuciones de probabilidad, análisis de la varianza, regresión y las pruebas de significación más comunes.

En definitiva, es una herramienta muy útil y presente en la gran mayoría de ordenadores, que permite la aplicación sencilla y rápida de las técnicas más comunes utilizadas en el análisis de datos. Además, es posible exportar e importar datos con otros programas y adaptar el formato de estos a nuestras necesidades.

STATGRAPHICS. Statistical Graphics System by Statistical Graphics Corporation. A Plus Ware Product. STSC

Es uno de los programas más populares en la actualidad, dada la sencillez del sistema de introducción de datos y de uso de las distintas herramientas estadísticas que posee, entre las cuales se incluyen:

- **Dibujo de funciones y estadística descriptiva**
  - Dibujo de funciones (X-Y, X-Y-Z, barras y circulares)
  - Métodos descriptivos (estadística general, histogramas, percentiles)
  - Estimación y prueba (de una, dos o más muestras)
  - Funciones de distribución (ajuste y dibujo de la función, probabilidades y valores críticos)
  - Análisis exploratorio de datos (gráficos Box-Whisker y de ramas)

- **Análisis de Regresión y ANOVA**

- **Procedimientos avanzados**
  - Análisis categórico de datos
  - Métodos multivariantes
  - Métodos no paramétricos

- **Diseño de experimentos**
- Diseño de superficies de respuesta
- Diseño de experimentos

- **Estadística multivariante**
  - Análisis de Componentes Principales (ACP)
  - Análisis del Factor
  - Análisis Cluster
  - Análisis Discriminante
  - Módulo de Series Temporales

La última versión (Statgraphics Plus para Windows®) incluye también una versión para MS-DOS®.

SPSS for Microsoft Windows

Es un programa clásico que en versiones anteriores requería de programación. La versión para Windows® mejora sensiblemente la facilidad de uso. Es uno de los paquetes estadísticos más completos en la actualidad. Se divide en un módulo base y una serie de módulos accesorios, de adquisición independiente.

Las principales herramientas incluidas en cada módulo son las siguientes:

- **Módulo base:**
  - Estadística fundamental
  - Frecuencias
  - Tabulaciones cruzadas
  - Pruebas *t* y no paramétricas
  - ANOVA
  - Regresión lineal, múltiple y correlaciones

- **Módulo de estadística profesional:**
  - Análisis del Factor
  - Análisis Cluster
  - Análisis Discriminante
  - Método de los K vecinos más próximos
  - Normalización multidimensional de datos

- **Módulo de estadística avanzada**
  - Regresión de Cox
  - Estimación Kaplan-Meier
  - Regresión logística
  - Análisis log-lineal
  - MANOVA
  - Regresión no lineal y no lineal limitada

- **Módulo de tendencias**
  - Estudio de Series Temporales

BMDP

Este es otro programa utilizado desde hace décadas, al igual que el SPSS®, y que requería en sus inicios de programación. Como en otros casos, se ha modernizado y adaptado al entorno Windows®.

Permite la aplicación de las siguientes técnicas:

- Estadística descriptiva
- Pruebas de significación
- Análisis de la Varianza
- Tablàs de frecuencias
- Regresión
- Gestión de datos

Statistica

Otro paquete estadístico bajo entorno Windows® que permite las siguientes técnicas:

- Estadística descriptiva
- Distribuciones
- Regresión
- Análisis Multivariante
- Series temporales
- Gráficos de control
- Diseño experimental

STAT-100 (BIOSOFT)

Este programa también trabaja bajo Windows® e incluye un tutorial muy útil para ayudarnos en su manejo.

Incluye un módulo de pruebas paramétricas y otro de pruebas no paramétricas, entre las que se pueden encontrar las más utilizadas (Kruskal-Wallis, U de Mann Whitney, etc.), transformación de datos, gestión y gráficos. En todos los casos el *output* que genera es sencillo y fácil de entender.

MATHEMATICA

Este potentísimo programa de cálculo simbólico tiene entre sus virtudes módulos específicos para estadística y ajuste de funciones. Si bien las opciones en lo que respecta al análisis de datos propiamente dicho no son tan amplias como en un paquete estadístico, las que están disponibles –como la Regresión no lineal, entre otras– son muy útiles debido a su elevada potencia de cálculo. Un punto en contra podría ser la complicación que representa a la hora de realizar representaciones gráficas.

Como se puede deducir de la Introducción, cualquier persona, científico o no, docto o novel en la ciencia de la estadística, es capaz de introducir los datos obtenidos a partir de un experimento u observación y aplicar cualquiera de los métodos estadísticos disponibles en su flamante ordenador. En no pocas ocasiones, ello desemboca en la aplicación de una prueba estadística equivocada a tenor de los objetivos que se persiguen. Otras veces se utilizarán métodos para los cuales los datos a analizar no cumplen los requisitos necesarios e indispensables para extraer información válida. La interpretación de los resultados es otro de los puntos más vulnerables al error en el proceso del análisis de datos.

