



UNED

Estadística básica > con R

Alfonso García Pérez

GRADO

Estadística básica con R

ALFONSO GARCÍA PÉREZ

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del Copyright, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamos públicos.

© Universidad Nacional de Educación a Distancia
Madrid 2013

Librería UNED: c/ Bravo Murillo, 38 - 28015 Madrid
Tels.: 91 398 75 60 / 73 73
e-mail: libreria@adm.uned.es

© Alfonso García Pérez

Todas nuestras publicaciones han sido sometidas a un sistema de evaluación antes de ser editadas

ISBN: 978-84-362-6093-9
Depósito legal: M. 20716-2011

Primera edición: mayo de 2010
Segunda reimpresión: mayo de 2013

Impreso en España - Printed in Spain
Imprime y encuaderna: Grafo, S.A..
Avda. Cervantes, 51, edif. 21, (Vizcaya)



Índice

1. Introducción al R	15
1.1. Introducción	15
1.2. El editor de objetos R	18
1.3. Datos en R	19
1.3.1. Vectores	20
1.3.2. Factores	21
1.3.3. Matrices	22
1.3.4. Estructuras de datos	25
1.3.5. Listas	26
1.3.6. Nombres a las filas y columnas de matrices y vectores .	27
1.4. Gráficos	27
1.4.1. Funciones gráficas de alto nivel	27
1.4.2. Funciones gráficas de bajo nivel	30
1.5. Otras cuestiones	30
1.6. Interfaz	30
1.7. Modificar y Crear Funciones	31
1.8. Librerías de R	34
1.9. Lecturas Recomendadas	36
2. Estadística Descriptiva	37
2.1. Introducción a la Estadística	37
2.1.1. Población e individuo	38
2.1.2. Muestras aleatorias	39
2.1.3. Variable aleatoria y Modelo probabilístico	40
2.1.4. Diferentes Estadísticas	41
2.2. Conceptos fundamentales de la Estadística Descriptiva	42
2.3. Distribuciones unidimensionales de frecuencias	45

2.3.1.	Representaciones gráficas de las distribuciones unidimensionales de frecuencias	49
2.3.2.	Medidas de tendencia central de caracteres cuantitativos	57
2.3.3.	Medidas de dispersión	64
2.3.4.	Medidas de asimetría	67
2.3.5.	Medidas de posición y dispersión con R	68
2.4.	Distribuciones bidimensionales de frecuencias	70
2.4.1.	Representaciones gráficas de las distribuciones bidimensionales de frecuencias	74
2.4.2.	Ajuste por mínimos cuadrados	78
2.4.3.	Precisión del ajuste por mínimos cuadrados	81
2.5.	Ejercicios de Autoevaluación	84
2.6.	Lecturas Recomendadas	86
3.	Probabilidad	87
3.1.	Introducción	87
3.2.	Espacio Muestral	89
3.3.	Conceptos de Probabilidad	91
3.4.	Propiedades elementales de la Probabilidad	93
3.5.	Asignación de Probabilidad en espacios muestrales discretos	96
3.6.	Modelo Uniforme	97
3.7.	Probabilidad condicionada	100
3.8.	Independencia de sucesos	101
3.9.	Teorema de la Probabilidad Total	101
3.10.	Teorema de Bayes	102
3.11.	Ejercicios de Autoevaluación	103
3.12.	Lecturas Recomendadas	104
4.	Modelos Probabilísticos	105
4.1.	Introducción	105
4.2.	Distribución de Probabilidad	106
4.2.1.	Funciones básicas de R en Probabilidades	112
4.3.	Variables aleatorias multivariantes	113
4.4.	Modelos unidimensionales discretos	114
4.4.1.	Distribución Binomial	114
4.4.2.	Distribución de Poisson	117
4.4.3.	Distribución Geométrica	119
4.4.4.	Distribución Hipergeométrica	120
4.4.5.	Distribución Binomial Negativa	121
4.5.	Modelos unidimensionales continuos	122
4.5.1.	Distribución Normal	122
4.5.2.	Distribución Uniforme	126

4.5.3.	Distribución Beta	126
4.5.4.	Distribuciones Gamma y Exponencial	127
4.5.5.	Distribución de Cauchy	127
4.6.	Modelos bidimensionales	127
4.6.1.	Distribución Normal bivalente	128
4.7.	Teorema Central del Límite	128
4.8.	Ejercicios de Autoevaluación	132
4.9.	Lecturas Recomendadas	133
5.	Estimadores. Distribución en el muestreo	135
5.1.	Introducción	135
5.2.	Método de la máxima verosimilitud	138
5.3.	Distribuciones asociadas a poblaciones normales	141
5.3.1.	Distribución χ^2 de Pearson	141
5.3.2.	Distribución t de Student	144
5.3.3.	Distribución F de Snedecor	146
5.4.	Estimación de la media de una población normal	149
5.5.	Estimación de la media de una población no necesariamente normal. Muestras grandes	150
5.6.	Estimación de la varianza de una población normal	153
5.7.	Estimación del cociente de varianzas de dos poblaciones normales independientes	154
5.8.	Estimación de la diferencia de medias de dos poblaciones normales independientes	156
5.9.	Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes . . .	159
5.10.	Datos apareados	160
5.11.	Tamaño muestral para una precisión dada	161
5.12.	Ejercicios de Autoevaluación	162
5.13.	Lecturas Recomendadas	164
6.	Intervalos de confianza	165
6.1.	Introducción	165
6.1.1.	Cálculo de Intervalos de Confianza con R	168
6.2.	Intervalo de confianza para la media de una población normal .	170
6.3.	Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes	172
6.4.	Intervalo de confianza para la varianza de una población normal	175
6.5.	Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes	177
6.6.	Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes	178

6.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes	181
6.8. Intervalos de confianza para datos apareados	182
6.9. Ejercicios de Autoevaluación	184
6.10. Lecturas Recomendadas	185
7. Contraste de hipótesis	187
7.1. Introducción y conceptos fundamentales	187
7.2. Contraste de hipótesis relativas a la media de una población normal	197
7.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes	201
7.4. Contraste de hipótesis relativas a la varianza de una población normal	210
7.5. Contraste de hipótesis relativas a las varianzas de dos poblaciones normales independientes	214
7.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes	219
7.7. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes	227
7.8. Contrastes de hipótesis para datos apareados	234
7.9. Ejercicios de Autoevaluación	235
7.10. Lecturas Recomendadas	236
8. Contrastes no paramétricos	237
8.1. Introducción	237
8.2. Pruebas χ^2	237
8.2.1. Pruebas χ^2 con R	239
8.2.2. Contraste de bondad del ajuste	240
8.2.3. Contraste de homogeneidad de varias muestras	249
8.2.4. Contraste de independencia de caracteres	253
8.3. Tests relativos a una muestra y datos apareados	258
8.3.1. El contraste de los signos	258
8.3.2. El contraste de los rangos signados de Wilcoxon	263
8.4. Tests relativos a dos muestras independientes	268
8.4.1. El contraste de Wilcoxon-Mann-Whitney	268
8.4.2. El contraste de la Mediana	272
8.5. Ejercicios de Autoevaluación	275
8.6. Lecturas Recomendadas	276

9. Análisis de la Varianza	277
9.1. Introducción	277
9.2. Análisis de la Varianza para un Factor: Diseño Completamente Aleatorizado	278
9.3. Análisis de la Varianza con R	283
9.4. Análisis de las condiciones	284
9.5. Comparaciones Múltiples	287
9.6. Comparaciones Múltiples con R	289
9.7. Ejercicios de Autoevaluación	290
9.8. Lecturas Recomendadas	292
10.Regresión Lineal y Correlación	293
10.1. Introducción	293
10.2. Modelo de la Regresión Lineal Simple	295
10.2.1. Interpretación de los coeficientes de regresión	297
10.3. Contraste de la Regresión Lineal Simple	298
10.3.1. Análisis de la variación explicada frente a la no explicada por la recta de regresión	299
10.3.2. Contraste de hipótesis para β_1	302
10.4. Regresión Lineal con R	304
10.5. Correlación Lineal	306
10.5.1. Estimación por punto de ρ	306
10.5.2. Contraste de hipótesis sobre ρ	307
10.6. Modelo de la Regresión Lineal Múltiple	308
10.6.1. Contraste de la Regresión Lineal Múltiple	310
10.7. Ejercicios de Autoevaluación	312
10.8. Lecturas Recomendadas	313
Bibliografía General	315
Soluciones a los Ejercicios de Autoevaluación	317
Obtención de R	321

**A los Tutores y Alumnos de la UNED
en reconocimiento a su esfuerzo diario**

Prólogo

Este texto es una introducción a los principales conceptos de la Estadística, término genérico que habitualmente se utiliza para englobar tanto a la *Estadística Descriptiva*, como al *Cálculo de Probabilidades*, como a la *Inferencia Estadística*.

En la *Estadística Descriptiva* se deja que los datos hablen por sí mismos, es decir se ordenan, representan, etc., de manera que puedan sugerirnos estructuras o modelos que los expliquen. Sirve además, como introducción de algunos conceptos de Cálculo de Probabilidades y de Inferencia Estadística.

En el *Cálculo de Probabilidades* se define y maneja la Probabilidad como medida de la incertidumbre que presentan los fenómenos aleatorios, fenómenos que son el objeto de estudio de la Estadística; se proponen, además, modelos que pueden regir estas experiencias aleatorias.

La *Inferencia Estadística* es, sin duda, la parte más interesante puesto que con ella se pueden obtener conclusiones de donde se extrajeron los datos, midiendo los posibles errores en términos de probabilidades.

Al estudio de estas tres partes es a lo que dedicaremos el texto pero, hoy en día, el uso del ordenador se convierte en indispensable, por lo que teníamos que incluir algún paquete estadístico que sirviera de ayuda a nuestro estudio. Elegimos el *Paquete Estadístico R* porque pensamos que es el idóneo en la ejecución de los Métodos Estadísticos. Además es gratuito. Al final del libro indicamos la dirección de Internet de dónde obtener R y damos indicaciones de cómo instalarlo en su ordenador.

Para estudiar el texto sólo se requiere una formación elemental de Matemáticas como la impartida en las modalidades de Bachillerato de Ciencias y Tecnología o Humanidades y Ciencias Sociales.

El texto está pensado para un cuatrimestre y como introducción a las partes antes mencionadas con objeto de que, al final de su estudio, el lector haya entendido los fundamentos y sepa utilizar los Métodos Estadísticos analizados.

Para que el texto sea lo más conciso posible, hemos publicado al margen del libro la *Addenda Fórmulas y Tablas Estadísticas* (citada como ADD) que sirve de resumen de algunas fórmulas utilizadas, incluyendo también tablas de

distribuciones, aunque las probabilidades asociadas a éstas se podrán calcular con R y evitar utilizar la Addenda ADD.

Aparecen al final de los capítulos algunos Ejercicios de Autoevaluación, cuya solución está al final de este libro, con objeto de que el lector valore si ha adquirido los contenidos del capítulo. En los textos *Problemas Resueltos de Estadística Básica* y *Ejercicios de Estadística Aplicada* aparecen numerosos Ejercicios de Autoevaluación.

Comenzamos el libro con un capítulo introductorio a R ya que en el resto del libro se irán resolviendo ejemplos con la ayuda de este software. De hecho, entendemos que una buena manera de aprender R es utilizarlo en situaciones concretas. Advertimos, no obstante, que no pretendemos enseñar *programación R* sino utilizar este paquete en la ejecución de Métodos Estadísticos. Si el lector está interesado en profundizar en este software, le recomendamos algunos textos al final del primer capítulo.

De igual manera, si el lector desea estudiar más Métodos Estadísticos de los que hay en el libro, le recomendamos los textos que se citan por sus acrónimos TA, *Métodos Avanzados de Estadística Aplicada. Técnicas Avanzadas*, y MR, *Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo*. De hecho, al final de cada capítulo aparecerán Lecturas Recomendadas por si quiere profundizar en los temas tratados en el capítulo, pero su lectura no es necesaria para seguir el desarrollo de este texto.

A continuación del capítulo sobre R se estudian tres capítulos sobre temas básicos: uno de Estadística Descriptiva y dos de Cálculo de Probabilidades.

En el Capítulo 5 se inicia verdaderamente la Inferencia Estadística con el estudio de los principales estadísticos a utilizar en las diversas situaciones planteadas. Los Intervalos de Confianza se estudian en el Capítulo 6 y los conceptos elementales de Tests de Hipótesis (la herramienta estadística más empleada, sin ninguna duda), se estudian en el Capítulo 7.

Los tests de hipótesis estudiados en el Capítulo 7 requieren habitualmente de la normalidad de los datos para poder ser utilizados. En el Capítulo 8 estudiamos tests, denominados no paramétricos, que no requieren de esta suposición.

Los dos últimos capítulos abordan las aplicaciones más comunes de la Inferencia Estadística; se trata del Análisis de la Varianza, estudiado en el Capítulo 9, y del Análisis de la Regresión, analizado en el Capítulo 10.

Quiero por último agradecer a mis compañeros de Departamento y al IUED la lectura detallada del texto que han ayudado a mejorar su primera versión.

Capítulo 1

Introducción al R

1.1. Introducción

R es un software *clónico* del paquete (no gratuito) S-Plus, que es un compendio de aplicaciones estadísticas que utilizan el lenguaje S, diseñado por la compañía AT&T's Bell Laboratories, en principio, para su uso interno. Por esta razón, casi todos los comandos utilizados aquí podrá emplearlos si dispone de S-Plus.

S-Plus consiguió una gran difusión en los últimos años del siglo XX y fueron dos profesores de la Universidad de Auckland (Nueva Zelanda), Ross Ihaka y Robert Gentleman los que elaboraron una versión reducida de S para tareas docentes. La R, inicial del nombre de pila de ambos profesores, sirvió de denominación al nuevo paquete estadístico.

En 1995 Martin Maechler les convenció para su distribución gratuita, estando disponible las primeras versiones piloto (denominadas con un 0 en el primer dígito) en 1999. Hoy en día, existen numerosas aportaciones (todas de libre distribución) programadas en R, las cuales se pueden obtener en la página web de donde se obtuvo R.

No estará de más hacer una advertencia. Nadie se responsabiliza de los resultados obtenidos con R, dado el carácter de libre distribución del software. No obstante, nosotros sí nos responsabilizamos de los cálculos que aparecen en este texto ya que han sido verificados por el autor.

Una última observación: al abrir R aparecerá la *línea de comandos*, que es aquella que comienza con el símbolo `>`. En los ejemplos del libro incluiremos este símbolo, el cual, lógicamente no debe ser tecleado si queremos reproducirlos.

Como dijimos en el Prólogo del libro, dado que vamos a analizar todos los ejemplos con el Paquete R, es conveniente empezar conociendo este software en profundidad. A ello dedicaremos este primer capítulo. Es muy interesante que,

mientras lo va leyendo, tenga abierto R para ir reproduciendo las instrucciones que aquí se indican.

Como ocurre con todos los paquetes estadísticos, R también utiliza un lenguaje propio. Toda instrucción que pueda ser ejecutada desde la línea de comandos se denomina *expresión*. Las expresiones se ejecutan con `Enter`.

Éstas pueden tener una longitud de más de una línea. Cuando se presiona `Enter` después de una expresión sintácticamente incompleta, ésta no se ejecuta ni se producen mensajes de error; aparece el *prompt* + al comienzo de una nueva línea de comandos invitándonos así a completar la expresión no concluida.

Los elementos básicos de R son los *objetos* y, por tanto, a ellos se referirán las expresiones de R. De hecho, los objetos son ficheros capaces de ser editados y, en su caso, ejecutados.

Un *objeto* es el resultado de ejecutar una expresión en la que aparece el operador `<-`. Dicho de otra forma, una expresión que nos interese guardar (por ejemplo para volver a utilizarla en otra ocasión o porque va a ser parte de otra expresión) puede ser *salvada* con el operador `<-`, también denominado operador *asignación*.

Por ejemplo, si queremos denominar a al número 1, ejecutaríamos la expresión

```
> a<-1
```

pudiendo hacer ahora

```
> a+a
[1] 2
```

Apuntemos que el número entre corchetes que aparece, `[1]`, indica solamente el lugar que ocupa el primer valor del resultado de ejecutar la expresión que le precede. Esto es especialmente útil cuando manejemos datos de gran dimensión.

El nombre asignado a un objeto debe empezar por una letra y puede incluir cualquier combinación de letras mayúsculas o minúsculas, números y puntos. Por ejemplo, dos nombres de objetos podrían ser `X` y `datos.nuevos`.

Como dijimos, los objetos obtenidos como resultado de ejecutar una expresión pueden ser utilizados como parte de una nueva expresión.

Los dos tipos de objetos más utilizados son el *dato*, sobre el que volveremos más adelante, y la *función*. Las funciones constan de un nombre seguido de dos paréntesis, `nombre()`; entre los paréntesis se incluyen sus *argumentos*. Si

sólo ejecutamos su nombre obtendremos su definición y si ejecuta `?nombre` obtendrá ayuda sobre su utilización.

El papel de las funciones es tan relevante que podemos afirmar que cuando ejecutamos R en realidad estaremos ejecutando funciones. Su importancia es tal que, de hecho, puede decirse que R es un *lenguaje funcional*, en el sentido de que sus *programas* se presentan como funciones escritas en su lenguaje. Pero lo más interesante es que, dada la flexibilidad de R, como veremos al final del capítulo, podemos crear nuestras propias funciones, añadiéndolas a las ya existentes, respondiendo así a nuestras necesidades particulares; éstas podrán combinar distintas funciones R que deseemos sean ejecutadas de forma conjunta.

Una de las funciones más sencillas y por medio de la cual *salimos* del programa es

```
> q()
```

Al ejecutarla, el ordenador nos preguntará si queremos conservar los cálculos que hayamos realizado en la sesión, mediante la pregunta *Guardar imagen de área de trabajo?* (*Save workspace image?* si eligió el idioma inglés en la instalación). Si respondemos Sí, al comenzar la sesión siguiente podremos volver a utilizar los resultados de la sesión recién finalizada.

Una observación que merece destacarse es que, desde esa línea de comandos podemos utilizar R como una potente calculadora matemática. Así, desde allí se pueden realizar las habituales operaciones matemáticas:

```
> 9*8
[1] 72
```

u obtener el valor de las funciones más conocidas como la potencial, la exponencial, la raíz cuadrada,

```
> 3^2
[1] 9
> exp(2)
[1] 7.389056
> sqrt(16)
[1] 4
```

También se pueden resolver sistemas de ecuaciones o hacer integración numérica y otras muchas aplicaciones matemáticas, pero R es un software especialmente creado para ejecutar Métodos Estadísticos, que el lector irá aprendiendo a manejar mientras lee este libro ya que son miles las funciones que



contiene R, agrupadas en lo que se denomina *librería*. Algunas de ellas no están en la versión que ha instalado sino en librerías adicionales que se pueden instalar, como veremos al final del capítulo.

Dos funciones que queremos ya destacar son, la función `objects`, mediante la cual podemos listar los objetos R existentes

```
> objects()
```

Y la función `rm`, utilizada para suprimir objetos; para ello debemos utilizar como argumentos suyos, los objetos a eliminar. Así, si queremos suprimir los objetos `dato1` y `función3`, ejecutaríamos la expresión

```
> rm(dato1,función3)
```

Finalizamos esta primera toma de contacto con R señalando que, para recuperar alguna instrucción ejecutada anteriormente, basta con pulsar la tecla  tantas veces como sea necesario hasta que la instrucción aparezca. Luego deberemos ejecutarla pulsando .

1.2. El editor de objetos R

Para crear o modificar objetos o funciones R, debemos utilizar las denominadas *funciones editoras* `edit` y, preferentemente, `fix`. Al ejecutarlas entraremos en el editor propio de R, o en el *Bloc de Notas* de Windows, o en el que hayamos establecido en las Preferencias del programa. Si los objetos a editar son datos, suele abrirse el Editor de Datos.

También pueden crearse funciones desde la línea de comandos. Si queremos crear una nueva función, por ejemplo $f(x) = 2x$, teclearemos desde la línea de comandos,

```
> f<-function(x){2*x}
```

habiendo tecleado entre las llaves la definición de la función f . Ahora, si queremos saber cuál es su valor en, por ejemplo, 12, ejecutaremos `f(12)`

```
> f(12)
[1] 24
```

Cualquier asignación que hagamos eliminará asignaciones previas con el mismo nombre. Es decir, si ahora denominamos *f* a otra función de R, la anterior asignación habrá sido eliminada. Por esta razón, si queremos modificar una función ya existente conservando a la vez ésta, deberemos crear primero la nueva con la expresión ya conocida

```
> nueva<-antigua
```

luego ejecutar

```
> fix(nueva)
```

la cual, una vez modificada, será salvada al salir del editor.

Un conjunto de instrucciones de R se denomina *script*, el cual puede ser guardado en el propio programa. Aquella parte que marquemos puede ser ejecutada con las teclas Ctrl-R.

1.3. Datos en R

El concepto de *dato* en R es más amplio que el utilizado en Estadística. De hecho, lo que en R se denomina dato es el resultado de ejecutar una expresión R, es decir, un tipo de objeto.

Así por ejemplo, un dato podrá ser una matriz en donde como primera columna aparecen los nombres de los individuos en los que se han observado las variables cuyos valores aparecen en las restantes columnas.

Cada tipo de dato tiene asociados determinados *atributos*; el más importante es su *modo*. Consideraremos cuatro clases de *modos*:

- *logical* (lógico): Modo binario en donde los valores posibles son T ó F (Verdadero o Falso).
- *numeric* (numérico): Modo en donde los valores posibles son números reales.
- *complex* (complejo): Modo en donde los valores posibles son números complejos.
- *character* (carácter): Modo en donde los valores posibles son caracteres separados por comillas.

Consideraremos cinco tipos diferentes de datos,

- *vector* (vector): Conjunto de elementos en un orden específico. Todos los elementos de un vector deben ser del mismo modo. Los más utilizados son los *vectores numéricos*, es decir, vectores cuyos elementos son números.
- *matrix* (matriz): Disposición bidimensional de elementos de un mismo modo.
- *factor* (factor): Vector cuyos elementos son valores procedentes de un número finito de *categorías*.
- *data frame* (estructura de datos): Disposición bidimensional de elementos cuyas columnas pueden estar formadas por elementos de distinto modo.
- *list* (lista): Expresión más general de dato, la cual puede contener colecciones arbitrarias de datos.

Aunque R tiene incorporado un Editor de Datos que permite visualizarlos y, en cierta medida, manejarlos de una forma algo más cómoda, siempre debemos tener presente que R considera diferentes tipos de datos que iremos describiendo en los siguientes apartados. Los más utilizados son el *vector*, la *matriz* y el *data frame*.

1.3.1. Vectores

El *vector* es el tipo de dato más utilizado en R, especialmente como argumento de funciones.

Como antes dijimos, todos los elementos de un vector deben ser del mismo modo. El otro atributo considerado en un vector es su *longitud*.

Si queremos conocer el modo o la longitud de un vector deben usarse las funciones *mode* y *length*.

La forma más sencilla de crear un vector es utilizando la función *c*. Por ejemplo, para crear el vector *x* formado por los números 1, 2 y 5, y conocer su modo y longitud ejecutaríamos la secuencia,

```
> x<-c(1,2,5)
> mode(x)
[1] "numeric"
> length(x)
[1] 3
```

Si los elementos de un vector son del modo *carácter*, debemos incluir dichos elementos no numéricos entre comillas. Así, ejecutando la expresión

```
> y<-c("Pepe", "Juan", "Luis Alfredo")
```

crearemos el vector y el cual tendrá tres elementos.

Obsérvese que los elementos de un vector pueden ser otros vectores. Por ejemplo, podríamos crear ahora el vector **z** formado por cinco elementos no numéricos de la forma

```
> z<-c("Felipe","Miguel Angel",y)
```

y ahora podemos hacer

```
> z
[1] "Felipe"    "Miguel Angel"  "Pepe"    "Juan"    "Luis Alfredo"

> mode(z)
[1] "character"

> length(z)
[1] 5
```

Una situación muy habitual es que tengamos nuestros datos en un fichero *ascii*, llamado, por ejemplo, *datos*. En ese caso, con objeto de crear un vector (no una matriz) debemos utilizar la función **scan**, direccionando el fichero *datos*. Así, si *datos* estuviera en un CD (de dirección *d:*), deberíamos ejecutar la expresión

```
> valores<-scan("d:\\datos")
```

con lo que habríamos creado el vector de datos **valores**. Lógicamente, si el *device* en donde están los datos no es *d*, sustituiríamos éste por el correspondiente. También deberíamos direccionar los datos si están en un subdirectorio. Por ejemplo, si estuvieran en el subdirectorio **curso** y los vamos a incorporar desde el *device* *d*, deberíamos ejecutar

```
> valores<-scan("d:\\curso\\datos")
```

1.3.2. Factores

El *factor* es un vector de datos no numéricos formado por datos procedentes de categorías; por ejemplo, datos obtenidos al anotar si el individuo es *hombre* o *mujer*.

1.3.3. Matrices

Como dijimos antes, una matriz es una disposición bidimensional en donde, al igual que ocurría con los vectores, todos los elementos deben ser del mismo *modo*.

Para crear una matriz utilizaremos la función `matrix` con dos argumentos, la función `c`, la cual tendrá a su vez como argumentos los datos a introducir, y el número de columnas que deberá tener la matriz. La matriz se construirá por columnas.

Por ejemplo, si queremos crear la matriz `ejemplo` en R a partir de la matriz

```
2 33 22 6
8 19 16 4
```

ejecutaríamos la expresión

```
> ejemplo<-matrix(c(2, 8, 33, 19, 22, 16, 6, 4),ncol=4)
```

Ahora podemos comprobar si lo hemos hecho bien ejecutando la expresión o sentencia que hemos marcado como (1)

```
> ejemplo                                     (1)
      [,1] [,2] [,3] [,4]
[1,]    2   33   22    6
[2,]    8   19   16    4
```

Podemos utilizar, en lugar del argumento `ncol`, el argumento `nrow`, el cual asigna el número de filas que deberá tener la matriz. No obstante, ésta se seguirá formando por columnas.

Otra posibilidad es utilizar ambos. En el caso de que queramos definir la matriz `z` con solamente cuatro de los ocho datos que teníamos en la matriz `ejemplo` anterior, ejecutaríamos la siguiente expresión

```
> z<-matrix(ejemplo,nrow=2,ncol=2)
```

Para comprobar el resultado obtenido ejecutaríamos

```
> z
```

obteniendo

```
> z
      [,1] [,2]
[1,]    2   33
[2,]    8   19
```

Observemos que R crea las matrices por columnas. Es decir, con los valores aportados por la función `c` va completando columnas. Si quisiéramos que la completara por filas, utilizaríamos el argumento `byrow=T`. Así, podemos ejecutar

```
> matrix(c(2, 8, 33, 19, 22, 16, 6, 4),ncol=4,byrow=T)
      [,1] [,2] [,3] [,4]
[1,]    2    8   33   19
[2,]   22   16    6    4
```

Observamos que algunos valores de los argumentos de las funciones se toman por defecto. Por esta razón es muy conveniente analizar los argumentos de la función que estemos utilizando con la opción `?funcion`.

Las matrices pueden ser de caracteres. Por ejemplo, el dato `personas`, formado por los seis individuos

Juan	Alfredo
Lupita	Enriqueta
Ernesto	Teodiselo

se obtendría mediante la secuencia

```
> personas <- matrix(c("Juan", "Lupita", "Ernesto", "Alfredo",
+ "Enriqueta", "Teodiselo"), ncol=2)
```

como comprobamos ejecutando (2)

```
> personas
      [,1] [,2]
[1,] "Juan" "Alfredo"
[2,] "Lupita" "Enriqueta"
[3,] "Ernesto" "Teodiselo"
(2)
```

Si tenemos dos o más vectores del mismo *modo* (es decir, numéricos o de caracteres) y además tienen la misma longitud, se pueden combinar para formar una matriz utilizando la función `cbind`.

Así por ejemplo, si tenemos un vector `x` con los consumos de veinte coches y un vector `y` con los kilómetros recorridos por esos mismos vehículos, se puede formar la matriz `w` de dimensión 20×2 mediante la expresión

```
> w <- cbind(x,y)
```

La función `cbind` une los vectores por columnas. De forma análoga se podría utilizar la función `rbind` para que los uniera por filas.

Además del *modo*, el otro atributo más importante de una matriz es su dimensión. Se puede averiguar mediante la función `dim`. Así, para averiguar la dimensión de la matriz `personas`, ejecutaríamos la expresión

```
> dim(personas)
[1] 3 2
```

que nos indica que es 3×2 .

Otra cuestión de interés en la construcción de matrices de datos es el nombre de las filas y columnas. Para poner nombre a las filas y columnas de una matriz se utiliza, dentro de la función `matrix`, el argumento `dimnames` el cual debe ser una lista de exactamente dos componentes, la primera de las cuales da los nombres de las filas de la matriz y la segunda la de los componentes. Así, si queremos poner nombres a las filas y las columnas de la matriz `ejem`

```
2  33  22  5
8  19  16  4
```

ejecutaríamos la expresión

```
> ejem<-matrix(c(2, 8, 33, 19, 22, 16, 5, 4), ncol=4,
+ dimnames=list(c("Individuo 1","Individuo 2"),
+ c("Hermanos","Edad","Peso","Escolaridad")))
```

Si queremos comprobar la operación realizada podemos ejecutar el nombre del nuevo objeto creado, obteniendo


```
> ejem
```

	Hermanos	Edad	Peso	Escolaridad
Individuo 1	2	33	22	5
Individuo 2	8	19	16	4

También es posible poner nombres a las filas y columnas de matrices ya creadas; por ejemplo, a la matriz anteriormente creada `z`

```
2 33
8 19
```

le podemos asignar nombres ejecutando la expresión

```
> dimnames(z) <- list(c("Individuo 1", "Individuo 2"),
+ c("Hermanos", "Edad"))
```

con lo que ejecutando `z` obtendríamos

```
> z
```

	Hermanos	Edad
Individuo 1	2	33
Individuo 2	8	19

Si queremos formar una matriz `A` (por ejemplo de 2 columnas) a partir de los datos de un fichero en un CD denominado `datos`, ejecutaríamos el siguiente comando, si el CD está en `d`:

```
> A<-matrix(scan("d:\\datos"),ncol=2)
```

1.3.4. Estructuras de datos

Las matrices (y los vectores) tienen una limitación importante: todos sus datos deben ser del mismo *modo*. Es decir, podemos tener *matrices numéricas*, o *matrices de caracteres* o *matrices lógicas*. No obstante, en la mayoría de las situaciones, nuestra *matriz de datos* asociada al experimento aleatorio que estamos considerando, contendrá datos de varios *modos* en diferentes columnas. En este caso, no podremos utilizar matrices, sino estructuras de datos o *data frames*.

Para crear *estructuras de datos* podemos utilizar dos funciones: Una, la función `data.frame`, la cual une al igual que la función `matrix`, objetos de varias clases, también por columnas.

Por otro lado, para leer datos procedentes de un fichero externo debemos utilizar la función `read.table`. Por ejemplo, si tenemos la siguiente matriz de datos en el fichero `d:datos`

	Peso	Talla	Sexo	Edad	EstaCivil
Ind1	65	1.65	F	45	Casado
Ind2	75	1.80	M	55	Casado
Ind3	80	1.95	M	47	Casado
Ind4	67	1.75	F	34	Soltero

podemos convertirlo en un data frame denominado `datos` ejecutando la siguiente expresión

```
> datos<-read.table("d:\\datos", header=T)
```

El argumento `header=T` es para indicar que queremos incorporar la primera línea de nombres de las variables.

1.3.5. Listas

El objeto-dato de R más flexible es la *lista*, el cual puede admitir datos de diferentes *modos* y de diferentes longitudes, e inclusive otras listas. Por esta razón, las listas son, habitualmente, el resultado final en donde se van incorporando los otros tipos de datos.

La mayoría de las funciones R que realizan un análisis estadístico, presentan sus resultados en una lista. Así, una lista podría estar constituida por los datos originales, los valores ajustados, los residuos, los estadísticos de contraste, el p-valor e inclusive hasta por el método empleado. Todo esto se combina en un sólo elemento: una lista.

Para crear una lista se utiliza la función `list` en donde cada uno de sus argumentos se convierte en una componente de la lista.

Por ejemplo, si queremos crear una lista formada por la estructura de datos `datos`, más arriba creada, y un vector formado por dos números, el 4 y el 5, ejecutaríamos `list(datos, c(4,5))` y R respondería

```
[[1]]:
  Peso  Talla Sexo Edad EstaCivil
Ind1 "65" "1.65" "F"  "45"  "Casado"
Ind2 "75" "1.80" "M"  "55"  "Casado"
```

```
Ind3 "80" "1.95" "M" "47" "Casado"
Ind4 "67" "1.75" "F" "34" "Soltero"
```

```
[[2]]:
[1] 4 5
```

1.3.6. Nombres a las filas y columnas de matrices y vectores

Mediante la función `names` podemos crear un *vector de nombres* de la misma longitud que el vector. Por ejemplo, si queremos crear un vector formado por el 7, el 4 y el 3, luego asignarle a esos elementos los nombres *Primer examen*, *Segundo examen* y *Tercer examen* y, por último, comprobar el resultado, tendríamos la siguiente secuencia

```
> z<-c(7,4,3)
> names(z)<-c("Primer examen","Segundo examen","Tercer examen")
> z
Primer examen Segundo examen Tercer examen
              7              4              3
```

Si sólo queremos conocer los nombres del vector `z`, ejecutaríamos la expresión `names(z)`.

1.4. Gráficos

La ventana de gráficos se abre de forma automática al ejecutar alguna función que los realice. Estas funciones se dividen en funciones gráficas de *alto nivel* y de *bajo nivel*. Con las primeras se obtiene un gráfico nuevo, mientras que con las segundas podemos hacer modificaciones de un gráfico ya existente. Además podemos controlar aspectos específicos de nuestros gráficos usando parámetros gráficos adicionales.

1.4.1. Funciones gráficas de alto nivel

Como antes dijimos, las funciones gráficas de alto nivel producen un gráfico totalmente nuevo, incluidos los ejes y sus etiquetas, borrando previamente el gráfico que pudiera existir en la ventana de gráficos.

Las funciones gráficas de alto nivel se dividen en varios grupos dependiendo, fundamentalmente, de lo que queramos representar. Si queremos representar funciones matemáticas, primero debemos crear un vector de valores correspondientes a las abscisas, es decir, al dominio de la función en los que va a ser evaluada ésta y, luego, realizar el gráfico deseado de pares de puntos utilizando la función de R, `plot`. Por ejemplo, para representar la función $f(x) = \sin(x)$ entre $-\pi$ y π , ejecutaríamos las siguientes dos expresiones

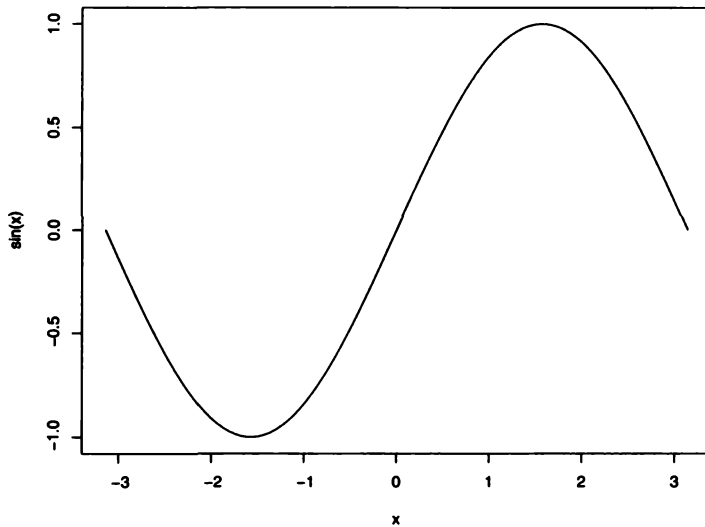


Figura 1.1 : Gráfico de la función seno

```
> x<-seq(-pi, pi, len=100)
> plot(x, sin(x), type="l")
```

con lo que aparecería en la ventana de gráficos la Figura 1.1. El argumento

`type="l"`

(en donde hemos utilizado la letra `l` y no el número 1) especifica que el gráfico de la función debe aparecer mediante trazos sólidos.

Si queremos representar pares de datos (x,y) mediante un diagrama de dispersión, simplemente utilizaríamos la función `plot`

```
> plot(x,y)
```

Ejemplo 1.1

Los siguientes pares de datos corresponden al *Peso de caucho* (en gramos) obtenido después de la vulcanización, variable Y , y la *Circunferencia de la corona del Guayule* (en cm.) de donde se obtuvo dicho caucho, variable X ,

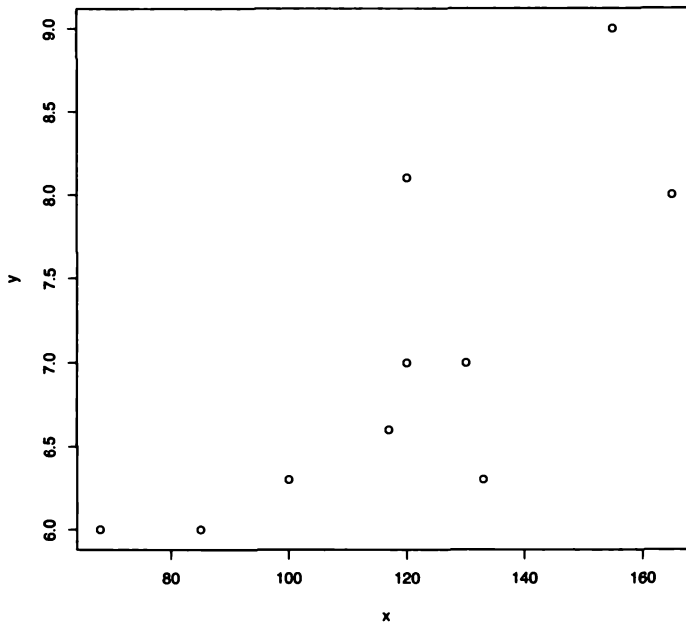


Figura 1.2 : Diagrama de dispersión

X	68	100	85	133	130	165	120	120	155	117
Y	6	6.3	6	6.3	7	8	7	6.3	9	6.6

Si queremos representar estos pares de datos en un diagrama de dispersión basta con que ejecutemos la siguiente secuencia de instrucciones (expresiones en terminología R), obteniendo el gráfico de la Figura 1.2.

```
> x<-c(68,100,85,133,130,165,120,120,155,117)
> y<-c(6,6.3,6,6.3,7,8,7,6.3,9,6.6)
> plot(x,y)
```

En el capítulo siguiente veremos que la función **plot** admite más argumentos, además de los pares de datos a representar, lo que nos permitirá obtener gráficos más interesantes.

