

# Índice

|  |           |
|--|-----------|
| <b>1. Introducción a la Estadística</b>  | <b>19</b> |
| 1.1. Introducción . . . . .  | 19        |
| 1.2. Población e individuo . . . . .   | 23        |
| 1.3. Muestras aleatorias . . . . .   | 24        |
| 1.4. Variable aleatoria y Modelo probabilístico . . . . .  | 25        |
| 1.5. Diferentes Estadísticas . . . . .   | 27        |
| 1.6. Tratamiento informático de los datos . . . . .  | 27        |
| <b>2. Estadística Descriptiva</b>  | <b>29</b> |
| 2.1. Introducción . . . . .  | 29        |
| 2.2. Conceptos fundamentales . . . . .   | 29        |
| 2.3. Distribuciones unidimensionales de frecuencias . . . . .                                    | 33        |
| 2.3.1. Representaciones gráficas de las distribuciones unidimensionales de frecuencias . . . . . | 37        |
| 2.3.2. Medidas de tendencia central . . . . .  | 43        |
| 2.3.3. Medidas de dispersión . . . . .   | 50        |
| 2.3.4. Medidas de asimetría . . . . .  | 53        |
| 2.4. Distribuciones bidimensionales de frecuencias . . . . .                                     | 54        |
| 2.4.1. Representaciones gráficas de las distribuciones bidimensionales de frecuencias . . . . .  | 58        |
| 2.4.2. Ajuste por mínimos cuadrados . . . . .  | 61        |
| 2.4.3. Precisión del ajuste por mínimos cuadrados . . . . .                                      | 63        |
| <b>3. Probabilidad</b>   | <b>69</b> |
| 3.1. Introducción . . . . .  | 69        |
| 3.2. Espacio Muestral . . . . .  | 70        |
| 3.3. Conceptos de Probabilidad . . . . .   | 73        |
| 3.4. Propiedades elementales de la Probabilidad . . . . .  | 75        |
| 3.5. Asignación de Probabilidad . . . . .  | 78        |

|   |            |
|---|------------|
| 3.6. Modelo Uniforme . . . . .  | 79         |
| 3.7. Probabilidad condicionada . . . . .  | 82         |
| 3.8. Independencia de sucesos . . . . .   | 83         |
| 3.9. Teorema de la Probabilidad Total . . . . .   | 83         |
| 3.10. Teorema de Bayes . . . . .  | 84         |
| <b>4. Modelos Probabilísticos</b>   | <b>87</b>  |
| 4.1. Introducción . . . . .   | 87         |
| 4.2. Distribución de Probabilidad . . . . .   | 88         |
| 4.3. Variables aleatorias multivariantes . . . . .  | 94         |
| 4.4. Modelos unidimensionales discretos . . . . .   | 95         |
| 4.4.1. Distribución Binomial . . . . .  | 95         |
| 4.4.2. Distribución de Poisson . . . . .  | 97         |
| 4.4.3. Distribución Geométrica . . . . .  | 98         |
| 4.4.4. Distribución Hipergeométrica . . . . .   | 99         |
| 4.4.5. Distribución Binomial Negativa . . . . .   | 100        |
| 4.5. Modelos unidimensionales continuos . . . . .   | 100        |
| 4.5.1. Distribución Normal . . . . .  | 100        |
| 4.5.2. Distribución Uniforme . . . . .  | 103        |
| 4.5.3. Distribución Beta . . . . .  | 103        |
| 4.5.4. Distribución Gamma . . . . .   | 104        |
| 4.5.5. Distribución de Cauchy . . . . .   | 104        |
| 4.6. Modelos bidimensionales . . . . .  | 104        |
| 4.6.1. Distribución Normal bivalente . . . . .  | 104        |
| 4.7. Teorema Central del Límite . . . . .   | 105        |
| 4.7.1. Teorema de Levy-Lindeberg . . . . .  | 105        |
| <b>5. Estimadores. Distribución en el muestreo</b>  | <b>107</b> |
| 5.1. Introducción . . . . .   | 107        |
| 5.2. Método de la máxima verosimilitud . . . . .  | 110        |
| 5.3. Distribuciones asociadas a poblaciones normales . . . . .                                    | 113        |
| 5.3.1. Distribución $\chi^2$ de Pearson . . . . .   | 113        |
| 5.3.2. Distribución $t$ de Student . . . . .  | 114        |
| 5.3.3. Distribución $F$ de Snedecor . . . . .   | 115        |
| 5.4. Estimación de la media de una población normal . . . . .                                     | 116        |
| 5.5. Estimación de la media de una población no necesariamente normal. Muestras grandes . . . . . | 118        |
| 5.6. Estimación de la varianza de una población normal . . . . .                                  | 120        |
| 5.7. Estimación del cociente de varianzas de dos poblaciones normales independientes . . . . .    | 121        |
| 5.8. Estimación de la diferencia de medias de dos poblaciones normales independientes . . . . .   | 123        |

