# BIOESTADÍSTICA
## SIN DIFICULTADES MATEMÁTICAS

## En busca de tesoros escondidos

**Luis Prieto Valiente**
**Inmaculada Herranz Tejedor**

# BIOESTADÍSTICA
## SIN DIFICULTADES MATEMÁTICAS

## En busca de tesoros escondidos

Análisis estadístico de datos en
investigación médica y sociológica

---

Para estudiantes y profesionales
Incluye 500 ejercicios y 600 preguntas de autoevaluación
Y ANEXO CON TODAS LAS SOLUCIONES


Madrid - Buenos Aires - México, D.F. - Bogotá

Ediciones Díaz de Santos

www.diazdesantos.es/ediciones (España)
www.diazdesantos.com.ar (Argentina)

ISBN: 978-84-7978-959-6
Depósito legal: M. 8.259-2010

Diseño de Cubierta: P55 Servicios Culturales
Fotocomposición: Estefanía Grimoldi
Impresión: FER Fotocomposición, S. A.

Este libro no hubiera sido posible
sin la presencia decisiva de:

*Blanca, Md, PhD.*
*Ana Belén y Luchi.*
*Jaime y Alfonso.*
*José Luis y Araceli.*

# Índice

# Introducción

## ¿QUÉ ES EL ANÁLISIS ESTADÍSTICO DE DATOS?

*Del salón en el ángulo oscuro,*
*de su dueño tal vez olvidada,*
*silenciosa y cubierta de polvo,*
*veíase el arpa.*
*¡Cuánta nota dormía en sus cuerdas,*
*como el pájaro duerme en la rama,*
*esperando la mano de nieve*
*que sabe arrancarla!*
*¡Ay —pensé— cuántas veces el genio*
*así duerme en el fondo del alma*
*y una voz como Lázaro espera*
*que le diga "levántate y anda"!*

### 1. El tesoro escondido

Quizá la más bella rima de Bécquer es un canto a los maestros en general, a todo el que ayuda a otro a saber más o hacer mejor, cuyo paradigma sería Ana Sullivan. Pero también alude a todos los que buscan y encuentran tesoros escondidos, como los especialistas en Análisis Estadístico de Datos, AED.

Los submarinistas bucean en los mares y encuentran galeones hundidos hace siglos con cofres repletos de joyas. Los buscadores de Alaska criban la arena de los ríos propicios y encuentran pepitas de oro. Miguel Ángel y Rodin separan del bloque de mármol los trozos periféricos no necesarios y encuentran *La Piedad, El David, El Pensador, El Beso*.

Los expertos en AED bucean en las bases de datos, criban grandes listas de números, desechan los no necesarios y encuentran relaciones no conocidas hasta ese momento, tales como:

- en los fumadores fallan más veces los implantes dentarios;
- los masajes oculares retrasan la aparición de "vista cansada";
- la aspirina disminuye notablemente el riesgo de enfermedad cardiovascular.

Por ejemplo, el archivo de Historias Clínicas de un gran hospital que tiene 200 000 expedientes y unos 300 datos de interés en cada uno, contiene 60 millones de datos. Esa gran colección de números encierra información muy útil, pero solamente usando las herramientas adecuadas se pueden desenterrar los tesoros escondidos en ella. La función del AED es sacar a la luz la información inmersa en bases de datos como los archivos de hospitales, censos de población, y registros de resultados de encuestas y estudios de todo tipo.

## 2. Los mismos métodos en todas las aplicaciones, la misma asignatura en todas las carreras universitarias

Todas las encuestas que evalúan las opciones políticas que elegirán los votantes, los productos que prefiere el consumidor, el alivio del dolor que produce un fármaco, el perjuicio que para la salud supone el sedentarismo… son analizadas con los mismos métodos.

Así, para calcular la ***media*** aritmética de la edad de un grupo de personas o de la puntuación obtenida en un test de inteligencia, el procedimiento es el mismo: sumar las cantidades de todas las personas y dividir por el número de ellas. Y para calcular el ***porcentaje*** de matrimonios que se divorcian o de accidentes de tráfico en que el conductor iba ebrio, se divide el número de casos en que se da esa característica por el total de ellos y se multiplica por cien.

El AED es básicamente el mismo en todas las áreas: medicina, sociología, economía, control de calidad... Por ello, en Occidente esta disciplina se imparte como obligatoria en prácticamente ***todas*** las licenciaturas universitarias y aunque el nombre varía según la carrera, en todas ellas el contenido es esencialmente el mismo.

Conocer los fundamentos lógicos, que no matemáticos, del AED es una necesidad para el científico y el profesional de este siglo, porque ello le permite entender aspectos importantes de la información que son inaccesibles al que ignore esos conceptos básicos. Al aprender estos métodos se está adquiriendo una herramienta útil en todos los campos de investigación.

## 3. Una herramienta cada vez más necesaria en la investigación científica, la gestión empresarial y el control de calidad

El auge del AED en los últimos decenios se debe a dos hechos:

a) El creciente desarrollo tecnológico permite generar información detallada masiva sobre los más diversos temas de interés.
b) El desarrollo de la informática permite analizar esa información con rapidez, precisión y bajo costo, inimaginables hace 40 años.

Aunque de lo dicho se desprende que el AED es herramienta necesaria para analizar la información que continuamente se genera en todas las ramas de saber y de la actividad social, son especialmente clarificadoras las palabras de Deming, padre del milagro industrial japonés: "Ningún recurso es tan escaso en las empresas como el conocimiento estadístico. No hay conocimiento que pueda contribuir tanto a mejorar la calidad, la productividad y la competitividad de las empresas como el de los métodos estadísticos". E Ishikawa, otra de la figuras claves en el desarrollo industrial nipón dice: "Las herramientas estadísticas básicas deben ser conocidas y utilizadas por todo el mundo en una empresa, desde la alta gerencia hasta los operarios en las líneas" (R. Romero y L Zúñiga. *Métodos Estadísticos en Ingeniería*. Ed. Uni. Politécnica de Valencia. 2005. Pág. 3).