Como consecuencia, es necesario, antes de sentarse delante de un ordenador para realizar un análisis estadístico de datos experimentales, los cuales la mayoría de las veces se han obtenido con grandes esfuerzos personales y materiales, adquirir unos conocimientos básicos que nos permitan aprovechar el trabajo realizado y no extraer conclusiones insuficientes o equivocadas o, en el peor de los casos, tirar por la borda toda una línea de investigación futura. Incluso después de un adecuado análisis de datos, debemos ser extremadamente cautos en la interpretación de los mismos, atendiendo siempre a los parámetros objetivos que nos ofrece la estadística, y teniendo en cuenta sus limitaciones.

1.2. La investigación experimental y la Ley de Probabilidad

La base de la investigación experimental son los experimentos, que pueden describirse como la suma de dos componentes: un fenómeno o sistema más o menos controlado por el científico, y un medio de observación e interacción, mediante el cual se registran una serie de observaciones y, que a su vez, permite modificar y controlar las variables del sistema.

Los fenómenos y sistemas que se estudian en cada una de las Ciencias Experimentales se ven afectados, a su vez, por muchos factores, la mayoría de los cuales no están controlados o identificados, y cuyos efectos entran en la categoría de los sucesos aleatorios o estocásticos.

Este tipo de sucesos se caracterizan porque es imposible predecir cuándo van a tener lugar. Sin embargo, si se repite la acción un número suficiente de veces, se puede observar una cierta regularidad que permite hacer predicciones en una serie larga de repeticiones. Esto llevó a los matemáticos del siglo XVIII a la idea de la regularidad estadística. En efecto, en poblaciones estadísticas grandes (muchas repeticiones), ciertos índices tienden a mantenerse casi constantes, con pequeñas fluctuaciones alrededor de un valor determinado.

Curiosamente, estas ideas se desarrollaron paralelamente en el tiempo en dos campos bien distintos, en el de los juegos de azar y en la observación experimental, sin que se llegaran a relacionar claramente. Un paso muy importante en el desarrollo de la Teoría de la Probabilidad fue la formulación de la similitud entre la regularidad estadística y la regularidad en los juegos de azar, afirmando que ambos fenómenos tenían una base común, es decir, una serie larga de tiradas de un dado no es más que una población estadística. En aquel tiempo, la mayoría de científicos experimentales no mostraron ningún interés por el tema, probablemente por desconocimiento. Una excepción fue Mendel, que interpretó acertadamente los resultados obtenidos en sus experimentos genéticos sobre la aparición de caracteres hereditarios en la descendencia como un fenómeno aleatorio.

Volviendo a la Teoría de la Probabilidad, se puede afirmar que, en una serie larga de repeticiones, la frecuencia relativa de un suceso (de un valor de la variable aleatoria) se aproxima a su probabilidad. Esta ley, llamada Ley Empírica del Azar, fue formalizada más tarde por J. Bernouilli (1712) en la ley débil de los grandes números y posteriormente por Borel (1900) en la ley fuerte de los grandes números.

Veamos un ejemplo extraído del campo de la genética sobre todo lo expuesto anteriormente: no se puede predecir el momento exacto en el que se producirá un error en la replicación del ADN de una célula durante su división, pero sí que se puede predecir qué proporción de errores tendrá lugar en un número suficientemente elevado de células.

En los experimentos científicos, el efecto de los factores que se observan y en ocasiones se controlan en algún grado, mantienen constantes una serie de características. Asimismo, se conocen una serie de otros factores que influyen en ellos o están relacionados de una u otra forma. La definición de fenómeno natural, en un sentido amplio, incluye los fenómenos que ocurren en la naturaleza sin ningún control del científico, otros en los que se puede ejercer un cierto control e incluso fenómenos provocados o forzados por el propio investigador.

Mientras que un físico no tiene ningún control sobre un tornado, un sistema experimental de simulación puede servir para estudiar corrientes de aire de una manera controlada. Incluso se pueden provocar fenómenos naturales, como mutaciones, mediante exposición de animales a radiaciones. A partir de la población sometida a un cierto tratamiento controlado, como por ejemplo exposición a rayos X, se tiene una cierta expectativa de que se produzcan mutaciones con una cierta frecuencia.

Por lo tanto, algunos de los parámetros que intervienen en los fenómenos de tipo aleatorio que tienen lugar en la naturaleza pueden ser controlados por el investigador. Así pues, de un sistema experimental se conoce, en alguna medida, el conjunto de resultados posibles y se tiene una esperanza de que ocurran ciertos sucesos. Sin embargo, otros factores y características, o bien no se pueden controlar o simplemente no se han identificado.

Según esto, la esperanza de que ocurra un suceso depende de la información disponible sobre el sistema en estudio. Como consecuencia, llegamos a una conclusión muy interesante: si conociéramos todos los factores (absolutamente todos) que intervienen en el sistema y las leyes que los rigen seríamos capaces de predecir cualquier suceso con precisión. Esto convierte el modelo determinista en un caso particular dentro del modelo aleatorio. En la figura 1.1 se presenta un esquema de este modelo.