1.4.2. Funciones gráficas de bajo nivel

Se utilizan para modificar gráficos ya existentes. Por tanto, al ser ejecutadas, no suprimen el gráfico existente en la ventana de gráficos.

Las funciones `points` y `lines` son las funciones gráficas de bajo nivel correspondientes a `type= p` y `type= l` respectivamente.

La función `abline` permite añadir una línea recta a un gráfico ya existente, especificando los valores de la ordenada en el origen y la pendiente. En el capítulo siguiente veremos algunas funciones más.

1.5. Otras cuestiones

Unas cuestiones adicionales que conviene conocer son que, por ejemplo, R distingue entre mayúsculas y minúsculas, luego dos expresiones pueden ser distintas aunque tengan las mismas letras en el mismo orden. Por el contrario, uno o varios espacios son interpretados de la misma manera.

R no ejecuta lo que haya en una línea detrás del símbolo `#`, por lo que, en ocasiones, se incluyen comentarios después de una expresión, comenzando éstos por `#`.

Un símbolo que aparece con frecuencia sobre todo en cuestiones relacionadas con Regresión, es el símbolo

~

que corresponde con el de código ASCII 126, con lo que se obtendrá manteniendo presionada la tecla `[Alt]` y, al mismo tiempo, tecleando en el teclado numérico el número 126; al soltar la tecla `[Alt]` obtendrá dicho símbolo.

1.6. Interfaz

El lector seguramente se estará preguntando si no es posible ejecutar R desde algún tipo de Interfaz (*Interface*) a base de ventanas y cuadros de diálogos, de manera similar a como lo hace, por ejemplo, el conocido paquete SPSS. La respuesta es que sí, efectivamente podría trabajar con un interfaz de forma que, en algunos casos, no necesite ejecutar sentencias R. Quede claro que se seguirán ejecutando en R pero será el interfaz quien lo haga; además, con éste sólo podrá ejecutar algunos métodos estadísticos muy básicos, razón por la cual no es la forma habitualmente elegida por los usuarios de R que suelen utilizar la línea de comandos.

Existen varios interfaces. Uno de ellos se denomina *DAS+R* y se puede obtener de

<http://www.statistik.tuwien.ac.at/StatDA/DASplusR>

Una vez hayamos *bajado* la carpeta .zip (denominada, salvo algún número al final, *DASplusR.zip*) se instala ejecutando, si la tenemos por ejemplo en c: ,

```
> install.packages(repos=NULL, "c://DASplusR.zip")
```

Pero el interfaz más habitualmente utilizado es *Rcmdr*, denominado *R-Commander*, y que se instala en un sesión de R (estando conectado a Internet) con la pestaña superior de la consola R, *Paquetes*.

Estando en una sesión de R (es decir, una vez *abierto* R), obtenemos el primer interfaz ejecutando

```
> library(DASplusR)
```

El interfaz *Rcmdr* se obtiene ejecutando

```
> library(Rcmdr)
```

1.7. Modificar y Crear Funciones

Cuando utilizamos R habitualmente estamos ejecutando funciones incorporadas a este paquete. Recordamos que todas las funciones dependen de unos *argumentos* que serían las correspondientes “variables” en términos de funciones matemáticas. Algunos valores de estos argumentos vienen pre-determinados y no es necesario asignarles un valor; por ejemplo, se puede obtener información sobre una función determinada, digamos *var*, ejecutando

```
> ?var
```

Como dijimos antes, las funciones se pueden modificar a nuestra conveniencia con las *funciones editoras* *fix* y *edit* (la primera de ellas lo que hace es llamar a la segunda pero al salir la modifica, luego es más recomendable).

Al modificar una función previamente incluida en R, habitualmente no querremos prescindir de ella, por lo que definiremos primero una nueva y modificaremos la nueva. Así por ejemplo, la función de R *var* calcula cuasi-varianzas muestrales y no varianzas muestrales (véase el Capítulo 2 si quiere

ahora su definición); es decir, divide la suma de los cuadrados de las diferencias por el tamaño de la muestra menos 1, $n - 1$, y no divide por n . Así, si incorporamos unos datos a R ejecutando (1) y luego ejecutamos (2), el valor obtenido, 4.66 es la cuasivarianza, puesto que la media de los cuadrados menos el cuadrado de la media, forma alternativa de calcular la varianza de unos datos, se obtienen ejecutando (3), comprobando que es igual a 4. De hecho, ejecutando (4) vemos que esto es así.

```
> x<-c(6,4,5,2,1,0,3) (1)
```

```
> var(x) (2)
```

```
[1] 4.666667
```

```
> mean(x^2)-mean(x)^2 (3)
```

```
[1] 4
```

```
> 6*var(x)/7 (4)
```

```
[1] 4
```

Por tanto, si queremos crear una nueva función, que denominaremos *varianza*, primero la creamos para no eliminar la existente ejecutando (5) y, a continuación, editamos esta nueva con (6) con lo que aparece lo que viene a continuación, que es todavía la función *var*

```
> varianza<-var (5)
```

```
> fix(varianza) (6)
```

```
function (x, y = NULL, na.rm = FALSE, use)
{
  if (missing(use))
    use <- if (na.rm)
      "complete.obs"
    else "all.obs"
  na.method <- pmatch(use, c("all.obs", "complete.obs", "pairwise.complete.obs"))
  if (is.data.frame(x))
    x <- as.matrix(x)
  else stopifnot(is.atomic(x))
  if (is.data.frame(y))
    y <- as.matrix(y)
  else stopifnot(is.atomic(y))
  .Internal(cov(x, y, na.method, FALSE))
}
```

Ahí podíamos pre-multiplicar la última línea por la longitud del vector menos 1 y dividir por esa longitud, es decir, incluir $(\text{length}(x)-1)/\text{length}(x)*$ obteniendo

```
> fix(varianza)
```



```
function (x, y = NULL, na.rm = FALSE, use)
{
  if (missing(use))
    use <- if (na.rm)
      "complete.obs"
    else "all.obs"
  na.method <- pmatch(use, c("all.obs", "complete.obs", "pairwise.complete.obs"))
  if (is.data.frame(x))
    x <- as.matrix(x)
  else stopifnot(is.atomic(x))
  if (is.data.frame(y))
    y <- as.matrix(y)
  else stopifnot(is.atomic(y))
  (length(x)-1)/length(x)*Internal(cov(x, y, na.method, FALSE))
}

> varianza(x)
[1] 4
```

Seguramente es más sencillo utilizar la función `var` en la definición de una nueva función, que podemos denominar `varianza2`, para lo cual no necesitamos, ni siquiera, *salvar* ninguna función previamente. Así, ejecutamos

```
> fix(varianza2)
```

apareciendo

```
function ()
{
}
```

Ahora sí que tenemos que programar de verdad una función. Primero debemos determinar el o los argumentos a considerar. En nuestro ejemplo podemos considerar los mismos cuatro argumentos de la función `var` o, si sólo la vamos a utilizar en muestras unidimensionales, incluir sólo una variable que podemos denominar como queramos, digamos `x`. Luego le decimos que calcule la verdadera varianza a partir de la cuasivarianza calculada con `var` tecleando

```
function(x)
{
  (length(x)-1)/length(x)*var(x)
}
```

al salvarla tendremos nuestra nueva función `varianza2`. El resultado debe ser el mismo por ambos caminos; lo comprobamos

```
> varianza(x)
[1] 4
> varianza2(x)
[1] 4
```

Esta segunda vía de crear funciones que incluyan nuevas funciones es la habitual de R. Lógicamente, para poder programar hay que conocer y manejar muchas funciones previamente.

Tenga en cuenta que las funciones de R que utilice en una nueva función, tienen que estar previamente en R. Si elige funciones muy específicas que no estuvieran, al invocarlas con la nueva función, se producirá un error.

Tenga cuidado con el editor. Si ha cometido algún error al programar, al cerrarlo se perderá todo lo que hubiera hecho, por lo que le aconsejamos copiarlo antes de cerrarlo.

Recuerde que al salir de R, debe decir que Sí quiere conservar los cambios porque, en caso contrario no le quedarán salvados. Si así lo hace, no tendrá que volver a repetir todo este proceso y tendrá una (en realidad si ha seguido los dos caminos, dos) nuevas funciones en su R.

1.8. Librerías de R

Una de las ventajas de R es que cada usuario puede crear funciones, según sus intereses y conveniencias, reuniéndolas todas en lo que se denomina un *paquete* o *librería*, el cual habitualmente pone al servicio de toda la comunidad científica, siempre de forma gratuita, acompañando a la librería de un pequeño manual de utilización. Esto aumenta la difusión de R que, a su vez, conduce a la creación de nuevas librerías aumentando la difusión de R y, así sucesivamente. Además, este proceso es muy rápido en el tiempo, lo que permite disponer de la forma de ejecutar nuevas técnicas a la vez que se estudian éstas.

Observemos que, de nuevo, nadie se responsabiliza de los cálculos que se obtengan con esas funciones por lo que siempre debemos, en nuestra opinión, verificar, analizar y estudiar lo qué hace cada función, antes de utilizarla. No obstante, si algún autor de prestigio la cita podemos confiar en que la haya chequeado.

Muchas librerías ya estarán incorporadas a la versión de R que utilicemos por lo que, si queremos comprobar si tenemos en nuestro programa una determinada librería, digamos *chemometrics*, ejecutaremos

```
> library(chemometrics)
```

Si el programa no dice nada es que la tenemos y la hemos *abierto* por lo que podemos utilizar las funciones que contiene. Apuntamos el hecho de que esta gran flexibilidad de R tiene el inconveniente de que suele aparecer una nueva versión de R cada pocos meses y en ésta algunas librerías habrán desaparecido integrándose a alguna nueva y algunas librerías nuevas habrán sido incorporadas. La flexibilidad y agilidad tienen un precio.

En algunas ocasiones tendremos claro qué librería queremos incorporar a R pero en otras ocasiones no. En estas últimas, una buena manera de actuar es buscar en Google el nombre (en inglés) del método que queramos utilizar precedido de una R entre corchetes. Por ejemplo, si queremos ejecutar Redes Neuronales con R, podemos buscar en Google

[R] neural networks

y el primer resultado que obtenemos en la búsqueda es la dirección

<http://stat.ethz.ch/R-manual/R-patched/library/nnet/html/nnet.html>

en donde lo primero que observamos es lo siguiente:

```
nnet {nnet} R Documentation

Fit Neural Networks

Description

Fit single-hidden-layer neural network, possibly with skip-layer connections.
```

es decir, que hay una librería denominada **nnet** que es el nombre que aparece entre corchetes a la izquierda, una de cuyas funciones denominada también **nnet**, ajusta, como dice, redes neuronales.

En la dirección

<http://lib.stat.cmu.edu/R/CRAN/web/packages/>

tenemos la relación de librerías “oficiales” en donde aparecerá ésta (cuidado con el orden alfabético porque distingue entre mayúsculas y minúsculas). *Pinchando* en su nombre tendremos, entre otras cosas, un fichero en pdf que nos da indicaciones de lo que hace.

Pero lo más interesante es que podemos incorporarla fácilmente a nuestro R. Para ello, con R abierto y conectados a Internet, desplegamos la pestaña superior **Paquetes** y elegimos la opción **Seleccionar espejo CRAN**; aquí elegimos, preferiblemente, algún lugar cercano a donde tengamos instalado el ordenador. Después, dentro de la misma pestaña **Paquetes**, elegimos la opción **Instalar paquete(s)** y allí seleccionamos el paquete que estamos buscando.

Bastará hacer esto una sola vez. Luego ya estará instalado en R como cualquier otro paquete y para abrirlo sólo tendremos que ejecutar la función `library`.

1.9. Lecturas Recomendadas

Braun W.J. y Murdoch, D.J. (2007). *A First Course in Statistical Programming*. Editorial Cambridge.

Dalgaard, P. (2002). *Introductory Statistics with R*. Editorial Springer.

García Pérez, A. (2008c). *Estadística Aplicada con R*. Editorial UNED. Colección: Varia.

Capítulo 2

Estadística Descriptiva

2.1. Introducción a la Estadística

Parece razonable comenzar el estudio de la Estadística hablando de los fenómenos que esta disciplina analiza. Se trata de los *Fenómenos Aleatorios* que son aquellos fenómenos o experiencias que, incluso en las mismas condiciones, pueden dar lugar a diferentes resultados.

Con la *Estadística Descriptiva* aprenderemos a describir los resultados de los fenómenos aleatorios y, lo que será mucho más interesante, con la *Inferencia Estadística* aprenderemos a entenderlos y predecirlos, comparando grupos por ejemplo, o estimando los valores más verosímiles, 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.

Además del interés que tiene la Estadística Descriptiva por sí misma, ésta nos servirá de introducción a las otras dos partes mencionadas. Así por ejemplo, las distribuciones de frecuencias relativas nos serán de utilidad en la definición de Probabilidad, o los Histogramas nos sugerirán posibles Distribuciones de Probabilidad, como veremos al considerar estos conceptos más adelante. Y también, las medidas de posición o dispersión aquí estudiadas, entre otras, se convertirán en estadísticos en Inferencia Estadística.

Pero comencemos con los fenómenos aleatorios. Sus ejemplos son innumerables: la aplicación de un determinado tratamiento médico a un grupo de personas 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. De hecho, 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 a la hora de sacar conclusiones.

2.1.1. 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 los enfermos españoles que padecen la enfermedad en estudio, o los enfermos del 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 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 habitualmente 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.

2.1.2. 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 cardíaco. Podríamos pensar que, cerrando los ojos y metiendo la mano en la jaula de los ratones, los cinco primeros seleccionados pueden 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 cardíaco, 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 continuación, 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.

2.1.3. Variable aleatoria y Modelo probabilístico

Habitualmente la situación que se plantea es la de un investigador que desea estimar 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 verosímilmente 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) que rija 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 estimar 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*.

2.1.4. Diferentes Estadísticas

Como 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.

2.2. Conceptos fundamentales de la Estadística Descriptiva

Comenzaremos definiendo algunos conceptos propios de la terminología de la Estadística Descriptiva.

Caracteres

Cada uno de los individuos de la población en estudio posee uno o varios *caracteres*. Así por ejemplo, si la población en consideración es la de los estudiantes de una determinada universidad, éstos poseerán una serie de caracteres, o si se quiere características, que permiten describirlo. Los caracteres en este ejemplo pueden ser “Facultad en la que está matriculado”, “Curso que sigue”, “Sexo”, “Edad”, etc. Precisamente la observación de uno o más de esos caracteres en los individuos de la muestra es lo que dará origen a los datos.

Los caracteres pueden ser de dos clases: *cuantitativos*, cuando son tales que su observación en un individuo determinado proporciona un valor numérico como medida asociada, como ocurre por ejemplo con los caracteres “Edad” o “Curso que sigue”, y *cualitativos*, cuando su observación en los individuos no suministra un número, sino la pertenencia a una clase determinada, como por ejemplo el “Sexo” o la “Facultad en la que está matriculado”.

Modalidades de los caracteres

Consideremos un carácter cualquiera como por ejemplo el “Gusto”. Este carácter, al ser observado en un individuo (una sustancia), puede presentar cuatro posibilidades, es decir, es posible percibir cuatro sensaciones diferentes: dulce, amargo, salado y ácido. Pues bien, a las posibilidades, tipos o clases que pueden presentar los caracteres las denominaremos *modalidades*.

Las modalidades de un carácter deben ser a la vez incompatibles y exhaustivas. Es decir, las diversas modalidades de un carácter deben cubrir todas las posibilidades que éste puede presentar y, además, deben ser disjuntas (un individuo no puede presentar más de una de ellas y debe presentar alguna de ellas).

Así, al estudiar algún carácter, como por ejemplo la Raza, el investigador deberá considerar todas las posibles modalidades del carácter (todas las posibles razas), con objeto de poder clasificar a todos los individuos que observe.

La matriz de datos

Habitualmente, la información primaria sobre los individuos, es decir, la forma más elemental en la que se expresan los datos es la de una matriz, en la que aparecen en la primera columna los individuos identificados de alguna manera y en las siguientes columnas las observaciones de los diferentes carac-

terres en estudio para cada uno de los individuos, tal y como aparece en la Tabla 2.1. Dicha matriz recibe el nombre de *matriz de datos*.

	Carácter 1	Carácter 2	...	Carácter p
Individuo 1	•	•	...	•
Individuo 2	•	•	...	•
...
Individuo n	•	•	...	•

Tabla 2.1

Así, los datos correspondientes a una investigación llevada a cabo para el estudio de una posible contaminación radioactiva en un determinado lugar produjeron como resultado la matriz de datos de la Tabla 2.2, en donde se recogen las observaciones de los caracteres “Edad”, “Sexo”, “Cáncer”, “Caída anormal del cabello” y “Profesión” en los 100 individuos seleccionados en la muestra.

	<i>Edad</i>	<i>Sexo</i>	<i>Cáncer</i>	<i>Caída cabello</i>	<i>Profesión</i>
Individuo 1	32	masculino	no	no	agricultor
Individuo 2	29	femenino	no	no	maestra
...
Individuo 100	61	masculino	sí	sí	agricultor

Tabla 2.2

En algunas ocasiones se reserva el nombre de matriz de datos a la obtenida de la anterior eliminando la primera columna.

Clases de datos

Es habitual denominar a los caracteres *variables estadísticas* o simplemente *variables*, calificándolas de cualitativas o cuantitativas según sea el correspondiente carácter, y hablar de los *valores de la variable* al referirnos a sus modalidades aunque, de hecho, solamente tendremos verdaderos valores numéricos cuando analicemos variables cuantitativas.

En ocasiones, con objeto de facilitar la toma de los datos, el investigador los agrupa en intervalos. Así por ejemplo, resulta más sencillo averiguar cuántos individuos hay en una muestra con una estatura, por ejemplo, entre 1'70 y 1'80 que medirlos a todos, en especial si tenemos marcas en la pared cada 10 cm.

Observemos, no obstante, que siempre se producirá una pérdida de información al agrupar los datos en intervalos y dado que hoy en día la utilización del ordenador es de uso corriente, un agrupamiento en intervalos es, en general, desaconsejable.

No obstante, por razones docentes admitiremos esta posibilidad, ya que precisamente el agrupamiento en intervalos traerá complicaciones adicionales en el cálculo de algunas medidas representativas de los datos.

Consideraremos, por tanto, tres tipos posibles de datos:

- I. Datos correspondientes a un carácter cualitativo.
- II. Datos sin agrupar correspondientes a un carácter cuantitativo.
- III. Datos agrupados en intervalos correspondientes a un carácter cuantitativo.

Agrupamiento en intervalos

Si tenemos la opción de poder agrupar los datos en intervalos, lo primero que debemos plantearnos (independientemente de lo que más arriba comentábamos) es la cuestión de cuántos y cuáles intervalos elegir.

Previamente daremos algunas definiciones. Si los intervalos o *clases*, como a veces se denominan, son:

$$[c_0 - c_1), [c_1 - c_2), \dots, [c_{j-1} - c_j), \dots, [c_{k-1} - c_k]$$

llamaremos *extremos* de la clase j -ésima a c_{j-1} y a c_j ; *amplitud* del intervalo a la diferencia de sus extremos, hablando de intervalos de amplitud constante o variable según tengan o no todos la misma amplitud y, por último, llamaremos *centro* o *marca de clase* correspondiente al intervalo j -ésimo al punto medio del intervalo; es decir, a $c'_j = (c_j + c_{j-1})/2$.

A lo largo del texto consideraremos que el dato c_j pertenece al intervalo $j + 1$, $j = 1, \dots, k - 1$, siendo el c_k del k -ésimo. Hacemos notar también que el primer y último intervalo generalmente tienen, respectivamente, el extremo inferior y superior indeterminados con objeto de incluir observaciones poco frecuentes.

Respecto a la cuestión que nos planteábamos al comienzo de este apartado, podemos considerar como regla general la de construir, siempre que sea posible, intervalos de amplitud constante, sugiriendo sobre el número k de intervalos a considerar el propuesto por Sturges (1926),

$$k = 1 + 3'322 \log_{10} n$$

siendo n el número total de datos.

Una vez determinado el número k de intervalos a considerar, y si es posible tomarlos de igual amplitud, ésta será

$$c = \frac{x_{(n)} - x_{(1)}}{k}$$

en donde $x_{(n)}$ es el dato mayor y $x_{(1)}$ el menor.

Ejemplo 2.1

Se midieron los niveles de colinesterasa en un recuento de eritrocitos en $\mu\text{mol}/\text{min}/\text{ml}$ de 34 agricultores expuestos a insecticidas agrícolas, obteniéndose los siguientes datos:

Individuo	Nivel	Individuo	Nivel	Individuo	Nivel
1	10'6	13	12'2	25	11'8
2	12'5	14	10'8	26	12'7
3	11'1	15	16'5	27	11'4
4	9'2	16	15'0	28	9'3
5	11'5	17	10'3	29	8'6
6	9'9	18	12'4	30	8'5
7	11'9	19	9'1	31	10'1
8	11'6	20	7'8	32	12'4
9	14'9	21	11'3	33	11'1
10	12'5	22	12'3	34	10'2
11	12'5	23	9'7		
12	12'3	24	12'0		

Tabla 2.3

Aplicando la fórmula de Sturges obtenemos,

$$k = 1 + 3'322 \log_{10} 34 = 1 + 3'322 \cdot 1'53148 = 6'08757$$

es decir, una sugerencia de 6 intervalos. Como el mayor valor es $x_{(34)} = 16'5$ y el menor $x_{(1)} = 7'8$, la longitud sugerida es

$$c = \frac{16'5 - 7'8}{6} = 1'45.$$

Parece, por tanto, razonable tomar como amplitud 1'5, obteniendo como intervalos en los que clasificar los datos

$$7'5 - 9, 9 - 10'5, 10'5 - 12, 12 - 13'5, 13'5 - 15, 15 - 16'5$$

2.3. Distribuciones unidimensionales de frecuencias

En este apartado consideraremos que tenemos datos correspondientes a un solo carácter, el cual, como antes dijimos llamaremos variable estadística y representaremos por X .

Llamaremos *frecuencia total* al número de datos n . Llamaremos *frecuencia absoluta* n_i de la modalidad M_i (valor x_i o intervalo I_i) de la variable X al número de datos que presentan la modalidad M_i (valor x_i o valor del intervalo I_i). Si existen k modalidades posibles, se verificará

$$\sum_{i=1}^k n_i = n_1 + n_2 + \dots + n_k = n.$$

Llamaremos *frecuencia relativa* f_i de la modalidad M_i (valor x_i o intervalo I_i) de la variable X al cociente $f_i = n_i/n$, verificándose,

$$\sum_{i=1}^k f_i = f_1 + f_2 + \dots + f_k = 1.$$

Llamaremos *frecuencia absoluta acumulada* N_i hasta la modalidad M_i (valor x_i o intervalo I_i) a la suma

$$N_i = n_1 + \dots + n_i = \sum_{j=1}^i n_j.$$

Claramente es $N_k = \sum_{j=1}^k n_j = n$.

Llamaremos *frecuencia relativa acumulada* F_i hasta la modalidad M_i (valor x_i o intervalo I_i) al cociente $F_i = N_i/n$, o lo que es lo mismo, a

$$F_i = f_1 + \dots + f_i = \sum_{j=1}^i f_j$$

siendo $F_k = \sum_{j=1}^k f_j = 1$.

Distribuciones unidimensionales de frecuencias

La tabla formada por las distintas modalidades (valores o intervalos) del carácter X y por las frecuencias absolutas (relativas, absolutas acumuladas o relativas acumuladas) recibe el nombre de *distribución de frecuencias absolutas* (*relativas*, *absolutas acumuladas* o *relativas acumuladas* respectivamente).

Tenemos, por tanto, para cada tipo de datos, cuatro distribuciones de frecuencias, obteniéndose a partir de una cualquiera de ellas las tres restantes, supuesto que se conoce la frecuencia total.

Las cuatro distribuciones de frecuencias se expresan en tablas como las siguientes dependiendo del tipo de datos que sean

Carácter cualitativo:

M_i	n_i	f_i	N_i	F_i
M_1	n_1	f_1	N_1	F_1
M_2	n_2	f_2	N_2	F_2
...
M_i	n_i	f_i	N_i	F_i
...
M_k	n_k	f_k	$N_k = n$	$F_k = 1$
	n	1		

Tabla 2.4

Carácter cuantitativo sin agrupar:

X_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
...
x_i	n_i	f_i	N_i	F_i
...
x_k	n_k	f_k	$N_k = n$	$F_k = 1$
	n	1		

Tabla 2.5

Carácter cuantitativo agrupado en intervalos:

I_i	n_i	f_i	N_i	F_i
I_1	n_1	f_1	N_1	F_1
I_2	n_2	f_2	N_2	F_2
...
I_i	n_i	f_i	N_i	F_i
...
I_k	n_k	f_k	$N_k = n$	$F_k = 1$
	n	1		

Tabla 2.6

Ejemplo 2.2

En un estudio sobre las razones por las que no fue completado un tratamiento de radiación seguido de cirugía en pacientes de cáncer de cabeza y cuello se obtuvieron los datos dados por la siguiente distribución de frecuencias absolutas,

<i>Causas</i>	n_i
Rehusaron cirugía	26
Rehusaron radiación	3
Empeoraron por una enfermedad ajena al cáncer	10
Otras causas	1
	40

Tabla 2.7

Las cuatro distribuciones de frecuencias serán,

<i>Causas</i>	n_i	f_i	N_i	F_i
Rehusaron cirugía	26	0'650	26	0'650
Rehusaron radiación	3	0'075	29	0'725
Empeoraron por una enfermedad ajena al cáncer	10	0'250	39	0'975
Otras causas	1	0'025	40	1
	40	1		

Tabla 2.8

Ejemplo 2.3

Tras encuestar a 25 familias sobre el número de hijos que tenían, se obtuvieron los siguientes datos,

X_i	n_i
0	5
1	6
2	8
3	4
4	2
	25

Tabla 2.9

Las cuatro distribuciones de frecuencias serán,

X_i	n_i	f_i	N_i	F_i
0	5	0'20	5	0'20
1	6	0'24	11	0'44
2	8	0'32	19	0'76
3	4	0'16	23	0'92
4	2	0'08	25	1
	25	1		

Tabla 2.10

Ejemplo 2.1 (continuación)

Los datos del Ejemplo 2.1, agrupados en los intervalos allí obtenidos, proporcionan las cuatro siguientes distribuciones de frecuencias,

I_i	n_i	f_i	N_i	F_i
7'5 – 9	3	0'088	3	0'088
9 – 10'5	8	0'236	11	0'324
10'5 – 12	10	0'294	21	0'618
12 – 13'5	10	0'294	31	0'912
13'5 – 15	1	0'029	32	0'941
15 – 16'5	2	0'059	34	1
	34	1		

Tabla 2.11

2.3.1. Representaciones gráficas de las distribuciones unidimensionales de frecuencias

La representación gráfica de una distribución de frecuencias depende del tipo de datos que la constituya.

Datos correspondientes a un carácter cualitativo

La representación gráfica de este tipo de datos está basada en la proporcionalidad de las áreas a las frecuencias absolutas o relativas. Veremos dos tipos de representaciones: el *Diagrama de Sectores* y el *Diagrama de Rectángulos*.

Diagrama de sectores:

Esta representación gráfica consiste en dividir un círculo en tantos sectores circulares como modalidades presente el carácter cualitativo, asignando un



Figura 2.1 : Gráfico de Sectores del Ejemplo 2.2

ángulo central a cada sector circular proporcional a la frecuencia absoluta n_i , consiguiendo de esta manera un sector con área proporcional también a n_i .

Así, los ángulos que corresponden a las cuatro modalidades del Ejemplo 2.2 serán,

Rehusaron cirugía:	234
Rehusaron radiación:	27
Empeoraron por una enfermedad ajena al cáncer:	90
Otras causas:	9

cuya representación gráfica sería la Figura 2.1. Ésta se obtendría con R, primero introduciendo los datos en (1) y, luego, ejecutando (2)

```
> x2<-c(26,3,10,1) (1)
> pie(x2) (2)
```

El problema es que, así, el ordenador elige unos colores arbitrarios y, lo que es más importante, denomina con simples números los sectores correspondientes a las clases que presenta la variable cualitativa. Si queremos que denomine de una manera concreta a los sectores, debemos crear primero un vector de nombres, es decir, un vector de caracteres, como hacemos en (3), pudiendo crear también un vector de colores en (4), obteniendo el gráfico deseado al ejecutar (5)

```
> n2<-c("reh. cirugía","reh. radia.,"empeoraron o.e.,"otras") (3)
```

```
> c2<-c(2,3,4,5) (4)
```

```
> pie(x2,labels=n2,col=c2) (5)
```

Si quisiéramos, además, poner título al gráfico podríamos utilizar otro argumento de la función `pie`, ejecutando (6) y obteniendo, finalmente, la Figura 2.1. Apuntamos el hecho de que este argumento `main`, así como `col` para dar color al gráfico, lo son de todas las funciones gráficas de R.

```
> pie(x2,labels=n2,col=c2,main="Causas") (6)
```

Diagrama de rectángulos:

Esta representación gráfica consiste en construir tantos rectángulos como modalidades presente el carácter cualitativo en estudio, todos ellos con base de igual amplitud. La altura se toma igual a la frecuencia absoluta o relativa (según la distribución de frecuencias que estemos representando), consiguiendo de esta manera rectángulos con áreas proporcionales a las frecuencias que se quieren representar.

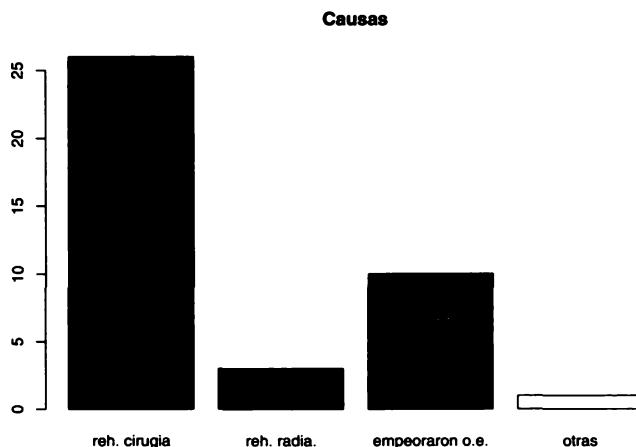


Figura 2.2 : Diagrama de Rectángulos del Ejemplo 2.2

La representación gráfica de la distribución de frecuencias absolutas del Ejemplo 2.2 sería la Figura 2.2. Ésta se obtuvo ejecutando (7), en donde la única variación con respecto a la función `pie`, es que `labels` no es un argumento de la función `barplot` sino que, como puede verse, el argumento correspondiente para añadir nombres a las clases es, `names`.

```
> barplot(x2,names=n2,col=c2,main="Causas") (7)
```

Datos correspondientes a un carácter cuantitativo agrupado en intervalos

Esta situación rara vez se presenta en Estadística porque la agrupación en intervalos implica pérdida de información: todos los datos del intervalo son tratados de igual manera al ser considerados iguales a la marca de clase, independientemente de los valores reales que tomaran. La razón fundamental de su uso hasta nuestros días fue la complejidad de manejar grandes cantidades de datos, problema resuelto con el uso habitual de los ordenadores. Por tanto, el agrupar datos es una opción no recomendada. Como es posible, no obstante, que, en algunas ocasiones los datos los tengamos en intervalos, vamos a indicar cómo representarlos.

La representación habitual es el *Histograma* en donde sobre cada intervalo se levanta un rectángulo con un área igual a la frecuencia, absoluta o relativa según la distribución que estemos considerando, por lo que hay que tener en cuenta si los intervalos tienen igual o distinta amplitud. La representación gráfica se consigue con la función `hist` aunque veremos que esta función está pensada para datos sin agrupar. El *Polígono de Frecuencias Acumuladas* podría realizarse de forma análoga a como se obtendrá la Función de distribución empírica en el siguiente apartado.

Ejemplo 2.1 (continuación)

Si los datos del Ejemplo 2.1 se agrupan en intervalos según la siguiente distribución de frecuencias,

I_i	n_i
7'5 – 9	3
9 – 10'5	8
10'5 – 12	10
12 – 13'5	10
13'5 – 15	1
15 – 16'5	2
	34

Tabla 2.12

para obtener el Histograma, primero introducimos en (1) las marcas de clase y en (2) las frecuencias absolutas. En (3), con la función `rep`, replicamos las marcas de clase, tantas veces como sea la frecuencia absoluta del intervalo del que es marca de clase obteniendo así los datos `col1` a representar. Vemos lo que comentábamos más arriba de la pérdida de información que se produce con la agrupación por intervalos: todos los datos de un intervalo son considerados iguales a la marca de clase.

Finalmente, en (4) indicamos cuáles queremos que sean los puntos de corte de los intervalos en la representación gráfica y en (5) los colores a utilizar en el gráfico, obteniendo el histograma al ejecutar (6).

De esta manera, el área del histograma no sumará 1 como habitualmente deseamos. Para conseguir esto, debemos utilizar el argumento `prob=T` ejecutando (7) y obteniendo la Figura 2.3 en donde de nuevo hemos puesto título al gráfico, utilizando el argumento `main`.

```

> m1<-c(8.25,9.75,11.25,12.75,14.25,15.75)           (1)
> n1<-c(3,8,10,10,1,2)                               (2)
> coli<-rep(m1,n1)                                     (3)
> d1<-c(7.5,9,10.5,12,13.5,15,16.5)                 (4)
> c1<-c(1,2,3,4,5,6)                                 (5)
> hist(coли,breaks=d1,col=c1)                         (6)
> hist(coли,breaks=d1,col=c1,main="Niveles de colinesterasa",prob=T) (7)

```

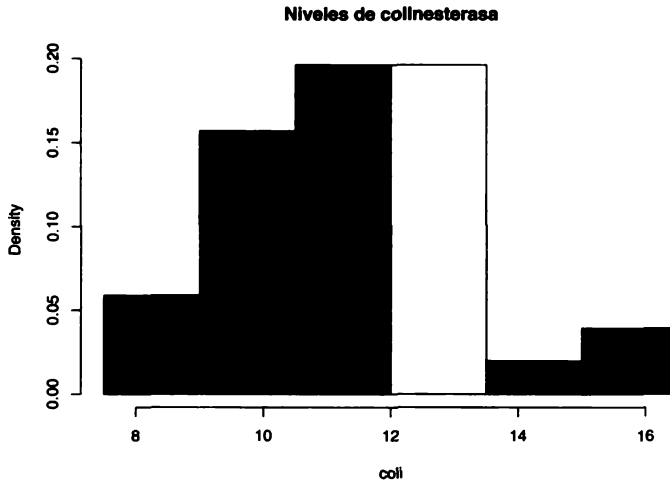


Figura 2.3 : Histograma del Ejemplo 2.1

Hacemos notar aquí que el orden de los argumentos dentro de las funciones es irrelevante siempre que le indiquemos cuál es el nombre del argumento al que damos valor. Es decir, obtendríamos exactamente la misma Figura 2.3 si en lugar de (7) ejecutáramos (8). Esta propiedad es válida en todas las funciones de R.

```

> hist(coли,breaks=d1,col=c1,prob=T,main="Niveles de colinesterasa") (8)

```

En este ejemplo todos los intervalos tenían la misma amplitud. Si no fuera así, sólo habría que modificar la longitud de los mismos. Pero, en todo caso, hay que decírselo al ordenador. Será mucho más sencillo cuando, en la siguiente sección sea el ordenador el que elija, básicamente, la forma de realizar el histograma.

Datos correspondientes a un carácter cuantitativo sin agrupar en intervalos

Ésta es la situación habitual que tendremos para un conjunto de datos cuantitativos. Las representaciones gráficas habituales serán, si son pocos los

valores distintos de la variable, el *Diagrama de barras*, con la misma filosofía del Diagrama de rectángulos antes estudiado y, si hay muchos valores distintos, el *Histograma*, o su versión modificada, el *Diagrama de hojas y ramas* (*stem and leaf plot*) inventado por John Tukey (1977). En el caso de frecuencias acumuladas la representación gráfica será el *Diagrama de Frecuencias acumuladas*, denominado *Función de distribución empírica* si las frecuencias acumuladas a representar son las relativas.