## 4. Los investigadores y profesionales aceptan con sumisión veredictos que no entienden

Más del 90% de los profesionales desconocen los conceptos básicos (de naturaleza lógica, no matemática) manejados en el AED, a los que se enfrentan ineludiblemente cuando leen informes técnicos y cuando tienen que comunicar sus propios resultados. Los cálculos del AED son para ellos una caja negra de la que salen números o frases incomprensibles que se ven obligados a usar por imperativo de la comunidad científica. Y creyendo que para entender esos conceptos necesitan habilidades matemáticas que no tienen, se resignan a no entenderlos y se refugian en recetas simplistas, repitiendo una y otra vez frases hechas que estorban más que ayudan, pues no las entiende ni quien las enuncia ni quien las lee. Ejemplos típicos son:

"La muestra de N = 250 es estadísticamente representativa"
"El resultado es estadísticamente significativo ($P < 0.05$)".

Estas son las estaciones del vía crucis que sufren muchos profesionales cuando planean un estudio, publican sus resultados o leen los publicados por otros investigadores

1º En primer lugar tiene que decidir el tamaño de la muestra que va a estudiar. Cree que hay fórmulas matemáticas para ello, las busca con esmero, no consigue entender cómo se aplican y suele tardar decenios en descubrir que en la decisión del tamaño de la muestra esas fórmulas tienen escaso, muchas veces nulo, peso.
2º Cualquier efecto de interés que encuentre en la muestra, por ejemplo, que una medicina cura más que otra, puede ser un artefacto debido a la Confusión creada por una tercera variable. Detectar factores de confusión es tarea obligada en toda investigación, pero para la mayoría de los investigadores es tarea imposible porque ignoran ese concepto.

3º Una vez identificado el resultado muestral correcto, hay que ver cuánta evidencia constituye a favor de la hipótesis investigada. De ello nos informa el valor P, pero el investigador no sabe interpretarlo e intenta usarlo para tomar decisiones donde no hay nada que decidir.

4º Finalmente, los investigadores ignoran incluso la existencia de la Inferencia Bayesiana, que debería formar parte esencial del análisis estadístico, no como alternativa, sino como complemento de los métodos clásicos.

Pero este estado de desconocimiento e insatisfacción puede ser revertido. Podemos revelarnos contra él y cambiar la caja negra por un pequeño conjunto de ideas claras que permitan interpretar correctamente los resultados del AED.

Este libro pretende colaborar a poner fin a tanta desinformación, enseñando a los investigadores esos conceptos básicos, de modo que puedan actuar cómoda y correctamente donde ahora actúan con inseguridad y muchas veces con error.

## 5. Una disciplina matemática al servicio de y entendible por los profesionales sin conocimientos matemáticos

El AED es, en sus fundamentos teóricos y en sus aplicaciones prácticas, una disciplina matemática, pero está al servicio de profesionales no matemáticos y *puede ser entendida y usada con eficiencia sin entrar en las razones matemáticas que la sustentan*. Los razonamientos lógicos del AED son los mismos que utiliza el hombre de la calle en la actividad cotidiana. No hay ningún proceso mental sofisticado que sea propio del análisis estadístico. Por ello pueden entenderlo todas las personas que se lo propongan, cualquiera que sea su formación previa. Ronald Fisher, la figura más destacada de la estadística teórica y aplicada de todos los tiempos dice:

"Las conclusiones lógicas que siguen a los cálculos estadísticos… son una cuestión exclusiva de la capacidad pensante de los humanos. Todas las personas inteligentes están igualmente capacitadas para ello y los estadísticos no tienen especial autoridad en ese aspecto" (Fisher, R. *The design of experiments*. Hafner Press. N.Y. 1935. Pág. 1-2).

## 6. El 5% de los métodos cubren el 80% de las necesidades

En este libro se ven los conceptos fundamentales y las técnicas más sencillas del AED, que integran el "Análisis Elemental". Este *Análisis Elemental* constituye del orden del 5% de esta disciplina, pero cubre más del 80% de los análisis requeridos en las investigaciones habituales.

Un símil fiel sería el número de palabras que se usan habitualmente en un idioma. El castellano, por ejemplo, tiene más de cien mil términos, pero con el 3% de ellos (unos 3 000) se confecciona más del 80% de las conversaciones habituales.

El AED contiene cientos de técnicas, algunas muy sofisticadas y potentes, que se aplican para resolver cuestiones específicas de gran interés. Pero la inmensa

mayoría de los análisis requieren en primer lugar el uso del *Análisis Elemental* y muchos estudios quedan básicamente completados con esas técnicas básicas.

## 7. El AED usa valores medios

El AED opera calculando valores *medios* de las variables implicadas en ciertos grupos de individuos. En el AED no interesa el valor que la variable toma en cada individuo, sino la *media* del grupo. Así, para diagnosticar y tratar correctamente a un enfermo, el médico debe saber cuanto fuma y su tensión arterial, TA. Pero para saber si fumar aumenta la TA, el médico tiene que conocer la media de la TA en el grupo de fumadores y el grupo de no fumadores.