Desarrollemos este extremo mediante un ejemplo: supongamos que conocemos todos los factores que pueden intervenir en la diferenciación y el desarrollo de células cancerígenas en un determinado tejido, así como las leyes que rigen estos factores y las relaciones entre ellos. En estas condiciones podríamos predecir con absoluta precisión cuándo se desarrollaría un cáncer en dicho tejido sometido a unas condiciones controladas estrictamente. Sin embargo, los científicos sólo conocen algunos de los factores de riesgo, y a partir de experiencias donde se controlan el mayor

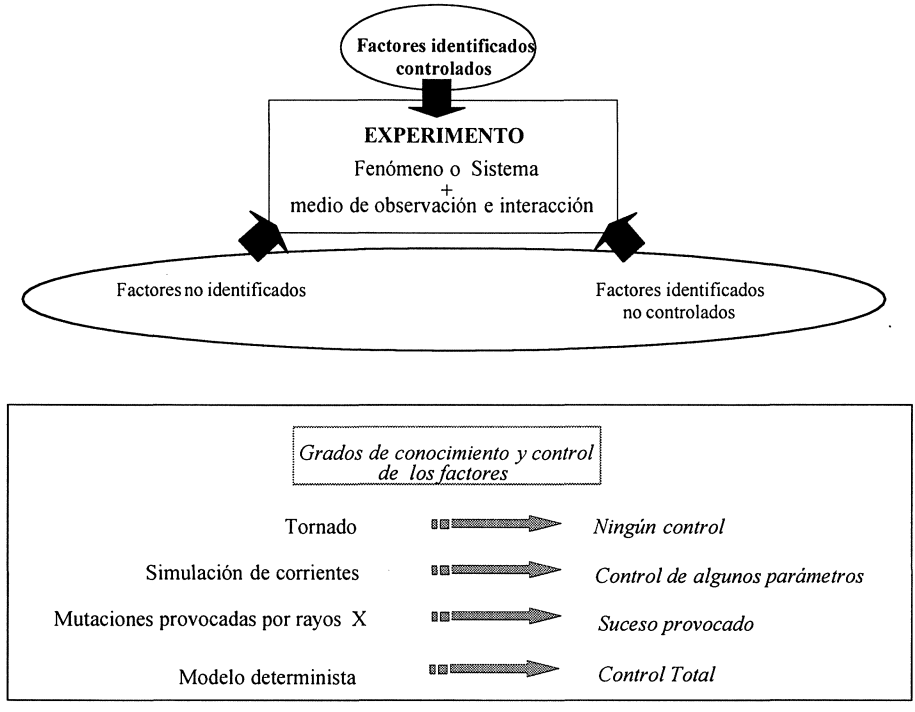


FIGURA 1.1. *Modelo Experimental*

número posible de variables conocidas se intenta conocer más del sistema en estudio, con el objeto de que algún día nos podamos acercar al caso ideal en el que todos los factores sean conocidos y las variables controladas.

Actualmente, dentro de las Ciencias Experimentales, hay una mayoría de sistemas que se rigen por modelos aleatorios, ya que es imposible controlar todos los factores que intervienen en un proceso físico, químico, y mucho menos biológico.

Las ideas vertidas en los párrafos anteriores hacen evolucionar la Estadística, ciencia que estudia el comportamiento de las variables aleatorias, hasta considerar que los fenómenos aleatorios se pueden medir de alguna forma, lo que permite desarrollar la Teoría de la Probabilidad como cualquier otra, aunque sea incapaz de predecir sucesos concretos con exactitud, mediante la llamada Axiomática de Kolmogorov. Los problemas en la aplicabilidad y la interpretación para diferentes fenómenos o sistemas son propios de la filosofía de la ciencia en la que se aplique esta teoría.

La aplicación de la Estadística al estudio de problemas en el ámbito de las Ciencias Experimentales (Biología, Geología, Química...) da lugar a disciplinas como la Biometría o la Quimiometría, las cuales han experimentado desde hace unos años un gran auge debido a su utilidad y a la implantación generalizada de la informática que, como ya se ha indicado, permite su fácil aplicación.

2. Estadística básica

En este capítulo se estudian los conceptos estadísticos básicos necesarios para poder entender las diferentes técnicas estadísticas que se tratan a lo largo de la obra, las cuales son de aplicación al estudio de las variables aleatorias.

La Estadística es la parte de las matemáticas que estudia el comportamiento de las variables aleatorias. Estas se definen como aquellas variables cuyos valores no están fijados de antemano, sino que cada uno de ellos tiene una probabilidad de que se produzca. Como se ha visto en el capítulo anterior, la mayoría de los factores que intervienen en los experimentos y en los sistemas experimentales, naturales o diseñados artificialmente por el hombre, se comportan como variables aleatorias, por lo que la evolución o el comportamiento de los sistemas en estudio es consecuencia de la suma de los efectos de muchas variables aleatorias. Debido a esto, es necesario conocer con profundidad este tipo de variables y debemos ser capaces de abstraer en nuestras mentes un sistema experimental como una serie de variables, descritas por los valores que toman a lo largo de un tiempo limitado de observación como consecuencia de los factores que les afectan.

En el desarrollo de este tema se obvia la descripción de conceptos básicos como exactitud, precisión, tipos de errores, etc., puesto que deben ser conocidos por cualquier científico experimental sea cual fuere su especialidad. Sin embargo, sí que se abordan otros conceptos, también básicos, que en ocasiones no son tratados en los textos con la suficiente amplitud, como los diferentes estimadores del valor medio y la dispersión, o conceptos como población, muestra o distribución, de primera importancia para entender todos los aspectos de los métodos estadísticos utilizados en el análisis de datos.