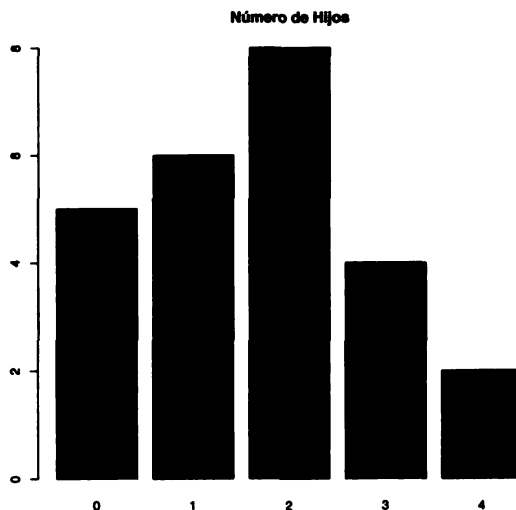


Figura 2.4 : Diagrama de barras del Ejemplo 2.3

Ejemplo 2.3 (continuación)

Como el número de valores distintos de la variable es sólo de cinco, la representación gráfica que procede es el diagrama de barras. Éste se puede obtener como un histograma con intervalos muy cortos, pero lo más simple es volver a utilizar la función `barplot`. Así, después de introducir los valores posibles en (1) y las frecuencias en (2), obtendremos la Figura 2.4 ejecutando (3)

```
> n3<-c(0,1,2,3,4)                                     (1)
> x3<-c(5,6,8,4,2)                                     (2)
> barplot(x3,names=n3,main="Número de Hijos")          (3)
```

Ejemplo 2.1 (continuación)

Si queremos hacer un histograma de ellos, dejando que el ordenador elija los puntos de corte, introduciríamos los datos como aparece en (1), y ejecutaríamos (2), obteniendo la Figura 2.5

```

> x1<-c(10.6,12.5,11.1,9.2,11.5,9.9,11.9,11.6,14.9,12.5,12.5,12.3,
+ 12.2,10.8,16.5,15,10.3,12.4,9.1,7.8,11.3,12.3,9.7,12,11.8,12.7,
+ 11.4,9.3, 8.6, 8.5, 10.1, 12.4, 11.1, 10.2)
> hist(x1)

```

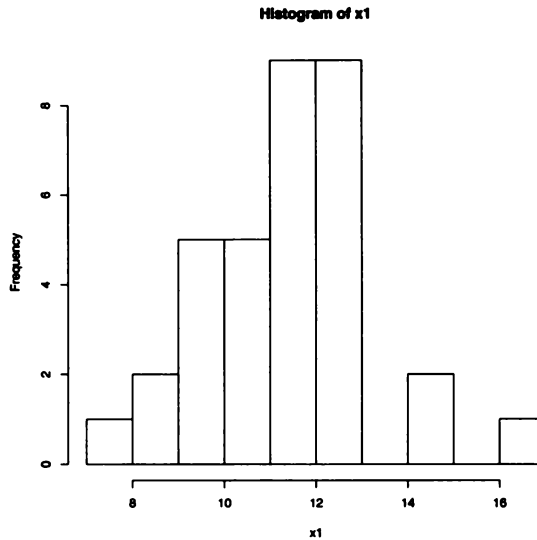


Figura 2.5 : Histograma del Ejemplo 2.1

Si queremos controlar los intervalos del histograma utilizaríamos el argumento **breaks** con un vector de puntos de corte, como antes. Si queremos simplemente fijar el número **n** de intervalos, pondríamos **breaks=n** y el ordenador suele elegir un número de intervalos similar a **n**.

El *Diagrama de hojas y ramas* se obtendría ejecutando la función **stem** en (3), apareciendo éste a continuación

```

> stem(x1)

```

The decimal point is at the |

```

 7 | 8
 8 | 56
 9 | 12379
10 | 12368
11 | 11345689
12 | 0233445557
13 |
14 | 9
15 | 0
16 | 5

```

Como se ve, el diagrama de hojas y ramas es un histograma o diagrama de barras girado, con la misma interpretación visual que éstos, pero con una característica adicional muy importante: del gráfico podemos recuperar las observaciones; así, en este ejemplo, si empezamos a leer el gráfico por arriba, vemos que las observaciones son, 7'8, 8'5, 8'6, ..., 16'5.

La Función de distribución empírica se obtiene con la siguiente combinación de funciones: primero creamos en (4) un vector igual al número de datos; en (5) le pedimos que represente los pares de datos (`sort(x1), (1:n)/n`) en donde `sort` es la función que ordena los datos, es decir, la que proporciona los estadísticos ordenados y $(1:n)/n$ es la ristra de valores de 1 a n , divididos por n . El valor `s` (s minúscula, no mayúscula) del argumento `type` se utiliza para que no haga un gráfico de puntos sino que los una con segmentos; el gráfico así creado es la Figura 2.6. (Los segmentos verticales, de hecho, sobran para que sea una verdadera función, pero la representación es útil.)

```
> n<-length(x1) (4)
```

```
> plot(sort(x1), (1:n)/n, type="s") (5)
```

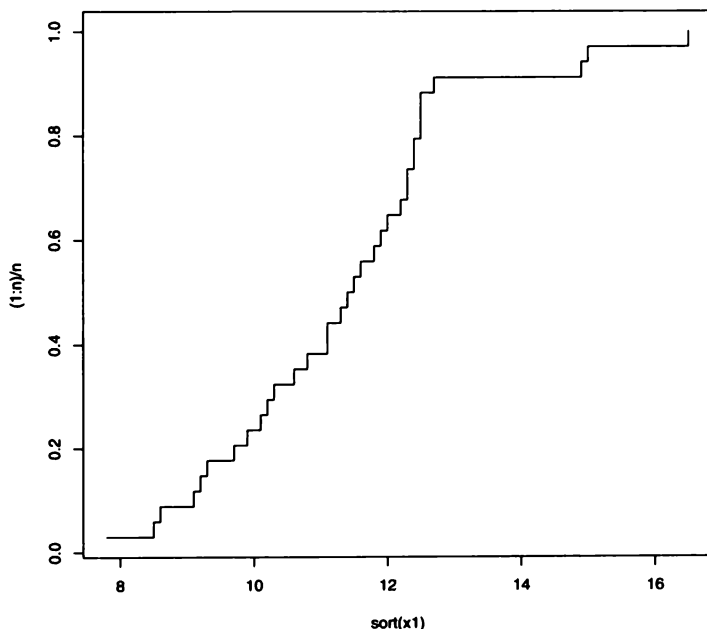


Figura 2.6 : Función de distribución empírica del Ejemplo 2.1

Si queremos que no aparezca en el eje de abscisas `sort(x1)` sino la variable, digamos `x`, y que en el eje de ordenadas del gráfico aparezca el rótulo **Función distribución empírica** ejecutaríamos la siguiente sentencia


```
> plot(sort(x1),(1:n)/n,type="s",xlab="x",ylab="Función distribución empírica")
```

Estos dos argumentos, `xlab` e `ylab` son también comunes a todas las funciones gráficas de R.

2.3.2. Medidas de tendencia central de caracteres cuantitativos

En esta sección definiremos una serie de medidas o valores que representan o resumen una distribución de frecuencias dada, siendo también útiles, por tanto, para realizar comparaciones entre distintas distribuciones de frecuencias. Estas medidas reciben el nombre de *promedios*, *medidas de posición* o *medidas de tendencia central* que, aunque alguna de ellas puede aplicarse a caracteres cualitativos, habitualmente lo son sobre caracteres cuantitativos.

Media aritmética

Llamando x_1, \dots, x_k a los datos distintos de un carácter cuantitativo en estudio, o las marcas de clase de los intervalos en los que se han agrupado dichos datos, y n_1, \dots, n_k a las correspondientes frecuencias absolutas de dichos valores o marcas de clase, llamaremos *media aritmética* de la distribución de frecuencias al valor

$$a = \frac{\sum_{i=1}^k x_i n_i}{n}$$

en donde n es la frecuencia total.

Ejemplo 2.3 (continuación)

La media aritmética de las veinticinco familias encuestadas será

$$a = \frac{42}{25} = 1'68$$

es decir, las familias encuestadas tienen como *número medio de hijos*, 1'68.

Ejemplo 2.1 (continuación)

La distribución de frecuencias del Ejemplo 2.1, utilizando las marcas de clase, será

x_i	n_i
8'25	3
9'75	8
11'25	10
12'75	10
14'25	1
15'75	2
	34

Tabla 2.13

la cual proporciona una media aritmética

$$a = \frac{\sum_{i=1}^6 x_i n_i}{n} = \frac{388'5}{34} = 11'426.$$

Mediana

La *mediana* es otra medida de posición, la cual se define como aquel valor de la variable tal que, supuestos ordenados los valores de ésta en orden creciente, la mitad son menores o iguales y la otra mitad mayores o iguales. Así, si en la siguiente distribución de frecuencias,

x_i	n_i	N_i
0	3	3
1	2	5
2	2	7
	7	

Tabla 2.14

ordenamos los valores en orden creciente,

0 0 0 1 1 2 2

el 1 será el valor que cumple la definición de mediana.

Lógicamente, en cuanto el valor de la frecuencia total sea ligeramente mayor, este procedimiento resulta inviable. Por esta razón, daremos a continuación una fórmula que permita calcularla. No obstante, será necesario distinguir los casos en los que los datos vengan agrupados de aquellos en los que vengan sin agrupar.

Datos sin agrupar:

La Figura 2.7, correspondiente a un diagrama de frecuencias absolutas acumuladas, recoge las dos situaciones que se pueden presentar

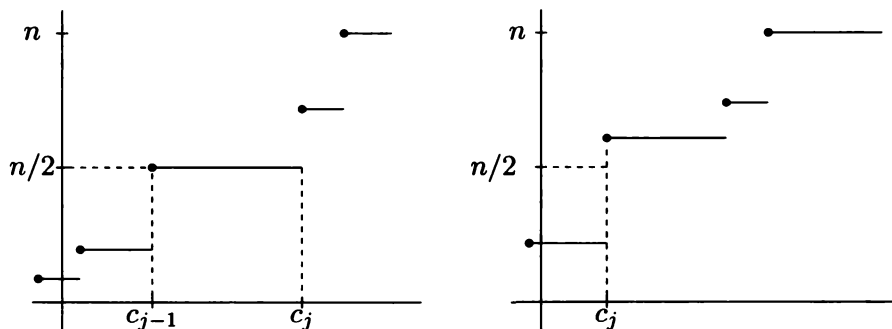


Figura 2.7

Si la situación es como la de la figura de la derecha, es decir, si

$$N_{j-1} < \frac{n}{2} < N_j$$

entonces la mediana es $M_e = c_j$.

Si la situación que se presenta es como la de la figura de la izquierda, entonces la mediana queda indeterminada, aunque en este caso se toma como mediana la media aritmética de los dos valores entre los que se produce la indeterminación; así pues, si

$$N_{j-1} = \frac{n}{2} < N_j$$

entonces la mediana es

$$M_e = \frac{c_{j-1} + c_j}{2}.$$

Ejemplo 2.3 (continuación)

La distribución de frecuencias acumuladas del Ejemplo 2.3 era

x_i	N_i
0	5
1	11
2	19
3	23
4	25

Tabla 2.15

y como es $n/2 = 12'5$ y en consecuencia

$$11 < 12'5 < 19$$

la mediana será $M_e = 2$.

Datos agrupados:

La Figura 2.8, correspondiente a un polígono de frecuencias absolutas acumuladas, nos plantea de nuevo dos situaciones diferentes a considerar

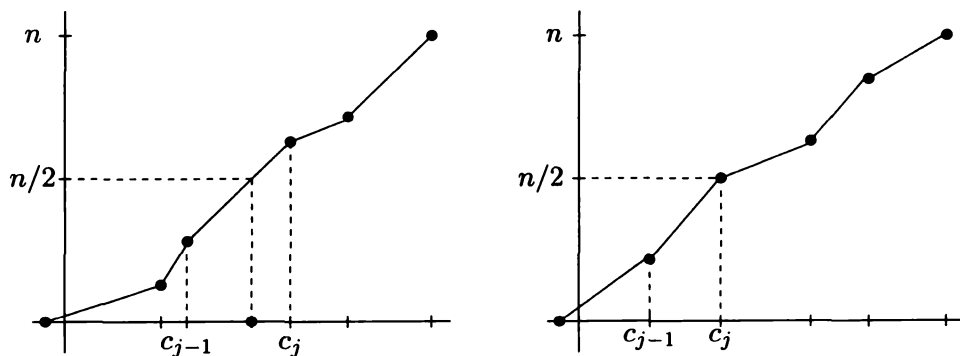


Figura 2.8

El más sencillo, el de la derecha, en el que existe una frecuencia absoluta acumulada N_j tal que $n/2 = N_j$, la mediana es $M_e = c_j$.

Si la situación es como la que se representa en la figura de la izquierda, en la que

$$N_{j-1} < \frac{n}{2} < N_j$$

entonces, la mediana, está en el intervalo $[c_{j-1} - c_j)$, es decir entre c_{j-1} y c_j , tomándose en ese caso, por razonamientos de proporcionalidad, como mediana el valor

$$M_e = c_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{n_j} \cdot a_j$$

siendo a_j la amplitud del intervalo $[c_{j-1} - c_j)$. (Véase, por ejemplo, Sánchez-Crespo y García-España, 1961, pág. 60-61.)

Ejemplo 2.1 (continuación)

La distribución de frecuencias del Ejemplo 2.1 era

I_i	n_i	N_i
7'5 – 9	3	3
9 – 10'5	8	11
10'5 – 12	10	21
12 – 13'5	10	31
13'5 – 15	1	32
15 – 16'5	2	34

Tabla 2.16

Al ser $n/2 = 17$ y estar

$$11 < 17 < 21$$

la mediana estará en el intervalo $[10'5 - 12)$, y aplicando la fórmula anterior, será

$$M_e = 10'5 + \frac{17 - 11}{10} \cdot 1'5 = 11'4.$$

Moda

La *moda* se define como aquel valor de la variable al que corresponde máxima frecuencia (absoluta o relativa). Para calcularla, también será necesario distinguir si los datos están o no agrupados.

Datos sin agrupar:

Para datos sin agrupar, la determinación del valor o valores (ya que puede haber más de uno) modales es muy sencilla. Basta observar a que valor le corresponde una mayor n_i . Ése será la moda.

Así en el Ejemplo 2.3, la simple inspección de la Tabla 2.9 proporciona como valor para la moda $M_d = 2$.

Datos agrupados:

Si los datos se presentan agrupados en intervalos es necesario, a su vez, distinguir si éstos tienen o no igual amplitud.

Si tienen amplitud constante c , una vez identificado el intervalo modal $[c_{j-1} - c_j)$, es decir el intervalo al que corresponde mayor frecuencia absoluta $n_j = \text{máx}\{n_1, \dots, n_k\}$, la moda se define, también por razones geométricas (Sánchez-Crespo y García-España, 1961, pág. 57-58), como

$$M_d = c_{j-1} + \frac{n_{j+1}}{n_{j-1} + n_{j+1}} \cdot c$$

Ejemplo 2.1 (continuación)

Este ejemplo presenta un caso de distribución bimodal, ya que tanto el intervalo $[10'5 - 12)$ como el $[12 - 13'5)$ tienen frecuencia absoluta máxima. Deberíamos aplicar, por tanto, para cada uno de los dos intervalos, la fórmula anterior, determinando así las dos modas de la distribución. No obstante, este ejemplo tiene además la peculiaridad adicional de ser ambos intervalos modales contiguos. En esta situación se considera la distribución unimodal, eligiendo como moda el extremo común, $M_d = 12$.

Si los intervalos tuvieran distinta amplitud a_j , primeros debemos *estandarizar* las frecuencias absolutas n_j , determinando los cocientes

$$l_j = \frac{n_j}{a_j}, \quad j = 1, \dots, k$$

y luego aplicar la regla definida para el caso de intervalos de amplitud constante a los l_j . Es decir, primero calcular el $l_j = \max\{l_1, \dots, l_k\}$ para determinar el intervalo modal $[c_{j-1} - c_j)$ y luego aplicar la fórmula

$$M_d = c_{j-1} + \frac{l_{j+1}}{l_{j-1} + l_{j+1}} \cdot a_j$$

siendo a_j la amplitud del intervalo modal $[c_{j-1} - c_j)$.

Ejemplo 2.4

Las frecuencias absolutas n_i y estandarizadas l_i (puesto que la longitud de los intervalos es distinta), correspondientes a puntuaciones obtenidas en un test psicológico fueron las siguientes,

I_i	n_i	l_i
0-20	8	0'4
20-30	9	0'9
30-40	12	1'2
40-45	10	2
45-50	9	1'8
50-60	10	1
60-80	8	0'4
80-100	4	0'2

Tabla 2.17

El intervalo modal es el 40 - 45 y la moda, por consiguiente,

$$M_d = 40 + \frac{1'8}{1'2 + 1'8} \cdot 5 = 43.$$

A diferencia de lo que ocurre con la media o con la mediana, sí es posible determinar la moda en el caso de datos cualitativos. Así, en el Ejemplo 2.2 puede afirmarse que la causa modal por la que no fue completado el tratamiento es $M_d = \text{rehusaron cirugía}$.

Cuantiles

Los cuantiles o cuantilas son las últimas medidas de posición que veremos. De hecho algunos autores las incluyen dentro de las medidas de dispersión al ser medidas de posición no centrales.

El *cuantil* $p_{r/k}$, $r = 1, 2, \dots, k - 1$ se define como aquel valor de la variable que divide la distribución de frecuencias, previamente ordenada de forma creciente, en dos partes, estando el $(100 \cdot r/k) \%$ de ésta formado por valores menores que $p_{r/k}$.

Si $k = 4$ los (tres) cuantiles reciben el nombre de *cuartiles*. Si $k = 10$ los (nueve) cuantiles reciben, en este caso, el nombre de *deciles*. Por último, si $k = 100$ los (noventa y nueve) cuantiles reciben el nombre de *centiles*.

Obsérvese que siempre que r y k mantengan la misma proporción (r/k) obtendremos el mismo valor. Así por ejemplo, el primer cuartil es igual al vigésimo quinto centil.

En este sentido, la mediana M_e es el segundo cuartil, o el quinto decil, etc.

Para el cálculo de los cuantiles de nuevo hay que considerar si los datos vienen o no agrupados en intervalos.

Datos sin agrupar:

Si los datos vienen sin agrupar y es

$$N_{j-1} < \frac{r}{k} \cdot n < N_j$$

el r -ésimo cuantil de orden k será $p_{r/k} = c_j$, valor al que corresponde la frecuencia absoluta acumulada N_j .

Si la situación fuera de la forma

$$N_{j-1} = \frac{r}{k} \cdot n < N_j$$

tomaríamos, en esta situación indeterminada,

$$p_{r/k} = \frac{c_{j-1} + c_j}{2}$$

Datos agrupados:

Si los datos se presentan agrupados y, para algún j , fuera

$$\frac{r}{k} \cdot n = N_j$$

el r -ésimo cuantil de orden k sería $p_{r/k} = c_j$.

Por último, si fuera

$$N_{j-1} < \frac{r}{k} \cdot n < N_j$$

el intervalo a considerar sería el $[c_{j-1} - c_j)$, al que corresponde frecuencia absoluta n_j y absoluta acumulada N_j , siendo entonces el cuantil el dado por la expresión,

$$p_{r/k} = c_{j-1} + \frac{\frac{r}{k} \cdot n - N_{j-1}}{n_j} \cdot a_j \quad r = 1, \dots, k - 1$$

en donde a_j es la amplitud del intervalo $[c_{j-1} - c_j)$.

Si el intervalo a considerar fuera el primero $c_0 - c_1$, se tomaría en la expresión anterior $N_{j-1} = 0$.

Ejemplo 2.3 (continuación)

Vamos a determinar el tercer cuartil. Como es

$$\frac{r}{k} \cdot n = \frac{3}{4} \cdot 25 = 18'75$$

y $11 < 18'75 < 19$, será $p_{3/4} = 2$.

Ejemplo 2.1 (continuación)

Vamos a determinar el séptima decil. De la Tabla 2.11 obtenemos

$$\frac{r}{k} \cdot n = \frac{7}{10} \cdot 34 = 23'8$$

y $21 < 23'8 < 31$, con lo que el intervalo a considerar será el $[12 - 13'5)$, y

$$p_{7/10} = 12 + \frac{23'8 - 21}{10} \cdot 1'5 = 12'42.$$

2.3.3. Medidas de dispersión

Las medidas de posición estudiadas en la sección anterior servían para resumir la distribución de frecuencias en un solo valor. Las *medidas de dispersión*, a las cuales dedicaremos esta sección, tienen como propósito estudiar lo concentrada que está la distribución en torno a algún promedio.

Estudiaremos las cuatro medidas de dispersión más utilizadas: *Recorrido*, *Varianza*, *Desviación típica* y *Coeeficiente de variación de Pearson*, estando definidas las tres primeras en unidades concretas y la cuarta en unidades abstractas.

Recorrido

Si x_{max} (también representado por $x_{(n)}$) es el dato mayor, o la última marca de clase si es que los datos vienen agrupados en intervalos, y x_{min} (ó $x_{(1)}$) el dato menor, o primera marca de clase, llamaremos *Recorrido* a

$$R = x_{max} - x_{min}$$

Así, en el Ejemplo 2.1 el recorrido es, si los datos viene aislados, $R = 16'5 - 8'5 = 8$ y, si vienen agrupados en intervalos (Tabla 2.11), $R = 15'75 - 8'25 = 7'5$.

Para los datos del Ejemplo 2.3 será $R = 4 - 0 = 4$.

La principal ventaja del recorrido es la de proporcionar una medida de la dispersión de los datos fácil y rápida de calcular. A veces se utiliza también el *Recorrido intercuartílico*, definido como la diferencia entre el tercer y primer cuartil.

Varianza

Denotando de nuevo por x_1, \dots, x_k los datos o las marcas de clase, llamaremos *Varianza* a

$$s^2 = \frac{1}{n} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - a^2$$

siendo a la media aritmética de la distribución.

Así en el Ejemplo 2.3 la varianza es

$$s^2 = 4'24 - (1'68)^2 = 1'4176$$

y en Ejemplo 2.1, considerando los datos agrupados en intervalos de la Tabla 2.13,

$$s^2 = 133'97 - (11'426)^2 = 3'42.$$

Al valor

$$S^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{n s^2}{n-1}$$

se le denomina *cuasivarianza*.

Desviación típica

La varianza tiene un problema, y es que está expresada en unidades al cuadrado. Esto puede producir una falsa imagen de la dispersión de la distribución. En su lugar suele utilizarse su raíz cuadrada, denominada *Desviación*

típica. Así, la desviación típica de la distribución de frecuencias del Ejemplo 2.3 es $s = 1'1906$ y la del Ejemplo 2.1 $s = 1'85$.

Coefficiente de variación de Pearson

La desviación típica sirve para medir de forma eficaz la dispersión de un conjunto de datos entorno a su media. Desgraciadamente, esta medida puede resultar engañosa cuando tratamos de comparar la dispersión de dos conjuntos de datos. Así por ejemplo, si tenemos dos grupos de personas de 11 y 25 años con medias y desviaciones típicas las dadas por la Tabla 2.18

	Peso medio	Desviación típica
11 años	40 Kgr.	2 Kgr.
25 años	50 Kgr.	2 Kgr.

Tabla 2.18

puede parecernos, al observar en ambos grupos una misma desviación típica, que ambos grupos de datos tienen la misma dispersión. No obstante, como parece lógico, no es lo mismo una variación de dos kilos en un grupo de elefantes que en uno de conejos. El *Coefficiente de variación de Pearson* elimina esa posible confusión al ser una medida de la variación de los datos pero en relación con su media (supuestamente mayor que cero). Se define como

$$V_p = \frac{s}{a} \cdot 100$$

siendo s y a respectivamente la desviación típica y la media aritmética de la distribución en estudio y en donde el factor 100 tiene como único objetivo el evitar operar con valores decimales.

De la definición de V_p se deduce fácilmente que aquella distribución a la que corresponda mayor coeficiente tendrá mayor dispersión.

En el ejemplo de la Tabla 2.18 al grupo de personas de 11 años le corresponde un coeficiente de variación de Pearson igual a

$$V_p = \frac{2}{40} \cdot 100 = 5$$

y al grupo de las personas de 25 años

$$V_p = \frac{2}{50} \cdot 100 = 4$$

lo que indica una mayor dispersión en el primer grupo.

Para el Ejemplo 2.3, V_p toma el valor

$$V_p = \frac{1'1906}{1'68} \cdot 100 = 70'869$$

y en el Ejemplo 2.1,

$$V_p = \frac{1'85}{11'426} \cdot 100 = 16'19.$$

2.3.4. Medidas de asimetría

Diremos que una distribución es *simétrica* cuando su mediana, su moda y su media aritmética coincidan. Claramente las distribuciones de los Ejemplos 2.1 y 2.3 no son, por tanto, simétricas.

Diremos que una distribución es *asimétrica a la derecha* si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda.

Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es *asimétrica a la izquierda*.

Existen varias medidas de la asimetría de una distribución de frecuencias. Aquí estudiaremos dos de ellas.

Coefficiente de asimetría de Pearson

El *coeficiente de asimetría de Pearson* se define como

$$A_p = \frac{a - M_d}{s}$$

siendo cero cuando la distribución es simétrica, positivo cuando existe asimetría a la derecha y negativo cuando existe asimetría a la izquierda.

En el Ejemplo 2.3, A_p es igual a

$$A_p = \frac{1'68 - 2}{1'1906} = -0'2688$$

indicando una ligera asimetría a la izquierda en la distribución de frecuencias correspondiente.

De la misma manera, para el Ejemplo 2.1 también se observa una ligera asimetría a la izquierda, al ser

$$A_p = \frac{11'426 - 12}{1'85} = -0'31.$$

De la definición se observa que este coeficiente sólo se podrá utilizar cuando la distribución sea unimodal. La otra medida de asimetría que veremos no presenta este inconveniente.

Coefficiente de asimetría de Fisher

El *coeficiente de asimetría de Fisher* se define como

$$A_f = \frac{\sum_{i=1}^k (x_i - a)^3 \cdot n_i}{n \cdot S^3}$$

siendo x_i los valores de la variable o las marcas de clase y $S = \sqrt{S^2}$, llamada a veces cuasidesviación típica.

La interpretación del coeficiente de Fisher es semejante a la del coeficiente de Pearson: se dice que la distribución es simétrica cuando vale cero, siendo el coeficiente positivo o negativo cuando exista asimetría a la derecha o izquierda respectivamente.

2.3.5. Medidas de posición y dispersión con R

Las principales medidas de posición y dispersión son la *Media*, obtenida con la función `mean`; la *Mediana*, cuyo valor lo obtenemos con `median`; la *Cuasivarianza* (no la *varianza*) para la que debemos ejecutar la función `var`; su raíz cuadrada, la *Cuasidesviación típica*, obtenida con `sd`, y los cuantiles, que se consiguen con `quantile`.

Un buen resumen de muchas de las medidas de posición se obtiene de una vez con la función `summary`.

No dejamos de mencionar que en el texto MR se enuncian muchas más, tales como la medias α -recortadas, α -Winsorizadas, etc., por lo que remitimos al lector a dicho texto para completar su conocimiento sobre ellas.

Un gráfico con el que podemos visualizar la dispersión y simetría de los datos, muy utilizado para analizar las suposiciones requeridas por el Análisis de la Varianza, es el *Diagrama de cajas* (*box-plot* inventado por Tukey (1977) y ejecutando por la función de R `boxplot`. Consiste en representar una caja en donde el lado inferior sea el primer cuartil, el superior el tercer cuartil, apareciendo dividida la caja por la mediana de los datos. Se añaden dos segmentos a la caja así formada para unirla al máximo y mínimo valor. Aquellos datos inferiores al primer cuartil menos 1'5 veces el recorrido intercuartílico, o superiores al tercer cuartil más 1'5 veces el recorrido intercuartílico se consideran anómalos y se representan por pequeños círculos fuera del diagrama de cajas.

Ejemplo 2.2 (continuación)

Las funciones antes mencionadas se aplican a un vector de datos numéricos, por lo que consideraremos el vector antes generado, `x1`. Para comprobar que lo tenemos, ejecutamos (1). (Recuerde que los números entre corchetes son indicaciones del lugar que ocupa el dato que sigue al corchete. Si su pantalla de ordenador es más grande o más pequeña esos números pueden ser distintos. Lo que debe ser igual a lo que aquí aparece es el vector de datos `x1`.) Ejecutando (2) y (3) obtenemos la media y la mediana. Con (4) obtenemos los principales cuantiles; si queremos uno en particular debemos utilizar el argumento `probs` para indicar qué cuantil queremos, como hacemos en (5).

Con (6) obtenemos la cuasivarianza muestral y con (7) su raíz cuadrada.

Como dijimos más arriba, con (8) obtenemos, de una vez, las principales medidas de posición.

El diagrama de cajas, dado por la Figura 2.9 se obtiene ejecutando (9). Se ve que el dato 16'5 es representado como anómalo.

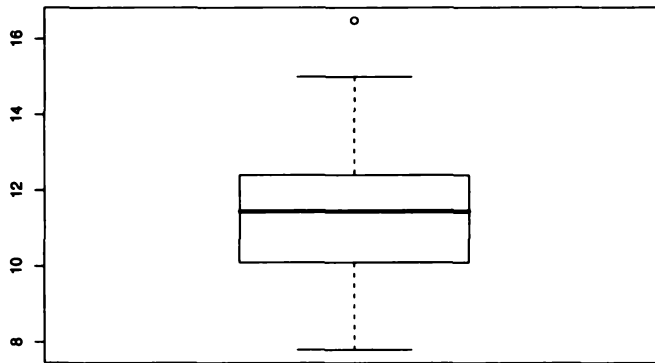


Figura 2.9: Diagrama de cajas del Ejemplo 2.2

```

> x1
[1] 10.6 12.5 11.1  9.2 11.5  9.9 11.9 11.6 14.9 12.5 12.5 12.3 12.2
[14] 10.8 16.5 15.0 10.3 12.4  9.1  7.8 11.3 12.3  9.7 12.0 11.8 12.7
[27] 11.4  9.3  8.6  8.5 10.1 12.4 11.1 10.2
> mean(x1)
[1] 11.35294
> median(x1)
[1] 11.45
> quantile(x1)
 0%   25%   50%   75%  100%
7.800 10.125 11.450 12.375 16.500
> quantile(x1,probs=0.25)
 25%
10.125
> var(x1)
[1] 3.514082
> sd(x1)
[1] 1.874588
> summary(x1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  7.80   10.12   11.45   11.35   12.38   16.50
> boxplot(x1,col=2)

```

2.4. Distribuciones bidimensionales de frecuencias

En esta sección estudiaremos la situación en la que los datos son observaciones de dos caracteres efectuadas en los individuos de una determinada población. Ambos caracteres pueden ser cuantitativos, como ocurre en el Ejemplo 2.5 con el *Peso* y la *Talla*, pueden ser cualitativos, como ocurre con el *Sexo* y el *Estado civil* en el Ejemplo 2.6, o uno cuantitativo y otro cualitativo, como ocurre con el Ejemplo 2.7 en el que se clasifica a los n individuos observados por el *Color de la piel* y el *Número de hijos*.

Ejemplo 2.5

Se observó el *Peso* y la *Talla* en 80 individuos, obteniéndose los siguientes datos,

<i>Talla</i>	1'50 – 1'60	1'60 – 1'70	1'70 – 1'80	1'80 – 1'90	1'90 – 2'00
<i>Peso</i>					
50 – 60	2	1	1	2	2
60 – 70	3	3	2	4	8
70 – 80	5	4	3	5	4
80 – 90	2	4	2	6	6
90 – 100	1	2	1	5	2

Tabla 2.19

Ejemplo 2.6

Se clasificaron 174 individuos de acuerdo con su *Sexo* y su *Estado civil*, obteniendo como resultado la siguiente tabla,

<i>Est. civil</i>	Soltero	Casado	Viudo	Separado ó Divorciado
<i>Sexo</i>				
Masculino	20	40	5	11
Femenino	29	38	11	20

Tabla 2.20

Ejemplo 2.7

Los datos sobre 173 personas en las que se observó el *Número de hijos* y la *Raza* vienen recogidos en la siguiente tabla de doble entrada,

<i>Raza</i>	Blanca	Negra	Mestiza
<i>Número de hijos</i>			
0	21	3	8
1	32	6	12
2	18	21	12
3	6	18	6
Más de 3	1	7	2

Tabla 2.21

En estas situaciones los datos se recogen en lo que de forma genérica se denomina una *Tabla de doble entrada* o *Tabla de contingencia*, cuya expresión general es la siguiente:

<i>Carácter B</i>		B_1	B_j			B_k	
<i>Carácter A</i>							
A_1		n_{11}	\cdots	n_{1j}	\cdots	n_{1k}	$n_{1\cdot}$
\dots		\dots	\dots	\dots	\dots	\dots	\dots
A_i		n_{i1}	\cdots	n_{ij}	\cdots	n_{ik}	$n_{i\cdot}$
\dots		\dots	\dots	\dots	\dots	\dots	\dots
A_l		n_{l1}	\cdots	n_{lj}	\cdots	n_{lk}	$n_{l\cdot}$
		$n_{\cdot 1}$	\cdots	$n_{\cdot j}$	\cdots	$n_{\cdot k}$	n

Tabla 2.22

en donde n_{ij} , denominada *frecuencia absoluta* del par (A_i, B_j) , representa el número de individuos, de entre los n , que poseen a la vez la modalidad A_i del carácter A y la modalidad B_j del carácter B .

Ésta es la forma habitual en la que se presentan los datos bidimensionales, aunque si dividimos las n_{ij} por n obtendremos una distribución bidimensional de frecuencias relativas f_{ij} , en donde sería

$$\sum_{i=1}^l \sum_{j=1}^k f_{ij} = 1.$$

Distribuciones marginales

Las tablas formadas con la primera y última columnas y con la primera y última filas de la Tabla 2.22

Carácter A	frec. absol.
A_1	$n_{1\cdot}$
\cdots	\cdots
A_i	$n_{i\cdot}$
\cdots	\cdots
A_l	$n_{l\cdot}$
	n

Tabla 2.23

Carácter B	frec. absol.
B_1	$n_{\cdot 1}$
\cdots	\cdots
B_j	$n_{\cdot j}$
\cdots	\cdots
B_k	$n_{\cdot k}$
	n

Tabla 2.24

se denominan *distribuciones marginales* (absolutas), respectivamente, de los caracteres A y B .

Ambas distribuciones marginales son distribuciones unidimensionales de frecuencias, por lo que les son aplicables todos los conceptos y representaciones gráficas desarrollados en la Sección 2.3, en particular los referentes a las

medidas de posición y dispersión, hablándose en este caso de *media marginal*, *varianza marginal*, etcétera, al referirnos a la media aritmética, varianza, ..., obtenidas a partir de la distribución de frecuencias unidimensional marginal correspondiente.

La interpretación de las distribuciones marginales es clara. Así, las distribuciones marginales del Ejemplo 2.6

<i>Sexo</i>	n_i	<i>Estado civil</i>	n_i
Masculino	76	Soltero	49
Femenino	98	Casado	78
	174	Viudo	16
		Sep/Div	31
			174

Tabla 2.25

Tabla 2.26

representan, la primera, cómo está distribuido el carácter *Sexo* entre los 174 individuos observados, al margen, es decir, prescindiendo del Estado civil que tengan (Tabla 2.25), y la segunda (Tabla 2.26), qué distribución tienen esos mismos datos considerando solamente el *Estado civil*, sin tener en cuenta su *Sexo*.

Este concepto de distribuciones marginales, así como el que veremos a continuación de distribuciones condicionadas, admiten una rápida generalización a más de dos variables, obteniendo en ese caso tantas distribuciones marginales como variables hayamos considerado.

Cuando existan más de dos variables en consideración también será posible agrupar un número m de ellas formando distribuciones marginales m -dimensionales. También podrían hacerse grupos de variables para conseguir distribuciones de frecuencias condicionadas *multivariantes*. En este texto introductorio no entraremos en ello, aunque básicamente no existe ninguna novedad conceptual en esta extensión.

Distribuciones condicionadas

Supongamos que con los datos del Ejemplo 2.7 estamos interesados en conocer la distribución del *Número de hijos* de los individuos de Raza blanca.

La distribución de frecuencias deseada consistirá en la tabla formada por la primera y segunda columnas de la Tabla 2.21,

Número de hijos/blanca	n_i
0	21
1	32
2	18
3	6
Más de 3	1
	<hr/> 78

Tabla 2.27

Análogamente, con respecto al Ejemplo 2.5, la distribución de frecuencias condicionada de los Pesos por individuos de Talla entre 1'80 y 1'90 será la dada por la Tabla 2.28

Peso/1'80 - 1'90	n_i
50 - 60	2
60 - 70	4
70 - 80	5
80 - 90	6
90 - 100	5
	<hr/> 22

Tabla 2.28

En general, la distribución (de frecuencias absolutas) condicionada del carácter A por la modalidad B_j del carácter B será,

A/B_j	
A_1	n_{1j}
...	...
A_i	n_{ij}
...	...
A_l	n_{lj}
	<hr/> $n_{.j}$

Tabla 2.29

y la del carácter B por la modalidad A_i del carácter A será,

B/A_i	
B_1	n_{i1}
...	...
B_j	n_{ij}
...	...
B_k	n_{ik}
	$n_{i.}$

Tabla 2.30

existiendo $k + l$ distribuciones condicionadas, si es que los caracteres B y A presentan, respectivamente, k y l modalidades cada uno.

De nuevo, éstas son distribuciones unidimensionales de frecuencias, por lo que se les podrá aplicar todas las definiciones, conceptos y representaciones gráficas estudiados en la Sección 2.3. En particular las medidas de posición y las de dispersión, hablándose entonces de *media condicionada*, *mediana condicionada*, *varianza condicionada*, etc.

2.4.1. Representaciones gráficas de las distribuciones bidimensionales de frecuencias

Como antes dijimos, las distribuciones marginales y condicionadas son distribuciones de frecuencias unidimensionales, y por tanto, su representación gráfica se ajustará a lo desarrollado en la Sección 2.3.1.

Por otro lado, de las distribuciones bidimensionales sólo consideraremos representaciones gráficas en el caso de que ambos caracteres sean cuantitativos, primero en el caso de que ambos vengan agrupados en intervalos y luego en el caso de que ambos tomen valores aislados.

Datos agrupados en intervalos correspondientes a un carácter cuantitativo

La representación habitual de este tipo de datos es un *Histograma tridimensional*, el cual se construye utilizando los mismos criterios que el histograma visto anteriormente.

Se utiliza un sistema de ejes coordenados en tres dimensiones, en donde los dos primeros ejes se reservan para las dos variables, representándose en altura la frecuencia, absoluta o relativa según la distribución que estemos representando.

Esto en el caso de que los intervalos de ambas variables tengan igual amplitud; si no, las alturas de los paralelepípedos deberán ser tales que su volumen resultante sea igual a la frecuencia.

Así, el histograma tridimensional correspondiente a los datos del Ejemplo 2.5 será el de la Figura 2.10, en el que se aprecian claramente los histogra-

mas unidimensionales correspondientes a las distribuciones unidimensionales condicionadas por cada una de las dos variables.