---

|  |            |
|--|------------|
| 5.9. Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes . . .                          | 126        |
| 5.10. Datos apareados . . . . .  | 127        |
| 5.11. Tamaño muestral para una precisión dada . . . . .  | 128        |
| <b>6. Intervalos de confianza</b>  | <b>131</b> |
| 6.1. Introducción . . . . .  | 131        |
| 6.2. Intervalo de confianza para la media de una población normal .  | 134        |
| 6.3. Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes . . . . .  | 136        |
| 6.4. Intervalo de confianza para la varianza de una población normal   | 138        |
| 6.5. Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes . . . . .   | 140        |
| 6.6. Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes . . . . .  | 141        |
| 6.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes . . . . .        | 143        |
| 6.8. Intervalos de confianza para datos apareados . . . . .  | 144        |
| <b>7. Contraste de hipótesis</b>   | <b>147</b> |
| 7.1. Introducción y conceptos fundamentales . . . . .  | 147        |
| 7.2. Contraste de hipótesis relativas a la media de una población normal . . . . .   | 156        |
| 7.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes . . . . .                                   | 160        |
| 7.4. Contraste de hipótesis relativas a la varianza de una población normal . . . . .  | 169        |
| 7.5. Contraste de hipótesis relativas a las varianzas de dos poblaciones normales independientes . . . . .   | 173        |
| 7.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes . . . . .                                     | 178        |
| 7.7. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes . . . . . | 185        |
| 7.8. Contrastes de hipótesis para datos apareados . . . . .  | 191        |
| <b>8. Análisis de la Varianza</b>  | <b>193</b> |
| 8.1. Introducción . . . . .  | 193        |
| 8.2. Un Factor: Diseño Completamente Aleatorizado . . . . .  | 195        |
| 8.2.1. Comparaciones Múltiples . . . . .   | 200        |
| 8.3. Un Factor: Diseño por Bloques Aleatorizados . . . . .   | 202        |

|  |            |
|--|------------|
| 8.4. Un Factor: Diseño de Cuadrado Latino . . . . .  | 209        |
| 8.5. Dos Factores: Diseño Completamente Aleatorizado . . . . .   | 217        |
| <b>9. Regresión lineal simple y Correlación</b>  | <b>225</b> |
| 9.1. Introducción . . . . .  | 225        |
| 9.2. Modelo de la Regresión Lineal Simple . . . . .  | 227        |
| 9.2.1. Interpretación de los coeficientes de la Regresión Lineal   | 229        |
| 9.3. Contraste de la Regresión Lineal Simple . . . . .   | 230        |
| 9.3.1. Análisis de la variación explicada frente a la no explicada<br>por la recta de regresión . . . . .      | 231        |
| 9.3.2. Contraste de hipótesis para $\beta_1$ . . . . .   | 234        |
| 9.4. Predicción de $Y$ para $X$ dada . . . . .   | 236        |
| 9.5. Estimación de la media de $Y$ para $X$ dada . . . . .   | 236        |
| 9.6. Correlación Lineal . . . . .  | 237        |
| 9.6.1. Estimación por punto de $\rho$ . . . . .  | 237        |
| 9.6.2. Contraste de hipótesis sobre $\rho$ . . . . .   | 238        |
| 9.6.3. Intervalos de confianza para $\rho$ . . . . .   | 241        |
| <b>10. Regresión lineal múltiple y Correlación</b>   | <b>243</b> |
| 10.1. Introducción . . . . .   | 243        |
| 10.2. Modelo de la Regresión Lineal Múltiple . . . . .   | 244        |
| 10.2.1. Notación matricial y resolución con R . . . . .  | 246        |
| 10.2.2. Interpretación de los coeficientes de la Regresión Lineal  | 247        |
| 10.3. Contraste de la Regresión Lineal Múltiple . . . . .  | 248        |
| 10.3.1. Análisis de la varianza explicada frente a la no explicada<br>por el hiperplano de regresión . . . . . | 248        |
| 10.3.2. Contrastes de hipótesis sobre los coeficientes de regre-<br>sión $\beta_i$ . . . . .                   | 251        |
| 10.4. Análisis de la Correlación . . . . .   | 254        |
| 10.4.1. Correlación múltiple . . . . .   | 255        |
| 10.4.2. Correlación parcial . . . . .  | 256        |
| <b>11. Análisis de la Covarianza</b>   | <b>259</b> |
| 11.1. Introducción . . . . .   | 259        |
| 11.2. Análisis de la Covarianza para un factor: Diseño Completamen-<br>te Aleatorizado . . . . .               | 260        |
| 11.3. Análisis de la Covarianza para dos factores: Diseño Completa-<br>mente Aleatorizado . . . . .            | 266        |
| <b>12. Pruebas <math>\chi^2</math></b>   | <b>279</b> |
| 12.1. Introducción . . . . .   | 279        |
| 12.2. Contraste de bondad del ajuste . . . . .   | 281        |