En primer lugar, se van a tratar los tipos de variables aleatorias que nos podemos encontrar en el trabajo experimental.

2.1. Tipos de variables aleatorias

Las variables aleatorias generadas en estudios experimentales se pueden clasificar de la siguiente forma:

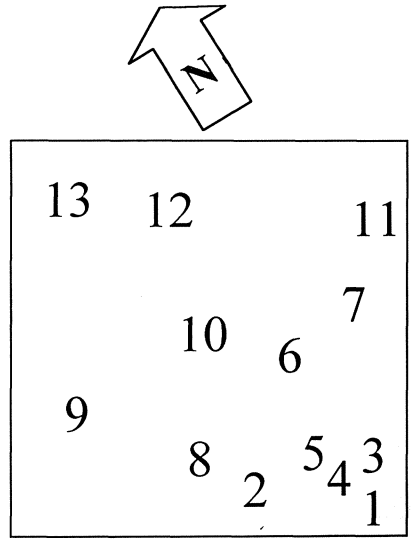
- Variables cuantificables
  - Continuas
  - Discontinuas
- Variables cualitativas
- Variables ordinales

Las **variables cuantificables** son aquellas que toman valores numéricos. Dentro de éstas, las **variables continuas** son las que pueden tomar un número infinito de valores entre dos números. Gran parte de las variables que se estudian en sistemas experimentales son **continuas,** como por ejemplo, longitudes, pesos, concentraciones de analitos, etc.

Las **variables discontinuas o discretas**, sólo pueden tomar valores numéricos fijos (en la mayoría de ocasiones números enteros), y no pueden tomar valores intermedios. Ejemplos típicos de variables **discontinuas** son las frecuencias (número de veces que aparece una determinada característica en una población) o los recuentos (número de colonias de bacterias).

Las **variables cualitativas** (o categóricas) son las variables que describen una cualidad que no puede medirse como en el caso anterior pero se puede expresar cualitativamente (normal o albino, vivo o muerto). En estos casos puede codificarse la variable: vivo=1, muerto=2. Una variable **cualitativa** es, por ejemplo en estudios toxicológicos, el registro de la mortalidad de organismos en un experimento que se puede describir como sano (1), afectado (2), muerto (3).

Por último, las **variables ordinales** aparecen cuando la característica en estudio no se puede medir pero sí ordenar. Ejemplo de variable **ordinal** es el orden de nacimiento de pupas de insectos en una determinada área, tal como se muestra en el siguiente esquema. Los números representan el orden en el tiempo en el que han nacido los insectos maduros situados en el lugar indicado.



Se debe tener en cuenta que en determinadas experiencias algunas variables continuas pueden ser codificadas como categóricas. Por ejemplo, el impacto de la con-

taminación sobre un enclave natural puede codificarse como nulo (0), bajo (1), medio (2) o alto (3), en este caso la variable puede considerarse continua, ya que describe el grado de degradación del área sin discontinuidades.

2.2. Población y muestra. Estimadores del valor central y la dispersión

Los estadísticos diferencian entre un parámetro referido a una población (n>30) y un parámetro referido a unas cuantas observaciones extraídas de una población (muestra).

En el caso de que debamos manejar una muestra, este conjunto limitado de medidas es nuestra única fuente de información acerca de la población, o lo que es lo mismo, del conjunto de *n* valores que conceptualmente toma una variable aleatoria. La información que debemos deducir acerca de la población estaría compuesta por estimaciones del valor medio, de la dispersión y, como veremos más adelante, la distribución de los valores, fundamentalmente.

En estudios ambientales, a menudo se obtienen el suficiente número de datos como para que los consideremos una población, por ejemplo, un número de medidas y pesadas de una población «biológica» de animales suficientemente elevado. Sin embargo, en otros casos, como en la determinación de la concentración de un contaminante en una matriz natural (por ejemplo, DDT en aves de una zona húmeda), a menudo debemos tomar decisiones a partir de un pequeño número de datos, ya que no es razonable realizar un número muy elevado de análisis (pues cada muestra significa la muerte de un ave) y de réplicas de cada análisis (por cuestiones económicas y de tiempo). Como consecuencia, deberemos estimar los parámetros que describan la población a partir de una pequeña muestra estadística. Uno de los intereses del científico será la obtención de valores que sean una buena estimación del valor central y la dispersión de la población cuyos indicadores serán diferentes según el tipo de variable aleatoria y el objetivo marcado. A continuación, se introducen los diferentes estimadores de estos parámetros.

Estimadores del valor central

- Media aritmética

Es el estimador del valor central más utilizado, su fórmula es:

$$\bar{x} = \frac{\left(x_1 + x_2 + \ldots + x_n\right)}{n} = \frac{1}{n}\sum_{n=1}^{n} x_n$$

Por otra parte, se debe tener en cuenta que un x_i anómalo transmite apreciablemente el error al valor estimado.

• Mediana

Se define como el valor que se sitúa en el centro tras ordenar las clases de frecuencias según la magnitud del valor que toma la variable. Es la estimación del valor central más utilizada cuando los valores no siguen una distribución normal.