Muchas de las frases con que describimos los más variados aspectos del mundo en torno se refieren a valores *medios*, aunque no lo digan explícitamente. Decimos "Fumar incrementa la TA", aunque haya fumadores con TA inferior a algunos no fumadores, porque la *media* de la TA es mayor en el colectivo de fumadores que en el de no fumadores. Decimos "Comer poco alarga la vida en los ratones", aunque alguno de ellos con dieta escasa haya muerto antes que otros con dieta abundante, porque la *media* de supervivencia fue mayor en el primer grupo que en el segundo. Del mismo modo, muchas de las afirmaciones que describen relaciones en el mundo vivo y el inerte se refieren a medias, aunque no se hace explícito. Y calculando promedios de unas variables en distintos grupos de individuos el AED puede detectar relaciones no conocidas previamente.

En el caso de variables tipo "sí" o "no" se evalúa la *proporción* de individuos con esa característica en los grupos de interés. Así, decimos "La obesidad favorece la diabetes", aunque haya obesos sin diabetes y delgados con ella, porque la *proporción* de diabéticos es mucho mayor en el colectivo de obesos que en el de no obesos.

## 8. Estadística Descriptiva e Inferencia Estadística

Pero cada vez que analizando unos datos concretos se detecta en ellos una relación interesante nos preguntamos si el hallazgo encontrado en la muestra analizada es una verdad general, válida para toda la población. Por ejemplo, si en la muestra analizada se encuentra que la media de la TA es mayor en el grupo de sedentarios que en el de deportistas, la cuestión es: ¿ocurre eso en la población general o es una anécdota de nuestra muestra? Puesto que generalmente observamos ***muestras*** que son solo una pequeña parte de la ***población*** que intentamos conocer, esa pregunta es constante en todos los campos de la actividad científica, del análisis de encuestas y del control de calidad. Por ello en el AED se distinguen dos fases claramente diferenciadas y complementarias:

- La *Estadística Descriptiva* describe las relaciones de interés en las muestras que estamos explorando, calculando medias y proporciones en ellas.
- La *Inferencia Estadística* valora en qué medida las relaciones encontradas en nuestros datos ocurren también en la población general.

## 9. La Estadística Descriptiva descubre las relaciones en la muestra

Calcula *medias* y *proporciones* en distintos subgrupos de individuos de las muestras analizadas. Un ejemplo típico sería: Entre 600 pacientes con cierta enfermedad tratados con el fármaco "**A**" curaron el **32%**, y entre otros 600 con la misma enfermedad tratados con el fármaco "**B**" curaron **50.3%**. La interpretación de estos resultados es obvia. Pero en otros casos la interpretación correcta de los resultados puede requerir reflexión cuidadosa, como se ve en estos ejemplos:

En una ciudad de 100 000 habitantes practican deporte (PD) 1 000, es decir, el 1%. Tras una campaña animando a los vecinos a que PD se consigue que lo hagan otros 1000. El alcalde dice que el número de personas que PD se ha duplicado, es decir, ha aumentado el **100%**. Pero la oposición dice que antes de la campaña eran sedentarias 99% y tras ella lo son 98%, es decir, el sedentarismo bajó solamente un 1%. Son dos enfoques muy diferentes de una misma realidad.

Como otro ejemplo veamos los resultados de un estudio que analiza el porcentaje de curaciones, "%+", obtenidas con cada uno de dos tratamientos, 'A' y 'B', para la misma enfermedad.

| | **A** | | | **B** | | |
|---|---|---|---|---|---|---|
| | Total | + | %+ | Total | + | %+ |
| Mujeres | 100 | 80 | **80**% | 500 | 300 | **60**% |
| Varones | 500 | 110 | **22**% | 100 | 2 | **2**% |
| Total | 600 | 190 | **32**% | 600 | 302 | **50.3**% |

En mujeres "A" cura el 80% y "B" cura el 60%. En varones "A" cura el 22% y "B" cura el 2%. Por tanto, "A" es más efectivo que "B" tanto en mujeres como en varones. Pero si no se considera el sexo y se comparan los resultados en "personas" (última fila de la tabla), se ve que "A" es peor que "B". Estos datos dejan confuso al investigador. Se pregunta cuál de los dos tratamientos es, según estas muestras, más efectivo, pues parece haber argumentos a favor de uno y a favor del otro. Aclarar este tipo de "contradicciones" es esencial para interpretar correctamente los datos y uno de los objetivos de la Estadística Descriptiva.

## 10. La Inferencia Estadística se basa en el cálculo de probabilidades

En biología, medicina, estudio de mercados y sociología hay pocas certezas. Las actitudes correctas no garantizan la respuesta buscada, solo aumentan la probabilidad de que ocurra. Suele pensarse que el cálculo de probabilidades implica notable complejidad matemática, pero la realidad es que en sus aplicaciones más frecuentes solo contiene cálculo matemático muy elemental y razonamiento lógico al alcance de todas las personas. Veamos con cuatro ejemplos algunas de sus aplicaciones.

1. El ciudadano de a pie observa admirado la vertiginosa ascensión de algunos personajes del mundo de las finanzas. Pero muchos de esos triunfadores no tienen ninguna habilidad especial, son solamente la consecuencia ne-

cesaria de las leyes del azar. Sea un millar de ejecutivos que teniendo un capital inicial de 100 000 € y mucha ambición deciden arriesgarlo todo en operaciones donde hay probabilidad 0.50 de duplicar su dinero en un año y 0.50 de perderlo todo. En el primer intento unos 500 de ellos perderán su capital inicial y los otros 500 lo duplican. Si estos últimos lo invierten todo en otra operación similar, 250 perderán todo y otros 250 tendrán 400 000 €. Si estos 250 repiten el proceso, unos 125 tendrán ahora 800 000 €. Hasta aquí los "supervivientes" se han arriesgado tres veces. El cálculo de probabilidades nos permite saber que si continúan haciéndolo hasta 8 veces, aproximadamente 4 de los 1000 iniciales van a tener éxito en los 8 intentos y su capital llegará a **25 600 000 €**, por azar e independientemente de sus habilidades. Son los triunfadores del momento. El público los mira con admiración. Pero ellos no han hecho nada especial. Todo el que conoce el cálculo elemental de probabilidades sabía desde el principio que 4 de ellos, aunque no se supiera cuales, superarían las 8 primeras apuestas.