|   |            |
|---|------------|
| 12.3. Contraste de homogeneidad de varias muestras . . . . .  | 289        |
| 12.4. Contraste de independencia de caracteres . . . . .  | 292        |
| <b>13. Estadística no paramétrica</b>   | <b>297</b> |
| 13.1. Introducción . . . . .  | 297        |
| 13.2. Contrastes relativos a una muestra y datos apareados . . . . .                                  | 298        |
| 13.2.1. El contraste de los signos . . . . .  | 298        |
| 13.2.2. El contraste de los rangos signados de Wilcoxon . . . . .                                     | 303        |
| 13.3. Contraste Kolmogorov-Smirnov de bondad del ajuste . . . . .                                     | 308        |
| 13.4. Contrastes relativos a dos muestras independientes . . . . .                                    | 311        |
| 13.4.1. El contraste de Wilcoxon-Mann-Whitney . . . . .   | 312        |
| 13.4.2. Contraste de Kolmogorov-Smirnov para dos muestras . . . . .                                   | 315        |
| 13.4.3. El contraste de la Mediana . . . . .  | 317        |
| 13.5. Contrastes relativos al Análisis de la Varianza . . . . .                                       | 319        |
| 13.5.1. Contraste de Kruskal-Wallis para el Diseño Completamente Aleatorizado con un factor . . . . . | 319        |
| 13.5.2. El contraste de Friedman para el Diseño por Bloques Aleatorizados con un factor . . . . .     | 322        |
| 13.6. Contraste de Spearman de independencia de variables aleatorias                                  | 324        |
| <b>14. Análisis de Datos</b>  | <b>329</b> |
| 14.1. Introducción . . . . .  | 329        |
| 14.1.1. Pasos a seguir en un Análisis de Datos . . . . .  | 330        |
| 14.2. Diagrama de hojas y ramas . . . . .   | 332        |
| 14.3. Análisis de la Simetría . . . . .   | 336        |
| 14.3.1. Gráfico de simetría . . . . .   | 337        |
| 14.3.2. Test basado en el coeficiente de asimetría . . . . .  | 341        |
| 14.4. Análisis de Normalidad . . . . .  | 342        |
| 14.4.1. Gráfico de normalidad . . . . .   | 342        |
| 14.4.2. Transformaciones Box-Cox . . . . .  | 344        |
| 14.4.3. Test basado en el coeficiente de curtosis . . . . .   | 345        |
| 14.5. Análisis de Homocedasticidad . . . . .  | 350        |
| 14.5.1. Gráfico de cajas . . . . .  | 352        |
| 14.5.2. Contrastes de igualdad de varianzas . . . . .   | 355        |
| <b>Fórmulas Estadísticas</b>  | <b>355</b> |
| <b>Tablas Estadísticas</b>  | <b>371</b> |
| <b>Bibliografía</b>   | <b>401</b> |

# Capítulo 1

## Introducción a la Estadística

### 1.1. Introducción

Con este capítulo el lector comienza un viaje que le llevará por los inmensos y sorprendentes senderos de los fenómenos naturales.

Con la Estadística podrá comprenderlos y explicarlos mejor, e incluso atreverse a hacer predicciones.

La razón de que la Estadística tenga un papel tan destacado en este ámbito es que prácticamente todos los fenómenos de la Naturaleza son de tipo aleatorio (no determinístico) siendo su lenguaje la Estadística puesto que, aunque las *palabras* que aquélla utiliza sean de índole matemático, las entiende articuladas como Estadística.

En este viaje aprenderá a describir los resultados de los fenómenos naturales con la *Estadística Descriptiva* y, lo que será mucho más interesante, con la *Inferencia Estadística* aprenderá a entenderla y predecirla, comparando por ejemplo grupos por ella formados, o estimando los valores más verosímiles que ella creó, u obteniendo intervalos en donde con gran confianza se encuentre un valor característico de la población analizada y, todo ello, midiendo y controlando nuestros posibles errores en términos de *Probabilidades*.