• Moda

Es el valor más frecuente de la población, y ésta es la información que nos aporta. Los valores anómalos no le afectan, ya que la clase con una frecuencia mayor sigue siendo la misma, a pesar de los valores extremos que puedan aparecer.

En la figura 2.1. se presentan ejemplos en los que se puede observar la relación entre los diferentes parámetros descritos. Mientras que en poblaciones simétricas los tres valores coinciden, en poblaciones con cierta asimetría se desplazan. Como consecuencia de esto, dependiendo de la distribución de las frecuencias de una población, los estimadores del valor medio pueden ser más o menos útiles a la hora de obtener información sobre ella. Más adelante se tratará esta cuestión con más profundidad.

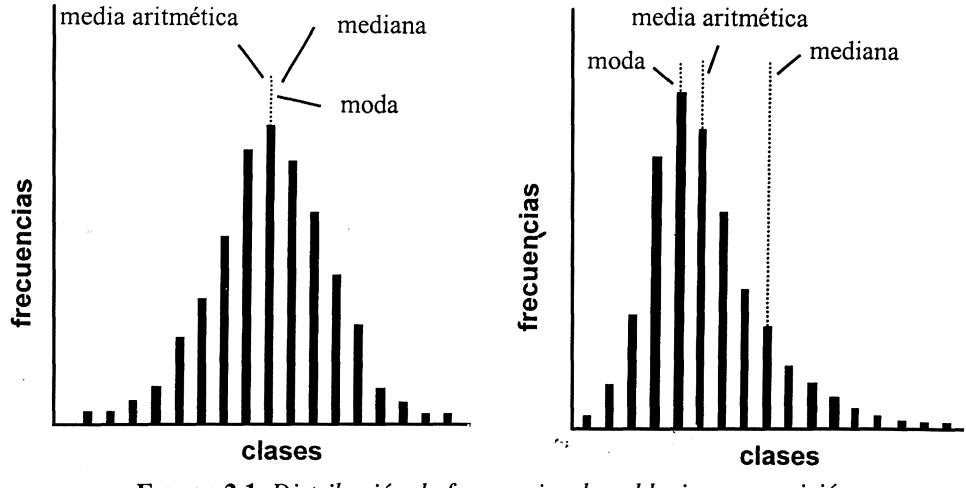


FIGURA 2.1. *Distribución de frecuencias de poblaciones y posición de los estimadores del valor medio*

• Media geométrica

En ocasiones, como veremos más adelante, es necesario transformar los datos en sus logaritmos. Si la media aritmética de estos se transforma, de nuevo, a la escala decimal, el valor obtenido no coincide con la media aritmética de los datos originales, y se denomina *media geométrica* (*GM*, del inglés *geometric mean*):

$$GM = anti\log\left(\frac{1}{n}\sum \log Y\right)$$

es decir, la media geométrica es la media de los logaritmos de la variable aleatoria Y transformada al sistema decimal, la cual también se puede expresar de la siguiente forma:

$$GM = \sqrt[n]{(y_1 \cdot y_2 \ldots y_n)}$$

Estimadores de la dispersión

• Desviación estándar

Este indicador de la dispersión tiene en cuenta todos los datos de la variable aleatoria, tanto los centrales como las colas, y es el más utilizado junto con la media aritmética para la presentación de datos de tipo cuantitativo.

Su fórmula para una población es:

$$s = \left[\frac{\sum (x_i - \bar{x})^2}{n}\right]^{\frac{1}{2}}$$

Como se puede observar, en el caso de que **n=1**, **s=0**, lo que es falso, por lo que para muestras con n<30 se debe utilizar la desviación estándar muestral:

$$s = \left[\frac{\sum (x_i - \bar{x})^2}{n-1}\right]^{\frac{1}{2}}$$

Otras manera de representar la desviación estándar es la desviación estándar relativa, la cual nos da una información más útil ya que está relacionada con la magnitud de la media aritmética. Normalmente, se expresa en forma de porcentaje, lo que también se denomina coeficiente de variación, si bien esta denominación no es muy apropiada, ya que en algunos casos mide una incertidumbre, no una variación. La fórmula de la desviación estándar relativa expresada en porcentaje sería:

$$d.e.r.(\%) = \frac{s}{\bar{x}} \cdot 100$$

El cuadrado de la desviación estándar es la varianza, que como veremos más adelante es un parámetro de importancia esencial en la mayoría de técnicas estadísticas.

• Error estándar de la media

A partir de una población estadística (n>30) podemos obtener directamente el valor medio, la dispersión y la forma en que se distribuye la población, ya que disponemos de un número suficiente de datos para obtener información exacta sobre la población conceptual de valores infinitos que puede tomar una variable aleatoria en estudio. Sin embargo, en otras ocasiones hemos constatado la imposibilidad de disponer de un número tan elevado de valores.

Pongamos el siguiente ejemplo: deseamos conocer la concentración de Pb en una muestra de agua. En este caso, la realización de más de 4 ó 5 réplicas del análisis no es práctica, y debemos buscar el valor medio de la población «concentración de Pb en la muestra de agua», o sea, el valor verdadero de la concentración de Pb a partir de estos valores.