2. En una ciudad de 4 millones de habitantes la enfermedad "E", que es mortal si no se diagnóstica a tiempo, afecta a 4 000 de ellos. Se pone en marcha un método diagnóstico precoz que da positivo en el 99% de los enfermos, pero también da positivo (equivocadamente) en el 20% de los sanos. Se somete a ese método a todos los ciudadanos y en los que da positivo son convocados a un estudio más completo para averiguar si realmente tienen E. La mayoría de los convocados están muy asustados e incluso han hecho testamento, pues asumen que tienen alta probabilidad de tener E. Pero los pocos que conocen el cálculo elemental de probabilidades saben que su probabilidad de tener E es solamente 5 por mil.
3. En un famoso concurso de la TV el presentador muestra al concursante tres cajas iguales cerradas. Dos están vacías, la otra contiene 20 000 €. El concursante elige una y si es la del dinero se lo queda. Tras la elección del concursante, el presentador (que sabe cuál es la caja buena), para dar más animación al proceso señala una de las cajas no elegidas, le dice que esa no es la buena y le pregunta si quiere reafirmarse en su elección inicial o prefiere cambiar a la otra. El concursante se debate en la duda. Se pregunta si elegir la otra caja supone aumentar, disminuir o dejar igual la probabilidad de acertar, pero no sabe la respuesta a esa pregunta. Solamente la saben los que conocen el cálculo elemental de probabilidades.
4. Como último ejemplo de Inferencia Estadística consideremos el caso de una enfermedad para la que no hay tratamiento y cura espontáneamente en el 20% de los casos. Un investigador cree que el producto "A" cura más del 20%, lo prueba en 8 enfermos y se le curan todos. Muchos miembros de la comunidad científica objetarán que tan buen resultado puede haberse presentado por azar, sin que "A" sea realmente efectivo, ya que es una muestra muy pequeña. Es una objeción muy razonable, pero los que conocen el cálculo elemental de probabilidades enseguida comprueban que podemos tener gran confianza en que "A" es realmente efectivo.

## EL LIBRO

Expone las ideas fundamentales aplicándolas siempre a situaciones concretas que entienden los estudiantes y profesionales de todas las disciplinas y sin utilizar más recursos matemáticos que las operaciones aritméticas elementales, de modo que sea accesible a todos, cualquiera que sea su nivel de formación en Ciencias. La explicación de cada nuevo concepto se sigue de numerosos ejemplos y ejercicios prácticos que deben ser resueltos para afianzar esos conocimientos. El énfasis se pone en los conceptos básicos que permiten entender los principios lógicos y procedimientos más comúnmente usados. No se entra en detalles que se usan con menor frecuencia y que el lector puede encontrar en libros más extensos, algunos de los cuales se reportan a continuación.

Para el lector interesado en realizar él mismo los cálculos más sencillos, se incluyen las fórmulas más usadas. Todos los conceptos fundamentales pueden entenderse sin conocerlas, pero conocerlas puede ayudar a entenderlos mejor y además permite aplicarlas a los propios datos sin depender de un programa informático.

Algunos capítulos incluyen uno o más apartados con explicaciones complementarias que amplían conceptos ya expuestos pero no son imprescindibles para seguir el hilo del razonamiento central. Su título aparece precedido por tres signos positivos (+++). El lector que quiera centrarse en las ideas básicas no necesita entrar en estos temas avanzados. Sin embargo, serán muy útiles al que desee adentrarse más en los fundamentos lógicos.

El ***Resumen*** al final de cada capítulo sintetiza las ideas fundamentales desarrolladas en él. Si el contenido de alguna frase del resumen no es claro para el lector, este debe buscar en el texto su explicación antes de pasar al siguiente capítulo.

La ***Encuesta de Autoevaluación*** de cada capítulo ayuda a repasar los conceptos vistos en él y a evaluar el nivel de conocimientos en ese tema. Se ha asimilado el tema cuando se han respondido correctamente más del 85% de las preguntas. Si la puntuación es inferior, se necesita revisar el capítulo.

El último apartado de algunos capítulos, ***Análisis con Ordenador***, propone al lector que realice con un paquete de programas estadísticos de su elección una serie de cálculos sobre la base de datos "DARIO", y le da los resultados que debe obtener. La presentación de esta base de datos se da a continuación y los datos están en la dirección de Internet www.metodologiadelainvestigacion.es.

### Contenidos peculiares del libro

Hay muchos libros —algunos muy buenos— que exponen los fundamentos del análisis de datos en ciencias biomédicas, sociales y económicas. Por ello los autores de éste se resistieron durante decenios a hacer uno más. Al fin se deciden a publicar este porque estiman que se diferencia claramente de otros en el enfoque de los siguientes temas:

– El Teorema de Bayes y su uso práctico.

- El Efecto de Confusión y su control con Análisis Estratificado.
- Las Distribuciones Estadísticas básicas.
- Los Diseños Caso-Control.
- El valor P de los tests estadísticos.
- El tamaño de Muestra.