Éstas son las tres grandes partes que componen lo que habitualmente se denomina Estadística: Estadística Descriptiva, Cálculo de Probabilidades e Inferencia Estadística, siendo esta última, sin ninguna duda, la más interesante.

#### Ejemplo 1.1

---

Muchas teorías sobre la esquizofrenia sugieren alteraciones en la actividad de una sustancia del sistema nervioso central denominada *dopamina*. Con objeto de analizar esta hipótesis se eligieron al azar 15 personas diagnosticadas de esquizofrenia (grupo de *psicóticos*) y 15 personas aparentemente normales (grupo de no *psicóticos*). A todas ellas se les extrajo una muestra de fluido cerebro-espinal y se anotó la actividad de la enzima *dopamina b-hidroxilasa* (DBH) obteniéndose los siguientes datos, en donde las unidades vienen expresadas en  $\text{nmol}/(\text{ml})(\text{h})/(\text{mg})$  de proteína:

|                      |        |        |        |        |        |        |        |        |
|----------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| <i>No psicóticos</i> | 0'0165 | 0'0145 | 0'0160 | 0'0130 | 0'0156 | 0'0114 | 0'0056 | 0'0123 |
|                      | 0'0133 | 0'0101 | 0'0147 | 0'0144 | 0'0136 | 0'0148 | 0'0135 |        |
| <i>Psicóticos</i>    | 0'0222 | 0'0245 | 0'0320 | 0'0350 | 0'0234 | 0'0203 | 0'0159 | 0'0165 |
|                      | 0'0266 | 0'0236 | 0'0178 | 0'0298 | 0'0213 | 0'0277 | 0'0412 |        |

¿Difiere la actividad DBH en estos dos grupos?

Como dijimos más arriba, uno de los propósitos de la Estadística, en concreto de la Estadística Descriptiva, es dejar que los datos hablen por sí mismos, representándolos, calculando algún valor que resuma a todos ellos, alguno que nos dé una idea de su dispersión, alguno que nos diga si la relación entre estos pares de datos es grande o pequeña, etc.

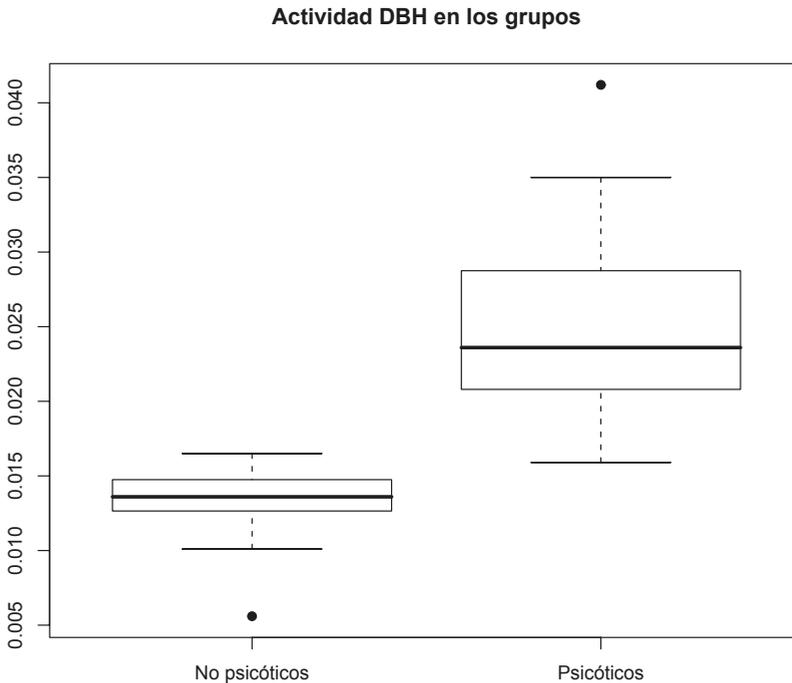


Figura 1.1 : Gráfico de Cajas

A ello dedicaremos el siguiente capítulo y el último, pero ya adelantamos que si hacemos un *Diagrama de Cajas* de estos pares de datos (gráfico que estudiaremos en el Capítulo 14), obtendremos la Figura 1.1 la cual nos indica cómo se distribuyen ambos grupos de datos sobre la imaginaria línea vertical que comienza con el nombre de cada grupo. Este gráfico nos muestra que hay

un valor en cada grupo (0'0056 en el primero y 0'0412 en el segundo), marcados en la figura como puntos negros, los cuales parecen estar alejados del resto de datos del grupo.