Se puede demostrar matemáticamente que si en lugar de datos individuales se representan medias obtenidas a partir de un número limitado de réplicas, éstas últimas pertenecen a una población «de medias» que tiene una distribución mucho más ajustada a la verdadera que los valores individuales, la cual se denomina distribución muestral de la media. La dispersión de la población de medias calculada como desviación estándar se denomina error estándar de la media (eem) y se relaciona con la desviación estándar de la población de valores individuales mediante la siguiente fórmula

$$eem = \frac{s}{\sqrt{n}}$$

Se puede deducir de la fórmula que a mayor número de datos individuales que intervengan en el cálculo de las medias, el eem será menor, es decir, la dispersión de los datos alrededor del valor verdadero será menor.

Existe una propiedad de la distribución muestral de la media muy importante en estadística y de gran repercusión en su aplicación a datos experimentales. Se puede demostrar que la distribución muestral de la media tiende a una Normal en el infinito, incluso en el caso de que la población original de datos individuales no sea una Normal, según el Teorema del Límite Central. Esto nos permite asumir la Normalidad de distribuciones de datos experimentales con *n* baja, siempre que no exista ningún factor que nos indique lo contrario, como veremos más adelante.

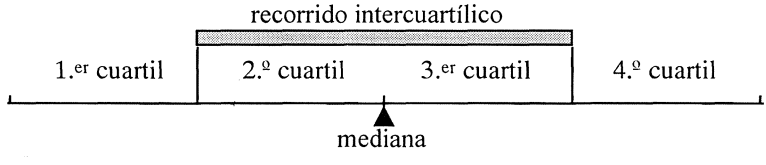
La importancia de este Teorema reside en que muchas pruebas estadísticas se basan en la Normalidad de la muestra.

• Recorrido

Es la diferencia entre los valores extremos, por lo que no considera los valores centrales y es muy sensible a la existencia de valores anómalos ya que estos, lógicamente, siempre se encuentran en los extremos.

• Recorrido intercuartílico

La mediana divide una muestra o población en dos partes iguales; si volvemos a dividir cada parte por el punto intermedio, se consigue dividir el recorrido de una variable en 4 partes o cuartiles. La diferencia entre el valor que separa el 1.er y el 4.º cuartil (*Lower* y *Upper quartile*) se denomina *recorrido intercuartílico,* tal como se muestra en el siguiente esquema:



Es fácil darse cuenta de que este parámetro amortigua el efecto de los valores extremos de la variable aleatoria, utilizándose en datos que no se distribuyen según una Normal, ya que en estos casos presenta una mayor robustez que otros indicadores de la dispersión.

Una muestra o población también se puede dividir en percentiles o, lo que es lo mismo, en 100 partes. En este caso, el grupo de datos de magnitud más baja que represente el 1% de la *n* de la muestra o población se situará en el 1.er percentil, y así sucesivamente. Los percentiles se utilizan principalmente en Ciencias Sociales.

2.3. Presentación de resultados

Siempre que se presente un resultado cuantitativo, este debe ir acompañado del error cometido en su medida. Lo más común es utilizar la media aritmética y la desviación estándar, como indicadores del valor central y la dispersión. Sin embargo, puesto que en ocasiones podemos utilizar otros estimadores, se debe especificar qué datos estamos presentando.

Por ejemplo, antes de expresar un resultado mediante la media aritmética y la desviación estándar relativa en porcentaje:

102±1

se debe indicar que los resultados se presentan como:

media ± desviación estándar relativa (%)

con el objeto de evitar confusiones. También es adecuado indicar el número de datos utilizados en los cálculos entre paréntesis.

A la hora de expresar este tipo de resultados también es muy importante utilizar las cifras que sean significativas. Supongamos que la calculadora nos da la siguiente media y desviación estándar a partir de 5 réplicas:

Media = 3.684653748 Desviación estándar = 0.745364763

Sería absurdo presentar estos valores, ya que en la medición hemos cometido error en la primera cifra decimal, por lo que debemos redondear hasta la primera cifra insegura, en este caso:

3.7±0.7

El redondeo de números acabados en 5 se debe hacer hacia el número par más cercano para evitar un sesgo siempre en el mismo sentido:

3.5→4 6.5→6

En estos ejemplos se ha utilizado el punto como separador de decimales según la norma anglosajona; en la española se deben separar por una coma. Algunos programas permiten especificar la opción que se prefiera, aunque la mayoría utilizan la norma anglosajona.

Por último, indicar que como norma general se debe redondear el resultado final, y no los resultados intermedios de los cálculos.

2.4. Propagación de errores aleatorios

Cuando calculamos el error cometido en una medición directa es fácil obtener una estimación del mismo mediante los indicadores estudiados anteriormente. Sin embargo, en el caso de que haya implicadas ecuaciones que incluyan parámetros con errores asociados, el problema se complica. Para conocer el error total cometido en estos casos, se debe estudiar la propagación de los errores a través de los cálculos realizados. A continuación, se presentan las fórmulas que permiten el cálculo del error final para diferentes tipos de ecuaciones.