## Lecturas sugeridas


- Armitage P, Berry G. *Statistical methods for medical researchers*. Blackwell. 1996.
- Box G.E., Hunter W.G., Hunter J.S. *Statistics for experimenters*. John Wiley. 2008.
- Martín Andrés A, Luna del Castillo JD. *Bioestadística + para las Ciencias de la Salud*. Norma-Capitel. 2004.
- Prieto L, Herranz I. *¿Qué significa 'estadísticamente significativo'? La falacia del criterio del 5% en la investigación científica*. Madrid. Díaz de Santos. 2005.
- Prieto L, Herranz I. *Bioestadística*. Madrid. Editorial universitaria. 2008.
- Romero R, Zúnica L. *Estadística*. Universidad de Valencia. 2004.
- Rothman K. *Modern Epidemiology*. Little Brown. Toronto. 1986.
- Snedecor G, Cochran WG. *Statistical Methods*. John Wiley and sons. 1982.
- García Márquez G. *Cien años de soledad*.
- Borges JL. *La Biblioteca de Babel* y *La Lotería de Babilonia*.
- Darío R. *La marcha triunfal*, *Sonatina*, *Canción de otoño en primavera*.

## LA BASE DE DATOS DARIO

**Al final de algunos capítulos se propone al lector una serie de cálculos sobre esta base de datos, para que los realice con un paquete estadístico de su elección.**

300 personas fueron tomadas al azar usando el censo de la población de Granada. En cada una se evaluó 20 variables relativas a su estado de salud, a los hábitos de vida que pueden influir en ella y a su opinión sobre cuestiones de ese ámbito. En este archivo se incluyen diez de ellas, para las 200 personas que aceptaron volver a medirse la TAD tras un día de ayuno total (salvo agua).

| **Variable** | | **Valores** |
|---|---|---|
| **Edad** | En años | |
| **Sexo** | | 1 = Varón, 2 = Mujer |
| **Deporte3** | Practica ejercicio con regularidad | 0 = No práctica, 1= Práctica moderada, 2 = Práctica intensa |
| **Dieta** | Hace dieta hipocalórica | 0 = No, 1 = Sí |
| **Yoga** | Practica yoga con regularidad | 0 = No, 1= Sí |
| **ResisInf** | Resistencia a las infecciones | 0 = Baja, 1 = Moderada,<br>2 = Media, 3 = Alta |
| **Ministro** | Qué opinión le merecería que se designase como Ministro de Sanidad a una persona ajena a todos los ámbitos de ese tema.<br>Escala de 0 (= muy mal) a 10 (= muy bien) | |
| **Sanidad** | Qué opinión tiene sobre la sanidad pública, en una escala de 0 (= muy mala) a 100 (= muy buena) | |
| **Tad** | Tensión arterial diastólica en mm de Hg, en condiciones basales | |
| **Tadpost** | Tensión arterial diastólica en mm de Hg, TRAS un día de ayuno total | |
| ***A partir de las variables anteriores genere usted las tres siguientes:*** | | |
| **Deporte2** | 0 = Sedentarios → No practica ejercicio con regularidad<br>1= Activos → practica ejercicio (son los '1' y ´2' de "Deporte3") | |
| **Diftad** | = Tadpost - Tad | |
| **Edad2** | 1 = menores de 40 años, 2 = 40 o más años | |

## ABREVIATURAS MÁS USADAS EN ESTE TEXTO

Cada una de las cuatro abreviaturas que siguen se usa cientos de veces en el libro.

| Abreviatura | Expresión completa, descripción y apartado en que se explica |
|---|---|
| **FR** | **Frecuencia Relativa** (singular o plural)<br>Relacionan el tamaño de una parte con el tamaño total de un colectivo.<br>Apartado 1.4 |
| **DF** | **Distribución de Frecuencias** (singular o plural)<br>Dice la cantidad o la FR de individuos que tienen cada valor de una variable o están incluidos en cada intervalo.<br>Apartado 1.5 |
| **Prob** | **Probabilidad** (singular o plural)<br>La FR de veces que aparece cierto resultado si un fenómeno aleatorio se repite millones y millones (un número indefinidamente grande) de veces.<br>Apartado 5.1 |
| **MA** | **Muestreo Aleatorio**<br>El hecho de extraer de una población millones y millones (un número indefinidamente grande) de muestras del mismo tamaño.<br>Apartado 6.3 |

# Capítulo 1

# ESTADÍSTICA DESCRIPTIVA. DISTRIBUCIONES DE FRECUENCIAS

## 1.1. Etapas de una investigación

En toda investigación científica pueden identificarse estas fases que siguen un orden cronológico evidente. Para llevar a cabo cada una de ellas es imprescindible que las anteriores estén correctamente realizadas.

1. Diseño del estudio
2. Recogida de la información (fase de campo)
3. Análisis de la información:
   *A. Tabulación de los datos*
   *B. Estadística Descriptiva*
   *C. Inferencia Estadística*
4. Discusión y conclusiones

**1. Diseño**

Es la planificación de todo lo que se va a hacer. Se confecciona un *Protocolo* en el que se justifica la necesidad de esa investigación, se demuestra su viabilidad y se explican los recursos y gastos que se emplearán. La elaboración del protocolo ayuda decisivamente a clarificar y ordenar las ideas, así como a prever las dificultades y prevenir sus soluciones. Además es necesario presentarlo a los comités científicos, económicos y éticos para obtener autorización y apoyo económico.