Esta cuestión de la posible presencia de datos anómalos en nuestra muestra es muy preocupante por la influencia que éstos pudieran tener en los resultados. Ante ellos caben dos opciones: quitarlos y considerar que tenemos 14 datos de cada grupo de pacientes o utilizar técnicas de Inferencia Estadística Robusta. El manejo de tales técnicas es el propósito del libro MR.

Si, por simplificar, quitamos aquí ambos datos y representamos los 14 restantes de cada grupo en lo que denominaremos en el Capítulo 2, un *Histograma*, obtendremos la Figura 1.2. Este gráfico parece indicarnos unos datos distribuidos (es decir, repartidos) de una forma acampanada.

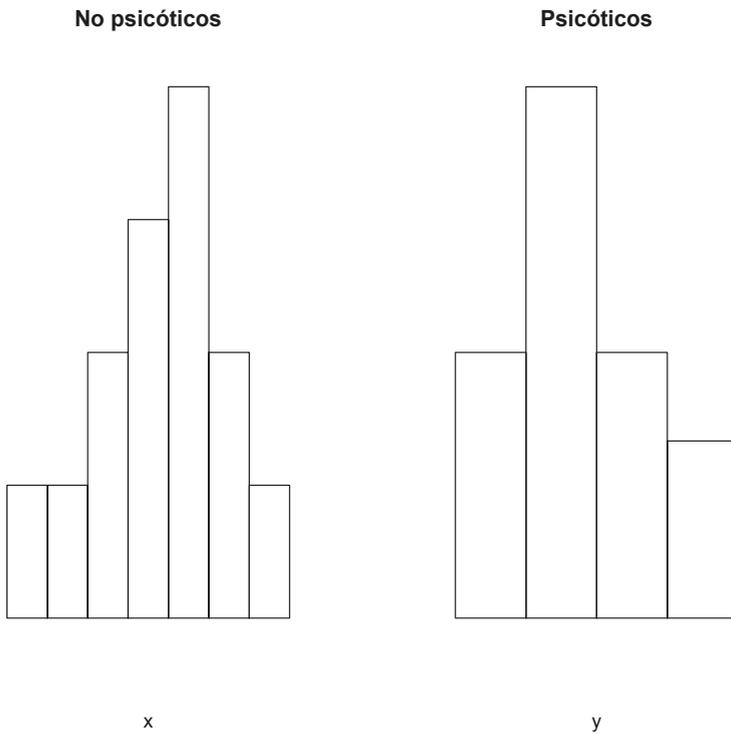


Figura 1.2 : Histogramas de los datos

Las técnicas de Inferencia Estadística requieren, para ser utilizadas, de determinadas suposiciones; la más habitual de todas, es que los datos se distribuyan según una *Distribución Normal*, es decir, que el modelo probabilístico que rige el fenómeno aleatorio de donde proceden los datos, tenga forma acampanada como sugieren estos gráficos. Aunque los conceptos de Probabilidad

y los Modelos Probabilísticos más importantes se estudiarán en los Capítulos 3 y 4, hemos representado sobre la figura anterior, las dos distribuciones de probabilidad que parece razonable admitir para dichos datos, obteniendo la Figura 1.3.