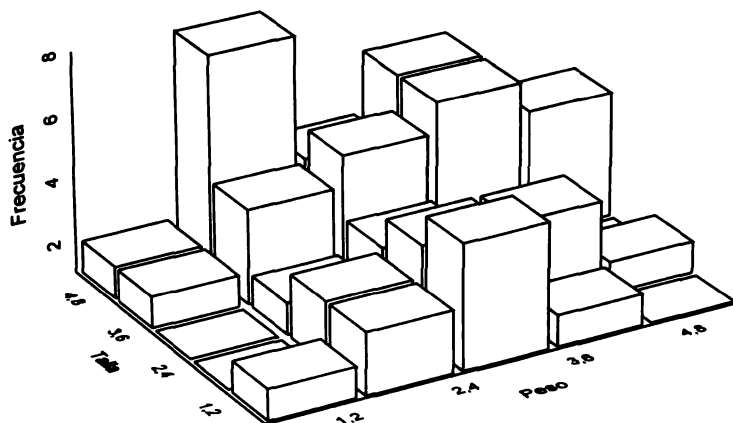


Figura 2.10: Histograma tridimensional del Ejemplo 2.5

Datos sin agrupar correspondientes a un carácter cuantitativo

La representación gráfica, denominada *Diagrama de barras tridimensional*, se hace utilizando también un sistema de ejes coordenados en tres dimensiones, levantando en cada par de valores (x_i, y_j) de la variable bidimensional (X, Y) , una barra de altura igual a su frecuencia (absoluta o relativa).

Ejemplo 2.8

Se preguntó a 34 profesores de Instituto el número de centros en los que habían estado anteriormente y el número de asignaturas que habían impartido, obteniéndose la siguiente distribución de frecuencias absolutas

Centros	Asignaturas					
	1	2	3	4	5	6
1	2	1	1	0	0	0
2	1	2	3	2	1	1
3	0	1	3	2	2	1
4	0	2	1	3	3	2

Tabla 2.31

Su representación gráfica será la Figura 2.11.

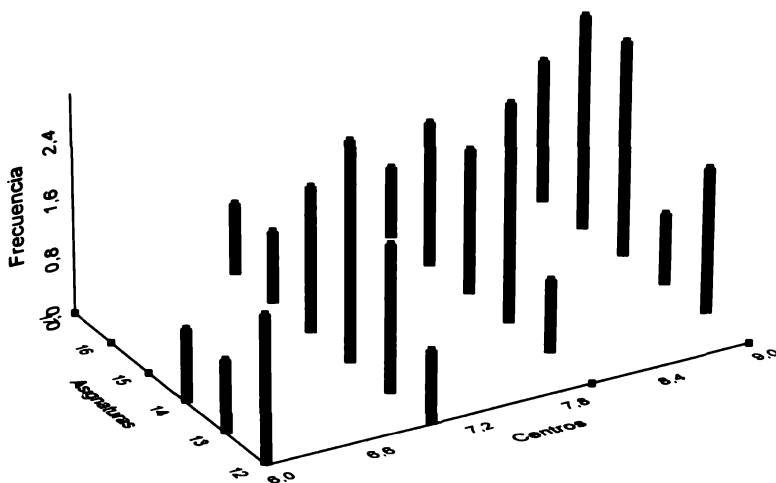


Figura 2.11: Diagrama de barras tridimensional del Ejemplo 2.8

No obstante, si no existen pares de valores repetidos, suele utilizarse el denominado diagrama de dispersión o *nube de puntos*, el cual consiste en representar en un sistema de ejes coordenados de dos dimensiones tantos puntos como datos, asignando a cada dato (x_i, y_j) el punto de coordenadas (x_i, y_j) .

Ejemplo 2.9

Tras preguntar a 20 personas con aficiones atléticas la marca que poseían en 100 metros lisos y las horas semanales que, por término medio, dedicaban a entrenar se obtuvieron los siguientes datos

Horas	21	32	15	40	27	18	26	50	33	51
Marca	13'2	12'6	13	12'2	15	14'8	14'8	12'2	13'6	12'6
Horas	36	16	19	22	16	39	56	29	45	25
Marca	13'1	14'9	13'9	13'2	15'1	14'1	13	13'5	12'7	14'2

Tabla 2.32

Su representación gráfica, dada por la Figura 2.12, se obtiene utilizando la función `plot(x,y)` después de incluir los datos como indicamos en (1) y (2), pero con esta función, podemos utilizar varios argumentos que indicamos a continuación invitando al lector a que los ejecute y a que los combine.

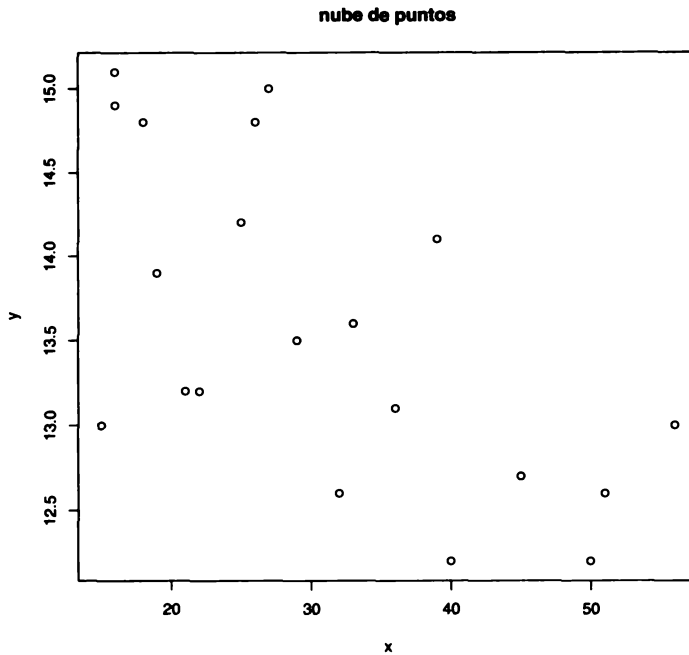


Figura 2.12: Nube de puntos del Ejemplo 2.9

```
> x<-c(21,32,15,40,27,18,26,50,33,51,36,16,19,22,16,39,56,29,45,25) (1)
```

$$y \leftarrow c(13.2, 12.6, 13, 12.2, 15, 14.8, 14.8, 12.2, 13.6, 12.6, 13.1, 14.9, 13.9, \quad (2)$$

+ 13.2, 15.1, 14.1, 13, 13.5, 12.7, 14.2)

```
> plot(x,y)
```

```
> plot(x,y,main="nube de puntos",col=3)      # pone título y color los puntos
```

```
> plot(x,y,xlim=c(inf,sup),ylim=c(inf,sup)) # limita el recorrido del gráfico
```

```
> plot(x,y,pch="2") # pone los puntos como un 2
```

```
> plot(x,y,pch=2) # pone los puntos como el símbolo
```

número 2. Hay del 0 al 18

```
> plot(x,y,xlab="abscisa",ylab="ordenada") # pone nombres a los ejes
```

```
> plot(x,y,xlab=" ",ylab=" ") # no pone ningún nombre a los ejes
```

```
> plot(x,y,axes=F) # no pone el marco al gráfico
```

2.4.2. Ajuste por mínimos cuadrados

Es un pensamiento común, la mayoría de las veces expresado de forma imprecisa, que el Peso y la Talla de los individuos de una población no son *independientes*, sino que por el contrario parece existir una determinada relación entre ellos, de forma que cuanto mayor sea la Talla de un individuo, mayor será su Peso.

También suele afirmarse que esta relación no es *funcional* en el sentido de que no se puede determinar una fórmula exacta que nos dé el Peso de un individuo *en función* de su Talla, sino que existen unos determinados valores entre los que razonablemente debería de estar su Peso.

La razón de tal idea se basa en la experiencia acumulada por las personas que ven una situación del tipo a la representada en la Figura 2.13, correspondiente a la nube de puntos del Peso y la Talla de 28 individuos.

Por otro lado, aunque no exista una ecuación que determine de forma exacta el Peso de una persona en función de su Talla, si fuéramos capaces de determinar una recta

$$y_{t_i} = \beta_0 + \beta_1 x_i$$

próxima a la nube de puntos (Figura 2.14), la Talla y_i de un individuo debería estar alrededor del valor que nos dé la recta y_{t_i} para un Peso x_i como el suyo.

Éste es el objetivo de la presente sección: determinar la ecuación de una recta $y_{t_i} = \beta_0 + \beta_1 x_i$, lo más próxima posible a una nube de puntos (x_1, y_1) , ..., (x_n, y_n) en el sentido de *mínimos cuadrados*.

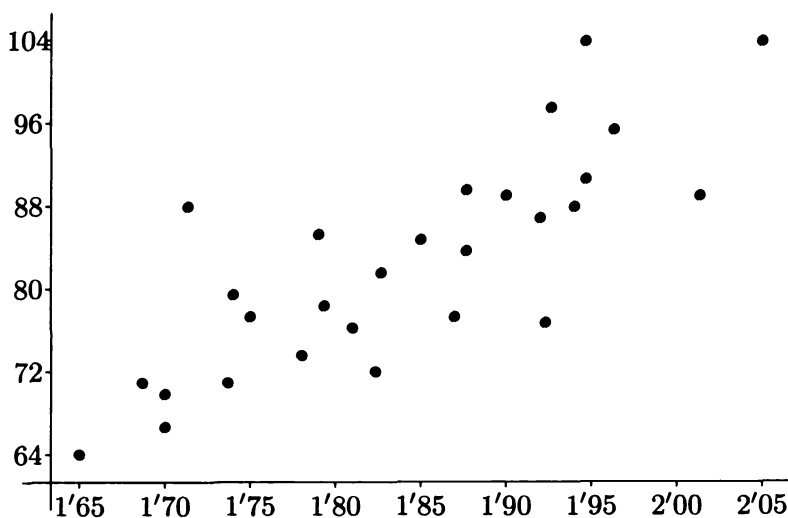


Figura 2.13

es decir, determinar los valores de β_0 y β_1 que hagan mínima la suma de los cuadrados de las desviaciones e_i entre los valores observados y_i y los teóricos dados por la recta y_{t_i}

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - y_{t_i})^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

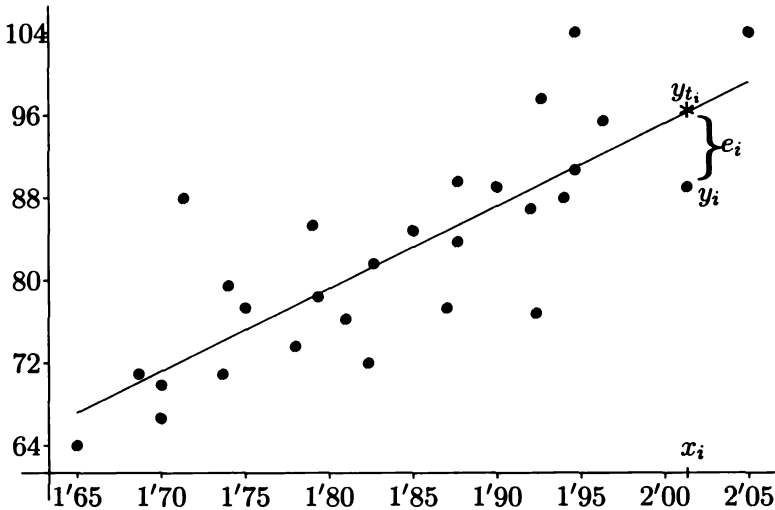


Figura 2.14

Matemáticamente este problema se resuelve considerando la ecuación anterior como una función de β_0 , derivando respecto a β_0 e igualando a cero dicha ecuación. A continuación, se considera la ecuación de la suma de los cuadrados como una función de β_1 , se deriva respecto a β_1 y se iguala a cero esta ecuación. Se obtiene así un sistema de dos ecuaciones con dos incógnitas de donde despejamos y obtenemos los valores para β_1 y β_0 ,

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

y

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}.$$

Ejemplo 2.9 (continuación)

La recta de ajuste para los datos de la Tabla 2.32 es

$$y = 15'05908 - 0'04785987 x$$

cuya representación gráfica sobre la nube de puntos es la Figura 2.15, obtenida ejecutando la función `lm`. Como luego vamos a representarla sobre la nube de puntos, la asignamos un nombre, `ajus`, al ejecutar (1). (Recuerde la Sección 1.5 en donde le enseñamos a obtener el símbolo que aparece entre la `y` y la `x`.)

Si queremos ver cuál es la recta obtenida, ejecutamos (2), obteniendo en (3) la ordenada en el origen (15'06 y la pendiente $-0'048$).

```
> ajus<-lm(y~x) (1)
```

```
> ajus (2)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

```
(Intercept)      x  
  15.05908    -0.04786 (3)
```

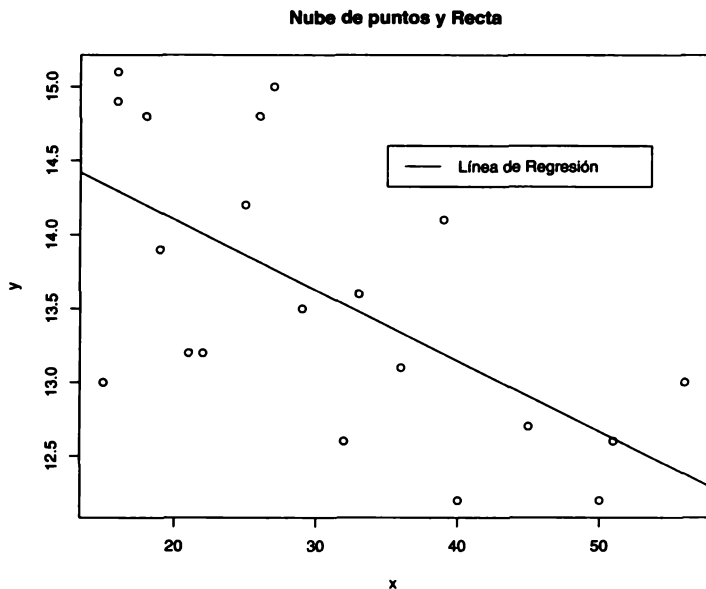


Figura 2.15: Nube de puntos y recta del Ejemplo 2.9

Podemos ahora añadirla sin más a la nube de puntos, ponerle diferentes colores y diferentes grosores y, hasta poner un rótulo al gráfico, con las siguientes instrucciones. Invitamos al lector a ejecutarlas y combinarlas.


```

> abline(ajus)                # añade la recta de regresión a la nube de puntos
> abline(ajus,col=2)          # pone color a la recta de regresión
> abline(15.06,-0.048,lty=2,col=4) # añade una recta de ordenada en el origen
                                   # 15.06, pendiente -0.048, grosor 2 y color 4
> legend(40,14.5,c("línea de regresión"),lty=c(1))
                                   # añade un rótulo en las coordenadas (40,14.5)

```

Destacamos cómo hemos podido añadir la recta simplemente dando su ordenada en el origen y su pendiente. Una posibilidad adicional es incluir una línea horizontal, *h*, en algún valor determinado *va1* de las ordenadas, y/o una línea vertical, *v*, en algún valor *va2* de las abscisas añadiendo a un gráfico ya existente la sentencia `abline(h=va1,v=va2)`; también se pueden poner colores. Nosotros hemos ejecutado la siguiente secuencia, además de (1), (2) y (3), para obtener la Figura 2.15,

```

> plot(x,y,main="Nube de puntos y Recta")
> abline(ajus,col=4)
> legend(35,14.6,c("Línea de Regresión"),lty=c(1),col=4)

```

2.4.3. Precisión del ajuste por mínimos cuadrados

La nube de puntos de la Figura 2.15 parece menos concentrada alrededor de su recta de ajuste que la de la Figura 2.14, lo que llevaría a pensar que la *predicción*

$$y = 15'05908 - 0'04785987 \cdot 60 = 12'19$$

de la marca que obtendría un aficionado que entrenara 60 horas semanales no es muy *fiable*.

La causa de esta falta de concentración puede ser que ambas variables no estén relacionadas linealmente (un atleta nunca llegaría a hacer una marca negativa por muchas horas que se entrenase).

Es probable que para este tipo de datos se ajustase *mejor* una función de tipo *exponencial* de la forma

$$y_{t_i} = a \cdot b^{x_i}$$

con $b < 1$. Es decir, se ajustase mejor a los datos

$$\{(x_i, \log y_i) : i = 1, \dots, n\}$$

una recta de la forma

$$\log y_{t_i} = A + B x_i$$

con pendiente $B = \log b$ negativa (y ordenada en el origen $A = \log a$). Valores que, dicho sea de paso, se obtendrán por las mismas expresiones que antes,

$$B = \log b = \frac{n \sum_{i=1}^n (x_i \log y_i) - (\sum_{i=1}^n x_i) (\sum_{i=1}^n \log y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

y

$$A = \log a = \frac{\sum_{i=1}^n \log y_i - B \sum_{i=1}^n x_i}{n}$$

obteniéndose trivialmente los valores de a y b a partir de los de A y B por las expresiones

$$a = \exp\{A\} \quad y \quad b = \exp\{B\}.$$

Es decir, que para éstos u otros datos no siempre será una recta la mejor función a determinar. En ocasiones será una *función potencial*

$$y_{t_i} = a \cdot x_i^b$$

la adecuada, lo que llevará a ajustar una recta

$$\log y_{t_i} = \log a + b \log x_i$$

a los datos $\{(\log x_i, \log y_i) : i = 1, \dots, n\}$.

Otras veces será necesario utilizar una *parábola*, o en general un *polinomio de grado n* , para conseguir un buen ajuste.

Necesitamos, pues, un valor que nos dé una medida de lo próxima que está la función que hemos ajustado a la nube de puntos de los datos; es decir, una medida de la *bondad del ajuste*.

Como el criterio que hemos utilizado para ajustar una función $y_{t_i} = f(x_i)$ a la nube de puntos $\{(x_i, y_i) : i = 1, \dots, n\}$ ha sido el de mínimos cuadrados, es decir, el de elegir como valores para los parámetros que definen la función f aquellos que minimicen la suma de cuadrados de las desviaciones

$$\sum_{i=1}^n (y_i - y_{t_i})^2$$

parece razonable que una vez determinados dichos parámetros, calculemos cuánto vale dicha suma de cuadrados para cada una de las funciones determinadas, eligiendo aquella para la que se obtenga un menor valor.

Este valor recibe el nombre de *Varianza Residual*

$$V_r = \frac{1}{n} \sum_{i=1}^n (y_i - y_{t_i})^2.$$

La función óptima sería, en principio, aquella para la que su varianza residual fuera cero; es decir, aquella que pasara por todos los puntos y_i .

No obstante, si esto se consigue utilizando una función muy complicada (siempre se puede elegir un polinomio de grado $n - 1$) el ajuste se considera inadecuado porque es preferible poder explicar el fenómeno en estudio (y poder hacer predicciones) con funciones lo más simples posible, incluso dividiendo el rango de las abscisas en intervalos adjuntos y ajustando rectas en ellos. (Puede verse el texto TA-Sección 11.4.1.)

Aunque a la hora de comparar el ajuste de los datos por dos funciones podemos utilizar la varianza residual, siendo mejor aquella para la que dicha varianza sea menor, es conveniente utilizar otro valor que permita decidir si un ajuste es o no adecuado en sí mismo (puede que uno sea mejor que otro aunque ambos sean muy malos).

Surge así el concepto de *Coficiente de Determinación* definido como

$$R^2 = 1 - \frac{V_r}{s_y^2}$$

siendo V_r la varianza residual y $s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - a_y)^2$ la varianza (marginal) de las y_i .

Este coeficiente está comprendido entre 0 y 1, hablándose de un buen ajuste en aquellos casos en los que R^2 esté cerca de 1, y de un mal ajuste en aquellos en los que sea cercano a 0. La valoración de lo que puede considerarse como *cerca* o *lejos* que está el coeficiente, deberá esperar hasta que aprendamos Inferencia Estadística.

Por último, veremos en esta sección un valor, relacionado con los anteriores en el caso de que se ajuste una recta, el cual adquirirá toda su importancia y significado en el Capítulo 10, ya que allí será utilizado para hacer inferencias sobre el grado o fuerza de la relación existente entre dos variables.

Se trata del *Coficiente de correlación lineal* de Pearson, definido como

$$r = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}$$

para el caso en que los n pares de datos vengan aislados, y que en el caso de que éstos aparezcan en forma de distribución bidimensional de frecuencias, la fórmula anterior resulta ser igual a

$$r = \frac{\sum_{i=1}^l \sum_{j=1}^k (x_i - a_x)(y_j - a_y) n_{ij} / n}{\sqrt{\frac{1}{n} \sum_{i=1}^l (x_i - a_x)^2 n_{i.}} \sqrt{\frac{1}{n} \sum_{j=1}^k (y_j - a_y)^2 n_{.j}}}$$

en donde a_x y a_y son las medias marginales.

Este coeficiente toma valores entre -1 y 1 . Estos dos valores extremos indicarían una relación funcional (la recta de mínimos cuadrados) entre los valores de X e Y y un valor igual a 0 indicaría que, mediante la recta de mínimos cuadrados, no vamos a poder explicar adecuadamente a la variable Y en función de X . Aunque estas situaciones extremas son poco frecuentes, valores de $|r|$ cercanos a 1 ó 0 permiten interpretar el grado o fuerza de la relación entre ambas variables, si es que es lineal su relación ya que, si no lo es, este coeficiente no nos será de utilidad.

Cuando el lector estudie los conceptos de Regresión y Correlación desde el punto de vista de la Inferencia Estadística entenderá mejor estos comentarios. Lo que si podemos indicar es que, si es $r > 0$, a medida que aumentemos los valores de las X aumentarán los de las Y , hablándose de *correlación positiva*, utilizando la expresión *correlación negativa* para cuando es $r < 0$.

Si se ha realizado el ajuste de una recta

$$y_{t_i} = \beta_0 + \beta_1 x_i$$

el coeficiente de determinación es igual al cuadrado del coeficiente de correlación, el cual se podrá calcular en este caso por la expresión

$$R^2 = (r)^2 = \frac{\beta_1^2 \left(\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2 / n \right)}{\sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 / n}.$$

Por último, digamos que para los datos del Ejemplo 2.9 el Coeficiente de correlación es $r = -0'6304$, y que, por tanto, el Coeficiente de determinación es $R^2 = 0'3974$. El coeficiente de correlación se obtiene con R ejecutando la función `cor`,