**2. Recogida de la información: recopilación de datos (fase de campo)**

La calidad de los datos es requisito imprescindible para la validez de la investigación. Es fundamental la *motivación* adecuada de todas las personas implicadas. Los "decretos" pueden conseguir que se obtenga información, pero no que esta sea de calidad. Un ejemplo muy ilustrativo es el incremento del rendimiento del

Departamento de Cálculo del Proyecto Manhattan cuando al ser encargado de ello R. Feynman consiguió motivarles informándoles del objetivo de su trabajo. Un segundo ejemplo más próximo en el espacio y en el tiempo es el Estudio Español de Malformaciones Congénitas (ECEMC), en el que desde hace 30 años más de un centenar de médicos recogen meticulosamente información sobre 200 variables de cada bebé malformado nacido en su centro y de otros tantos controles sanos.

**3. Análisis de la información**

Comienza recopilando la información más relevante en forma de Tablas. Continúa haciendo Estadística Descriptiva de lo observado. Y culmina haciendo Inferencia Estadística para ver hasta qué punto los resultados encontrados en la *muestra* pueden extrapolarse a la *población*.

**4. Discusión y conclusiones**

Se elaboran teniendo en cuenta los conocimientos previos que había sobre el tema y los resultados obtenidos en esta investigación.

Esta disciplina, Análisis de Datos, cubre el tercer punto. Los cuatro primeros capítulos tratan de la Tabulación y la Estadística Descriptiva. El resto del libro se dedica a la Estadística Inferencial.

## 1. 2. Tabulación de datos

A lo largo de la fase de campo de un estudio la información de los distintos individuos implicados se va recogiendo en *Historias Clínicas* (estudios hospitalarios), en *Libretas de Laboratorio* (investigaciones básicas con animales o cultivos) o en *Encuestas Sanitarias* (estudios epidemiológicos). Las características evaluadas en cada uno de los individuos se denominan *variables*.

El análisis de esa información comienza con la *Tabulación* de los datos, que consiste en disponerlos en tablas donde cada fila corresponde a un individuo y cada columna corresponde a una variable. El siguiente esquema representa una tabla con datos de 200 individuos en los que se ha recogido, entre otras, las variables: sexo, grupo sanguíneo, número de caries, peso y edad. Para cada individuo, cada variable toma un valor concreto. Por ejemplo, el primer individuo, identificado con el número 1 es un hombre (codificamos con "1" a los hombres y con "2" a las mujeres), tiene grupo sanguíneo "AB", tiene 3 caries, pesa 65.9 kg y tiene 18 años.

| **0BInd. nº** | **Sexo** | **Fuma** | **Grupo S.** | **. . .** | **Caries** | **TA** | **1BEdad** |
|---|---|---|---|---|---|---|---|
| **1** | 1 | 2 | AB | | 3 | 65.9 | 18 |
| **4** | 1 | 2 | A | | 3 | 65.0 | 42 |
| **8** | 2 | 2 | 0 | | 2 | 59.4 | 25 |
| **---** | --- | | --- | | --- | --- | --- |
| | | | | | | | |
| **200** | 1 | 2 | AB | | 1 | 83.7 | 49 |

En general, las variables se clasifican en:

- **Cualitativas**, si recogen alguna cualidad no numérica del individuo. Se llaman *dicotómicas* si presentan solamente dos posibilidades y *politómicas* si presentan varias posibilidades (ordenables o no). Los valores de este tipo de variables suelen registrarse con *códigos* numéricos, que no indican cantidad, sino un convenio, como por ejemplo 1 = Varón, 2 = Mujer.
- **Cuantitativas**, si recogen una información numérica. Pueden ser *discretas* si toman valores enteros, como número de hijos o número de caries (1, 2, 3,… caries) o *continuas* si pueden tomar cualquier valor dentro de un intervalo, como edad, peso, tensión arterial…

| | |
|---|---|
| CUALITATIVA | **Dicotómica.** Sexo: 1=Varón, 2=Mujer |
| | **Politómica.** Grupo Sanguíneo: 1=A, 2=B, 3=AB, 4=O |
| | **Politómica Ordenada.** Higiene: 1=Nada, 2=Poca, 3=Media, 4=Mucha |
| CUANTITATIVA | **Discreta.** Número de Caries |
| | **Continua.** Concentración de ácido úrico, Edad, TA ... |

La clasificación de los individuos en cada variable tiene que ser exhaustiva (todo individuo debe tener un grupo en el que encaje) y excluyente (cada individuo encaja solamente en un grupo).

Las tablas así creadas contienen toda la información de interés, pero en muchos casos el investigador no está interesado en el valor que toma una variable en cada individuo, sino en la proporción de ellos con cierto valor o en la media de esa variable en un grupo. Una de las variables de nuestra tabla indica si fuman o no y la otra su tensión arterial, TA. Para diagnosticar y tratar a cada sujeto el médico necesita conocer su TA y cuánto fuma. Pero para investigar si fumar favorece la aparición de hipertensión hay que formar dos grupos de personas, los que fuman y los que no, y en cada uno de ellos ver la media de la TA o la proporción de hipertensos.

Por ello el siguiente paso en el análisis de los datos es realizar la *Estadística Descriptiva*, que tiene por objeto resumir el comportamiento de cada una de las variables en el grupo estudiado y en cada subgrupo que sea de interés. Este resumen se hará de distinta forma dependiendo del tipo de variable.

En todas ellas se pueden calcular *Distribuciones de Frecuencia* y en las cuantitativas se puede, además, calcular *medias y desviaciones*.

## 1. 3. Frecuencias Relativas, FR

Las frecuencias relativas, FR, relacionan el tamaño de una parte con el tamaño total de un colectivo. Si decimos que en Oviedo hay **4 000** obesos y en Madrid hay **16 000,** es claro que en la capital de España hay muchos más casos que en la de Asturias y la cantidad de recursos necesarios para asistir a estos pacientes es mucho mayor en la primera que en la segunda.