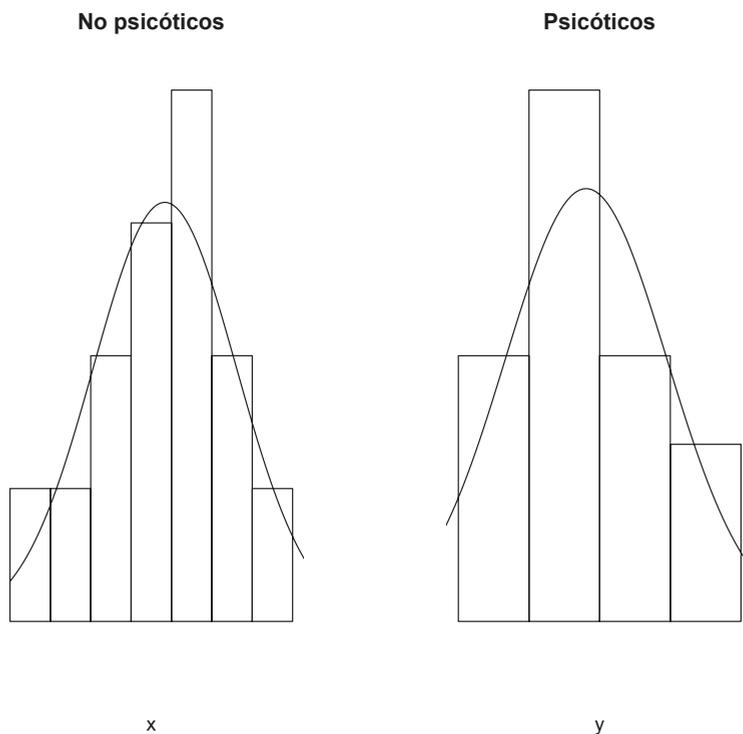


Figura 1.3 : Histogramas con curvas normales

Una vez que la Estadística Descriptiva nos ha mostrado los datos tal cual y con ella hemos deducido que parece razonable admitir distribuciones de tipo normal para los datos, podemos hacer uso de la Inferencia Estadística para, por ejemplo, contrastar si puede admitirse que la actividad DBH es igual en ambos grupos o es significativamente mayor en el grupo de psicóticos. Los diferentes Métodos Estadísticos serán estudiados en los Capítulos 5 al 13 del libro.

Con objeto de precisar un poco más, diremos que los *fenómenos* o *experimentos aleatorios* es decir, aquellos a los que se pueden aplicar las técnicas de la Estadística, se caracterizan, en contraposición con los deterministas, como aquellos que en las mismas condiciones iniciales pueden dar lugar a diferentes resultados.

La aplicación de un determinado tratamiento médico a un grupo de per-

sonas elegidas al azar es un ejemplo de experimento aleatorio en Medicina.

La utilización de tres tipos diferentes de abono en unas parcelas elegidas al azar es otro ejemplo en Agricultura.

La selección aleatoria de individuos de una población con objeto de investigar cuántos hijos menores de 18 años tienen como media en esa población es otra experiencia aleatoria en Sociología.

La selección de personas a la salida de los colegios electorales (elegidos al azar) para adelantar los resultados de unas elecciones debería de realizarse, también, mediante un experimento aleatorio para poder sacar conclusiones fiables.

Los ejemplos de experimentos aleatorios son innumerables. De hecho, como dijimos al comienzo del capítulo, casi sin excepción, todos los fenómenos o experiencias que se realizan en Biología, Sociología, Medicina, Psicología, Química, Física, etcétera, pueden ser calificados de aleatorios. Inclusive, hoy en día, las excavaciones arqueológicas se sirven de las técnicas de la Estadística Aplicada a la hora de sacar conclusiones (véase García y Cabrero, 2009b).

## 1.2. Población e individuo

Los fenómenos aleatorios se presentan en un mundo real formado por *individuos*, en los que se observa el fenómeno aleatorio en estudio. El conjunto de todos los individuos recibe el nombre de *población*.

Así, el conjunto de pacientes de los que se seleccionan aquellos que van a ser sometidos a tratamiento, constituyen la población en el primer ejemplo mencionado al final de la sección anterior, siendo cada uno de ellos el individuo. Las parcelas forman la población del segundo ejemplo, etc.

Como se ve, los términos *población* e *individuo*, no deben ser entendidos necesariamente en un sentido de población humana y persona humana, sino, respectivamente, como colectivo del que queremos sacar conclusiones y como elemento o unidad que compone la población.

Una cuestión muy importante es la de determinar con precisión lo que constituye la población ya que de ella se elegirán unos cuantos individuos con objeto de obtener conclusiones acerca de toda la población.

Así, en el primer ejemplo, puede considerarse como población la de los enfermos españoles que padecen la enfermedad en estudio, o la de los enfermos al mal en estudio en todo el mundo, o alguna de las dos anteriores pero con individuos que tienen una edad comprendida entre dos valores determinados. La definición de lo que constituye la población depende del experimentador y de la naturaleza del problema que se investiga. No obstante, una vez definida, de ella se tomarán las observaciones y se deberán sacar las conclusiones.

Al conjunto de individuos que elegimos de la población lo denominaremos

*muestra.*

Insistimos en que es muy importante el fijar la población con toda precisión, ya que solamente la obtención de una muestra representativa de la población permitirá obtener conclusiones fiables sobre ella.

Habitualmente la muestra representativa se obtendrá por un procedimiento aleatorio (es decir, de azar), lo cual permitirá medir y controlar los posibles errores en términos de probabilidades, pero insistimos en que lo importante es obtener una muestra representativa de la población sea o no por un procedimiento aleatorio. La ventaja de utilizar un mecanismo de azar es que éste nos garantiza que la muestra será representativa, mientras que con otros procedimientos, en general, no tendremos tal garantía. Sin embargo, una población suficientemente homogénea puede soslayar este mecanismo; así todos tenemos experiencias de situaciones en las que al ir a comprar un producto y pedir información sobre él (es decir, que nos enseñen una *muestra* del producto que queremos comprar) el dependiente elige un *individuo* de forma subjetiva como representativo de la *población en estudio* y nosotros consideramos que éste es lo suficientemente representativo de la *población* como para poder decidir sobre su adquisición.