• En el caso de combinaciones lineales:

$$y = k + k_1 \cdot a + k_2 \cdot b + \ldots\ldots (k_i = constante)$$

la varianza de y, es decir, la varianza de una suma o diferencia de medidas independientes, tiene la propiedad de ser igual a la suma de las varianzas:

$$s_y = \sqrt{(k_a \cdot s_a)^2 + (k_b \cdot s_b)^2 + (k_c \cdot s_c)^2 + \ldots}$$

Por ejemplo, la adición de un volumen de valorante con una bureta implica la lectura del volumen antes y después de su adición. Supongamos que ajustamos a la marca 0 el volumen de líquido en la bureta para comenzar una valoración; después de la valoración leemos en la escala 20.5 ml, el volumen resultante en este caso es la diferencia 20.5-0=20.5 ml. Si la precisión de la bureta es de 0.01 ml, la desviación estándar sería:

$$s = \sqrt{(0.01)^2 + (0.01)^2} = 0.014$$

• La varianza de ecuaciones de la forma:

$$y = \frac{k \cdot a \cdot b}{c \cdot d}$$

se puede calcular mediante la fórmula:

$$\frac{s_y}{y} = \sqrt{\left(\frac{s_a}{a}\right)^2 + \left(\frac{s_b}{b}\right)^2 + \left(\frac{s_c}{c}\right)^2 + \left(\frac{s_d}{d}\right)^2}$$

En este caso, la suma de los cuadrados de las desviaciones estándar relativas de las medidas independientes es igual al cuadrado de la desviación estándar relativa del resultado.

Calculemos *y* para valores de desviación estándar relativa de las medias independientes: *a=0.1, b=2, c=0.8, d=1*. El resultado sería:

$$\frac{s_y}{y} = \sqrt{0.1^2 + 2^2 + 0.8^2 + 1^2} = 2.4$$

es decir, ligeramente superior a la desviación estándar relativa de la medida que presenta una mayor imprecisión. Esto nos indica que para conseguir un resultado *y* más preciso debemos intentar mejorar la precisión de la medida con mayor error en su determinación (en este caso *d*), ya que esto mejorará la precisión del resultado final.

• En el caso de ecuaciones más complejas de la forma general:

$$y = f(x)$$

el error de *y* se puede calcular mediante *derivación logarítmica*. Veamos un ejemplo: deseamos determinar la superficie de un círculo y conocemos su radio R, magnitud que llevará asociada el error cometido en su medición. La fórmula es:

$$S = p \cdot R^2$$

puesto que el cuadrado de R es el producto de dos variables que no son independientes, no podemos utilizar la fórmula para expresiones multiplicativas.

En primer lugar, debemos sacar logaritmos neperianos:

$$\ln S = \ln p + 2 \ln R$$

A continuación derivamos:

$$\frac{dS}{S} = \frac{d\pi}{\pi} + \frac{2dR}{R}$$

Puesto que π es una constante, la expresión nos queda:

$$\frac{dS}{S} = \frac{2dR}{R}$$

o sea, el error relativo de S es igual a $2/R$ (la derivada de $2 \cdot \ln R$) multiplicado por el error de R. La expresión general sería:
dada una función $z = f(x,y,....)$
el error de z será:

$$\varepsilon_z = \left[\frac{df}{dx}\right] \cdot \varepsilon_x + \left[\frac{df}{dy}\right] \cdot \varepsilon_y + ...$$

En algunas ocasiones, también se puede derivar simplemente la función para poder calcular el error propagado.

Respecto a los errores sistemáticos, debido a que tienen un signo definido, la propagación es similar a la de los errores aleatorios pero cambiando los diferenciales por incrementos con signo. En los siguientes capítulos se muestra la manera de detectar y eliminar este tipo de errores.

2.5. Distribución de las variables aleatorias

Habitualmente, los estimadores del valor medio y de la dispersión que se utilizan son la media aritmética y la desviación estándar relativa. Aunque la desviación estándar proporciona una medida de la dispersión de un conjunto de resultados alrededor de la media, no indica la forma en la que están distribuidos. Para determinar la distribución de los valores que toma una variable aleatoria, es necesario conocer un gran número de datos. Estos se pueden representar situando en el eje de abscisas las clases, es decir, los intervalos en los que se pueden agrupar los datos, y en el eje de ordenadas las frecuencias de clase, o lo que es lo mismo, el número de veces que aparecen los valores incluidos en los intervalos seleccionados.

Veamos un ejemplo: se analiza la concentración de Pb en una muestra de agua 40 veces, los resultados son los siguientes (en ng/mL):

41.5 41.3 41.7 41.8 41.0 41.1 41.1 41.9 41.5 41.6 41.4 42.1 42.3 42.5 42.5 42.9 43.5 43.6 44.2 45.0 44.9 40.6 40.9 40.8 40.5 39.9 39.8 39.9 38.5 38.6 37.9 36.2 46.5 41.2 41.5 40.6 40.3 40.7 40.9 43.1

Como hemos visto anteriormente, cuando se desea representar gráficamente una población compuesta por numerosos datos, es necesario agruparlos en intervalos o «clases», representando el número de veces que aparecen los valores que pertenecen a cada clase (frecuencia de clase). En el ejemplo, la tabla de clases y frecuencias sería:

Clase	frecuencia
36	1
37	1
38	2
39	3
40	8
41	13
42	5
43	3
44	2
45	1
46	1

En la figura 2.2 se muestra la representación de la frecuencia de clase para cada intervalo.