```
> cor(x,y)
[1] -0.6304069
```

2.5. Ejercicios de Autoevaluación

Ejercicio 2.1

Se preguntó a 100 personas en la ciudad suiza de Ginebra cuál era el país de procedencia de su padre, obteniéndose los siguientes datos en porcentajes:

País	
<i>Suiza</i>	0'30
<i>España</i>	0'15
<i>Italia</i>	0'10
<i>Francia</i>	0'20
<i>Austria</i>	0'10
<i>Alemania</i>	0'10
<i>Otro</i>	0'05

Analizar descriptivamente estos datos.

Ejercicio 2.2

Se analizó el IVA que se aplicaba, en un momento determinado, en diversos países europeos, por la compra de obras de arte. Los resultados obtenidos fueron los siguientes:

País	
<i>España</i>	0'16
<i>Italia</i>	0'20
<i>Bélgica</i>	0'06
<i>Holanda</i>	0'06
<i>Alemania</i>	0'07
<i>Portugal</i>	0'17
<i>Luxemburgo</i>	0'06
<i>Finlandia</i>	0'22

Analizar descriptivamente estos datos.

Ejercicio 2.3

En un estudio sobre el número de personas que formaban las bandas de "gangsters" en el Chicago de 1927, se obtuvieron los siguientes datos sobre 825 de dichas bandas:

<i>Tamaño banda</i>	frecuencia	<i>Tamaño banda</i>	frecuencia
3 – 6	37	41 – 51	51
6 – 11	198	51 – 76	26
11 – 16	191	76 – 101	25
16 – 21	149	101 – 201	25
21 – 26	79	201 – 501	11
26 – 31	46	501 – 1000	2
31 – 41	55		

Analizar descriptivamente estos datos.

Ejercicio 2.4

Los siguientes pares de datos corresponden al *Peso en gramos de caucho* obtenido (después de la vulcanización), variable Y , y la *circunferencia en cm. de la corona del guayule* de donde se obtuvo dicho caucho, variable X ,

X	68	100	85	133	130	165	120	120	155	117
Y	6	6'3	6	6'3	7	8	7	8'1	9	6'6

Determinar la recta de mínimos cuadrados y el coeficiente de correlación.

Ejercicio 2.5

Se cree que existe una relación lineal en los homínidos entre el volumen encefálico (en c.c.) del recién nacido, X , y el del adulto, Y . Para ello se eligieron al azar cuatro especies de homínidos obteniéndose los siguientes datos (Bermúdez, 2010, pág. 69)

	Especie			
	Chimpancé	Australopiteco	Homo Ergaster	Homo Sapiens
X	160	180	300	380
Y	400	450	800	1350

Determinar la recta de mínimos cuadrados y el coeficiente de correlación.

2.6. Lecturas Recomendadas

Tukey, J.W. (1977). *Exploratory Data Analysis*. Editorial Addison-Wesley.

Capítulo 3

Probabilidad

3.1. Introducción

La Teoría de la Probabilidad constituye la base o fundamento de la Estadística ya que las inferencias que hagamos sobre la población o poblaciones en estudio se moverán dentro de unos márgenes de error, controlado en términos de probabilidades.

La Probabilidad que habitualmente manejaremos en Estadística vendrá ligada a un Modelo Probabilístico, como los que veremos en el siguiente capítulo, modelo el cual suponemos rige nuestro fenómeno aleatorio en estudio. Por tanto, cuando asumamos un Modelo, tendremos establecido, de una vez, todo el esquema probabilístico que necesitamos. Pero como en muchas ocasiones vamos a tener que operar con probabilidades, por lo que es muy conveniente conocer sus principales propiedades así como conocer los elementos que acompañan al Modelo admitido como válido.

Vamos a denominar *Nivel I* al nivel en donde están los elementos básicos: el conjunto de individuos analizados, que denominaremos *Espacio Muestral*, Ω , y la *probabilidad* P con la que éstos van a ser seleccionados para dar lugar a la muestra. P va a estar definida no sólo sobre Ω sino sobre el conjunto de todos los subconjuntos posibles de Ω ya que no sólo tiene sentido hablar de la probabilidad de extraer para nuestra muestra a, digamos, Juan o José, sino también tiene sentido hablar de la probabilidad de extraer, por ejemplo, a todos los varones. Este conjunto, sobre el que va a estar definida P , lo denominaremos *Espacio de Sucesos* \mathcal{A} . Tenemos, por tanto, los tres elementos básicos del Cálculo de Probabilidades: (Ω, \mathcal{A}, P) . En situaciones más generales, el espacio muestral podrá ser más complejo y el Espacio de Sucesos requerirá una estructura más sofisticada que recibirá el nombre de σ -álgebra; no obstante, siempre tendremos los tres elementos básicos anteriores que reciben el nombre de *Espacio Probabilístico*.

Pero, como dijimos anteriormente, estaremos interesados en observar alguna característica X en los individuos del Espacio Muestral, por ejemplo, su Estatura; esta característica X es una variable estadística, como sabemos por el capítulo anterior, pero que, al estar sometida a la aleatoriedad del muestreo, no permite conocer su valor con seguridad sino con una cierta probabilidad, dando lugar a lo que, en el capítulo siguiente denominaremos *variable aleatoria*, siendo un determinado *Modelo Probabilístico* el que regirá su comportamiento. De momento observemos que, mediante X , pasamos del Nivel I caracterizado por la tripleta (Ω, \mathcal{A}, P) , a un *Nivel II* al ser X una transformación de Ω en los números reales \mathbb{R}

$$X : \Omega \longrightarrow \mathbb{R}$$

es decir, a cada individuo seleccionado le observamos, por ejemplo, su Estatura, pudiendo hablar ahora, no ya en términos del Nivel I de la probabilidad de seleccionar a tal o cual individuo, sino, en términos del Nivel II, de la probabilidad de que la Estatura de los individuos de la población (mejor dicho, de la Estatura de un individuo elegido al azar de la población) esté, por ejemplo, entre 1'70 y 1'80,

$$P\{1'70 < X < 1'80\}.$$

Según sea el Espacio Muestral Ω y según sea P , así se *distribuirá* X , es decir, según sea la Población y la ley P que rige la selección aleatoria de los individuos de la Población en estudio, la variable aleatoria X seguirá uno u otro *Modelo Probabilístico*. Aquí es donde entramos nosotros. Desde un punto de vista práctico no nos va a interesar mucho el Nivel I de la situación que estemos estudiando (aunque ahí estará siempre), sino precisar qué variable X estamos observando y cuál es el Modelo Probabilístico que la rige.

Añadamos que, si estamos observando una sola característica, es decir X es *unidimensional*, hablaremos de Estadística Univariante, mientras que cuando observemos más de una característica en los individuos seleccionados (los p valores de una fila de la Matriz de Datos) estaremos dando lugar a la Estadística Multivariante.

En este capítulo estudiaremos el Nivel I antes mencionado por lo que el lector interesado exclusivamente en Inferencia Estadística puede pasar ya al Capítulo 4 puesto que el resto de secciones de este capítulo están elaboradas desde un punto de vista algo más abstracto. Por otro lado, aquel lector que quiera profundizar más aún en este Nivel I, puede continuar con el libro de Quesada y García (1988).

3.2. Espacio Muestral

Como ya apuntábamos en la Introducción del primer capítulo, la Estadística, y por tanto el Cálculo de Probabilidades, se ocupan de los denominados *fenómenos o experimentos aleatorios*.

El conjunto de todos los resultados posibles diferentes de un determinado experimento aleatorio se denomina *Espacio Muestral* asociado a dicho experimento y se suele representar por Ω . A los elementos de Ω se les denomina *sucesos elementales*.

Así por ejemplo, el espacio muestral asociado al experimento aleatorio consistente en el lanzamiento de una moneda es $\Omega = \{Cara, Cruz\}$; el espacio muestral asociado al lanzamiento de un dado es $\Omega = \{1, 2, 3, 4, 5, 6\}$, siendo *Cara* y *Cruz* los sucesos elementales asociados al primer experimento aleatorio y 1, 2, 3, 4, 5 y 6 los seis sucesos elementales del segundo experimento aleatorio; el espacio muestral asociado a la selección aleatoria de individuos de una población determinada, será el conjunto de personas de esa población y, los sucesos elementales, dichos individuos.

Aunque se pueden considerar varios tipos de Espacios Muestrales, no sólo los finitos (o numerables) sino también los continuos (por ejemplo, un intervalo), Ω no es más que un conjunto abstracto de puntos, por lo que el lenguaje, los conceptos y propiedades de la teoría de conjuntos constituyen un contexto natural en el que desarrollar el Cálculo de Probabilidades.

Si \mathcal{A} el conjunto de las partes de Ω , es decir, el conjunto de todos los subconjuntos de Ω (o como dijimos antes, esa estructura denominada σ -álgebra), cualquier elemento de \mathcal{A} , contendrá una cierta incertidumbre, por lo que trataremos de asignarle un número entre 0 y 1 como medida de su incertidumbre. En Cálculo de Probabilidades dichos subconjuntos reciben en el nombre de *sucesos*, siendo la medida de la incertidumbre su probabilidad.

Por tanto, asociado a todo experimento aleatorio existen tres conjuntos: El espacio muestral Ω , la clase de los sucesos, es decir, el conjunto de los elementos con incertidumbre asociados a nuestro experimento aleatorio \mathcal{A} , y una función $P : \mathcal{A} \rightarrow [0, 1]$, la cual asignará a cada suceso (elemento de \mathcal{A}) un número entre cero y uno como medida de su incertidumbre.

Advertimos no obstante, que la elección del espacio muestral asociado a un experimento aleatorio, incluso en los casos más elementales, no tiene por qué ser única, sino que dependerá de qué sucesos elementales queramos considerar como distintos y del problema de la asignación de la probabilidad sobre esos sucesos elementales.

Ejemplo 3.1

Consideremos el experimento aleatorio consistente en extraer una bola al azar de una urna compuesta por tres bolas rojas, dos blancas y una verde. Podemos considerar como espacio muestral

$$\Omega_1 = \{\omega_1, \omega_2, \omega_3\}$$

en donde sea $\omega_1 = \text{bola roja}$, $\omega_2 = \text{bola blanca}$ y $\omega_3 = \text{bola verde}$, aunque también podíamos haber considerado como espacio muestral el conjunto

$$\Omega_2 = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$

en donde $\omega_i = \text{bola roja}$, $i = 1, 2, 3$, $\omega_i = \text{bola blanca}$, $i = 4, 5$ y $\omega_6 = \text{bola verde}$, haciendo las bolas distinguibles.

Ambos pueden ser considerados espacios muestrales del experimento descrito, eligiendo el que más nos convenga, por ejemplo, a la hora de asignar la probabilidad a los sucesos elementales de uno u otro espacio muestral.

Respecto a la clase de los sucesos \mathcal{A} , es natural que ésta tenga una estructura tal que permita hablar no sólo de sucesos sino también de su unión, intersección, diferencia, complementario, etcétera, debiendo ser la clase \mathcal{A} , en consecuencia, cerrada a dichas operaciones entre “conjuntos” (entre sucesos). Ésta es la situación del conjunto de las partes cuando Ω es finito o inclusive numerable (caso, por ejemplo, del espacio muestral asociado al experimento aleatorio consistente en lanzar una moneda hasta que salga cara por primera vez). En otras ocasiones en las que Ω sea un conjunto continuo (por ejemplo, cuando estudiamos el tiempo que tarda un isótopo radioactivo en volverse inestable), deberá ser \mathcal{A} un conjunto estrictamente más pequeño que el conjunto de las partes de Ω .

En todo caso podemos pensar en \mathcal{A} como en el conjunto que contiene todos los elementos de interés, es decir, todos los sucesos a los que les corresponde una probabilidad.

Apuntemos además algunas peculiaridades del Cálculo de Probabilidades respecto a la teoría de conjuntos. Aquí, el conjunto vacío \emptyset recibe el nombre de *suceso imposible*, definido como aquel subconjunto de Ω que no contiene ningún suceso elemental y que corresponde a la idea de aquel suceso que no puede ocurrir.

De forma análoga, el espacio total Ω recibe el nombre de *suceso seguro* al recoger dicha denominación la idea que representa.

Llamaremos *sucesos incompatibles* a aquellos cuya intersección sea el suceso imposible.

Por último, digamos que la inclusión de sucesos, $A \subset B$, se interpreta aquí como que siempre que se cumpla el suceso A se cumple el B ; por ejemplo, siempre que *salga el 2* (suceso A) sale *par* (suceso B).

Ejemplo 3.2

El espacio probabilístico asociado al experimento aleatorio consistente en el lanzamiento de un dado, tendrá como espacio muestral $\Omega = \{1, 2, 3, 4, 5, 6\}$ y como espacio de sucesos el conjunto de las partes por ser Ω finito, el cual contiene 2^6 elementos,

$$\mathcal{A} = \{ \emptyset, \{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{1, 6\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{2, 6\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{4, 5\}, \{4, 6\}, \{5, 6\}, \{1, 2, 3\}, \{1, 2, 4\}, \{1, 2, 5\}, \{1, 2, 6\}, \{1, 3, 4\}, \{1, 3, 5\}, \{1, 3, 6\}, \{1, 4, 5\}, \{1, 4, 6\}, \{1, 5, 6\}, \{2, 3, 4\}, \{2, 3, 5\}, \{2, 3, 6\}, \{2, 4, 5\}, \{2, 4, 6\}, \{2, 5, 6\}, \{3, 4, 5\}, \{3, 4, 6\}, \{3, 5, 6\}, \{4, 5, 6\}, \{1, 2, 3, 4\}, \{1, 2, 3, 5\}, \{1, 2, 3, 6\}, \{1, 2, 4, 5\}, \{1, 2, 4, 6\}, \{1, 2, 5, 6\}, \{1, 3, 4, 5\}, \{1, 3, 4, 6\}, \{1, 3, 5, 6\}, \{1, 4, 5, 6\}, \{2, 3, 4, 5\}, \{2, 3, 4, 6\}, \{2, 3, 5, 6\}, \{2, 4, 5, 6\}, \{3, 4, 5, 6\}, \{1, 2, 3, 4, 5\}, \{1, 2, 3, 4, 6\}, \{1, 2, 3, 5, 6\}, \{1, 2, 4, 5, 6\}, \{1, 3, 4, 5, 6\}, \{2, 3, 4, 5, 6\}, \Omega \}.$$

Obsérvese que este conjunto contiene los sucesos sobre los que habitualmente se tiene incertidumbre, como por ejemplo que salga un número par, $\{2, 4, 6\}$, o un número mayor que cuatro, $\{5, 6\}$, o simplemente que salga un seis, $\{6\}$, y que como se ve es cerrado respecto de las operaciones entre conjuntos.

El último elemento del espacio probabilístico es la probabilidad que, como antes dijimos, está definida sobre \mathcal{A} , asignando a cada suceso un número entre 0 y 1. Éste es el objetivo de la siguiente sección.

3.3. Conceptos de Probabilidad

En la sección anterior vimos que a cada suceso A le corresponde su probabilidad $P(A)$, pero, ¿este número viene dado?, ¿es un número desconocido?, ¿le tenemos que calcular nosotros?

En los casos más sencillos bastará con *asignar* la probabilidad a los sucesos elementales de un experimento aleatorio. La probabilidad de los demás sucesos se podrá calcular utilizando las propiedades que más adelante veremos.

En los casos más complicados (que habitualmente se corresponderán con las situaciones reales) asignaremos un *Modelo Probabilístico* al experimento en cuestión, como ideal que creemos corresponde a la situación en estudio, ideal que se puede chequear inferencialmente. Más adelante hablaremos de la asignación de probabilidades; ahora analicemos brevemente los conceptos que se han desarrollado a lo largo de la Historia, con el propósito de formalizar las ideas intuitivas que desde el origen del hombre siempre existieron sobre la Probabilidad, aunque no llegaran a formalizarse hasta comienzos del siglo XX.

Concepto Frecuentista

Es un hecho, empíricamente comprobado, que la frecuencia relativa de un suceso tiende a estabilizarse cuando la frecuencia total aumenta.

Surge así el *concepto frecuentista* de la probabilidad de un suceso como un número ideal al que converge su frecuencia relativa cuando la frecuencia total tiende a infinito.

Así, solemos afirmar que la probabilidad de que salga un seis al tirar un dado es $1/6$ porque al hacer un gran número de tiradas su frecuencia relativa es aproximadamente ésa.

El problema radica en que, al no poder repetir la experiencia infinitas veces, la probabilidad de un suceso ha de ser aproximada por su frecuencia relativa para un n suficientemente grande, y ¿cuán grande es un n grande? O, ¿qué hacer con aquellas experiencias que sólo se pueden repetir una vez?

Concepto Clásico

Está basado en el concepto de resultados igualmente verosímiles y motivado por el denominado *Principio de la razón insuficiente*, el cual postula que si no existe un fundamento para preferir una, entre varias posibilidades, todas deben ser consideradas equiprobables.

Así, en el lanzamiento de una moneda perfecta la probabilidad de cara debe ser igual que la de cruz y, por tanto, ambas iguales a $1/2$.

De la misma manera, la probabilidad de cada uno de los seis sucesos elementales asociados al lanzamiento de un dado debe ser $1/6$.

Laplace recogió esta idea y formuló la regla clásica del cociente entre casos favorables y casos posibles, supuestos éstos igualmente verosímiles.

El problema aquí surge porque en definitiva *igualmente verosímil* es lo mismo que *igualmente probable*, es decir, se justifica la premisa con el resultado. Además ¿qué ocurre cuando estamos considerando un experimento donde no se da esa simetría?, o, ¿qué hacer cuando el número de resultados posibles es infinito?

Concepto Subjetivo

Se basa en la idea de que la probabilidad que una persona dé a un suceso debe depender de su juicio y experiencia personal, pudiendo dar dos personas distintas probabilidades diferentes a un mismo suceso.

Estas ideas pueden formalizarse, y si las opiniones de una persona satisfacen ciertas relaciones de consistencia, se puede llegar a definir una probabilidad para los sucesos.

El principal problema a que da lugar esta definición es que dos personas diferentes pueden dar probabilidades diferentes a un mismo suceso.

Definición formal de Probabilidad

Los anteriores conceptos de lo que debería ser la probabilidad de un suceso, llevaron a Kolmogorov a dar una definición axiomática de probabilidad. Es decir, a introducir rigor matemático en el concepto de probabilidad, de forma que se pudiera desarrollar un teoría sólida sobre el concepto definido.

Así, llamaremos *Probabilidad* a una aplicación

$$P : \mathcal{A} \mapsto [0, 1]$$

tal que

- *Axioma 1:* Para todo suceso A de \mathcal{A} sea $P(A) \geq 0$.
- *Axioma 2:* Sea $P(\Omega) = 1$.
- *Axioma 3:* Para toda colección de sucesos incompatibles, $\{A_i\}$ con $A_i \cap A_j = \emptyset$, $i \neq j$ debe ser

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

que el lector puede interpretar fácilmente, aquí y en lo que sigue, aunque no tenga todavía muy claro los conceptos de unión y suma infinitos.

Obsérvese que esta definición no dice cómo asignar las probabilidades ni siquiera a los sucesos elementales. Sólo dice que cualquier asignación que hagamos debe verificar estos tres axiomas para que pueda llamarse Probabilidad.

3.4. Propiedades elementales de la Probabilidad

Toda probabilidad cumple una serie de propiedades, las cuales se obtienen como consecuencia de los axiomas que debe de cumplir. A continuación vamos a demostrar las más importantes:

(a) $P(\emptyset) = 0$.

En efecto: Si consideramos la sucesión infinita

$$\{A_i\}_{i=1}^{\infty} = \{A, \emptyset, \emptyset, \dots\}$$

es

$$\bigcup_{i=1}^{\infty} A_i = A$$

por lo que, por el Axioma 3, deberá ser

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

es decir,

$$P(A) = P(A) + \sum_{i=2}^{\infty} P(A_i)$$

de donde se deduce que $P(A_i) = P(\emptyset)$, para todo $i = 2, \dots$, no debe sumar nada, es decir, debe ser

$$P(\emptyset) = 0.$$

(b) Se cumple la *aditividad finita* para sucesos incompatibles. Es decir,

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

si $A_i \cap A_j = \emptyset$, $i \neq j$.

En efecto: Basta considerar la sucesión

$$\{A_i\}_{i=1}^{\infty} = \{A_1, \dots, A_n, \emptyset, \emptyset, \dots\}$$

y aplicar de nuevo el Axioma 3 y luego la propiedad anterior, quedando

$$P\left(\bigcup_{i=1}^n A_i\right) = P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^n P(A_i) + 0$$

es decir, la propiedad deseada.

(c) La probabilidad del complementario de un suceso A es

$$P(A^*) = 1 - P(A).$$

En efecto: Aplicando primero el Axioma 2 y luego la aditividad finita acabada de demostrar, será

$$P(A \cup A^*) = P(\Omega) = 1$$

y

$$P(A) + P(A^*) = 1$$

de donde se obtiene la propiedad propuesta.

(d) Si dos sucesos son tales que $A \subset B$, entonces es $P(A) \leq P(B)$.

En efecto: B se puede poner de la forma

$$B = A \cup (B - A)$$

con lo que, por la aditividad finita de la probabilidad, será

$$P(B) = P(A) + P(B - A).$$

La propiedad enunciada se tendrá ahora, por el Axioma 1, como consecuencia de ser $P(B - A) \geq 0$.

(e) Si dos sucesos no son incompatibles, la probabilidad de su unión debe calcularse por la siguiente regla:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

En efecto: Los sucesos A y B se pueden escribir como unión de sucesos disjuntos de la forma,

$$A = (A \cap B) \cup (A \cap B^*) \quad , \quad B = (A \cap B) \cup (A^* \cap B)$$

con lo que, por la propiedad de aditividad finita antes demostrada, será

$$P(A) = P(A \cap B) + P(A \cap B^*) \quad \text{y} \quad P(B) = P(A \cap B) + P(A^* \cap B)$$

es decir,

$$P(A \cap B^*) = P(A) - P(A \cap B) \quad \text{y} \quad P(A^* \cap B) = P(B) - P(A \cap B).$$

Como, por otro lado, $A \cup B$ se puede expresar como unión disjunta de la forma

$$A \cup B = (A \cap B) \cup (A \cap B^*) \cup (A^* \cap B)$$

su probabilidad será

$$P(A \cup B) = P(A \cap B) + P(A \cap B^*) + P(A^* \cap B)$$

y, sustituyendo los valores antes calculados para los dos últimos sumandos, quedará

$$P(A \cup B) = P(A \cap B) + P(A) - P(A \cap B) + P(B) - P(A \cap B)$$

o en definitiva,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

como queríamos demostrar.

3.5. Asignación de Probabilidad en espacios muestrales discretos

Por las propiedades demostradas en la sección anterior es suficiente conocer la probabilidad de los sucesos elementales ya que, entonces, se podrá determinar la de cualquier otro suceso.

Así, en el Ejemplo 3.2, si la probabilidad de obtener un 1 es p_1 , la de un 3 p_2 y la de un 5 p_3 , la del suceso *obtener un número impar* será, por la propiedad (b), $p_1 + p_2 + p_3$.

Es decir, el problema radica en *asignar una probabilidad a los sucesos elementales*: asignar un número entre 0 y 1 a cada uno de los sucesos elementales, de tal forma que su suma sea 1.

En principio, cualquier asignación que cumpla los tres axiomas mencionados en la definición de probabilidad es válida. No obstante, el propósito del Cálculo de Probabilidades, como soporte de la Estadística, es el de construir un esquema matemático que refleje de la forma más exacta posible el fenómeno aleatorio real que estemos estudiando, por lo que la asignación de probabilidad que hagamos debe ser lo más ajustada posible a la realidad que estamos observando. Así, en el Ejemplo 3.2 la asignación razonable será,

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6.$$

En otras ocasiones, la observación del mismo fenómeno en otra población semejante a la que estamos estudiando, o inclusive en la objeto de estudio en un tiempo anterior, permitirá obtener una distribución de frecuencias a partir de la cual asignar una probabilidad.

Ejemplo 3.3

Un estudio sobre el color de los ojos en niños recién nacidos de una población determinada dio la siguiente distribución de frecuencias relativas

Color	f_i
Azules	0'05
Verdes	0'02
Castaños	0'69
Negros	0'24

Supuesto que no consideremos la componente genética que esta característica tiene, no teniendo en cuenta el color de ojos de los padres, podríamos considerar esta distribución de frecuencias relativas como una buena aproximación de la probabilidad y decir, por ejemplo, que la probabilidad que tiene un recién nacido de esta población de tener los ojos claros es

$$P\{\text{Ojos Claros}\} = P\{\text{Azules}\} + P\{\text{Verdes}\} = 0'05 + 0'02 = 0'07.$$

A veces es precisamente la asignación de la probabilidad la que determina

el Espacio Muestral. Así, en el Ejemplo 3.1, si consideramos como Espacio Muestral

$$\Omega_2 = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$$

en donde era $\omega_i = \text{bola roja}$, $i = 1, 2, 3$, $\omega_i = \text{bola blanca}$, $i = 4, 5$ y $\omega_6 = \text{bola verde}$, los seis sucesos elementales pueden ser considerados como equiprobables, siendo en ese caso, $P(\omega_i) = 1/6$, mientras que si consideramos como espacio muestral

$$\Omega_1 = \{\omega_1, \omega_2, \omega_3\}$$

en donde sea $\omega_1 = \text{bola roja}$, $\omega_2 = \text{bola blanca}$ y $\omega_3 = \text{bola verde}$, los sucesos dejan ya de ser equiprobables, por lo que, en una situación más compleja, la elección de un espacio muestral en donde los sucesos elementales sean equiprobables puede ser más adecuada.

Aquí, por las propiedades estudiadas en la sección anterior, es equivalente utilizar Ω_2 con sucesos elementales equiprobables, que utilizar Ω_1 con $P(\omega_1) = 1/2$, $P(\omega_2) = 2/6$ y $P(\omega_3) = 1/6$.

Sin embargo, advertimos ya, que ni la mayoría de los fenómenos aleatorios que se observan en la naturaleza admiten un esquema tan sencillo, ni que, como veremos en el siguiente capítulo, será necesario detallar esta asignación en los sucesos elementales en la mayoría de las situaciones reales, es decir, analizar el que hemos denominado en la Introducción, Nivel I. Se podrá actuar en una forma más *encapsulada* (es decir, trabajar en el Nivel II), asignando de forma global un *Modelo Probabilístico* a la característica que estemos estudiando, el cual recibe el nombre de *Distribución de Probabilidad*. No obstante, en esa modelización global que hagamos de la realidad, siempre será posible descender hasta la probabilidad que tiene asociada.

Anunciamos también que la asignación que hagamos, tanto en el Nivel I como en el Nivel II, podrá ser *contrastada* con las observaciones que hagamos de nuestro experimento aleatorio mediante la Inferencia Estadística, de forma que podamos estar razonablemente seguros de nuestras conclusiones.

Dentro de las posibles asignaciones de probabilidad existe una que destaca, tanto por ser de las más utilizadas como por obtenerse de ella interesantes propiedades. Se trata del denominado Modelo Uniforme.

3.6. Modelo Uniforme

En esta sección estudiaremos un caso particular muy importante, el cual se corresponde con una situación en la que los sucesos elementales del espacio muestral (que suponemos finito) puedan ser considerados como equiprobables.

Ejemplo 3.4

Consideremos el experimento aleatorio consistente en lanzar una moneda al aire. En el espacio muestral asociado, $\Omega = \{Cara, Cruz\}$, ambos sucesos elementales pueden considerarse como equiprobables.

Ejemplo 3.5

Si seleccionamos al azar una carta de una baraja española, los cuarenta sucesos elementales correspondientes a las cuarenta cartas, pueden ser considerados como equiprobables, estando de nuevo ante un esquema de modelo uniforme.

Ejemplo 3.6

Supongamos el experimento aleatorio consistente en lanzar dos dados al aire. De nuevo estaremos ante un modelo uniforme.

Ejemplo 3.7

Consideremos el experimento aleatorio consistente en lanzar al aire una moneda dos veces. El espacio muestral que razonablemente vendrá asociado será, $\Omega = \{(C, C), (C, X), (X, X)\}$, siendo C y X , respectivamente, la *cara* y la *cruz* de la moneda.

En este espacio muestral los sucesos no son equiprobables, aunque puede conseguirse esta simetría si consideramos como espacio muestral $\Omega = \{(C, C), (C, X), (X, C), (X, X)\}$.

En todos estos casos de modelos uniformes, en especial en aquellos que el espacio muestral es finito, $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$, el cálculo de las probabilidades de los sucesos resulta sencillo ya que, al ser los sucesos elementales incompatibles y equiprobables, será

$$1 = P(\Omega) = P(\omega_1) + \dots + P(\omega_n) = n \cdot P(\omega_i)$$

con lo que $P(\omega_i) = 1/n, \forall i = 1, \dots, n$. Por tanto, si un suceso A es unión de k sucesos elementales, será

$$P(A) = \frac{k}{n} = \frac{\text{casos favorables a } A}{\text{casos posibles}}$$

con lo que, en definitiva, el cálculo de probabilidades de sucesos en un modelo uniforme, se limita a contar el número de casos favorables a dicho suceso y el número de casos posibles.

No obstante, dicho cómputo no resultará siempre fácil por lo que es conveniente tener presente las fórmulas de las *variaciones*, *combinaciones* y *permutaciones*, ya que éstas facilitarán el cálculo.

Si de un grupo de N elementos tomamos n , y nos importa el orden de los n elementos seleccionados, tendremos *variaciones* y si no nos importa el orden, tendremos *combinaciones*. Además, si admitimos la posibilidad de que entre estos n pueda haber elementos repetidos, hablaremos, respectivamente, de *variaciones* y de *combinaciones con repetición*.

Por último, si solamente queremos contar el número posible de reordenaciones de un conjunto de elementos, hablaremos de *permutaciones con o sin repetición* dependiendo de que admitamos o no la posibilidad de que haya elementos repetidos.

Las fórmulas son:

Variaciones de N elementos tomados de n en n

$$V_{N,n} = N \cdot (N - 1) \cdot \dots \cdot (N - n + 1)$$

Variaciones con repetición de N elementos tomados de n en n

$$RV_{N,n} = N^n$$

Combinaciones de N elementos tomados de n en n

$$C_{N,n} = \binom{N}{n} = \frac{N!}{n! (N - n)!}$$

Combinaciones con repetición de N elementos tomados de n en n

$$RC_{N,n} = \binom{N + n - 1}{n} = \binom{N + n - 1}{N - 1} = \frac{(N + n - 1)!}{n! (N - 1)!}$$

Permutaciones de N elementos

$$P_N = N! = N \cdot (N - 1) \cdot \dots \cdot 2 \cdot 1$$

Permutaciones con repetición de N elementos, uno de los cuales se repite n_1 veces, otro n_2 veces, ..., otro n_r veces

$$RP_N^{n_1, \dots, n_r} = \frac{N!}{n_1! \cdot n_2! \cdot \dots \cdot n_r!}$$

Ejemplo 3.8

Una enciclopedia en seis volúmenes es colocada en una estantería de forma aleatoria. La probabilidad de que resulte colocada de forma correcta, supuesto que esto signifique empezar a contar por la izquierda, será

$$P(A) = \frac{\text{casos favorables}}{\text{casos posibles}} = \frac{1}{6!}.$$

3.7. Probabilidad condicionada

Mediante un espacio probabilístico damos una formulación matemática a un fenómeno aleatorio que estamos observando. Parece por tanto razonable que si observamos algo que aporta información a nuestro fenómeno aleatorio, esto deba de alterar el espacio probabilístico de partida.

Por ejemplo, la extracción de una bola de una urna con tres bolas blancas y dos negras, puede formalizarse con un espacio probabilístico en el que los sucesos elementales sean las cinco bolas y donde la probabilidad sea uniforme sobre estos cinco sucesos elementales, es decir, igual a $1/5$.

Si extraemos una bola de la urna, es decir, si observamos el suceso $A = \text{bola negra}$, y no la devolvemos a la urna, es razonable que el espacio probabilístico cambie en el sentido no sólo de que ahora ya habrá únicamente cuatro sucesos elementales, sino que además la función de probabilidad deberá cambiar en orden a recoger la información que la observación del suceso A nos proporcionó.

Es decir, en el nuevo espacio probabilístico deberá hablarse de probabilidad *condicionada* por el suceso A , de forma que se recojan hechos tan evidentes como que ahora la probabilidad (condicionada) de obtener negra se habrá reducido y habrá aumentado la de blanca.

Las propiedades vistas en el capítulo anterior para las distribuciones de frecuencias condicionadas llevan a la siguiente definición

Definición

Dado un espacio probabilístico (Ω, \mathcal{A}, P) y un suceso $B \in \mathcal{A}$ tal que $P(B) > 0$, llamaremos *probabilidad condicionada* del suceso A por el suceso B a

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

A partir de esta definición podemos deducir que

$$P(A \cap B) = P(A/B) \cdot P(B)$$

y como los sucesos A y B pueden intercambiarse en la expresión anterior, será (lógicamente si es $P(A) > 0$),

$$P(A \cap B) = P(A/B) \cdot P(B) = P(B/A) \cdot P(A)$$

por lo que tenemos una expresión más para calcular la probabilidad condicionada

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}.$$

3.8. Independencia de sucesos

Existen situaciones en las que la información suministrada por la ocurrencia de un suceso B no altera para nada el cálculo de la probabilidad de otro suceso A . Son aquellas en las que el suceso A es *independiente* de B . Es decir, cuando

$$P(A/B) = P(A).$$

Como entonces, por la última expresión de la probabilidad condicionada, es

$$P(B/A) = \frac{P(A) \cdot P(B)}{P(A)} = P(B)$$

y, por tanto, se podría decir que también B lo es de A , hablaremos de *sucesos independientes* cuando esta situación ocurra. La definición formal que se da a continuación implica estas dos situaciones.

Definición

Dos sucesos A y B de un mismo espacio probabilístico (Ω, \mathcal{A}, P) se dicen independientes cuando

$$P(A \cap B) = P(A) \cdot P(B)$$

3.9. Teorema de la Probabilidad Total

En el cálculo numérico de probabilidades tiene una gran aplicación práctica el siguiente resultado.

Teorema 3.1

Sea un espacio probabilístico (Ω, \mathcal{A}, P) y $\{A_n\} \subset \mathcal{A}$ una partición de sucesos de Ω . Es decir,

$$\bigcup_n A_n = \Omega \quad \text{y} \quad A_i \cap A_j = \emptyset \quad \text{para todo } i \neq j.$$

Entonces, para todo suceso $B \in \mathcal{A}$ es

$$P(B) = \sum_n P(B/A_n) \cdot P(A_n).$$

Resultado que se puede parafrasear diciendo que la probabilidad de un suceso que se puede dar de varias maneras es igual a la suma de los productos de las probabilidades de éste en cada una de esas maneras, $P(B/A_n)$, por las probabilidades de que se den estas maneras, $P(A_n)$.

Ejemplo 3.9

Una población está formada por tres grupos étnicos: A (un 30 %), B (un 10 %) y C (un 60 %). Además se sabe que el porcentaje de personas con *ojos claros* en cada una de estas poblaciones es, respectivamente, del 20 %, 40 % y 5 %. Por el Teorema de la Probabilidad Total, la probabilidad de que un individuo elegido al azar de esta población tenga ojos claros es

$$\begin{aligned} P(\text{ojos claros}) &= P(A) \cdot P(\text{ojos claros}/A) + P(B) \cdot P(\text{ojos claros}/B) + \\ &+ P(C) \cdot P(\text{ojos claros}/C) = 0'3 \cdot 0'2 + 0'1 \cdot 0'4 + 0'6 \cdot 0'05 = 0'13. \end{aligned}$$

3.10. Teorema de Bayes

El siguiente teorema es un resultado con una gran carga filosófica detrás, el cual mide el cambio que se va produciendo en las probabilidades de los sucesos a medida que vamos haciendo observaciones. Paradójicamente a su importancia, su demostración no es más que la aplicación de la definición de probabilidad condicionada seguida de la aplicación del teorema de la probabilidad total.

Teorema 3.2

Sea un espacio probabilístico (Ω, \mathcal{A}, P) , $\{A_n\} \subset \mathcal{A}$ una partición de sucesos de Ω y $B \in \mathcal{A}$ un suceso con probabilidad positiva. Entonces, para todo suceso A_i es

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_n P(A_n) \cdot P(B/A_n)}.$$

Este teorema tiene una interpretación intuitiva muy interesante. Si las cosas que pueden ocurrir las tenemos clasificadas en los sucesos A_i de los cuales conocemos sus probabilidades $P(A_i)$, denominadas *a priori*, y se observa un suceso B , la fórmula de Bayes nos da las probabilidades *a posteriori* de los sucesos A_i , ajustadas o modificadas por B .

Ejemplo 3.10

Supongamos que tenemos una urna delante de nosotros de la cual sólo conocemos que, o es la urna A_1 con 3 bolas blancas y 1 negra, o es la urna A_2 con 3 bolas negras y 1 blanca. Con objeto de obtener más información acerca de cuál urna tenemos delante, realizamos un experimento consistente en extraer una bola de la urna desconocida. Si suponemos que la bola extraída resultó blanca, suceso B , y, *a priori*, ninguna de las dos urnas es más verosímil que la otra, es decir, creemos que es $P(A_1) = P(A_2) = 1/2$, entonces la fórmula de Bayes nos dice que las probabilidades *a posteriori* de cada urna son

$$P(A_1/1B) = \frac{3}{4} \quad \text{y} \quad P(A_2/1B) = \frac{1}{4}$$

habiendo alterado de esta forma nuestra creencia sobre la urna que tenemos delante: antes creíamos que eran equiprobables y ahora creemos que es tres veces más probable que la urna desconocida sea la A_1 .

Pero, ¿qué ocurrirá si extraemos otra bola? Lógicamente, en la fórmula de Bayes deberemos tomar ahora como probabilidades *a priori* las calculadas, $3/4$ y $1/4$, pues éstas son nuestras creencias sobre la composición de la urna, antes de volver a realizar el experimento.

Si suponemos que la bola no fue reemplazada (se deja para el lector el caso de reemplazamiento), y sale una bola negra, suceso $2N$, la fórmula de Bayes nos devolvería la incertidumbre inicial, ya que sería

$$P(A_1/2N) = \frac{1}{2} \quad \text{y} \quad P(A_2/2N) = \frac{1}{2}.$$

Si hubiera salido blanca, la fórmula de Bayes, al igual que la lógica, también sería concluyente,

$$P(A_1/2B) = 1 \quad \text{y} \quad P(A_2/2B) = 0.$$

La utilización de la fórmula de Bayes, es decir, la utilización de distribuciones de probabilidad *a posteriori* como modelos en la estimación de parámetros, al recoger ésta tanto la información muestral, $P(B/A_i)$, como la información *a priori* sobre ellos, $P(A_i)$, constituye una filosofía inferencial en gran desarrollo en los últimos años denominada *Inferencia Bayesiana*.

3.11. Ejercicios de Autoevaluación

Ejercicio 3.1

Hallar la probabilidad de un suceso sabiendo que el cuadrado de esta probabilidad menos el cuadrado de la probabilidad del suceso complementario es igual a $1/9$.

Ejercicio 3.2

Se selecciona al azar una carta de una baraja española de 40 cartas. Considerando los sucesos A = "Obtener una espada", B = "Obtener un caballo" y C = "Obtener el rey de oros", calcular la probabilidad del suceso $A \cup B$ y del suceso $A \cup C$.

Ejercicio 3.3

En un estudio sobre un examen de tres partes para estudiantes de un determinado colegio, un psicólogo ha establecido que dichos estudiantes tienen una probabilidad $0'4$ de resolver con éxito la primera parte, $0'6$ de resolver con éxito la segunda habiendo pasado con éxito la primera y $0'85$ de superar con éxito la tercera habiendo resuelto con éxito las dos primeras. ¿Cuál es la probabilidad de que un estudiante, elegido al azar de entre los del mencionado colegio, supere las tres pruebas con éxito?

Ejercicio 3.4

Se extraen al azar dos cartas de una baraja española de 40 cartas. Calcular la probabilidad de sean dos ases, si:

- La primera carta se devolvió al mazo de cartas antes de sacar la segunda.

b) La primera carta no se devolvió al mazo de cartas antes de sacar la segunda.

Ejercicio 3.5

Supongamos que una prueba médica para diagnosticar la presencia de un determinado virus, da positivo en el 99 % de los casos que se aplica a personas que posean el virus, y que da negativo en el 97 % de los casos que se aplica a personas que no lo poseen. Si se cree que la probabilidad de que una persona elegida al azar tenga el virus es 0'05, ¿cuál es la probabilidad de que una persona tenga realmente el virus cuando la prueba le haya dado positiva?

3.12. Lecturas Recomendadas

- Feller W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones*. Volumen I. Editorial Limusa.
- Haigh, J. (2003). *Matemáticas y Juegos de Azar: Jugar con la Probabilidad*. Editorial Tusquets.

Capítulo 4

Modelos Probabilísticos

4.1. Introducción

Como hemos indicado con anterioridad, la situación que el investigador tiene planteada habitualmente es la de analizar una determinada variable X en los individuos de una población. Ésta puede ser *unidimensional*, como por ejemplo el número de pulsaciones por minuto en individuos hipertensos, o la renta anual de los habitantes de la zona euro, o la ansiedad en un grupo de pacientes, etcétera; o puede ser *multidimensional*, si se observan varias variables unidimensionales en cada individuo de la población, como por ejemplo observar el peso y la talla (variable bidimensional), o los ingresos anuales, el número de coches y el número de hijos de una unidad familiar (variable tridimensional).

Más en concreto, el investigador tendrá el propósito de estudiar alguna característica relacionada con dicha variable, como su *media* (renta media, peso medio, etc.) tratando, bien de inferir un valor para dicho *parámetro poblacional*, bien de construir un *intervalo de confianza* para él, o bien tratando de decidir, mediante un *contraste de hipótesis*, entre dos conjuntos de posibles valores del parámetro.

Estas técnicas inferenciales serán estudiadas con detalle en los próximos capítulos, y veremos allí que sus resultados dependerán de la *distribución de probabilidad* o *modelo probabilístico* supuesto como ley que rige el fenómeno aleatorio en estudio. Por eso, en este capítulo estudiaremos un *catálogo de Modelos Probabilísticos* analizando sus principales propiedades y viendo cuáles son los fenómenos aleatorios típicos a los que asociar estos modelos.

Aquí no entraremos a valorar si un modelo determinado se ajusta bien o no a unos datos; eso se tratará en el Capítulo 8.

Antes de pasar a analizar algunos de los Modelos más utilizados en las siguientes secciones, estudiaremos los elementos que permiten caracterizarlos.

4.2. Distribución de Probabilidad

Supongamos una población constituida por 50 millones de individuos. Como estudiamos en el capítulo anterior, la selección aleatoria de los individuos de esta población puede formalizarse, a Nivel I, mediante un espacio probabilístico (Ω, \mathcal{A}, P) en el que el espacio muestral esté constituido por los individuos de la población

$$\Omega = \{\omega_1 = \text{Abad Abad}, \dots, \omega_{50,000,000} = \text{Zurdo Zamora}\}$$

y tal que sobre el conjunto \mathcal{A} de los sucesos esté definida una probabilidad P , de forma que todos los sucesos elementales sean equiprobables: Modelo Uniforme.

Habitualmente estaremos interesados en alguna característica de la población más que en la población misma. Así, es habitual desear conocer el peso medio de la población o la estatura media, etcétera, interesándonos, por tanto, no los individuos ω_i , sino una función suya $X(\omega_i)$ como por ejemplo su peso.

Es decir, habitualmente estaremos interesados no en el espacio probabilístico, sino en una transformación suya (el Nivel II), tal que no sólo nos dé los valores de la característica en estudio para los individuos de la población,

$$X : \Omega \longrightarrow \mathbb{R}$$

sino que conserve la probabilidad P , aglutinando la nueva P_X las probabilidades de los sucesos elementales ω_i a los que corresponda el mismo valor mediante X ,

$$P_X(A) = P\{\omega \in \Omega : X(\omega) \in A\} = P\{X^{-1}(A)\}.$$

Así por ejemplo, si en la población existiesen 20 millones de individuos con un peso entre 60 y 75 kilos, la transformación X debe ser tal que

$$P_X\{[60, 75]\} = P\{\omega \in \Omega : 60 \leq X(\omega) \leq 75\} = \frac{2}{5}.$$

La función X recibe el nombre de *variable aleatoria* y P_X el de su *distribución de probabilidad*.

Evidentemente, sobre un espacio probabilístico es posible definir muchas variables aleatorias. Cuando se consideran a la vez varias de ellas, X_1, \dots, X_p , de forma que en los individuos de la población se observan varios caracteres, queda constituido lo que se denomina una *variable aleatoria multidimensional*, o *vector aleatorio* $X = (X_1, \dots, X_p)$.

Nada impide que los sucesos elementales del espacio muestral Ω sean números reales, por lo que, en ese caso, la aplicación identidad es la variable aleatoria natural a considerar.

En otras ocasiones, aunque los sucesos elementales no sean numéricos, la variable aleatoria a estudiar resulta obligada. Tanto es así que en ocasiones se identifican a los sucesos elementales con los valores de ésta.

Ejemplo 4.1

Consideremos el experimento aleatorio del lanzamiento de un dado. El espacio muestral es

$$\Omega = \left\{ \begin{array}{|c|} \hline \cdot \\ \hline \end{array}, \begin{array}{|c|} \hline \cdot \cdot \\ \hline \end{array}, \begin{array}{|c|} \hline \cdot \cdot \cdot \\ \hline \end{array}, \begin{array}{|c|} \hline \cdot \cdot \cdot \cdot \\ \hline \end{array}, \begin{array}{|c|} \hline \cdot \cdot \cdot \cdot \cdot \\ \hline \end{array}, \begin{array}{|c|} \hline \cdot \cdot \cdot \cdot \cdot \cdot \\ \hline \end{array} \right\}$$

y la probabilidad igual a $1/6$ sobre cada uno de ellos.

La variable aleatoria a considerar de forma natural es $X = \text{número de puntos de la cara superior del dado}$. Es tan evidente la consideración de tal variable que en el Ejemplo 3.2 incluso reemplazamos los sucesos elementales por los valores de dicha variable aleatoria. La distribución de probabilidad de X es, para $x = 1, \dots, 6$,

$$P_X(\{x\}) = P\{\omega : X(\omega) = x\} = \frac{1}{6}$$

Asociada a toda variable aleatoria existe una función $F(x)$, denominada *función de distribución* de X , la cual va midiendo la probabilidad acumulada por X hasta el punto x . Es decir

$$F(x) = P\{\omega \in \Omega : X(\omega) \leq x\}.$$

Esta función tiene la propiedad de caracterizar la distribución de probabilidad de X , P_X . Es decir, a partir de una de ellas se obtiene la otra, siendo habitualmente más cómodo trabajar con la función de distribución.

Ejemplo 4.1 (continuación)

La función de distribución de X será una función en escalera que salta $1/6$ en los valores de la variable,

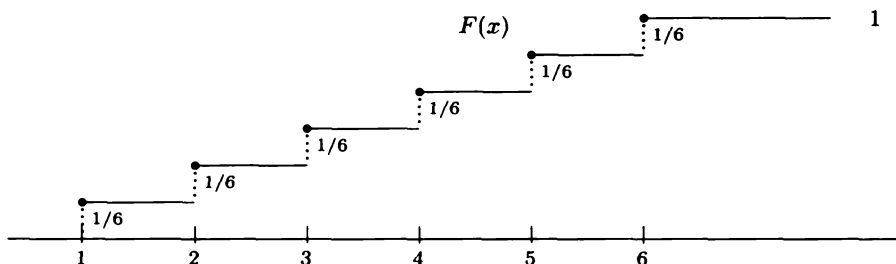


Figura 4.1

Si una variable aleatoria toma valores aislados, como ocurría en el ejemplo anterior, se denomina *discreta*. Si por el contrario puede tomar cualquier valor

de un intervalo, como por ejemplo ocurre con el peso, o la talla, la variable aleatoria recibe el nombre de *continua*. Estos calificativos se aplican también a su distribución, hablando de *distribuciones discretas* o *continuas*.

De la misma definición se deduce que la función de distribución de una variable aleatoria discreta es una función en escalera como la de la Figura 4.1, mientras que la correspondiente a una variable continua es una función continua como la de la Figura 4.2.

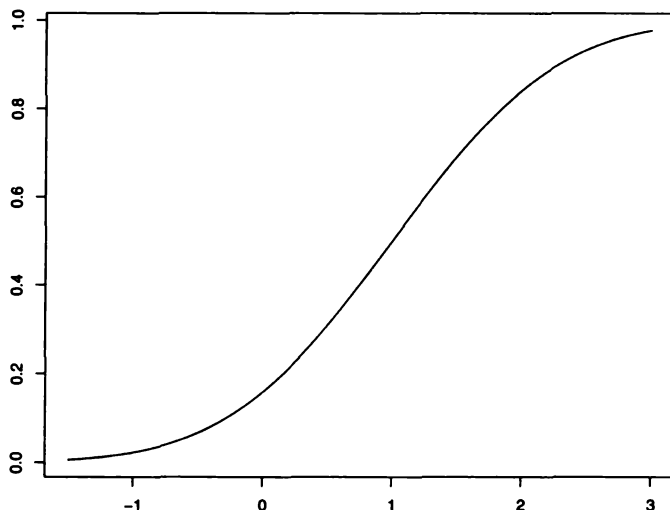


Figura 4.2

A partir de las propiedades de las probabilidades se puede deducir que las funciones de distribución son

1. No decrecientes.
2. Continuas por la derecha.
3. $\lim_{x \rightarrow -\infty} F(x) = 0$ y $\lim_{x \rightarrow \infty} F(x) = 1$.

Las variables aleatorias discretas X , las cuales hemos visto tienen una función de distribución en escalera, tienen asociadas una función, denominada *función de masa*, $p_X(x)$, la cual da la probabilidad de los valores de dicha variable aleatoria; es decir,

$$p_X(x) = P_X(\{x\}) = P\{\omega : X(\omega) = x\}.$$

Por la definición de función de distribución, las funciones de masa y de distribución de una variable aleatoria discreta están relacionadas por las expresiones:

$$p_X(x) = F(x) - F(x-)$$

en donde $F(x-)$ es el límite por la izquierda de F en x . Se ve, por tanto, que la función de masa recoge el valor del salto de la función de distribución, e inversamente,

$$F(x) = \sum_{y \leq x} p_X(y).$$

De manera análoga, las variables aleatorias continuas X tienen asociada una función, denominada *función de densidad*, $f_X(x)$, la cual indica la *velocidad* a la que crece su función de distribución, siendo

$$f_X(x) = \frac{d}{dx} F(x)$$

e inversamente,

$$F(x) = \int_{-\infty}^x f_X(y) dy.$$

lo que implica, por la Propiedad 3 de las funciones de distribución, que sea $\int_{-\infty}^{\infty} f_X(y) dy = F(\infty) = 1$.

(Las denominadas *integrales impropias*, como estas en las que aparecen los símbolos $-\infty$ o $+\infty$, pueden interpretarse como $\int_{-\infty}^x f_X(y) dy = \lim_{b \rightarrow -\infty} \int_b^x f_X(y) dy$. Se han incluido en el texto para completar el concepto que se estudia pero no tendrá que calcular el lector ninguna de ellas.)

Así pues, la distribución de una variable aleatoria se puede caracterizar por su distribución de probabilidad, por su función de distribución, o por su función de masa o densidad (esta última según sea discreta o continua):

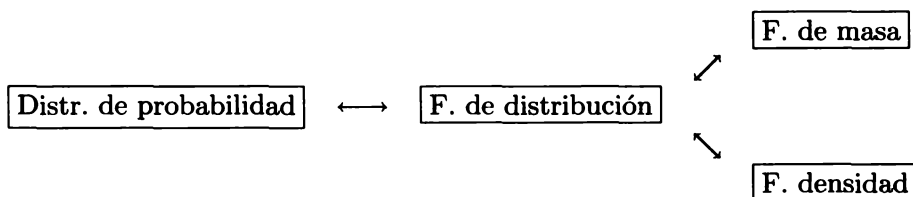


Figura 4.3

Estas funciones de masa y de densidad tienen una interpretación clara a partir, respectivamente, del diagrama de barras y del histograma: son *modelos teóricos* de donde proceden los datos que el investigador examina. De

ahí que se hable de que la variable en estudio tiene una determinada *Distribución de Probabilidad*, o mejor aún, un determinado *Modelo Probabilístico*. Éste habrá que suponerlo con objeto de hacer inferencias sobre X y, como veremos más adelante, si nuestros datos presentan —por ejemplo en el caso continuo— un histograma tal que cuando las bases de los rectángulos que lo forman tienden a cero a medida que la frecuencia total aumenta, la curva resultante se ajusta bien al modelo supuesto, las inferencias que hagamos serán aceptables. En caso contrario deberemos cambiar el modelo.

En las Secciones 4.4 y 4.5 estudiaremos algunos de los modelos más importantes.

Características de una distribución de probabilidad

Dado que los modelos de probabilidad representan, básicamente, un *ideal* de las distribuciones de frecuencias estudiadas en el Capítulo 2, tendrán, al igual que éstas, una medidas de posición, de dispersión, etc. Aquí sólo nos centraremos en una de posición y dos de dispersión, definiéndolas primero para el caso discreto y luego para el continuo.

Dada una variable aleatoria discreta X , con función de masa p_X , llamaremos *media* o *esperanza* de X a la suma de los valores que toma por las probabilidades con que los toma

$$\mu_X = E[X] = \sum_x x p_X(x)$$

y *varianza* de X a

$$\sigma_X^2 = V(X) = \sum_x (x - \mu_X)^2 p_X(x).$$

Dada una variable aleatoria continua X , con función de densidad f_X , llamaremos *media* o *esperanza* de X a la integral

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

y *varianza* de X a

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx.$$

En ambos casos, llamaremos *desviación típica* de X a la raíz cuadrada de la varianza:

$$\sigma_X = D(X) = \sqrt{\sigma_X^2}$$

teniendo estas medidas las misma interpretación que tenían en el Capítulo segundo.

Ejemplo 4.1 (continuación)

Esta distribución de probabilidad es de tipo discreto puesto que toma valores aislados. Como dijimos más arriba, su media será igual a los valores que toma por las probabilidades con que los toma,

$$E[X] = \sum_{x=1}^6 x p_X(x) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3'5.$$

La varianza será igual a los valores que toma la variable, menos la media acabada de calcular al cuadrado, por las probabilidades con que los toma,

$$\begin{aligned} V(X) &= \sum_x (x - \mu_X)^2 p_X(x) = (1 - 3'5)^2 \cdot \frac{1}{6} + (2 - 3'5)^2 \cdot \frac{1}{6} + (3 - 3'5)^2 \cdot \frac{1}{6} + (4 - 3'5)^2 \cdot \frac{1}{6} \\ &+ (5 - 3'5)^2 \cdot \frac{1}{6} + (6 - 3'5)^2 \cdot \frac{1}{6} = 6'25 \cdot \frac{1}{6} + 2'25 \cdot \frac{1}{6} + 0'25 \cdot \frac{1}{6} + 0'25 \cdot \frac{1}{6} + 2'25 \cdot \frac{1}{6} + 6'25 \cdot \frac{1}{6} = \frac{17'5}{6} = 2'92. \end{aligned}$$

Una forma alternativa de calcular la varianza es utilizando la expresión

$$V(X) = E[X^2] - (E[X])^2$$

la cual es válida no sólo para distribuciones discretas sino también para continuas.

La *media de los cuadrados* será, en el caso de distribuciones discretas

$$E[X^2] = \sum_x x^2 p_X(x)$$

es decir, valores que toma la variable, al cuadrado, por las probabilidades con que los toma, y en el caso de distribuciones de tipo continuo con función de densidad f_X , la integral

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx.$$

En el ejemplo que nos ocupa sería

$$E[X^2] = \sum_x x^2 p_X(x) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} = \frac{91}{6} = 15'17$$

con lo que la varianza sería, utilizando la fórmula anterior,

$$V(X) = 15'17 - 3'5^2 = 2'92.$$

La desviación típica sería la raíz cuadrada de la varianza, es decir,

$$\sigma_X = D(X) = \sqrt{2'92} = 1'71.$$

No debe perderse de vista que estas características de la distribución de X no son más que *idealizaciones* de las medidas de posición y dispersión de la distribución de frecuencias de las observaciones (ver Capítulo 2), cuyo histograma/diagrama de barras, de frecuencias relativas sugiere cuál debe ser la distribución de probabilidad de la variable en estudio.

4.2.1. Funciones básicas de R en Probabilidades

Existen cuatro funciones básicas en R que permiten determinar, respectivamente, los valores de la Función de Distribución, la Función de Masa o Densidad (según se trate de una distribución discreta o continua), los Cuantiles y la obtención de Números Aleatorios de determinados Modelos Probabilísticos. En concreto, se trata de las funciones

`pdistribu(x,par)` con la que calculamos el valor de la Función de Distribución del modelo *distribu* en el punto x . Es decir, $F(x)$, siendo F la función de distribución de *distribu*.

`ddistribu(x,par)` con la que calculamos el valor de la función de masa o densidad de la distribución *distribu* en el punto x .

`qdistribu(p,par)` con la que podemos calcular el p -cuantil de la distribución *distribu*. Es decir, $F^{-1}(p)$, siendo F la función de distribución de *distribu*.

`rdistribu(n,par)` mediante la que podemos conseguir n valores obtenidos al azar según el modelo *distribu*.

El segundo (o incluso tercer) argumento utilizado en las cuatro funciones anteriores, `par`, quiere indicar que es ahí en donde deberemos incluir el parámetro o parámetros de la distribución considerada.

Así, si queremos obtener, por ejemplo, el valor de la función de distribución de una distribución normal (que definiremos con precisión un poco más adelante) de parámetros 1 y 2, $N(0, 1)$, en el punto $x = 1.5$, ejecutaríamos (1) obteniendo en (2) el valor $F(x) = 0.5987$.