### 1.3. Muestras aleatorias

Como antes hemos dicho, el trabajar con muestras aleatorias es una cuestión de suma importancia a la hora de obtener buenas conclusiones, ya que la muestra será la *materia prima* a utilizar en la elaboración de las inferencias, y solamente de buena materia prima se pueden obtener buenos productos.

Como ejemplo de lo importante que es tener una muestra aleatoria, supongamos que queremos asignar diez ratones a dos grupos con objeto de determinar los efectos de un cierto medicamento en el ritmo cardiaco. Podríamos pensar que, cerrando los ojos y metiendo la mano en la jaula de los ratones, los cinco primeros seleccionados pueden pasar a formar el grupo experimental y los cinco restantes el grupo control, habiendo obtenido así una muestra aleatoria.

Sin embargo, este procedimiento de selección tendría el inconveniente de haber asignado al grupo experimental los cinco ratones más torpes o de mayor peso, ya que fueron los que antes se dejaron atrapar. Si el peso o la agilidad están relacionados con el ritmo cardiaco, es probable que exista sesgo en la elección de nuestros grupos, y por tanto en nuestras conclusiones, ya que como antes dijimos, éstas se basarán en la muestra seleccionada.

Una forma de evitar estos problemas en el ejemplo mencionado, consistiría en asignar un número a cada uno de los ratones y seleccionar al azar, a con-

tinuación, cinco bolas de una urna que contuviera diez bolas numeradas del uno al diez. Una vez clasificados los ratones en dos grupos se calificaría a uno de ellos de grupo control, lanzando una moneda al aire.

Este laborioso proceso se simplifica notablemente con la utilización de programas que generan *números aleatorios* aunque en estos siempre está presente la arbitrariedad del inicio o *semilla* de la elección.

De todas formas, en los trabajos de campo, la selección aleatoria es más complicada, por lo que deberemos admitir que un muestreo aleatorio es un ideal que el investigador debe esforzarse en conseguir y que probablemente nunca llegará a alcanzar completamente. La propia Inferencia Estadística proporciona técnicas que permiten chequear si la muestra obtenida puede considerarse como aleatoria o no.

## 1.4. Variable aleatoria y Modelo probabilístico

Habitualmente la situación que se plantea es la de un investigador que desea conocer el valor de alguna característica de la población en estudio, como por ejemplo la estatura media de la población española, o la ganancia de peso después de aplicar a los pacientes un tratamiento médico, o el determinar el intervalo en el que, con gran probabilidad, se encuentre dicha característica, o alternativamente poder decidir si, por ejemplo, dicha estatura media es tal o cual valor.

Es decir, la situación que se presenta es la de una característica o valor poblacional objeto de investigación, al que denominaremos parámetro poblacional o simplemente *parámetro*, estando éste asociado a una variable en estudio. Así, en el ejemplo anterior asociado a la variable estatura, el parámetro en estudio era la estatura media, mientras que en el segundo lo era la ganancia media de peso, asociado a la variable diferencia de los pesos antes y después de aplicar el tratamiento.

En el capítulo cuarto veremos que, desde un punto de vista técnico, esta *variable* en estudio que deberemos identificar en el experimento que estemos realizando, se corresponde con lo que matemáticamente se denomina *variable aleatoria*  $X$  y que, como aquí, de forma habitual denominaremos simplemente variable.

Con objeto de hacer *inferencias* sobre el parámetro en estudio, es decir, o bien poder llegar a dar un valor como estimación suya (*estimación por punto*), o bien dar un intervalo numérico en el que verosíblemente se encuentre (*estimación por intervalos de confianza*), o bien poder decidir si puede considerarse razonable un valor u otro para dicho parámetro (*contraste de hipótesis*), el investigador selecciona al azar de la población unos cuantos individuos, digamos

$n$ , los cuales, como antes dijimos, constituyen la *muestra*, siendo  $n$  el *tamaño muestral*, en los que se observará la variable en estudio: peso, talla, etc.

Se obtendrán así  $n$  realizaciones de la variable aleatoria en estudio  $X$ , que representaremos por  $(X_1, \dots, X_n)$ , entendiéndose cada  $X_i$ ,  $i = 1, \dots, n$  como el valor que toma la variable en estudio (peso, talla, etc.) en el individuo seleccionado al azar en el  $i$ -ésimo lugar.

En este libro sólo consideraremos la situación en la que cada individuo es seleccionado de forma independiente e idéntica a como lo son los demás. Matemáticamente esto significa que las  $n$  variables aleatorias son lo que se dice *independientes e idénticamente distribuidas*.