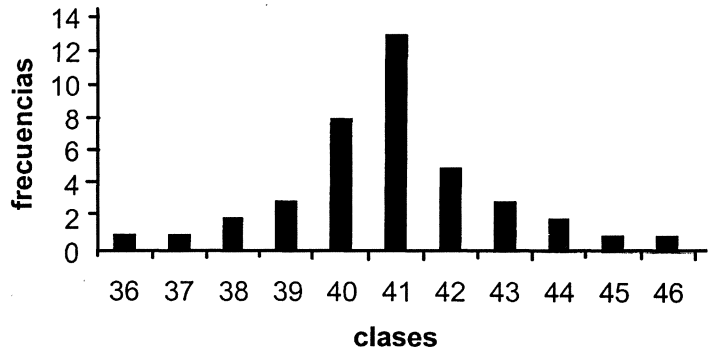


FIGURA 2.2. *Representación de los datos en clases* vs. *frecuencias*

Este ejemplo nos permite comentar algunos aspectos sobre el tratamiento de datos. La existencia de factores aleatorios en el proceso experimental hace que las réplicas realizadas varíen con una cierta amplitud alrededor del valor verdadero de concentración de Pb en la muestra de agua.

Se deben tener en cuenta los siguientes aspectos:

- Cuanto mayor sea el número de réplicas realizadas *(n)*, más nos aproximamos al valor verdadero, es decir, la estimación del valor central coincidirá en mayor medida con la verdadera concentración.

- No es práctico realizar 100 réplicas de un análisis. En la realidad nosotros realizaremos 4, 5 ó 6 réplicas a lo sumo a partir de las cuales calcularemos la media aritmética y la desviación estándar.
- No es recomendable realizar sólo 3 réplicas ya que, en ese caso, un valor anómalo es difícilmente reconocible.
- Cualquier dato experimental cuantitativo lleva intrínseca una incertidumbre que se debe presentar junto a él. La manera más usual es presentar la media y la desviación estándar.

Si consideramos una *n* que tiende a infinito y que la amplitud de cada intervalo de clase tiende a 0, obtenemos la curva representada en la figura 2.3, que constituye la denominada *función de densidad de probabilidad.* En la mayoría de los casos se obtiene la **distribución Normal o Gaussiana,** ya que es la que siguen los valores de las variables aleatorias estudiadas en gran parte de los ensayos experimentales. En general, cuando las variaciones se deben a factores aleatorios obtendremos una distribución Normal. Un ejemplo claro es el expuesto anteriormente. En un análisis químico, los resultados obtenidos analizando repetidas veces la misma muestra varían respecto al valor verdadero debido a los errores aleatorios intrínsecos a cualquier procedimiento experimental.

Cuando se representan las frecuencias relativas en el eje de ordenadas, el área bajo la función de densidad de probabilidad (f(x)) es igual a 1, ya que representa la

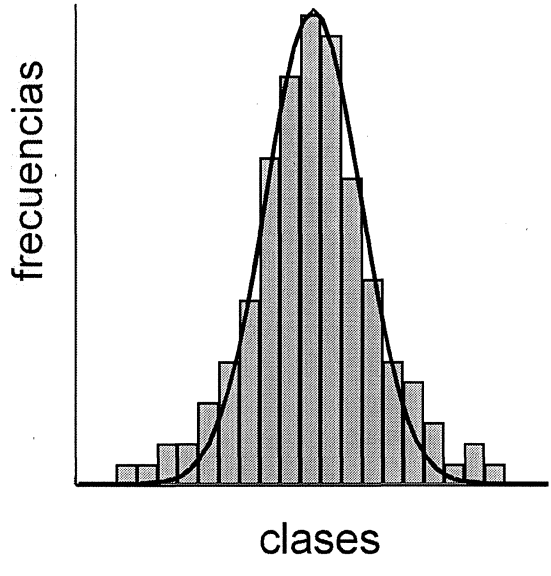


FIGURA 2.3. *La Distribución Normal*

probabilidad de que la variable tome cada uno de los valores posibles representados en el eje de abscisas, o lo que es lo mismo:

$$\int_{-\infty}^{+\infty} f(x) \cdot dx = 1$$

La probabilidad de que una variable aleatoria se encuentre entre los valores x_1 y x_2 será:

$$P(x) = \int_{x_1}^{x_2} f(x) \cdot d(x)$$

La deducción de la función f(x) es compleja y cae fuera de los objetivos de este libro, por lo que, directamente, la curva de distribución normal f(x) se puede expresar de la siguiente manera:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Donde μ es el valor medio y σ la desviación estándar de la función de probabilidad. Cuando una variable x sigue una distribución Normal de media μ y varianza σ^2 se describe como:

$$x \rightarrow N(\mu, \sigma^2)$$

En general, las letras griegas μ y σ se utilizan para representar los parámetros de la población (valor verdadero y dispersión como desviación estándar), mientras que los signos x y s se utilizan cuando nos referimos a una muestra estadística (media muestral y desviación estándar muestral).

Como se ha indicado anteriormente, esta es la distribución que seguirán los valores de gran parte de las variables aleatorias que podemos obtener en estudios experimentales.

En la figura 2.4. se muestran diferentes formas de la distribución Normal según la magnitud de μ y σ.