```
> pnorm(1.5,1,2) (1)
[1] 0.5987063 (2)
```

O, si como es muy habitual en Estadística, queremos determinar el valor de una abscisa de una normal $N(0, 1)$ que deje a la derecha un área de probabilidad 0.025, (es decir, el cuantil 0.975 habitualmente representado por $z_{0.025}$), ejecutaríamos (3) obteniendo el resultado 1.96 en (4)

```
> qnorm(0.975,0,1) (3)
[1] 1.959964 (4)
```

En lugar de *distribu*, en las cuatro funciones de R antes mencionadas podemos utilizar los modelos probabilísticos que estudiaremos en las secciones siguientes.

4.3. Variables aleatorias multivariantes

Hasta ahora se han considerado solamente variables aleatorias unidimensionales, en el sentido de que en cada suceso elemental ω_i del espacio muestral, solamente observábamos una característica $X(\omega_i)$.

Una *variable aleatoria multivariante* (X_1, \dots, X_p) no es más que un vector de variables aleatorias unidimensionales, pudiendo generalizarse los conceptos vistos hasta ahora.

Centrándonos en el caso de una variable aleatoria bidimensional, (X, Y) , los conceptos vistos en el Capítulo 2 para distribuciones de frecuencias bidimensionales, pueden *idealizarse* dando lugar a *distribuciones marginales* de probabilidad, *distribuciones condicionadas* de probabilidad, etc.

Si idealizamos, respectivamente, el diagrama de barras e histograma tridimensionales de la Sección 2.4.1 correspondientes a distribuciones de frecuencias relativas, mediante una *función de masa bidimensional*

$$p_{XY}(x, y) = P\{X = x, Y = y\}$$

y mediante una *función de densidad bidimensional* $f_{XY}(x, y)$, tendremos caracterizada la distribución de probabilidad de una variable aleatoria bidimensional (X, Y) discreta o continua respectivamente, para la cual también tendrán sentido las características poblacionales tales como las medias marginales (por ejemplo la de X)

$$\mu_X = \begin{cases} \sum_x x p_X(x) = \sum_x x \sum_y p_{XY}(x, y) & \text{Caso discreto} \\ \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x, y) dy & \text{Caso continuo} \end{cases}$$

la *covarianza poblacional*

$$\mu_{11} = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p_{XY}(x, y) & \text{Caso discreto} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f_{XY}(x, y) dx dy & \text{Caso continuo} \end{cases}$$

y el *coeficiente de correlación poblacional*, ρ , definido por

$$\rho = \frac{\mu_{11}}{\sigma_X \sigma_Y}$$

en donde σ_X y σ_Y son las desviaciones típicas marginales.

(Observemos que las *integrales dobles* que han aparecido no son más que las habituales integrales, calculadas primero respecto a una de las variables y luego respecto de la otra, pero no se preocupe el lector que no tendrá que calcular ninguna en el texto.)

Independencia de variables aleatorias

Un concepto de gran interés es el de independencia de variables aleatorias. Éste se corresponde con la idea de observaciones independientes. Así, si anotamos el número x_1 que obtenemos al lanzar un dado, el valor x_2 obtenido al volver a lanzarlo es *independiente* de obtenido en la primera tirada. Esta idea se traduce en que la probabilidad de obtener el par (x_1, x_2) será $P\{(X_1, X_2) = (x_1, x_2)\} = P\{X_1 = x_1\} \cdot P\{X_2 = x_2\} = 1/36$.

Esta propiedad es la que utilizamos para caracterizar a las variables aleatorias independientes, a través de sus densidades o masas. Diremos que las variables aleatorias discretas $\{X_1, \dots, X_n\}$ son *independientes*, si y sólo si la función de masa conjunta es el producto de las funciones de masa marginales,

$$p_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \prod_{i=1}^n p_{X_i}(x_i).$$

Análogamente, diremos que las variables aleatorias continuas $\{X_1, \dots, X_n\}$ son *independientes*, si y sólo si la función de densidad conjunta es el producto de las marginales,

$$f_{(X_1, \dots, X_n)}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i).$$

4.4. Modelos unidimensionales discretos

En esta sección estudiaremos las características más importantes de los modelos probabilísticos discretos más destacados.

4.4.1. Distribución Binomial

Esta distribución modeliza el *número de éxitos* en experimentos denominados, de forma genérica, *Pruebas de Bernoulli*. Estas pruebas consisten en la realización de ensayos repetidos e independientes, existiendo en cada ensayo solamente dos resultados posibles —denominados de forma genérica *éxito* y *fracaso*— y manteniéndose constante la probabilidad éxito a lo largo de los ensayos. Por ejemplo, si se está estudiando el Sexo en el nacimiento, se puede denominar *éxito* a nacer Mujer y *fracaso* a nacer Hombre. Cada nacimiento puede modelizarse como una Prueba de Bernoulli y el número de hembras en

un grupo determinado en bebés recién nacidos sería la variable a modelizar con una *Distribución Binomial*.

Si el número de pruebas de Bernoulli que se realizan es n (n nacimientos) y la probabilidad de éxito en cada una de ellas es p (probabilidad de nacer hembra), la variable de interés X es el *número de éxitos en las n pruebas*, siendo la función de masa de esta distribución,

$$p_X(x) = P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x} \quad , \quad x = 0, 1, \dots, n$$

en donde deben ser $n \geq 1$ y $0 < p < 1$. En este caso diremos que X sigue una distribución (o tiene un modelo) binomial de parámetros n y p y lo representaremos por $X \sim B(n, p)$.

Su media y su varianza son, respectivamente,

$$E[X] = np \quad \quad V(X) = np(1-p).$$

Ejemplo 4.2

Supongamos que en una determinada población la probabilidad de que un bebé recién nacido sea niña es 0'48.

El estudio de la variable *número de niñas* de entre n bebés recién nacidos en esa población, puede modelizarse mediante una distribución binomial, en la que cada nacimiento sea considerado como una prueba de Bernoulli y el suceso *éxito* sea *nacer niña*, es decir, $X \sim B(n, 0'48)$.

De esta forma, admitiendo que la probabilidad p de nacer niña se mantiene constante de un nacimiento a otro, se podrán determinar fácilmente probabilidades como la de que en 5 nacimientos de esa población elegidos al azar haya 2 niñas; ésta probabilidad será

$$P\{X = 2\} = p_X(2) = \binom{5}{2} (0'48)^2 (0'52)^3 = 0'324.$$

El cálculo de probabilidades asociadas a distribuciones binomiales se hace hoy en día con R, como veremos a continuación. No obstante, en ADD, *Tabla 1*, vienen recogidos los valores de la función de masa para algunos valores del parámetro p .

Obsérvese que los valores de p mayores que 0'5 nunca aparecerán, ya que en ese caso, por las propiedades de los números combinatorios, si representamos aquí por P_p la función de masa de una $B(n, p)$, al ser

$$P_p\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{n-x} (1-p)^{n-x} p^x = P_{1-p}\{X = n-x\}$$

se calcula la función de masa en $n-x$ de una binomial $B(n, 1-p)$, siendo $1-p < 0'5$.

Así por ejemplo, si queremos calcular la probabilidad $P\{X = 2\}$ con $X \sim B(6, 0'55)$, será

$$P_{0'55}\{X = 2\} = P_{0'45}\{X = 4\} = 0'1861.$$

El caso particular de que sólo se considere una prueba de Bernoulli, es decir, de que sea $X \sim B(1, p)$ recibe el nombre de *Distribución de Bernoulli*.

Para calcular probabilidades binomiales con R, debemos utilizar el comando `binom`. Así,

```
> pbinom(x,n,p) # valor de la función de distribución en x de la binomial(n,p)
> dbinom(x,n,p) # valor de la función de masa en x de la binomial(n,p)
> qbinom(q,n,p) # cuantil de orden q de la binomial(n,p)
# (hasta él la probabilidad acumulada es q)
> rbinom(m,n,p) # muestra aleatoria de tamaño m de la binomial(n,p)
```

El uso de estas funciones nos permite, no sólo el calcular probabilidades, sino también dibujar las funciones de masa o distribución.

Por ejemplo, si queremos representar la función de masa de una distribución binomial $B(5, 0'33)$ ejecutaríamos la siguiente secuencia para obtener la Figura 4.4. Destacamos que en el argumento `type` de la función `plot` hemos utilizado la opción `h` para obtener un gráfico del tipo histograma.

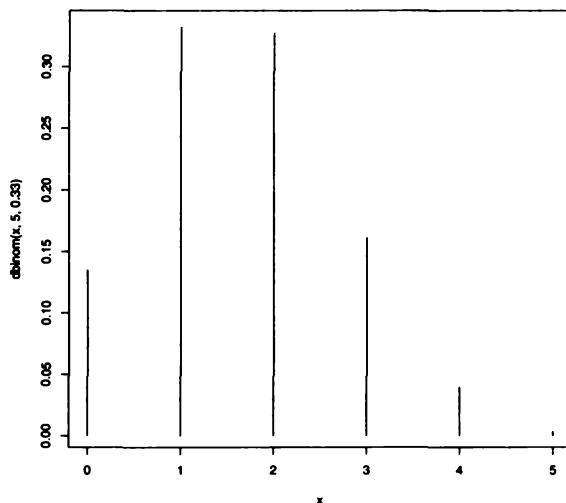


Figura 4.4: Función de masa de la $B(5, 0'33)$

```
> x<-seq(0,5,len=6)
> plot(x,dbinom(x,5,0.33),type="h")
```

Si queremos representar la función de distribución de la $B(5, 0.33)$ ejecutaríamos la siguiente secuencia obteniendo la Figura 4.5, en donde la **s** en el argumento **type** de la función **plot** es minúscula, no mayúscula. Aunque los trazos verticales sobran en el dibujo, éste es muy interesante.

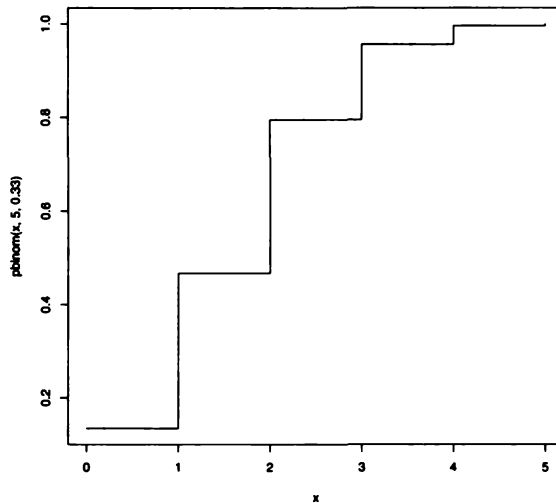


Figura 4.5: Función de distribución de la $B(5, 0.33)$

```
> x<-seq(0,5,len=6)
> plot(x,pbinom(x,5,0.33),type="s")
```

4.4.2. Distribución de Poisson

Esta distribución fue introducida por Simeon Denis Poisson en 1837 y se utiliza, por lo general, para modelizar el número de veces (0, 1, 2, ...) que ocurren sucesos raros, como por ejemplo, el número de incendios por año en una compañía de seguros, o el número de suicidios por año, o el número de nacimientos múltiples por año. Ladislaus von Bortkiewicz la utilizó en 1898 para modelizar el número anual de muertes por coces de caballo en el ejército prusiano. En 1910, Rutherford y Geiger la consideraron como modelo para el número de partículas emitidas por una sustancia radiactiva. Durante la Segunda Guerra Mundial, el ejército británico la utilizó (véase el Ejemplo 4.3) para demostrar que los alemanes lanzaban las bombas al azar y no a un

objetivo determinado. La función de masa de la variable *número de éxitos* de esta distribución es

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

siendo $\lambda > 0$ el único parámetro del que depende esta distribución. Si una variable aleatoria X tiene como modelo probabilístico una *Distribución Poisson* de parámetro λ , lo expresaremos poniendo $X \sim \mathcal{P}(\lambda)$.

De hecho, este parámetro es su media y su varianza, $E[X] = \lambda = V(X)$, coincidencia nada habitual en los modelos probabilísticos y que puede servir como indicación de que una variable puede ser modelizada por esta distribución.

Al igual que pasaba con la distribución binomial, existen tablas de la distribución de Poisson que nos dan su función de masa para distintos valores del parámetro. En ADD, *Tabla 2*, aparecen los más utilizados.

Aproximación de la distribución binomial por la de Poisson

Un hecho interesante es que la función de masa de la distribución binomial $B(n, p)$ converge, cuando n crece, a la función de masa de una distribución de Poisson(λ) con $\lambda = np$.

Por tanto, cuando queramos calcular probabilidades binomiales con n grande, podremos utilizar las tablas de la distribución de Poisson.

Esta aproximación es buena cuando p es muy pequeño con respecto a np y a su vez np es también muy pequeño con respecto a n . Como indicación de estas cantidades, se toma como buena la aproximación cuando al menos es $np < 5$ y $p < 0.1$. El hecho de que la distribución de Poisson se obtenga como límite de unas pruebas de Bernoulli cuando p es muy pequeño, es la razón por la que se asocia a la distribución de Poisson con un modelo para *sucesos raros*.

Ejemplo 4.3

Una situación habitual en la que se aplica la distribución de Poisson es la de un número n muy grande de *celdas* (por ejemplo cuadrículas en las que se divide una zona geográfica) de manera que en cada celda es lo suficientemente pequeña para que haya un *éxito* o ninguno, con probabilidad p de *éxito* pequeña.

Pues bien, suponiendo que los *éxitos* se reparten al azar sobre las cuadrículas (hipótesis más delicada de admitir y que determinará una mayor adecuación del modelo a la realidad cuanto más verosímil sea) la variable aleatoria X , *número de éxitos de la cuadrícula*, seguirá una distribución de Poisson $\mathcal{P}(np)$. De hecho, una binomial $B(n, p)$ aproximada por una Poisson(np).

Esto es lo que ocurrió durante la Segunda Guerra Mundial cuando el ejército alemán bombardeaba intensamente Londres desde Calais con las bombas V2. El área central de Londres de 144 km² fue dividida en 576 cuadrados de 0.25 km² cada uno. Como habían caído en total 537 bombas V2 en esa zona, el número medio observado de bombas por cuadrado era de $537/576 = 0.932$. Si los alemanes disparaban al azar ("siguiendo una distribución de Poisson"), no habría mucha diferencia entre las bombas observadas en cada cuadrado y las

esperadas según una distribución de Poisson(0'932). Los valores observados y los esperados (si fuera cierta esa distribución de Poisson) vienen recogidos en la siguiente tabla (la última frecuencia viene dada porque en una cuadrícula cayeron 7 bombas)

Número de V2 en un cuadrado	Frecuencia absoluta observada (cuadrículas con x impactos)	Frecuencia absoluta esperada $n \cdot P_\lambda(X = x)$ $576 \cdot P_{0'932}(X = x)$
x	n_i	
0	229	226'81
1	211	211'39
2	93	98'50
3	35	30'60
4	7	7'13
5 ó más	1	1'57
	576	576

Aunque para poder concluirlo adecuadamente debemos utilizar las herramientas inferenciales que se estudian en el Capítulo 8, ya podemos apreciar que no parece haber mucha diferencia entre ambas columnas y concluir que los alemanes no *apuntaban* sino que las bombas caían al azar (eso sí, siguiendo un modelo probabilístico Poisson(0'932)). (¿O tenía quizá el ejército alemán un sistema tan sofisticado de lanzamiento de misiles que le permitía enviar las bombas según el modelo probabilístico que eligiera?)

Para ejecutar con R aplicaciones de esta distribución, el comando que debemos utilizar es `pois`. Así

```
> ppois(x,a)  # valor de la función de distribución en x de la Poisson(a)
> dpois(x,a)  # valor de la función de masa en x de la Poisson(a)
> qpois(p,a)  # cuantil de orden p de la Poisson(a)
> rpois(n,a)  # muestra aleatoria de tamaño n de la Poisson(a)
```

4.4.3. Distribución Geométrica

La *Distribución Geométrica* de parámetro p es un modelo que se asocia también con pruebas de Bernoulli, es decir, del tipo éxito/fracaso con probabilidad de éxito p aunque modelizando ahora la variable *número de fallos antes del primer éxito*. Su función de masa es

$$p_X(x) = (1 - p)^x p, \quad x = 0, 1, 2, \dots$$

en donde debe ser $0 < p \leq 1$. Se expresa con $X \sim \text{Geom}(p)$.

Su media y su varianza son respectivamente

$$E[X] = \frac{1 - p}{p} \quad V(X) = \frac{1 - p}{p^2}.$$

Ejemplo 4.2 (continuación)

El ejemplo más habitual que se modeliza con esta distribución es el del *número de niños antes de tener la primera niña* (y su análogo con *niño*). Si se supone de nuevo que la probabilidad de nacer niña es 0'48, la probabilidad de, por ejemplo, tener 3 niños antes de nacer la primera niña será

$$p_X(3) = 0'52^3 \cdot 0'48 = 0'06749.$$

Advertimos para finalizar que, en algunas ocasiones, se modeliza el número de pruebas necesarias para que aparezca el primer éxito en lugar del número de fallos, pero ambos esquemas son equivalentes.

El comando a utilizar en R es `geom`. Así

```
> pgeom(x,p) # función de distribución en x de la geométrica(p)
> dgeom(x,p) # función de masa en x de la geométrica(p)
> qgeom(q,p) # cuantil de orden q de la geométrica(p)
> rgeom(n,p) # muestra aleatoria de tamaño n de la geométrica(p)
```

4.4.4. Distribución Hipergeométrica

Este modelo, *Distribución Hipergeométrica*, se utiliza para situaciones que se adaptan al siguiente esquema: Se supone una caja con N piezas de las cuales D son defectuosas y $N - D$ no defectuosas. Se extraen sin reemplazamiento n piezas (o las n de una vez) de la caja y estamos interesados en modelizar el *número de defectuosas extraídas en las n seleccionadas*.

El cálculo de probabilidades nos da el valor de esta probabilidad que será la función de masa de este modelo,

$$p_X(x) = \frac{\binom{D}{x} \binom{N-D}{n-x}}{\binom{N}{n}} \quad \text{máx}\{0, n - N + D\} \leq x \leq \text{mín}\{n, D\}.$$

El que una variable X siga esta distribución lo expresaremos como $X \sim \text{Hiper}(D, N, n)$.

Su media y su varianza son respectivamente

$$E[X] = \frac{Dn}{N} \quad V(X) = \frac{D(N-D)n(N-n)}{N^2(N-1)}.$$

Ejemplo 4.4

Un fábrica que produce lotes de 20.000 tornillos cree que en su producción hay un 0'1 % de defectuosos en cada lote. Para conseguir un certificado de calidad, la fábrica debe pasar

un estricto control de calidad en el que se extraen 20 tornillos de un lote, concediéndose el certificado si, como mucho, aparece un tornillo defectuoso entre los seleccionados.

Antes de solicitar el certificado de calidad, se quiere calcular la probabilidad de pasar dicho test, por lo que el estadístico de la empresa modeliza la variable $X = \text{número de tornillos defectuosos en cada lote examinado}$, con el modelo $X \rightsquigarrow \text{Hiper}(20, 20000, 20)$ por ser $p = D/N$, es decir, $0'001 = D/20000$ es decir, $D = 20$.

La probabilidad de no pasar el test será

$$P\{X > 1\} = 1 - P\{X \leq 1\} = 0'000178$$

tan pequeña que decide solicitar el certificado.

(Observamos cómo la probabilidad es una medida de la incertidumbre, la cual nos puede servir, como en este caso, para tomar decisiones pero advertimos que, cuando se haga el test de certificación, o en general, se realice el experimento aleatorio, saldrá un tornillo defectuoso o ninguno o muchos. La probabilidad es una guía para, antes de seleccionar la muestra, analizar —medir— lo que parece puede ocurrir. Una vez que se extraiga la muestra ya no tendrá sentido hablar de la probabilidad de algo porque ya no habrá nada aleatorio de lo que hablar. Lo que tenga que ocurrir habrá ocurrido.)

El comando a utilizar en R es `hyper`. Así

```
> phyper(x,D,N-D,n)  # función de distribución en x de la hipergeométrica(D,N,n)
> dhyper(x,D,N-D,n)  # función de masa en x de la hipergeométrica(D,N,n)
> qhyper(p,D,N-D,n)  # cuantil de orden p de la hipergeométrica(D,N,n)
> rhyper(n,D,N-D,n)  # muestra aleatoria de tamaño n de la hipergeométrica(D,N,n)
```

4.4.5. Distribución Binomial Negativa

Esta distribución de probabilidad es una generalización de la distribución geométrica antes estudiada. Con una *Distribución Binomial Negativa* modelizamos de nuevo un experimento de Bernoulli, del tipo éxito/fracaso con probabilidad de éxito p , pero analizando ahora la variable $X = \text{número de fallos antes del éxito } n\text{-ésimo}$. La función de masa de este modelo, expresado de la forma $X \rightsquigarrow BN(n, p)$, será

$$p_X(x) = \binom{n+x-1}{n-1} (1-p)^x p^n \quad x = 0, 1, 2, \dots$$

en donde debe ser $n > 0$ y $0 < p \leq 1$.

Ejemplo 4.2 (continuación)

Si queremos ahora calcular la probabilidad de tener tres niños antes de la segunda niña, el modelo adecuado será el de una $BN(2, 0'48)$ siendo la probabilidad buscada,

$$p_X(3) = \binom{2+3-1}{2-1} (1-0'48)^3 0'48^2 = 4 \cdot 0'1406 \cdot 0'2304 = 0'1296.$$

El comando a utilizar en R es `nbinom`. Así

```
> pnbinom(x,n,p)  # función de distribución en x de la binomial negativa(n,p)
> dnbinom(x,n,p)  # función de masa en x de la binomial negativa(n,p)
> qnbinom(q,n,p)  # cuantil de orden q de la binomial negativa(n,p)
> rnbinom(n,n,p)  # muestra aleatoria de tamaño n de la binomial negativa(n,p)
```

4.5. Modelos unidimensionales continuos

Estudiaremos algunos modelos continuos. En García Pérez (2010) se estudian otros específicos del Área de la Salud.

4.5.1. Distribución Normal

Esta distribución fue propuesta por primera vez como modelo probabilístico por De Moivre en 1733, obteniéndola como límite de la distribución binomial. Aunque Laplace la obtuvo también en 1774 como límite de la distribución hipergeométrica, la referencia más utilizada en relación con la distribución que nos ocupa es la de Laplace (1814) y Gauss (1809) en donde la utilizaron en el análisis de los errores en Astronomía y Geodesia.

Su nombre se debe, sin embargo, a Quetelet y es curioso resaltar que la gran importancia de esta distribución como modelo probabilístico se debía, en gran medida, a que los matemáticos pensaban que este modelo era el habitual (es decir, el *normal*) para explicar los fenómenos naturales porque pensaban era un hecho comprobado experimentalmente pero, por otro lado, los usuarios de la Estadística, creían que era un teorema matemático y que por eso, debían suponerla como modelo habitual. Hoy en día se considera un modelo probabilístico más aunque gran parte de la Estadística está desarrollada suponiendo esta distribución como modelo, siendo los Métodos Robustos (véase el texto MR) una buena alternativa para hacer inferencias sin esta suposición.

La *Distribución Normal* se define como aquella distribución cuya función de densidad es

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \quad , \quad -\infty < x < \infty$$

donde debe ser $\sigma > 0$, de representación gráfica la Figura 4.6 para el caso $\mu = 0$ y $\sigma = 1$.

Como se ve en la expresión de la función de densidad, la distribución normal depende de dos parámetros, μ y σ , que la caracterizan, por lo que se expresa de la forma $X \sim N(\mu, \sigma)$.

Se puede demostrar que si $X \sim N(\mu, \sigma)$ entonces es $E[X] = \mu$ y $V(X) = \sigma^2$.

El caso de la distribución normal de media $\mu = 0$ y desviación típica $\sigma = 1$, es de singular importancia; de ahí que, en ocasiones, a la $N(0, 1)$ se la denomine *normal estándar*.

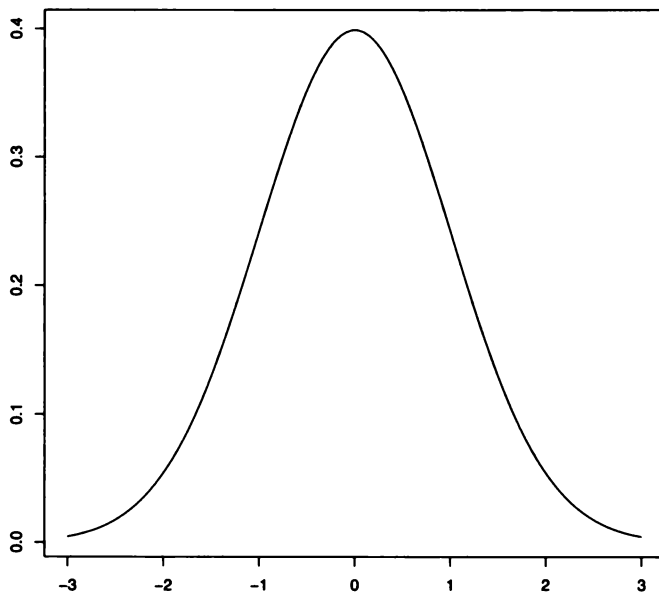


Figura 4.6: Función de densidad de la $N(0, 1)$

La relación existente entre una variable $X \sim N(\mu, \sigma)$ y una $Z \sim N(0, 1)$ es muy sencilla:

$$Z = \frac{X - \mu}{\sigma} \quad \text{o bien} \quad X = \mu + \sigma Z$$

El paso de una $N(\mu, \sigma)$ a una $N(0, 1)$ se denomina *tipificación*, y es muy importante en Estadística ya que la función de distribución de una normal —la cual nos permite calcular probabilidades de áreas bajo la curva normal— no admite una expresión explícita, sino que es necesario acudir a una tablas calculadas a tal efecto.

No obstante, por el proceso de tipificación antes mencionado, no será preciso dar tablas de todas las normales, sino solamente de la $N(0, 1)$. Éste es el contenido de la *Tabla 3* de ADD, la cual va dando las probabilidades cola de la normal estándar.

La simetría de la distribución normal respecto su media y el que el área bajo la curva normal sea 1, permite calcular todo tipo de áreas bajo dicha curva. Aconsejamos hacer un dibujo en el se represente el área que queremos calcular, con objeto de facilitar los cálculos.

Ejemplo 4.5

Si $Z \sim N(0, 1)$, buscando adecuadamente en la *Tabla 3* de ADD obtenemos los siguientes valores:

$$P\{Z < 2'03\} = 1 - P\{Z > 2'03\} = 1 - 0'0212 = 0'9788.$$

$$P\{Z < -0'3\} = P\{Z > 0'3\} = 0'3821.$$

$$P\{Z > -1'39\} = 1 - P\{Z < -1'39\} = 1 - P\{Z > 1'39\} = 1 - 0'0823 = 0'9177.$$

$$P\{-1'2 < Z < 1'05\} = P\{Z < 1'05\} - P\{Z < -1'2\} = P\{Z < 1'05\} - P\{Z > 1'2\} = \\ = 1 - 0'1469 - 0'1151 = 0'738.$$

$$P\{1'68 < Z < 3'36\} = P\{Z > 1'68\} - P\{Z > 3'36\} = 0'0461.$$

$$P\{-1'2 < Z < -0'03\} = P\{0'03 < Z < 1'2\} = 0'3729.$$

Ejemplo 4.6

Si $X \sim N(3, 2)$, las probabilidades correspondientes a esta distribución se determinan, primero, tipificando y después por la búsqueda de la probabilidad tipificada en la *Tabla 3* de ADD. Así por ejemplo,

$$P\{X < 1'5\} = P\{(X - \mu)/\sigma < (1'5 - 3)/2\} = P\{Z < -0'75\} = 0'2266.$$

Ejemplo 4.7

En ocasiones, la probabilidad que queremos buscar no está en la *Tabla 3*. De hecho esta tabla no es más que un breve resumen de la distribución normal estándar.

Esto ocurre especialmente si queremos utilizarla al revés, tratando de determinar la abscisa correspondiente a una probabilidad. No obstante, podemos calcular, aproximadamente, estas probabilidades intermedias mediante una interpolación lineal con la cual, de hecho, aproximamos por una recta el trozo de curva que no conocemos, con el consiguiente error, que podrá ser mayor cuanto mayor sean los valores entre los que se interpola.

Así, si queremos conocer el z tal que $P\{Z > z\} = 0'01$, al ser $P\{Z > 2'32\} = 0'0102$ y $P\{Z > 2'33\} = 0'0099$, será $z = 2'326667$.

Digamos por último en relación con esta distribución, que si es $X \sim N(\mu, \sigma)$, el coeficiente de asimetría es $E[(X - \mu)^3]/\sigma^3 = 0$ y el de apuntamiento o curtosis es $E[(X - \mu)^4]/\sigma^4 = 3$, valores que, calculados en una muestra, permiten un rápido análisis de si los datos se distribuyen como una normal. Más bien al revés: si no se obtienen valores muestrales cercanos a 0 y 3 respectivamente, los datos no van a seguir una distribución normal.

El comando a utilizar en R es `norm`. Así

```
> pnorm(x,a,b) # función de distribución en x de la normal(a,b)
```

```

> dnorm(x,a,b)  # función de densidad en x de la normal(a,b)
> qnorm(p,a,b)  # cuantil de orden p de la normal(a,b)
> rnorm(n,a,b)  # muestra aleatoria de tamaño n de la normal(a,b)

```

Si no indicamos nada, R toma por defecto los valores $a = 0$, $b = 1$.

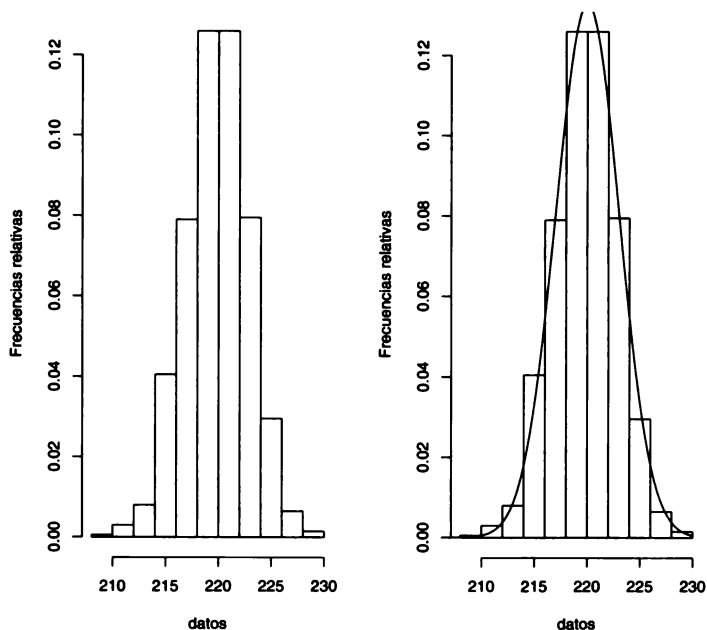


Figura 4.7: Histograma con la función de densidad de la $N(220,3)$

Además del cálculo de probabilidades de distribuciones de probabilidad, otra de las utilidades de R es la de permitir analizar visualmente si unos datos parecen seguir una determinada distribución de probabilidad. Así, el gráfico de la izquierda de la Figura 4.7 es un histograma de las frecuencias relativas de la concentración, en partes por millón, de bifenil policlorado PCB (un agente contaminante industrial) en huevos de pelícano elegidos al azar en la isla californiana de Anacapa, situada frente a la ciudad de Los Ángeles.

Se cree que esta variable puede seguir una distribución normal $N(220, 3)$, en base a estudios anteriores. Para comprobar gráficamente si esta afirmación parece razonable, ejecutaríamos la siguiente secuencia de comandos, para superponer la función de densidad de esa normal sobre el histograma anterior, obteniendo el gráfico de la derecha de la Figura 4.7.