La realización de las observaciones en los individuos de la muestra dará origen a los *datos*.

Los valores posibles de cada variable aleatoria junto con las probabilidades con los que los toma, se denomina distribución o ley de probabilidad de la variable aleatoria en estudio, o más brevemente *modelo probabilístico*. Lo que habitualmente haremos será suponer un modelo probabilístico que creemos sea un reflejo lo más exacto posible del fenómeno aleatorio que estemos estudiando, obteniendo así una simplificación de éste y además la posibilidad de utilizar técnicas matemáticas con objeto de conseguir reglas de actuación a la hora de hacer inferencias con los datos.

Será necesario, por tanto, que a la hora de estudiar un fenómeno aleatorio, el investigador identifique la variable o variables en estudio, así como que suponga una ley de probabilidad (es decir, un modelo) el cual rijan dicha variable en estudio. Esta ley de probabilidad habitualmente estará completamente determinada salvo uno o dos parámetros, siendo el objetivo de la Estadística hacer inferencias sobre ellos. Así por ejemplo, el deseo de averiguar la estatura media de los españoles se puede *modelizar* admitiendo que la variable en estudio  $X$  (estatura del individuo seleccionado al azar de la población) sigue una ley de probabilidad o modelo probabilístico en forma de campana simétrica, en donde el centro de simetría sea la estatura media (parámetro) a investigar.

Dedicaremos el cuarto capítulo a insistir sobre este proceso de *idealización o modelización*, el cual es muy importante, puesto que, de hecho, permite pasar de la realidad tangible que el investigador tiene en su experimento del *mundo real*, al *mundo matemático* que le permitirá utilizar las técnicas y resultados de la Estadística. Es necesario, no obstante, saber con precisión lo que el investigador tiene y desea conocer, para poder utilizar con precisión estas potentes técnicas. Finalmente también será necesario tener suficientes conocimientos y experiencia para poder *interpretar* correctamente los resultados que obtenga, pasando así del *mundo matemático* en el que estaba al *mundo real*.

## 1.5. Diferentes Estadísticas

Como ya hemos dicho más arriba, el propósito de la *Inferencia Estadística* es el de obtener conclusiones de la población en estudio en base a la muestra obtenida de ella, mientras que el objetivo de la *Estadística Descriptiva* es el de, dados los datos, ordenarlos, simplificarlos, resumirlos, clasificarlos, etcétera, determinando de esta manera un conjunto de valores que, además de proporcionar una rápida impresión de sus principales características, permitan hacer comparaciones con otros conjuntos de datos.

En la Estadística Descriptiva no se hacen suposiciones extrañas a los datos, como puede ser la de un modelo probabilístico poblacional; se deja que los datos *hablen por sí mismos*.

Por el contrario, las técnicas de la Inferencia Estadística requerirán de suposiciones ajenas a los datos (simetría en la distribución modelo, población normal, etcétera).

Existe aún una tercera posibilidad: utilizar información *a priori* sobre el parámetro a la hora de hacer nuestras inferencias. Esta situación, denominada *Inferencia Bayesiana*, no será tratada aquí, porque, además de requerir más suposiciones y de apenas estar incluida en los habituales Paquetes Informáticos de Estadística, implica unas consecuencias difíciles de admitir: si esta información *a priori* es subjetiva (como habitualmente ocurre) y, en consecuencia, cambia de un investigador a otro, las conclusiones estadísticas también lo harán y, por tanto, se perderá la objetividad de las conclusiones estadísticas obtenidas. Esta incongruencia es, además, considerada como meritoria por los usuarios de esta técnica.

Por tanto, en este texto solamente abordaremos el estudio de la *Estadística Descriptiva*, al que dedicaremos el capítulo segundo, así como el análisis de conceptos y técnicas elementales de *Inferencia Estadística* no Bayesiana.

## 1.6. Tratamiento informático de los datos

Los conceptos que vamos a estudiar en este texto serán ilustrados con numerosos ejemplos prácticos, con objeto de que puedan ser entendidos y aplicados por el lector con la ayuda de una simple calculadora.

Muchos de ellos podían haber sido resueltos con la ayuda de algún paquete estadístico. No obstante no hemos querido subordinar el libro en tal sentido por dos razones: la primera porque éste es, fundamentalmente, un libro en el que se exponen, con numerosos ejemplos, los conceptos básicos de la Estadística, siendo el ordenador una herramienta eficaz la cual permite una enorme velocidad y una gran capacidad de cálculo. Pero lo más importante en Estadística,