Pero la **frecuencia relativa**, **FR**, de ese problema es mayor en Oviedo, ya que entre los 200 000 ovetenses los obesos son 4 000 / 200 000 = 2 / 100 = 0.02, es decir, dos por cien o veinte por mil. Entre los 4 millones de habitantes de Madrid los obesos son 16 000 / 4 000 0000 = 16 / 4 000 = 0.004, es decir, 4 cada mil. La tabla siguiente resume estos datos:

| | Frecuencia Absoluta | Núm. de habitantes | FR Proporción | FR Porcentaje | FR Tanto por mil |
|---|---|---|---|---|---|
| Oviedo | 4 000 | 200 000 | 0.02 | 2 % | 20 por 1 000 |
| Madrid | 16 000 | 4 000 000 | 0.004 | 0.4 % | 4 por 1 000 |

La FR se puede dar como *proporción* (por ejemplo, 0.02) o como *porcentaje* (2%). Pero si son FR muy bajas es más útil darlas como *tanto por mil* o *por diez mil* u otra cantidad pertinente. Por ejemplo, si en Madrid hay 80 hemofílicos, la frecuencia relativa es "80 entre 4 millones", es decir, 80 / 4 000 000 = 0.00002 ó 0.002%. Pero esa cifra se entiende mejor expresándola como "2 por cien mil" ó "20 por millón".

El porcentaje se calcula multiplicando la proporción por 100, es decir, corriendo la coma (hoy día el punto) dos lugares a la derecha. Si de un total de 80 personas 32 son enfermos, dividiendo 32 por 80 obtenemos la proporción de enfermos, que multiplicada por 100 nos da el porcentaje:

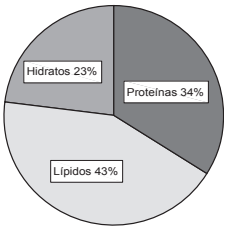
$$\text{Proporción: } 32 / 80 = 0.40 \quad , \quad \text{Porcentaje: } 0.40 \cdot 100 = 40\%$$

Tanto en Estadística Descriptiva como en la Inferencial se usa constantemente las Frecuencias Relativas. La mayoría de las veces se las refiere como "FR".

## 1. 4. Distribuciones de Frecuencias, DF

El comportamiento de una variable cualitativa en un grupo de individuos se resume dando su **Distribución de Frecuencias, DF**, que consiste en anotar la cantidad de individuos que tienen cada valor de la variable. Por ejemplo, si de cada individuo se recoge el tipo de dieta que sigue (hay tres posibilidades) la columna del medio de la siguiente tabla da la frecuencia absoluta de cada dieta:

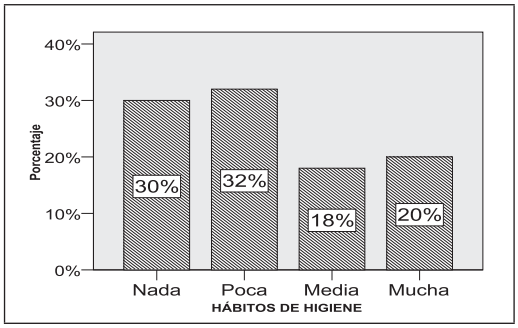
| | Frecuencia | FR |
|---|---|---|
| Proteínas | 68 | 0.34 |
| Lípidos | 86 | 0.43 |
| Hidratos | 46 | 0.23 |
| Total | 200 | 1 |



De los 200 individuos estudiados, 68 siguen dieta rica en proteínas, 86 dieta rica en lípidos y 46 dieta rica en hidratos. Estas cantidades forman la ***DF Absolutas*** (número de individuos en cada una de las categorías).

La **DF Relativas** indica la **FR** de individuos para cada valor de la variable. En nuestro ejemplo 68 / 200 = 0.34 o 34% sigue dieta rica en proteínas, 43% en lípidos y 23% en hidratos.

Estas cantidades pueden representarse gráficamente en diagramas de sectores, donde el ángulo de cada sector es proporcional a la frecuencia del valor que representa. También puede presentarse en un diagrama de barras, donde la altura de cada una es proporcional a las frecuencias de cada clase




Este diagrama de barras representa la distribución de frecuencias relativa de la variable "Hábitos de Higiene" recogida en 200 individuos: el 30% presentaba nada de higiene, el 32% poca, el 18% una cantidad media y el 20% mucha.

- Si la variable en estudio es cuantitativa con pocos valores, su comportamiento se resume también dando su DF de la misma forma que con las cualitativas. Con variables cuantitativas podemos, además, calcular las llamadas *Frecuencias Acumuladas*, que indican la cantidad o FR de individuos con valor **igual o menor** que uno dado. Por ejemplo:

| Num de Caries | Frecuencia absoluta | FR | FR acumulada |
|---|---|---|---|
| 0 | 40 | 0.20 | 0.20 |
| 1 | 10 | 0.05 | 0.25 |
| 2 | 50 | 0.25 | 0.50 |
| 3 | 80 | 0.40 | 0.90 |
| 4 | 20 | 0.10 | 1 |

El 25% (20+5) tiene una caries o menos, el 50% (25+25) tiene 2 caries o menos,...