```
> z<-seq(200,230,len=100)
> lines(z,dnorm(z,220,3),type="l")
```

Del gráfico se puede deducir que es muy adecuado suponer este modelo para la variable en estudio.

4.5.2. Distribución Uniforme

La *Distribución Uniforme*, de parámetros (a, b) , es un modelo que asigna, de forma continua, igual probabilidad a todas las partes del intervalo (a, b) en el que está definida. Su función de densidad es

$$f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b$$

y lo representaremos por $X \rightsquigarrow Unif(a, b)$.

Su media y su varianza son, respectivamente,

$$E[X] = \frac{a+b}{2} \quad V(X) = \frac{(b-a)^2}{12}.$$

El comando a utilizar en R es `unif`. Así

```
> punif(x,a,b) # función de distribución en x de la uniforme(a,b)
> dunif(x,a,b) # función de densidad en x de la uniforme(a,b)
> qunif(p,a,b) # cuantil de orden p de la uniforme(a,b)
> runif(n,a,b) # muestra aleatoria de tamaño n de la uniforme(a,b)
```

Si no indicamos nada, R toma por defecto los valores $a = 0$, $b = 1$.

4.5.3. Distribución Beta

La *Distribución Beta* de parámetros (a, b) tiene por función de densidad,

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

en donde Γ es la función $\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt$, definida para $p > 0$.

Deben ser $a > 0$ y $b > 0$ y su media y su varianza son, respectivamente,

$$E[X] = \frac{a}{a+b} \quad V(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

El comando a utilizar en R es `beta`. Así

```
> pbeta(x,a,b) # función de distribución en x de la beta(a,b)
> dbeta(x,a,b) # función de densidad en x de la beta(a,b)
> qbeta(p,a,b) # cuantil de orden p de la beta(a,b)
> rbeta(n,a,b) # muestra aleatoria de tamaño n de la beta(a,b)
```

4.5.4. Distribuciones Gamma y Exponencial

La *Distribución Gamma* de parámetros (a, b) tiene por función de densidad,

$$f_X(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b} \quad , \quad x > 0$$

en donde deben ser $a > 0$ y $b > 0$. Su media y su varianza son, respectivamente,

$$E[X] = ab \quad V(X) = ab^2.$$

El caso particular de $a = 1$ se denomina *Distribución Exponencial* y, tanto esta distribución como su generalización, la distribución gamma, son distribuciones muy habituales en la modelización de tiempos de supervivencia; es decir, de tiempos que transcurren hasta que un determinado suceso, como la muerte de un paciente o el fallo de una componente electrónica, acontece.

El comando a utilizar para calcular probabilidades con R es `gamma`. Así

```
> pgamma(x,a,b)  # función de distribución en x de la gamma(a,b)
> dgamma(x,a,b)  # función de densidad en x de la gamma(a,b)
> qgamma(p,a,b)  # cuantil de orden p de la gamma(a,b)
> rgamma(n,a,b)  # muestra aleatoria de tamaño n de la gamma(a,b)
```

Si no especificamos el valor del parámetro, R toma $b = 1$.

4.5.5. Distribución de Cauchy

La *Distribución de Cauchy* de parámetros (a, b) tiene por función de densidad,

$$f_X(x) = \frac{b}{\pi} \frac{1}{b^2 + (x - a)^2} \quad , \quad -\infty < x < \infty$$

en donde debe ser $b > 0$, mientras que a puede ser cualquier número real.

El comando a utilizar en R es `cauchy`. Así

```
> pcauchy(x,a,b)  # función de distribución en x de la Cauchy(a,b)
> dcauchy(x,a,b)  # función de densidad en x de la Cauchy(a,b)
> qcauchy(p,a,b)  # cuantil de orden p de la Cauchy(a,b)
> rcauchy(n,a,b)  # muestra aleatoria de tamaño n de la Cauchy(a,b)
```

Si no le damos valores a los parámetros, R toma por defecto, $a = 0$ y $b = 1$.

4.6. Modelos bidimensionales

En esta sección sólo consideraremos la distribución normal bivalente, la cual es de tipo continuo, y que nos da idea del caso multivariante.

4.6.1. Distribución Normal bivalente

Una variable aleatoria bidimensional (X, Y) se dice que sigue una distribución normal bivalente de medias (μ_1, μ_2) y de varianzas-covarianzas $(\sigma_1^2, \sigma_2^2, \mu_{11})$, si su función de densidad es

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot \exp \left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

si $-\infty < x < \infty$, $-\infty < y < \infty$.

En este caso, las medias y varianzas marginales son $E[X] = \mu_1$, $V(X) = \sigma_1^2$, $E[Y] = \mu_2$, $V(Y) = \sigma_2^2$ y la covarianza $\mu_{11} = \rho\sigma_1\sigma_2$.

Un caso especialmente interesante es el de $\mu_1 = \mu_2 = 0$ y $\sigma_1 = \sigma_2 = 1$, cuya representación gráfica es la Figura 4.8.

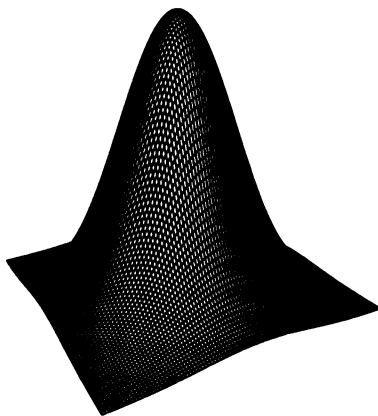


Figura 4.8

4.7. Teorema Central del Límite

Como dijimos al estudiar la distribución normal, este modelo recibió esta denominación porque en el siglo XIX se creía que era el modelo habitual de la mayoría de los fenómenos de la naturaleza. Esta idea, no obstante, tiene su fundamento en el siguiente resultado:

Si X_1, X_2, \dots es una sucesión de variables aleatorias independientes, idénticamente distribuidas y con varianza común σ^2 finita, entonces la variable aleatoria

$$\frac{X_1 + X_2 + \dots + X_n - n\mu}{\sqrt{n}\sigma}$$

en donde μ es la media común, tiene como distribución asintótica una $N(0, 1)$.

Es decir, que si tenemos un gran número de observaciones independientes X_1, X_2, \dots , sea cual sea su distribución común (mientras tenga varianza finita, situación habitual en los modelos considerados), para n suficientemente grande podemos aproximar la distribución de la variable aleatoria anterior por una $N(0, 1)$, calculando probabilidades de esta variable mediante las tablas de la normal estándar.

Obsérvese que dividiendo por n en la expresión anterior, podemos expresar el resultando diciendo que para n grande es

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

siendo $\bar{x} = (X_1 + \dots + X_n)/n$ la media aritmética de las observaciones (que más adelante denominaremos *media muestral*). Es decir, que la distribución de \bar{x} es, aproximadamente,

$$\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

aunque los datos no procedan de distribuciones normales.

Existen otros teoremas, también conocidos como del límite, que dan otras condiciones para conseguir la convergencia, aunque su análisis se sale de los objetivos de este libro.

Una aplicación directa de dicho teorema es la aproximación de una distribución binomial $X \rightsquigarrow B(n, p)$ por una normal,

$$X \approx N\left(np, \sqrt{np(1-p)}\right).$$

Esta aproximación se admite como válida cuando $np > 5$ si es que $p \leq 0'5$, y cuando $n(1-p) > 5$ si es que $p > 0'5$.

Una de las grandes posibilidades que tiene R es la de permitirnos *experimentar* con las distribuciones de probabilidad, pudiendo por ejemplo, hacer experimentos con el Teorema Central del Límite.

Ejemplo 4.8

Acabamos de decir que una binomial $B(n, p)$ se puede aproximar cuando n es grande, en determinadas condiciones, por una $N(np, \sqrt{np(1-p)})$. En concreto, esta aproximación se

admite como buena cuando sea $np > 5$, si es que es $p \leq 0.5$, y cuando sea $n(1 - p) > 5$ si es que es $p > 0.5$. Podemos experimentar con R estas dos posibilidades (cuando n es grande), ejecutando, por ejemplo, (1), con lo que obtenemos una muestra de tamaño 1000 de una binomial $B(50, 0.5)$. En (2) representamos el histograma (en color) de estos datos. En (3) construimos un vector de abscisas, representando con (4), para estas abscisas, la densidad normal sobre el histograma anterior. Obtenemos así la Figura 4.9.

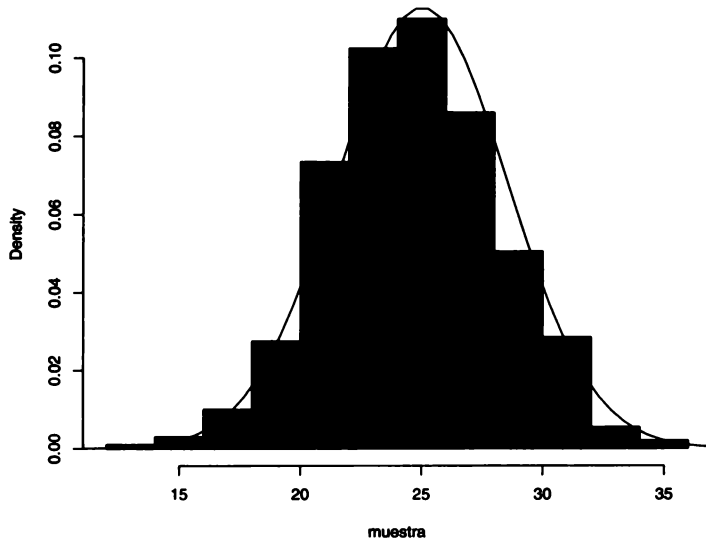


Figura 4.9: Histograma de datos binomiales con aproximación normal

Varias observaciones: primero que, cuando el lector replique (1) obtendrá muestras distintas cada vez, puesto que, en cada ocasión, estará seleccionando 1000 datos al azar de una binomial $B(50, 0.5)$, por lo que su gráfico puede diferenciarse algo de nuestra Figura 4.9. Segundo, la normal por la que aproximamos será la de media $np = 50 \cdot 0.5 = 25$ y desviación típica $\sqrt{np(1 - p)} = \sqrt{50 \cdot 0.5 \cdot 0.5} = 3.535$; si cambiamos los parámetros de la binomial, deberán cambiarse también los parámetros de la normal a los que se aproxima bien o mal la binomial. Esto sucederá en el otro experimento que sigue a éste. Tercero, cada normal que sirve de aproximación tendrá unos valores de abscisas en los que moverse, de manera que la elección de abscisas establecida en (3) debe revisarse si cambiamos los parámetros de la binomial a aproximar.

```
> muestra<-rbinom(1000,50,0.5) (1)
```

```
> hist(muestra,prob="T",col=3) (2)
```

```
> y<-seq(0,50,len=100) (3)
```

```
> lines(y,dnorm(y,25,3.535),col=2) (4)
```

Invitamos al lector a que realice experimentos del tipo anterior, variando los parámetros de la binomial. Por ejemplo, si ejecuta la secuencia de comandos siguiente, la aproximación

será mala. Obsérvese cómo, al haber modificado los parámetros de la binomial, hemos hecho lo propio con los de la normal y los del recorrido de ésta.

```
> muestra<-rbinom(1000,5,0.5)
> hist(muestra,prob="T",col=3)
> y<-seq(0,5,len=100)
> lines(y,dnorm(y,2.5,1.118),col=2)
```

Ejemplo 4.9

Pero esta aproximación proporcionada por el Teorema Central del Límite no sólo es válida para la binomial, sino para la media de datos procedentes de cualquier distribución que tenga varianza finita (y, por tanto, media finita). En concreto, el teorema nos dice que si tenemos una muestra X_1, \dots, X_n de datos procedentes de un modelo de media μ y desviación típica σ , entonces

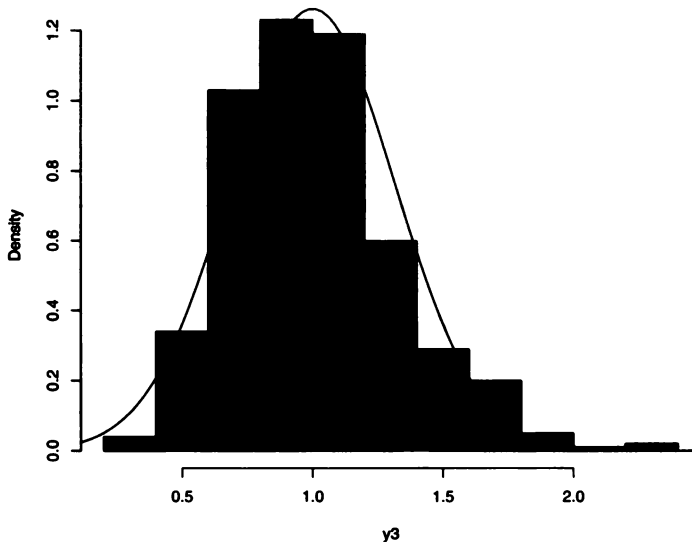


Figura 4.10: Histograma de datos $\text{gamma}(1,1)$ con aproximación normal

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

seguirá aproximadamente (si n es grande) una distribución $N(0,1)$. Esta propiedad es utilizada con mucha frecuencia en Inferencia Estadística. Veamos cómo R nos visualiza esta aproximación con datos procedentes de una $\text{Gamma}(1,1)$, distribución que tiene como media

1 y como desviación típica 1, con lo que la media aritmética de datos procedentes de esta distribución, deberá tener, aproximadamente, una distribución $N(1, 1/\sqrt{n})$.

Para visualizar esta aproximación ejecutamos la secuencia siguiente de comandos, en donde aprendemos a utilizar alguna función nueva de R. Primero vamos a obtener muestras de tamaño $n = 4$, con lo que la aproximación no será buena,

```
> y1<-rgamma(5000,1,1) (1)
```

```
> y2<-matrix(y1,ncol=4) (2)
```

```
> y3<-rowMeans(y2) (3)
```

```
> hist(y3,prob=T,col=3)
```

```
> y<-seq(0,2.5,len=100)
```

```
> lines(y,dnorm(y,1,0.5),col=2)
```

En (1) hemos obtenido 5000 datos procedentes de la $\text{gamma}(1,1)$. En (2) hemos creado una matriz de $5000/4=1250$ muestras de una $\text{gamma}(1,1)$. En (3) obtenemos un vector de 1250 medias muestrales de tamaño $n = 4$ de una $\text{gamma}(1,1)$, procediendo con los tres últimos comandos a realizar las representaciones gráficas.

Con la secuencia de los comandos siguientes, al ser las muestras de tamaño $n = 10$, ya obtenemos una buena aproximación, mostrada en la Figura 4.10, en donde hemos modificado los parámetros de la normal a aproximar.

```
> y2<-matrix(y1,ncol=10)
```

```
> y3<-rowMeans(y2)
```

```
> hist(y3,prob=T,col=3)
```

```
> lines(y,dnorm(y,1,0.316),col=2)
```

4.8. Ejercicios de Autoevaluación

Ejercicio 4.1

Calcular la media y la desviación típica de la variable aleatoria cuya función de distribución es la siguiente:

$$F(x) = \begin{cases} 0 & \text{si } x < -1 \\ 1/3 & \text{si } -1 \leq x < 0 \\ 8/15 & \text{si } 0 \leq x < 0'5 \\ 1 & \text{si } x \geq 0'5 \end{cases}$$

Ejercicio 4.2

Se lanza un dado 8 veces seguidas. Calcular la probabilidad de obtener un *dos* en más de una ocasión.

Ejercicio 4.3

Si X es una variable aleatoria con distribución normal de media 2 y desviación típica 4, determinar de forma razonada las siguientes probabilidades

- a) $P\{|X| < 6\}$
- b) $P\{|2 - X| < 4\}$.

Ejercicio 4.4

Al estudiar la duración en días, X , de una determinada componente eléctrica, se admitió para tal variable aleatoria una distribución exponencial con tiempo medio de vida 1000 días. Se pide:

- a) Determinar la función de distribución de X .
- b) Calcular la desviación típica de X .
- c) Calcular la probabilidad de que la componente eléctrica dure más de 1500 días.
- d) Si una componente eléctrica, con el modelo exponencial aquí admitido, ha durado ya 600 días, ¿cuál es la probabilidad de que dure más de 1500 días?

Ejercicio 4.5

El uno por ciento de los niños sufre efectos secundarios tras la administración de un determinado antibiótico. Si éste fue aplicado a seis niños, determinar la probabilidad de que

- a) Ninguno padezca efectos secundarios.
- b) Lo padezca más de un niño.
- c) Si se suministrase el antibiótico a 1000 niños, ¿cuál sería el número medio de niños con efectos secundarios?
- d) Calcular la probabilidad de que, de esos mil niños, padezcan efectos secundarios más de 15.

4.9. Lecturas Recomendadas

- Johnson, N.L., Kemp, A.W. y Kotz, S. (2005). *Univariate Discrete Distributions*. Tercera Edición. Editorial Wiley.
- Johnson, N.L., Kotz, S. y Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Volumen I. Segunda Edición. Editorial Wiley.
- Johnson, N.L., Kotz, S. y Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Volumen II. Segunda Edición. Editorial Wiley.

Capítulo 5

Estimadores. Distribución en el muestreo

5.1. Introducción

Como ya indicamos en el Capítulo 2, el objetivo de la Inferencia Estadística es el obtener conclusiones sobre una población mediante la observación de una parte de la misma denominada *muestra*.

Concretamente, si estamos interesados en una característica o parámetro poblacional asociado a una variable aleatoria observable X —como por ejemplo su media $\theta = E[X]$ — siendo la distribución de X (es decir, el modelo supuesto para X) en parte conocida —por ejemplo $N(\theta, 1)$ —, el propósito de la Inferencia Estadística es el obtener conclusiones sobre θ en base a una muestra aleatoria simple de X , es decir, en base a n observaciones de X , X_1, \dots, X_n , entendiéndose cada X_i , $i = 1, \dots, n$ como el valor de la variable X en el individuo seleccionado al azar en el lugar i -ésimo (por ejemplo la talla del individuo i -ésimo).

El proceso de selección en una muestra aleatoria simple —piense por ejemplo en una selección con reemplazamiento de bolas de una urna— conlleva el que las X_i sean variables aleatorias independientes e idénticamente distribuidas con distribución común la de X .

Así pues, formalmente una muestra aleatoria simple de una variable aleatoria X es una variable aleatoria n -dimensional (X_1, \dots, X_n) cuyas variables aleatorias unidimensionales que la componen —realizaciones de X — son independientes y con la misma distribución —la de X .

Por tanto, si X es continua con función de densidad f , la función de densidad conjunta de (X_1, \dots, X_n) será

$$f(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i) = f(x_1) \cdot \dots \cdot f(x_n)$$

y si X es discreta con función de masa p , la función de masa conjunta será

$$p(x_1, \dots, x_n) = \prod_{i=1}^n p(x_i).$$

Esta distribución de (X_1, \dots, X_n) se denomina *distribución muestral* y dado que la situación habitual que se plantea en Inferencia es la de ser la distribución de X no totalmente conocida, sino dependiente de algún parámetro θ desconocido (el cual puede ser multivariante), la distribución muestral también dependerá de θ , haciéndose referencia explícita de éste en su expresión: si es de tipo continuo, se expresa como f_θ y si es de tipo discreto, como p_θ .

Ejemplo 5.1

Sea (X_1, \dots, X_n) una muestra aleatoria simple de una población $N(\mu, \sigma)$ siendo $\theta = (\mu, \sigma)$ un parámetro desconocido. Ésta es la forma habitual en la que suele expresarse la extracción de una muestra de una población en la que se estudia una variable modelizada como una $N(\mu, \sigma)$ en la que tanto la media como la desviación típica son desconocidas.

Como este modelo tiene por función de densidad

$$f_{\mu, \sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

la distribución muestral tendrá como función de densidad conjunta

$$f_{\mu, \sigma}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\mu, \sigma}(x_i) = \frac{1}{(\sigma\sqrt{2\pi})^n} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}.$$

Ejemplo 5.2

Sea (X_1, \dots, X_n) una muestra aleatoria simple de una distribución $\mathcal{P}(\lambda)$, siendo λ un parámetro desconocido.

La distribución muestral tendrá como función de masa conjunta,

$$p_\lambda(x_1, \dots, x_n) = \prod_{i=1}^n p_\lambda(x_i) = \frac{e^{-n\lambda} \lambda^{x_1 + \dots + x_n}}{x_1! \cdots x_n!}.$$

Otra cuestión que también enunciamos en el Capítulo 2 es la estimación. Como allí dijimos, estaremos interesados bien en asignar —o mejor dicho inferir— un valor numérico al parámetro θ (*estimación por punto*), o bien en inferir un conjunto de valores plausibles para θ (*estimación por intervalos*).

de confianza y contraste de hipótesis). En este proceso será imprescindible contar más que con la muestra, con una función cuya $T(X_1, \dots, X_n)$ denominada *estimador o estadístico*.

Así, si θ es la media de la población, parece razonable utilizar la *media muestral*

$$T(X_1, \dots, X_n) = \bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

en su estimación, entendida ésta como función —media aritmética— de los valores que observemos.

De hecho, los estimadores muestrales contruidos por analogía de las medidas descriptivas estudiadas en el Capítulo 2, son, en general, buenos estimadores de los correspondientes parámetros poblacionales.

Así, la *varianza muestral*

$$s^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{x})^2$$

es un buen estimador de la varianza poblacional $\sigma^2 = V(X)$.

El *coeficiente de correlación muestral*

$$r = \frac{n \sum_{i=1}^n X_i Y_i - (\sum_{i=1}^n X_i) (\sum_{i=1}^n Y_i)}{\sqrt{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \sqrt{n \sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2}}$$

lo es del coeficiente de correlación poblacional ρ , etc.

Observemos, no obstante, que los estimadores $T(X_1, \dots, X_n)$ son variables aleatorias, ya que son funciones (T) de la muestra que es aleatoria (no se sabe qué valores se van a obtener, luego no se sabe, por ejemplo, cual será su media aritmética).

Por tanto, al ser $T(X_1, \dots, X_n)$ una variable aleatoria, tendrá una distribución de probabilidad que denominaremos *distribución en el muestreo* de T , la cual es de fundamental en Inferencia, ya que nos dará la *distribución* de los valores de T . Es decir, cuáles son los más probables, o cuál es su dispersión, etc.

Así, un estimador cuya media sea el parámetro a estimar: $E[T] = \theta$ —denominado *centrado* o *insesgado*— es deseable, puesto que esta propiedad nos expresa una *cancelación* entre los valores mayores y menores que toma, respecto al parámetro.

Un estimador con poca varianza nos indicará una mayor probabilidad de obtención de valores cercanos al parámetro.

La elección del estimador adecuado a cada problema que se esté estudiando es una cuestión delicada, además de no existir en general solución única, saliéndose, en todo caso, de los límites de este libro.

En la siguiente sección estudiaremos un método general para determinarlos, el cual suele conducir a estimadores con buenas propiedades, al menos para muestras grandes.

Además, desde la sección cuarta veremos, para cada una de las situaciones habituales que se suelen plantear, el estimador que razonablemente debe utilizarse, así como su distribución en el muestreo.

Advertimos, no obstante, que estas secciones no son más que indicativas, ya que en ellas no se han tenido en cuenta otras consideraciones como la existencia de datos anómalos, o la debilidad en la suposición del modelo poblacional, o la de independencia, etc.

Un análisis profundo del problema concreto que se esté estudiando es siempre necesario, así como recomendable la consulta a un estadístico profesional, de la misma manera que una enciclopedia médica, por muy buena que ésta sea, nunca podrá sustituir al médico en el análisis de una dolencia.

5.2. Método de la máxima verosimilitud

Supongamos una urna compuesta por bolas blancas y negras de la que sólo sabemos que la proporción de blancas es $p = 1/2$ ó $p = 1/3$.

De ella extraemos dos bolas con reemplazamiento resultando una blanca y otra negra.

La idea del *Método de la Máxima Verosimilitud* consiste en dar como estimación del parámetro aquel valor —de entre los posibles— que haga máxima la probabilidad del suceso observado, es decir, de la muestra obtenida.

Así, en nuestro ejemplo, si fuera $p = 1/2$ entonces la probabilidad de obtener una bola blanca y otra negra sería $0'5$, mientras que si fuera $p = 1/3$, dicho suceso tendría probabilidad $4/9 < 0'5$. Por tanto, el método de la máxima verosimilitud propone dar como estimación puntual de p el valor $\hat{p} = 1/2$.

Como hemos dicho, el método de la máxima verosimilitud propone como estimador del parámetro aquel que maximice la probabilidad del suceso observado, es decir de la muestra observada. Es decir, aquel que maximice la función de masa o densidad de la muestra observada, $p_\theta(x_1, \dots, x_n)$ ó $f_\theta(x_1, \dots, x_n)$.

Pero al decir de la *muestra observada* estamos diciendo que, en esa función, los valores x_1, \dots, x_n están fijos y lo que en realidad hacemos variar es θ con objeto de maximizar la función.

Para resaltar este hecho, a la función de probabilidad de la muestra —función de masa o de densidad— la representaremos por $L(\theta)$, y la denomi-

naremos *función de verosimilitud* de la muestra,

$$L(\theta) = \begin{cases} p_{\theta}(x_1, \dots, x_n) & \text{si es discreta la variable} \\ f_{\theta}(x_1, \dots, x_n) & \text{si es continua la variable.} \end{cases}$$

El método de la máxima verosimilitud propone como estimador de θ aquel $\hat{\theta}$ que maximice la función de verosimilitud,

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Como el máximo de una función y el de su logaritmo se alcanzan en el mismo punto, habitualmente determinaremos el $\hat{\theta}$ tal que

$$\log L(\hat{\theta}) = \max_{\theta} \log L(\theta).$$

El cálculo de este máximo se determina de la forma habitual en la que se determinan los máximos de una función. Por tanto, en muchas ocasiones aunque no siempre, este máximo se calculará derivando respecto al parámetro o parámetros, igualando a cero y despejando.

Ejemplo 5.3

Sea (X_1, \dots, X_n) una muestra aleatoria simple de una distribución $\mathcal{P}(\lambda)$, siendo λ un parámetro desconocido.

La función de verosimilitud y su logaritmo, serán respectivamente,

$$L(\lambda) = p_{\lambda}(x_1, \dots, x_n) = \frac{e^{-n\lambda} \lambda^{x_1 + \dots + x_n}}{x_1! \cdots x_n!}$$

$$\log L(\lambda) = -n\lambda + \log \lambda \sum_{i=1}^n x_i - \log \prod_{i=1}^n x_i!.$$

La derivada respecto a λ igualada a cero —denominada ecuación de verosimilitud,

$$\frac{d}{d\lambda} \log L(\lambda) = 0$$

será

$$-n + \frac{1}{\lambda} \sum_{i=1}^n x_i = 0$$

de donde despejando, se obtiene como estimador de máxima verosimilitud para λ

$$\hat{\lambda} = \frac{\sum_{i=1}^n X_i}{n}$$

es decir, la media muestral \bar{x} .

Ejemplo 5.4

Sea (X_1, \dots, X_n) una muestra aleatoria simple de una población $N(\mu, \sigma)$ siendo la media μ y la varianza σ^2 dos parámetros desconocidos.

La función de verosimilitud será

$$L(\theta) = L(\mu, \sigma^2) = f_{\mu, \sigma^2}(x_1, \dots, x_n) = \frac{1}{(\sigma^2 2\pi)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}$$

de logaritmo

$$\log L(\mu, \sigma^2) = -\frac{n}{2} \log \sigma^2 + \log \left(\frac{1}{\sqrt{2\pi}} \right)^n - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

Las derivadas de esta función, primero respecto a μ y luego respecto a σ^2 , igualadas a cero, proporcionan las siguientes ecuaciones de verosimilitud

$$\begin{cases} +\frac{1}{2\sigma^2} \sum_{i=1}^n 2(x_i - \mu) = 0 \\ -\frac{n}{2} \frac{1}{\sigma^2} + \frac{2 \sum_{i=1}^n (x_i - \mu)^2}{4(\sigma^2)^2} = 0 \end{cases}$$

sistema de dos ecuaciones con dos incógnitas, μ y σ^2 , de las que despejando se obtienen como soluciones $\hat{\mu} = \bar{x}$ y $\hat{\sigma}^2 = s^2$.

Obsérvese que si hubiera sido conocida la varianza σ^2 , el proceso anterior, con una incógnita y por tanto una ecuación de verosimilitud, hubiera dado como estimador de μ , la media muestral.

Por otro lado, si fuera conocida μ y desconocida la varianza, el método de la máxima verosimilitud proporcionaría como estimador de ésta

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Por ejemplo, si fuera $\mu = 0$ el estimador de σ^2 sería

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2.$$

No obstante, para determinar probabilidades como la de que el estimador $\hat{\sigma}^2$ subestime al parámetro, $P\{\hat{\sigma}^2 < \sigma^2\}$, necesitaremos conocer la distribución en el muestreo de $\hat{\sigma}^2$, es decir, la distribución de la suma de cuadrados de normales $N(0, 1)$.

Los estimadores de los parámetros de poblaciones normales tienen unas distribuciones en el muestreo muy peculiares, y dado que el caso de población normal —es decir que la variable X en observación siga una distribución normal— es muy destacado en Inferencia, al ser ésta una suposición habitual, vamos a dedicar la siguiente sección al estudio de las distribuciones de probabilidad que tendrán los mencionados estimadores.

5.3. Distribuciones asociadas a poblaciones normales

En esta sección estudiaremos tres distribuciones de probabilidad continuas del estilo a las analizadas en la Sección 4.5, pero con la peculiaridad de ser distribuciones en el muestreo de los principales estimadores —recuérdese que un estimador es una variable aleatoria— utilizados en las inferencias sobre los parámetros de una población normal.

5.3.1. Distribución χ^2 de Pearson

Sean X_1, \dots, X_n , n variables aleatorias independientes, cada una de las cuales sigue una distribución $N(0, 1)$.

Llamaremos *distribución χ^2 de Pearson* a la distribución de la variable aleatoria suma de los cuadrados de las n variables $N(0, 1)$

$$Y = X_1^2 + X_2^2 + \dots + X_n^2.$$

El subíndice n de la χ_n^2 corresponde al número de variables aleatorias independientes cuya suma forma la χ_n^2 y se denomina *grados de libertad* de la variable. Tiene que ser, por tanto, un número entero positivo.

La χ_n^2 es una distribución continua cuya función de densidad es

$$f(y) = \frac{y^{n/2-1} e^{-y/2}}{2^{n/2} \Gamma(n/2)} \quad y > 0$$

siendo su media $E[Y] = n$ y su varianza $V(Y) = 2n$.

No obstante, su función de densidad, no es muy utilizada en Inferencia. La que sí es utilizada es su función de distribución o, más en concreto, uno menos ella, $P\{\chi_n^2 > p\}$, es decir, la probabilidad cola. Con objeto de obtener estas probabilidades cola, en la *Tabla 4* de ADD, aparecen algunos de sus valores; o mejor dicho, las abscisas a las que corresponden probabilidades cola p , expresadas por líneas según los grados de libertad n .

Como ocurría con las tablas de la normal, mediante interpolación lineal o mejor aún, utilizando R, podremos obtener valores intermedios.

Ejemplo 5.5

Directamente de las tablas y por interpolación se obtienen respectivamente los valores

$$P\{\chi_{10}^2 > 7'267\} = 0'7 \quad y \quad P\{\chi_{21}^2 > 30\} = 0'09377$$

aunque habitualmente las tablas se utilizarán para determinar abscisas de probabilidades dadas. Así, la abscisa $\chi_{15;0'05}^2$, es decir, el punto tal que una distribución χ_{15}^2 deja a la derecha un área —una probabilidad cola— de 0'05 es $\chi_{15;0'05}^2 = 25$.

Al igual que antes, por interpolación se pueden calcular probabilidades intermedias. Así, $\chi_{23;0'2}^2 = 29'015$.

Obsérvese que en la tabla sólo aparecen valores de la χ_n^2 hasta $n = 30$ grados de libertad. La razón es que para mayor número de grados de libertad, esta distribución se aproxima por una normal. En concreto, se verifica que

$$\sqrt{2\chi_n^2} \approx N(\sqrt{2n-1}, 1).$$

Ello permite calcular probabilidades (y buscar abscisas) para grados de libertad mayores de 30. Así,

$$\begin{aligned} P\{\chi_{41}^2 > 24'5\} &= P\left\{\sqrt{2\chi_{41}^2} - \sqrt{2 \cdot 41 - 1} > \sqrt{2 \cdot 24'5} - \sqrt{2 \cdot 41 - 1}\right\} \approx \\ &\approx P\{Z > 7 - 9\} = P\{Z > -2\} = 1 - P\{Z > 2\} = 1 - 0'0228 = 0'9772 \end{aligned}$$

en donde $Z \sim N(0, 1)$.

Si queremos utilizar R en la cálculo de probabilidades relacionadas con esta distribución, el comando a utilizar es `chisq`

```
> pchisq(x,n)    # función de distribución en x de la Chi-cuadrado con n grados
> dchisq(x,n)    # función de densidad en x de la Chi-cuadrado con n grados
> qchisq(p,n)    # cuantil de orden p de la Chi-cuadrado con n grados
> rchisq(m,n)    # muestra aleatoria de tamaño m de la Chi-cuadrado con n grados
```

Análisis experimental de las convergencias de la χ^2

R no sólo se utiliza para calcular probabilidades de distribuciones. También sirve para realizar experimentos que permitan visualizar el comportamiento límite de las distribuciones de probabilidad.

Así, por un lado, sabemos que la distribución χ_n^2 se aproxima a una normal cuando n aumenta por el Teorema Central del Límite; en concreto, que

$$\frac{\chi_n^2 - n}{\sqrt{2n}} \approx N(0, 1).$$

Esto lo vemos gráficamente si ejecutamos la siguiente secuencia de comandos:

```
> x<-seq(0,40,len=100)
> plot(x,dchisq(x,4),type="l",col=1)
```

```
> lines(x,dchisq(x,9),type="l",col=2)
> lines(x,dchisq(x,13),type="l",col=3)
```

aunque, acabamos de ver, que es mejor utilizar la siguiente aproximación:

$$\sqrt{2\chi_n^2} - \sqrt{2n-1} \approx N(0,1)$$

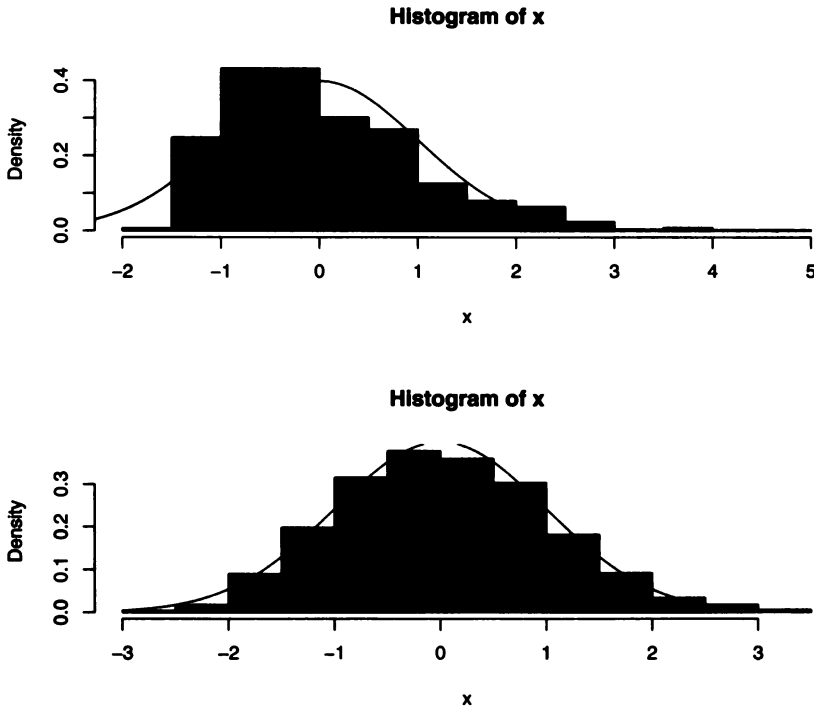


Figura 5.1 : Aproximaciones normales a la χ^2

Para ver gráficamente que esta segunda aproximación es mejor que la dada por el Teorema Central de Límite, representemos ambas aproximaciones; primero, con $n = 5$, ejecutando

```
> par(mfrow=c(2,1))
> y<-seq(-3,3,len=100)
> x<-(rchisq(1000,5)-5)/sqrt(2*5)          # primera aproximación
> hist(x,prob="T",col=2)
> lines(y,dnorm(y),col=3)
> x<-sqrt(2*rchisq(1000,5))-sqrt(2*5-1)    # segunda aproximación
> hist(x,prob="T",col=2)
```

```
> lines(y,dnorm(y),col=3)
```

en donde para este pequeño tamaño muestral vemos ya que la segunda es mejor que la primera aproximación en la Figura 5.1, donde la primera aproximación viene dada por la figura superior y la segunda por la inferior.

Si aumentamos el tamaño de la muestra a $n = 10, 20, 30$, iremos comparando ambas aproximaciones. Por ejemplo, para $n = 30$ utilizaríamos la siguiente secuencia.