- Si la variable en estudio es cuantitativa discreta con muchos valores o continua se pueden hacer **intervalos** y contar el número de individuos en cada uno para hacer la correspondiente DF. Los intervalos deben cubrir todo el rango de posibles valores y no solaparse. No hay razones matemáticas ni biológicas para determinar la anchura de los intervalos. El investigador agrupa los datos como sea más útil a su estudio. Una misma variable puede ser agrupada con distintos criterios en distintos momentos. Por ejemplo, con la edad pueden formarse dos grandes grupos, niños y adultos. Y en otro momento pueden agruparse como niños, jóvenes, maduros y viejos, esta-

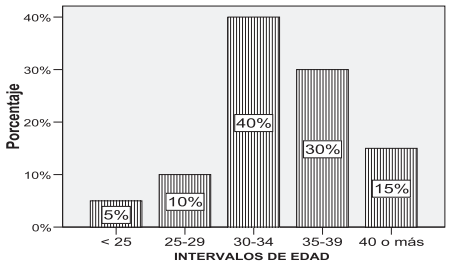
bleciendo los puntos de corte pertinentes. O pueden hacerse intervalos por decenios: de 0 a 9, de 10 a 19… de 90 a 99.

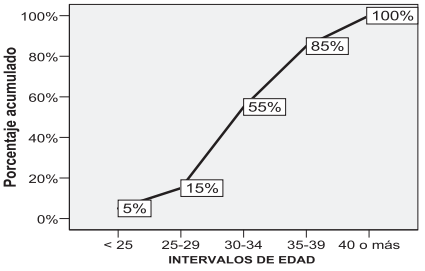
También se pueden representar las frecuencias absolutas y/o relativas de cada intervalo en diagramas de barras o sectores, y las frecuencias acumuladas en diagramas barras y de líneas.

La tabla que sigue da la DF de la edad de 200 individuos:

**INTERVALOS DE EDAD**

| | | | Porcentaje |
|---|---|---|---|
| Menos de 25 años | 10 | 5,0 | 5,0 |
| 25-29 años | 20 | 10,0 | 15,0 |
| 30-34 años | 80 | 40,0 | 55,0 |
| 35-39 años | 60 | 30,0 | 85,0 |
| 40 o más años | 30 | 15,0 | 100,0 |
| Total | 200 | 100,0 | |




La DF acumulada —columna a la derecha— dice que el 15% de los individuos tiene 29 o menos años, el 55% tiene 34 o menos y el 85% tiene menos de 40.

Muchos investigadores tienden a tabular las variables continuas acordando una partición en intervalos y registrando para cada individuo el intervalo al que pertenece. Ello supone una notable pérdida de información que debe evitarse. Si, por ejemplo, la edad se sustituye por la década a la que pertenece, un paciente de 41 años es indistinguible de uno de 49, lo que disminuiría la posibilidad de detectar relaciones de interés. Por ello lo razonable es tabular los valores originales y en la fase de análisis agrupar por los intervalos que interesan en cada momento.

## 1. 5. Percentiles

Asociado al concepto de FR Acumulada está el de **Percentil**. En una DF de una variable cuantitativa el percentil que le corresponde a cada valor es el porcentaje de individuos con valor igual o menor que aquel. En la DF del número de caries, 1 carie es el percentil 25, 2 caries es el percentil 50,... En la DF de la edad, 39 años es el percentil 85. Si en la DF de la estatura de un colectivo, 177 cm es el percentil 90, quiere decir que el 90% de los individuos mide 177 o menos.

La idea de percentil es muy útil al ubicar un individuo en el contexto del grupo al que pertenece y esta ubicación relativa suele ser más útil que el valor numérico original. A unos padres preocupados por el escaso desarrollo del niño no les ayuda saber que en un test de capacidad mental alcanzó **220** puntos. Inmediatamente preguntan cuáles son las puntuaciones propias de los niños normales. Lo que necesitan es saber en qué percentil está su hijo. Si resulta ser el percentil **3**, es mala noticia, puesto que solo el 3% de los niños de esa edad tienen esa capacidad o menos. Si 220 puntos es el percentil 47 o el 55, es una noticia tranquilizadora, puesto que el niño está en la zona media. Y si 220 puntos es el percentil 99, de cada cien niños solo uno le supera. Los percentiles se usan constantemente en la Inferencia Estadística.

A un paciente no le dice mucho saber que su tensión arterial es 120. Lo útil para él es saber a qué percentil corresponde esa cifra. Si es el percentil 97 quiere decir que de su sexo y edad, solo 3 de cada 100 le superan. Es una cifra *relativamente* muy alta, y por ello preocupante.

En muchas ocasiones para analizar los datos hay que calcular "proporción de una proporción". Si, por ejemplo, en un colectivo son enfermos el 10% y el 50% de los enfermos han sido operados, es claro que los operados son el 5% del colectivo (la mitad del 10%), pero con cifras menos sencillas se requiere un cálculo con el que algunos investigadores no están familiarizados. La siguiente sección lo explica.

## 1. 6. Calculando porcentajes de porcentajes

### A. Cálculo del Total de individuos si se conoce el porcentaje

Si sabemos que de un total de 80 personas el 60% son varones, para saber la cantidad de varones se multiplica el total por la proporción: Cantidad de varones = $80 \cdot 0.60 = 48$ varones

** **Ejercicio 1.1:** En un grupo de 300 personas el 20% son fumadores. ¿Cuántos fumadores hay?

** **Ejercicio 1.2:** En un grupo de 120 personas el 30.83% son bebedores. ¿Cuántos bebedores hay?

### B. Cálculo del porcentaje de un porcentaje

El porcentaje de un porcentaje se calcula multiplicando las proporciones. Si el 60% de un colectivo son varones y el 10% de los varones tienen adenoma de próstata, ¿qué porcentaje de personas tiene adenoma de próstata? Son el 10% del 60% = $0.10 \cdot 0.60 = 0.06$ o 6%

** **Ejercicio 1.3:**

a. En otra población el 70% de los habitantes son mujeres y el 80% de ellas han parido, ¿qué porcentaje de la población ha parido?
b. Si el 40% de las mujeres que parieron lo hicieron con cesárea, ¿qué porcentaje de la población sufrió cesárea?