```
> par(mfrow=c(2,1))
> y<-seq(-3,3,len=100)
> x<-(rchisq(1000,30)-30)/sqrt(2*30)      # primera aproximación
> hist(x,prob="T",col=2)
> lines(y,dnorm(y),col=3)
> x<-sqrt(2*rchisq(1000,30))-sqrt(2*30-1)  # segunda aproximación
> hist(x,prob="T",col=2)
> lines(y,dnorm(y),col=3)
```

5.3.2. Distribución t de Student

Como veremos en la siguiente sección, en poblaciones normales $N(\mu, \sigma)$, la distribución en el muestreo de la media muestral \bar{x} es también normal, aunque dependiente de σ , siendo este parámetro habitualmente desconocido.

Este hecho hace inviable el cálculo de probabilidades relacionadas con dicha media muestral, a menos que las muestras sean lo suficientemente grandes como para poder aplicar el teorema central del límite.

W.S. Gosset encontró empíricamente, en 1903, una solución a este problema. La distribución resultante se conoce como t de Student, pseudónimo con el que firmaba sus trabajos (Student, 1908).

Si X, X_1, \dots, X_n son $n+1$ variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma)$, llamaremos *distribución t de Student* a la distribución de la variable aleatoria

$$T = \frac{X}{\sqrt{\frac{1}{n} \sum_{i=1}^n X_i^2}}.$$

El número de variables aleatorias independientes del denominador recibe el nombre de *grados de libertad* de la t de Student, por lo que deberá ser un número entero positivo. A esta distribución se la suele representar por t_n .

Dividiendo arriba y abajo por σ vemos que una t_n es el cociente entre una normal $N(0, 1)$ y la raíz cuadrada de una χ_n^2 dividida por sus grados de

libertad, si ambas distribuciones son independientes.

La distribución t_n es de tipo continuo siendo su función de densidad,

$$f(y) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2} \quad -\infty < y < \infty$$

la cual tiene como un aspecto semejante a una $N(0, 1)$. De hecho, para muestras grandes converge a una $N(0, 1)$.

Se puede demostrar que si $Y \sim t_n$, entonces es $E[Y] = 0$ y $V(Y) = n/(n-2)$.

Al igual que pasaba con la distribución χ_n^2 , el cálculo de probabilidades relacionadas con esta distribución está tabulado en la *Tabla 5* de ADD, en donde aparecen diversas probabilidades cola para diferentes grados de libertad n .

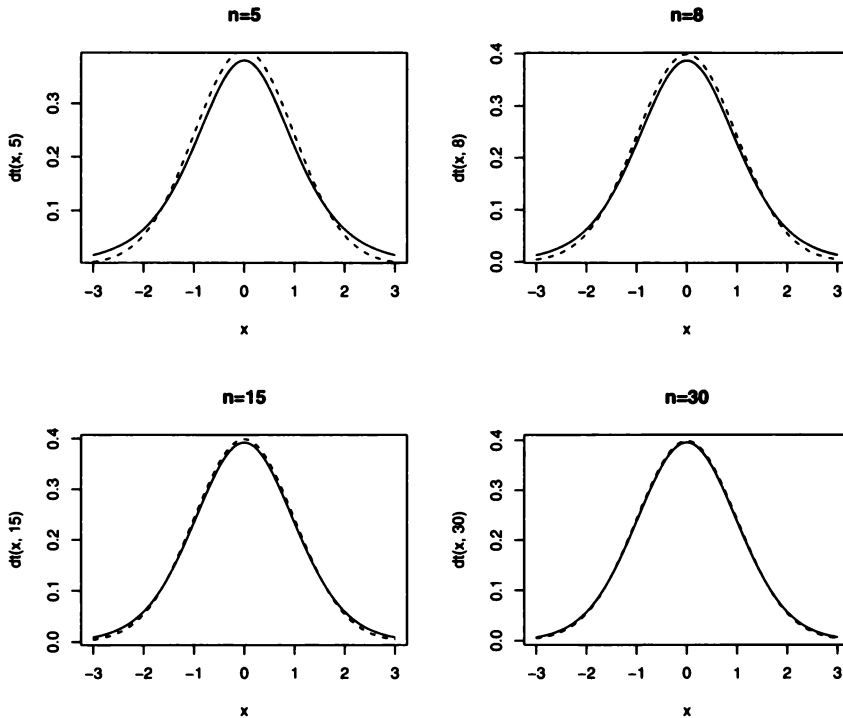


Figura 5.2 : Aproximaciones normales a la t de Student

Como antes dijimos para n grande t_n converge a una $N(0, 1)$. Esas probabilidades límite aparecen en la última línea de la tabla.

Al igual que ocurría con las tablas de las distribuciones antes mencionadas, habrá que utilizar interpolación lineal para calcular valores intermedios.

Obsérvese también que la función de densidad es simétrica (es decir, $f(-y) = f(y)$) lo que permite calcular probabilidades para abscisas negativas utilizando la *Tabla 5*. Así por ejemplo,

$$P\{t_{24} < -1'318\} = P\{t_{24} > 1'318\} = 0'1.$$

Si queremos calcular probabilidades con R, el comando a utilizar en R es `t`

```
> pt(x,n)    # función de distribución en x de la t-Student con n grados
> dt(x,n)    # función de densidad en x de la t-Student con n grados
> qt(p,n)    # cuantil de orden p de la t-Student con n grados
> rt(m,n)    # muestra aleatoria de tamaño m de la t-Student con n grados
```

Análisis experimental de la convergencia de la *t* de Student

Dijimos más arriba que esta distribución converge a una $N(0, 1)$ cuando n aumenta. Podemos comprobar esto experimentalmente ejecutando la siguiente secuencia de instrucciones, la cual dio origen a la Figura 5.2, en la cual hemos ido representando la densidad de la *t* de Student con $n = 5, 8, 15$ y 30 grados de libertad y sobre-impressionando en cada caso, la densidad de la normal estándar.

```
> par(mfrow=c(2,2))
> x<-seq(-3,3,len=100)
> plot(x,dt(x,5),type="l",col=4,main="n=5")
> lines(x,dnorm(x),col=3,lty=2)
> plot(x,dt(x,8),type="l",col=4,main="n=8")
> lines(x,dnorm(x),col=3,lty=2)
> plot(x,dt(x,15),type="l",col=4,main="n=15")
> lines(x,dnorm(x),col=3,lty=2)
> plot(x,dt(x,30),type="l",col=4,main="n=30")
> lines(x,dnorm(x),col=3,lty=2)
```

5.3.3. Distribución *F* de Snedecor

Una distribución relacionada con la estimación del cociente de varianzas de dos poblaciones normales es la denominada *F* de Snedecor.

Sean X_1, \dots, X_{n_1} , Y_1, \dots, Y_{n_2} , $n_1 + n_2$ variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma)$. Llamaremos *distribución F de Snedecor* a la distribución de la variable aleatoria

$$F = \frac{\frac{1}{n_1} \sum_{i=1}^{n_1} X_i^2}{\frac{1}{n_2} \sum_{i=1}^{n_2} Y_i^2}.$$

El número de sumandos del numerador y denominador recibe el nombre de *grados de libertad* de la F de Snedecor, por lo que se suele representar esta distribución como F_{n_1, n_2} . Lógicamente, estos números deben ser enteros positivos.

Si en F dividimos arriba y abajo por σ , vemos que una F_{n_1, n_2} es el cociente entre dos χ^2 independientes, cada una de ellas dividida por sus grados de libertad.

Esta distribución también es de tipo continuo cuya función de densidad, de aspecto semejante a la χ^2 . Su función de densidad es

$$f(y) = \frac{\Gamma(n_1/2 + n_2/2)}{\Gamma(n_1/2) \Gamma(n_2/2)} \frac{n_1^{n_1/2} n_2^{n_2/2} y^{n_1/2-1}}{(n_1 y + n_2)^{(n_1+n_2)/2}} \quad y > 0.$$

apareciendo las probabilidades cola de esta distribución en ADD *Tabla 6*.

Un hecho importante en relación con la búsqueda de abscisas de esta distribución, es que, como por definición es

$$F_{n,m} = \frac{1}{F_{m,n}}$$

si representamos por $F_{m,n;p}$ el valor de la abscisa de la función de densidad de una F de Snedecor con (m, n) grados de libertad que deja a la derecha una área de probabilidad p ,

$$P\{F_{m,n} > F_{m,n;p}\} = p$$

entonces es

$$F_{n,m;1-p} = \frac{1}{F_{m,n;p}}$$

Si queremos calcular con R probabilidades relacionadas con esta distribución, el comando a utilizar es `f`. Así

```
> pf(x,n1,n2)  # función de distribución en x de la F(n1,n2)
> df(x,n1,n2)  # función de densidad en x de la F(n1,n2)
> qf(p,n1,n2)  # cuantil de orden p de la F(n1,n2)
> rf(n,n1,n2)  # muestra aleatoria de tamaño n de la F(n1,n2)
```

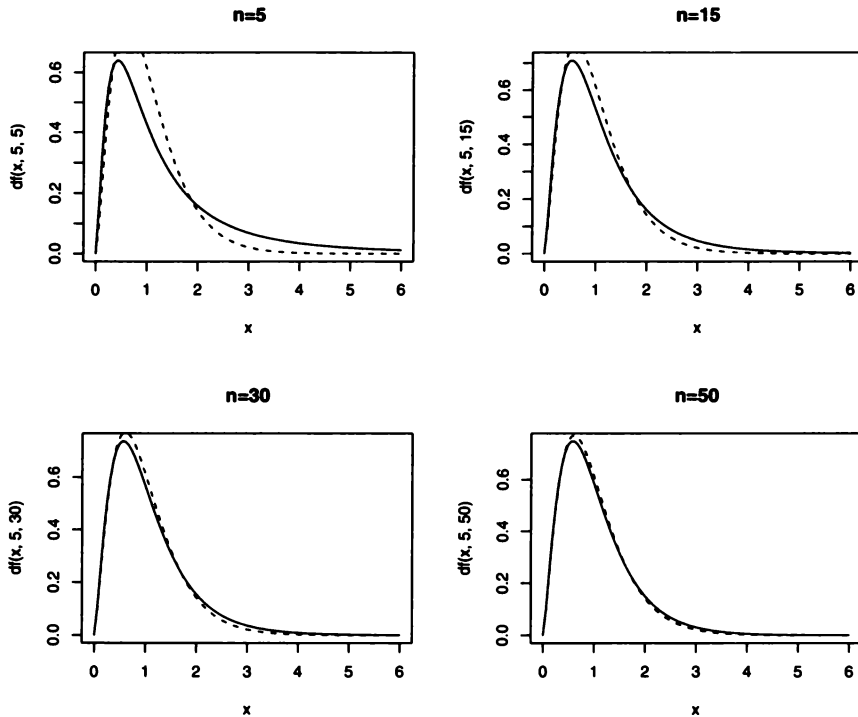


Figura 5.3 : Aproximaciones de la F de Snedecor

Análisis experimental de la convergencia de la F de Snedecor

Se puede demostrar que una $F_{(m,n)}$ converge cuando los segundos grados de libertad aumentan (es decir, en este caso, cuando n aumenta) al cociente de una χ_m^2/m . Para comprobar esto experimentalmente, ejecutamos la siguiente secuencia de comandos en donde una F de Snedecor $F_{(5,n)}$ se aproxima a una $\chi_5^2/5$ según vemos en la Figura 5.3.

```
> par(mfrow=c(2,2))
> x<-seq(0,6,len=100)
> plot(x,df(x,5,5),type="l",col=3,main="n=5")
> lines(x,5*dchisq(5*x,5),col=4,lty=2)
> plot(x,df(x,5,15),type="l",col=3,main="n=15")
> lines(x,5*dchisq(5*x,5),col=4,lty=2)
> plot(x,df(x,5,30),type="l",col=3,main="n=30")
> lines(x,5*dchisq(5*x,5),col=4,lty=2)
> plot(x,df(x,5,50),type="l",col=3,main="n=50")
> lines(x,5*dchisq(5*x,5),col=4,lty=2)
```

5.4. Estimación de la media de una población normal

En esta sección estudiaremos cuál debe ser el estimador a utilizar para estimar la media μ , cuando para la variable en estudio X se supone como modelo una $N(\mu, \sigma)$, así como su distribución en el muestreo, cuestión de gran interés más adelante cuando determinemos los intervalos de confianza y los contrastes de hipótesis óptimos, correspondientes a μ .

En ADD aparecen resumidos los resultados que se irán obteniendo para cada situación, lo que permitirá una rápida consulta.

Antes de empezar a obtener resultados, damos a continuación un teorema clave en dicha obtención.

Teorema de Fisher

Sea X_1, \dots, X_n una muestra aleatoria simple de una población $N(\mu, \sigma)$. Entonces, si \bar{x} y S^2 son, respectivamente, la media y cuasivarianza muestrales se tiene que

$$(a) \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

$$(b) \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2.$$

$$(c) \bar{x} \text{ y } \frac{(n-1)S^2}{\sigma^2} \text{ son independientes.}$$

A partir de este teorema obtenemos los siguientes resultados

σ conocida

Cuando la varianza poblacional es conocida, es razonable utilizar la media muestral \bar{x} para estimar μ . Su distribución en el muestreo es una normal de media μ y desviación típica σ/\sqrt{n} . Es decir,

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

σ desconocida

Si σ es desconocida, el resultado anterior sigue siendo válido, pero de poco nos va a servir al depender la distribución de \bar{x} de este parámetro desconocido.

Precisamente para esta situación fue para la que Student construyó su distribución.

Como una t de Student es el cociente entre una $N(0, 1)$ y la raíz cuadrada de una χ^2 dividida por sus grados de libertad, si es que ambas distribuciones son independientes, del teorema de Fisher obtenemos que

$$\frac{\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \rightsquigarrow t_{n-1}$$

de donde simplificando se obtiene que

$$\boxed{\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}}$$

Obsérvese que, como para muestras grandes —digamos $n > 30$ — la distribución t de Student se aproxima por una $N(0, 1)$, las probabilidades del cociente anterior se buscarán en las tablas de la normal en esos casos. De ahí que, en ocasiones, se hagan comentarios de que la distribución t se utiliza en el caso de muestras pequeñas.

Ejemplo 5.6

Se supone que la longitud craneal de los individuos de una población sigue una distribución normal con una desviación típica de 12'7 mm. Si elegimos de esa población al azar 10 individuos, la probabilidad de que la media de esa muestra difiera de la poblacional en más de 4'4 mm. será

$$P\{|\bar{x} - \mu| > 4'4\} = P\{|Z| > 1'1\} = 2 \cdot 0'1357 = 0'2714$$

por ser

$$\frac{\bar{x} - \mu}{12'7/\sqrt{10}} \rightsquigarrow N(0, 1).$$

Si hubiera sido desconocida la varianza poblacional y la muestra nos hubiera dado una cuasidesviación típica $S = 12$, la probabilidad buscada sería,

$$P\{|\bar{x} - \mu| > 4'4\} = P\{|t_9| > 1'1595\} = 2 \cdot P\{t_9 > 1'1595\} = 2 \cdot 0'1447 = 0'2894$$

al tener que utilizar una t de Student, por ser la varianza poblacional desconocida y las muestras pequeñas,

$$\frac{\bar{x} - \mu}{S/\sqrt{10}} \rightsquigarrow t_9$$

5.5. Estimación de la media de una población no necesariamente normal. Muestras grandes

En esta sección estudiaremos la situación en la que el modelo que se supone para la variable en estudio no es normal, o al menos no estamos lo suficientemente seguros de que lo sea como para poder utilizar los resultados de la sección anterior.

Hacemos notar que hay que considerar varias situaciones. Si no se conoce el modelo y las muestras son suficientemente grandes se deben utilizar los resultados del siguiente apartado.

Si no se conoce el modelo y las muestras no se pueden considerar grandes, o éstas son de datos *cualitativos* en el sentido de ser por ejemplo ordenaciones de individuos en un test de inteligencia, los métodos a utilizar son *no paramétricos*, los cuales se estudiarán en el Capítulo 8.

Otra cosa es que la población no sea normal pero sea conocida. En estos casos habrá que utilizar métodos específicos de la población en cuestión, determinando primero el estimador adecuado al parámetro en estudio, por ejemplo mediante el método de la máxima verosimilitud, y luego calculando su distribución en el muestreo.

Vemos en esta sección dos situaciones de este tipo, en las que para muestras grandes se obtienen distribuciones límite normales.

Población no necesariamente normal

Si no conocemos o no queremos suponer un modelo determinado para la variable en estudio, siempre que ésta tenga varianza finita σ^2 , podemos utilizar el teorema central del límite obteniendo para muestras suficientemente grandes, digamos $n > 30$, que

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Esto claro está, si σ es **conocido**, ya que si no lo es, tampoco nos servirá este resultado.

Si el tamaño muestral es algo mayor, digamos $n > 100$, podemos sustituir la varianza por un estimador suyo, obteniendo en el caso de que σ sea **desconocido** que

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \approx N(0, 1)$$

Población binomial

Si estamos interesados en estimar una proporción poblacional p , como por ejemplo la de alérgicos al polen de las acacias en España, es razonable establecer un modelo binomial para la variable dicotómica en estudio con p con probabilidad de éxito $X \sim B(1, p)$.

En este caso, las observaciones X_1, \dots, X_n serán unos o ceros según presenten o no los n individuos de la muestra la característica en estudio —en el ejemplo sean alérgicos o no.

Utilizando el método de la máxima verosimilitud se puede deducir que el estimador de p es la *proporción muestral*

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

siendo su distribución en el muestreo tal que $n\hat{p} \sim B(n, p)$.

Por las buenas propiedades asintóticas que tienen los estimadores de máxima verosimilitud, se puede demostrar que si las muestras son suficientemente grandes, digamos $n > 100$, es

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}} \approx N(0, 1)$$

Población Poisson

Si se admite un modelo $X \sim \mathcal{P}(\lambda)$, el estimador de máxima verosimilitud para λ basado en una muestra aleatoria simple de tamaño n de X , era la media muestral \bar{x} , el cual tiene una distribución en el muestreo tal que $n\bar{x} \sim \mathcal{P}(n\lambda)$.

No obstante, para muestras grandes, digamos $n > 100$, es posible utilizar su distribución asintótica, que es

$$\frac{\bar{x} - \lambda}{\sqrt{\bar{x}/n}} \approx N(0, 1)$$

Ejemplo 5.7

Con objeto de estimar los niveles de hierro en la sangre de los varones adultos sanos, se obtuvo una muestra de tamaño 100 que proporcionó una cuasidesviación típica de 15 microgramos por cada 100ml de sangre. La probabilidad de que la media de esa misma muestra difiera de la media poblacional en más de 3 microg/100ml será

$$P\{|\bar{x} - \mu| > 3\} = P\{|Z| > 2\} = 0.0456.$$

Ejemplo 5.8

Se sabe que el 8 % de los peces tropicales de los arrecifes coralinos de una zona geográfica son de una determinada especie.

La probabilidad de que entre 150 peces capturados en el área haya más del 10 % de dicha especie será,

$$P\{\hat{p} > 0'1\} = P\{Z > 0'90\} = 0'1841$$

al poder suponer que la proporción muestral sigue una distribución normal

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

con $p = 0'08$, por ser las muestras suficientemente grandes.

Obsérvese que, como suponemos en este ejemplo que p es conocido, no es necesario utilizar una estimación suya en la distribución de \hat{p} .

5.6. Estimación de la varianza de una población normal

Al igual que en las secciones anteriores, habrá que distinguir la situación en la que la media es conocida de la que no lo es.

μ desconocida

Si la media μ es desconocida, el teorema de Fisher ya nos indicaba que el estimador de la varianza en este supuesto debía ser la cuasivarianza muestral S^2 ya que, entre otras razones, al ser

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

será

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$$

es decir, $E[S^2] = \sigma^2$, lo cual supone que S^2 posee una propiedad deseable en los estimadores.

Por tanto, si μ es desconocida el estimador a utilizar para la estimar la varianza σ^2 es S^2 con distribución en el muestreo

$$\boxed{\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2}$$

μ conocida

Si μ es conocida, el caso anterior nos sugiere que el estimador a considerar sea similar al allí considerado pero utilizando, en lugar de la media muestral, la media poblacional μ ya que es conocida. Es decir, parece razonable utilizar el estimador

$$\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2.$$

Además, como las X_i siguen distribución $N(\mu, \sigma)$, seguirá $(X_i - \mu)/\sigma$ una distribución $N(0, 1)$, con lo que será

$$\boxed{\frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^2} \sim \chi_n^2}$$

por definición de la distribución χ_n^2 .

Ejemplo 5.9

Calcular la probabilidad de que en un recuento de glóbulos blancos en individuos de una muestra aleatoria simple de tamaño 10, la cuasivarianza muestral sobrestime a la varianza poblacional en más de un tercio de su valor, suponiendo que el número de glóbulos blancos sigue una distribución normal.

La probabilidad pedida será

$$P\{S^2 > \sigma^2 + \sigma^2/3\} = P\{\chi_9^2 > 12\} = 0'2333.$$

5.7. Estimación del cociente de varianzas de dos poblaciones normales independientes

En esta sección estudiaremos la distribución en el muestreo de los estimadores utilizados en inferencias sobre el cociente de las varianzas de dos poblaciones normales independientes.

Así, supondremos que X_1, \dots, X_{n_1} una muestra aleatoria simple, de tamaño n_1 , de una $N(\mu_1, \sigma_1)$ y que Y_1, \dots, Y_{n_2} una muestra aleatoria simple, de tamaño n_2 , de una $N(\mu_2, \sigma_2)$, siendo ambas independientes y con medias muestrales, respectivamente, \bar{x}_1 y \bar{x}_2 .

Como la cuasivarianza muestral es un buen estimador de la varianza poblacional, parece razonable estimar σ_1^2/σ_2^2 mediante el cociente S_1^2/S_2^2 .

No obstante, con objeto de hacer inferencias sobre el cociente de las varianzas poblacionales, es necesario conocer la distribución en el muestreo del estimador utilizado en dichas inferencias. De hecho, ésta es la razón de utilizar el cociente de las cuasivarianzas muestrales en lugar de otra función suya como por ejemplo su diferencia.

Aplicando el teorema de Fisher a cada una de las dos poblaciones, sabemos que es

$$\frac{\sum_{i=1}^{n_1} (X_i - \bar{x}_1)^2}{\sigma_1^2} = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi_{n_1-1}^2$$

y

$$\frac{\sum_{j=1}^{n_2} (Y_j - \bar{x}_2)^2}{\sigma_2^2} = \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi_{n_2-1}^2$$

siendo además ambas χ^2 independientes.

Por tanto, será

$$\frac{\frac{1}{n_1 - 1} \frac{\sum_{i=1}^{n_1} (X_i - \bar{x}_1)^2}{\sigma_1^2}}{\frac{1}{n_2 - 1} \frac{\sum_{j=1}^{n_2} (Y_j - \bar{x}_2)^2}{\sigma_2^2}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F_{n_1-1, n_2-1}$$

al ser el cociente de dos χ^2 independientes divididas por sus grados de libertad, una F de Snedecor.

Por tanto, si las medias poblacionales μ_1 y μ_2 son desconocidas será

$$\boxed{\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F_{n_1-1, n_2-1}}$$

Si fueran μ_1 y μ_2 conocidas, como una χ^2 es la suma de cuadrados de normales $N(0, 1)$ independientes y

$$\sum_{i=1}^{n_1} \left(\frac{X_i - \mu_1}{\sigma_1} \right)^2 \rightsquigarrow \chi_{n_1}^2 \quad \text{y} \quad \sum_{j=1}^{n_2} \left(\frac{Y_j - \mu_2}{\sigma_2} \right)^2 \rightsquigarrow \chi_{n_2}^2$$

siendo además ambas distribuciones independientes por proceder las observaciones de poblaciones independientes, tendremos, por los mismos razonamientos anteriores, que

$$\frac{\frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \mu_1)^2}{\frac{1}{n_2} \sum_{j=1}^{n_2} (Y_j - \mu_2)^2} \sim F_{n_1, n_2}$$

Aunque los papeles de la primera y segunda poblaciones son intercambiables, suele tomarse como población 1, la de mayor cuasivarianza muestral, debido a que se obtienen mejores inferencias con $S_1^2 > S_2^2$, por la asimetría de la distribución F de Snedecor.

Ejemplo 5.10

Un investigador supone que los niveles de vitamina A en dos poblaciones humanas independientes se distribuyen normalmente con el mismo nivel medio y varianzas iguales $\sigma_1^2 = \sigma_2^2$. Extraída una muestra aleatoria de cada población de tamaños $n_1 = 10$ y $n_2 = 12$ respectivamente, se obtuvieron como cuasivarianzas muestrales los valores $S_1^2 = 955$ y $S_2^2 = 415'2$. ¿Qué probabilidad habría de haber observado un desequilibrio entre las cuasivarianzas muestrales mayor del obtenido?

Como si las varianzas poblacionales son iguales, el cociente S_1^2/S_2^2 sigue una distribución $F_{9,11}$, la probabilidad pedida será,

$$P\left\{\frac{S_1^2}{S_2^2} > 2'3\right\} = P\{F_{9,11} > 2'3\} = 0'09696.$$

5.8. Estimación de la diferencia de medias de dos poblaciones normales independientes

Un resultado de cálculo de probabilidades es que si tenemos dos normales independientes, su suma (o diferencia) es una normal con media la suma (o diferencia) de las medias y desviación típica la raíz cuadrada de la suma de las varianzas.

La situación que nos ocupa en esta sección, es la de dos muestras independientes, la muestra X_1, \dots, X_{n_1} de una $N(\mu_1, \sigma_1)$ y la muestra Y_1, \dots, Y_{n_2} procedente de una $N(\mu_2, \sigma_2)$.

Aplicando el teorema de Fisher a cada una de las muestras se obtiene que es

$$\bar{x}_1 \sim N(\mu_1, \sigma_1/\sqrt{n_1}) \quad y \quad \bar{x}_2 \sim N(\mu_2, \sigma_2/\sqrt{n_2})$$

siendo ambas normales independientes, por proceder de poblaciones independientes. Por tanto, será

$$\bar{x}_1 - \bar{x}_2 \rightsquigarrow N \left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

por lo que, si σ_1 y σ_2 son conocidas, será

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0, 1)$$

Si σ_1 y σ_2 son desconocidas y las muestras son pequeñas —digamos $n_1 + n_2 \leq 30$ — habrá que recurrir a una t de Student.

No obstante, en el caso que nos ocupa habrá que diferenciar dos situaciones, admitiendo, la mayoría de las veces, una de las dos por medio de un contraste de hipótesis, el cual estudiaremos en el Capítulo 7.

(a) σ_1 y σ_2 se suponen iguales.

Aplicando el teorema de Fisher a cada una de las dos poblaciones independientes, tenemos,

independientes	
Población 1	Población 2
<ul style="list-style-type: none"> • $\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi_{n_1-1}^2$ • $\bar{x}_1 \rightsquigarrow N \left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}} \right)$ • independientes 	<ul style="list-style-type: none"> • $\frac{(n_2 - 1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi_{n_2-1}^2$ • $\bar{x}_2 \rightsquigarrow N \left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}} \right)$ • independientes

de donde se deduce que debe ser

$$\left\{ \begin{array}{l} \bullet \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi_{n_1+n_2-2}^2 \\ \bullet \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \rightsquigarrow N(0, 1) \\ \bullet \text{ independientes} \end{array} \right.$$

Si al construir la t de Student —como cociente entre la $N(0, 1)$ y la χ^2 independientes— utilizamos la suposición de ser $\sigma_1 = \sigma_2$, quedará

$$\boxed{\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$$

(b) σ_1 y σ_2 **no se suponen iguales**.

En este caso, al construir la t de Student, no se puede llevar a cabo la simplificación antes realizada. De hecho, esta situación no está resuelta completamente, existiendo varias aproximaciones a los grados de libertad f de la distribución t_f del cociente

$$\boxed{\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_f}$$

Aquí consideraremos la *aproximación de Welch* que define a f como el entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Ejemplo 5.11

Un equipo de investigadores está tratando de averiguar si existen o no diferencias significativas en los niveles de ácido úrico del suero de pacientes hospitalizados por una dolencia cardíaca y el resto de los pacientes de un determinado hospital.

Con este propósito se obtuvo una muestra aleatoria simple de $n_1 = 10$ pacientes con problemas cardíacos y $n_2 = 10$ enfermos sin ellos, para las que se obtuvieron unas cuasivarianzas muestrales de $S_1^2 = 45$ y $S_2^2 = 43$.

Suponiendo que los niveles de ácido úrico en ambas poblaciones siguen distribuciones normales, la probabilidad de que la diferencia de las medias muestrales estime la diferencia de medias poblacionales con un error mayor de 2mg/100ml será, si suponemos la varianzas poblacionales iguales,

$$\begin{aligned} P\{|\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)| > 2\} &= P\left\{|t_{18}| > \frac{2 \sqrt{18} \sqrt{100}}{\sqrt{9 \cdot 45 + 9 \cdot 43} \sqrt{10 + 10}}\right\} = \\ &= P\{|t_{18}| > 0'6742\} = 2 \cdot 0'257256 = 0'5145. \end{aligned}$$

5.9. Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

Si las muestras son lo suficientemente grandes como para poder aplicar el teorema central del límite, digamos $n_1 + n_2 > 30$ en el primer caso que sigue y $n_1 + n_2 > 100$ en el segundo y tercero, además de ser en los tres casos n_1 aproximadamente igual a n_2 , tendremos que,

Si σ_1 y σ_2 son conocidas

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0, 1)$$

Si σ_1 y σ_2 son desconocidas

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0, 1)$$

Poblaciones binomiales

Si además de ser las muestras grandes, las poblaciones independientes son binomiales: $X_1 \rightsquigarrow B(1, p_1)$ y $X_2 \rightsquigarrow B(1, p_2)$, será

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}\right)$$

Ejemplo 5.12

Con objeto de estudiar si existen diferencias significativas en la longitud (en mm.) del cuerpo de dos poblaciones de *rana pipiens*, geográficamente aisladas, se tomó una muestra aleatoria simple de individuos machos en cada una de las dos poblaciones obteniéndose los siguientes resultados

$$\begin{aligned} n_1 &= 65 & , & \quad \bar{x}_1 = 75 & , & \quad S_1^2 = 225 \\ n_2 &= 35 & , & \quad \bar{x}_2 = 79 & , & \quad S_2^2 = 195 \end{aligned}$$

La probabilidad de que la diferencia de las medias muestrales estime la diferencia de medias poblacionales con un error mayor de 2mm será,

$$P\{|\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)| > 2\} \approx P\{|Z| > 0'665\} = 2 \cdot 0'253 = 0'506.$$

5.10. Datos apareados

En esta sección se estudia el caso en que tengamos pares de datos dependientes $(X_1, Y_1), \dots, (X_n, Y_n)$. Situación que se da, por ejemplo, cuando se quiere estudiar la eficacia de una dieta de adelgazamiento, para lo cual, una vez elegidos n pacientes al azar de la población de personas sometidas a la dieta en estudio, se determina su peso antes de iniciar el tratamiento, X_i y después de finalizarlo, Y_i .

En este esquema no podemos suponer ambas poblaciones X e Y como independientes, sino formando una población —una variable aleatoria— bidimensional. Por ejemplo una normal bidimensional en el que el factor de dependencia se valora, en cierta forma, por el coeficiente de correlación poblacional ρ .

La manera de abordar este tipo de datos consiste en definir la variable diferencia $D = X - Y$, la cual ya es unidimensional, y aplicar a los n datos, supuestamente obtenidos de ella, $X_1 - Y_1, \dots, X_n - Y_n$, los resultados deducidos en las Secciones 5.4, 5.5 y 5.6.

Así por ejemplo, si admitimos normalidad en la población. Es decir, si admitimos que

$$D \rightsquigarrow N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}\right)$$

o mas brevemente, $D \rightsquigarrow N(\mu_d, \sigma_d)$, con

$$\mu_d = \mu_1 - \mu_2 \quad \text{y} \quad \sigma_d = \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

en el caso de que σ_d sea desconocida y las muestras sean pequeñas, tendremos que

$$\frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} \rightsquigarrow t_{n-1}$$

en donde es

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i) = \bar{x} - \bar{y} \quad y \quad S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - Y_i - \bar{d})^2.$$

Ejemplo 5.13

A doce pacientes elegidos al azar se les suministró una nueva dieta, combinada con unos determinados ejercicios físicos, con objeto de estudiar su efectividad en la reducción de los niveles de colesterol en sangre.

Para ello se anotó su nivel X_i , antes del comienzo del programa, e Y_i , después de finalizado el mismo, obteniéndose los siguientes resultados, los cuales proporcionaron un valor de $S_d^2 = 535'06$

Individuo	X_i	Y_i	$D_i = X_i - Y_i$
1	201	200	1
2	231	236	-5
3	221	216	5
4	260	233	27
5	228	224	4
6	237	216	21
7	326	296	30
8	235	195	40
9	240	207	33
10	267	247	20
11	284	210	74
12	201	209	-8

La probabilidad de que, supuesto que el efecto de la dieta sea nulo, estimemos una diferencia mayor que 14'7 será

$$P_{\mu_d=0}\{|\bar{d}| > 14'7\} = P\{|t_{11}| > 2'201\} = 0'05.$$

5.11. Tamaño muestral para una precisión dada

En Estadística es frecuente que se quieran determinar resultados con una determinada precisión, medida ésta en términos de probabilidad.

Así, en el Ejemplo 5.6 se puede estar interesado en determinar el tamaño muestral para que el error en la estimación sea menor que 2, de forma que esto

ocurra en el 95 % de las veces que muestreemos, es decir, de forma que haya probabilidad 0'95 de que esa diferencia sea así de pequeña.

Es decir, determinar n de forma que sea

$$P\{|\bar{x} - \mu| < 2\} = 0'95.$$

Esto, en resumidas cuentas, lo que implica es una ecuación de la que despejar n .

Así, la expresión anterior es equivalente a

$$P\left\{|Z| < \frac{2}{12'7/\sqrt{n}}\right\} = 0'95$$

en donde $Z \sim N(0, 1)$, por ser

$$\frac{\bar{x} - \mu}{12'7/\sqrt{n}} \sim N(0, 1)$$

Como por otro lado es $P\{|Z| < 1'96\} = 0'95$, deberá ser

$$\frac{2}{12'7/\sqrt{n}} = 1'96$$

es decir, $n = 154'9$. Tomando $n = 155$ obtendremos la precisión deseada.

En general, la situación que se tendrá dependerá del problema que se trate, por lo que aquí no obtendremos expresiones del tamaño muestral en las diversas situaciones posibles. Del enunciado del problema es de donde se obtendrá, para cada caso, una ecuación en n y, utilizando la distribución en el muestreo del estadístico que aparezca en la ecuación, se despejará la incógnita n .

5.12. Ejercicios de Autoevaluación

Ejercicio 5.1

Los saltamontes de la región africana de Asyut se caracterizan por tener una longitud media de 2 cm., pudiendo admitirse una distribución normal para la longitud de tales ortópteros. Elegida una muestra aleatoria de 20 de ellos, sus longitudes en cm. fueron las siguientes:

1'90, 1'85, 2'01, 1'95, 2'05, 2'00, 1'97, 2'02, 1'89, 2'01,

2'05, 1'95, 1'87, 2'05, 1'97, 1'85, 2'02, 1'95, 1'93, 2'05

Utilizando estos datos, se pide:

- La estimación de máxima verosimilitud de la desviación típica poblacional σ .
- Calcular la probabilidad de que el estimador de máxima verosimilitud de σ subestime el verdadero valor de dicho parámetro.

Ejercicio 5.2

Estudios anteriores han demostrado que puede admitirse una distribución de Poisson para el número de hembras, en una determinada región, de la mosca tropical americana (*Dermatobia hominis*) la cual se caracteriza por poner sus huevos en un mosquito, pasando las larvas de la mosca a la piel de la persona cuya sangre ha chupado el mosquito.

Examinada la región en cuestión en 10 días elegidos al azar, se obtuvo el siguiente número de moscas hembra de la citada especie:

2, 1, 3, 3, 4, 2, 1, 2, 3, 2

Se pide:

- Estimar la ley de probabilidad que rige el mencionado número de moscas hembra de la región en estudio.
- Determinar la estimación de máxima verosimilitud del número medio de moscas hembra de la región en cuestión.
- Calcular el número mínimo de días que debe de muestrearse en la región en estudio, para que la diferencia entre el número medio de moscas hembra de la región y su estimación, se diferencien en menos de una, con probabilidad 0'99.

Ejercicio 5.3

Por razones aún desconocidas, el porcentaje p de esquizofrénicos en todos los países es, de forma invariable, del 1 %. Determinar el tamaño de muestra necesario para que el porcentaje de esa muestra difiera en términos absolutos de p en menos de 0'003 con probabilidad 0'9, suponiendo que dicho tamaño muestral va a resultar grande.

Ejercicio 5.4

Se cree que los niveles de CO_2 en el planeta siguen una distribución de probabilidad dada por la siguiente función de densidad

$$f_{\theta}(x) = \frac{\exp\{x/\theta\}}{\theta(e-1)}, \quad 0 < x < \theta.$$

Con objeto de estimar el parámetro de esta distribución, se midió en 13 lugares el nivel de dióxido de carbono en partes por millón, obteniéndose los siguientes datos:

263, 298, 305, 342, 332, 333, 333, 333, 345, 339, 340, 342, 404

Determinar la estimación de máxima verosimilitud del parámetro θ .

Ejercicio 5.5

Se sometió a 9 personas a un curso intensivo de Estadística de dudosa eficacia, anotándose el nivel de conocimientos de estos nueve alumnos antes del comienzo del curso, X , y una vez finalizado éste, Y . Los resultados obtenidos por los nueve estudiantes fueron los siguientes:

X_i	7	6	5	3	6	2	6	5	7
Y_i	8	6	4	6	7	6	5	6	7

Admitiendo para X e Y distribuciones normales de igual media, calcular la probabilidad de que repitiendo el curso con una nueva muestra también de 9 alumnos, se obtuviera una diferencia de medias muestrales mayor que la obtenida en ésta (es decir, se mejoraran los resultados del curso realizado), suponiendo que, en esa nueva muestra, la cuasivarianza muestral será la misma que en el experimento realizado.

5.13. Lecturas Recomendadas

Mood, A.M., Graybill, F.A. y Boes, D.C. (1974). *Introduction to the Theory of Statistics*. Editorial McGraw-Hill.

Capítulo 6

Intervalos de confianza

6.1. Introducción

En el capítulo anterior estudiamos los principales estimadores que, razonablemente, deberíamos utilizar en cada una de las situaciones que allí se planteaban, situaciones que vienen marcadas por las estructuras probabilísticas concretas supuestas en cada caso: un modelo normal, la varianza conocida, etc.

Así, si queremos estimar la talla media, θ , de los individuos de una determinada población, supuesto que ésta sigue una distribución normal $N(\theta, 1)$, vimos en el capítulo anterior que razonablemente deberemos utilizar la media muestral \bar{x} para estimar θ , conociendo además su distribución en el muestreo.

Los estimadores allí estudiados, estimaban parámetros poblacionales *puntualmente*, es decir, eran funciones de la muestra que hacían corresponder a cada valor de ésta un único número, de entre los posibles valores que puede tomar el parámetro, el cual habitualmente será una buena estimación de éste.

No obstante, rara vez coincidirá esta *estimación puntual* con el desconocido valor del parámetro. Es decir, rara vez la media de la muestra seleccionada al azar será tal que $\bar{x} = \theta$.

Sin duda, es mucho más interesante concluir la inferencia con un intervalo de posibles valores del parámetro —al que denominaremos *Intervalo de Confianza*—, de manera que, antes de tomar la muestra, el desconocido valor del parámetro se encuentre en dicho intervalo con una probabilidad todo lo alta que deseemos.

Así por ejemplo, es mucho más deseable concluir afirmando que la media poblacional θ está entre $\bar{x} - 0'1$ y $\bar{x} + 0'1$, con probabilidad 0'99, que diciendo que la media muestral \bar{x} es un buen estimador de θ .

Con objeto de aumentar la precisión de la inferencia, serán deseables intervalos de confianza lo más cortos posible.

No obstante, la longitud del intervalo de confianza dependerá de lo alta que queramos sea la probabilidad con la que dicho intervalo —cuyos extremos son aleatorios— cubra a θ . Así, si queremos determinar dos puntos a y b tales que

$$P\{\bar{x} - a < \theta < \bar{x} + b\} = 1 - \alpha$$

al suponer que $X \rightsquigarrow N(\theta, 1)$, será

$$\frac{\bar{x} - \theta}{\sigma/\sqrt{n}} \rightsquigarrow N(0, 1).$$

Como puede demostrarse que entre todos los a y b que cumplen la ecuación anterior, el intervalo más corto es aquel en el que $a = b$, escribiremos la ecuación anterior de la forma

$$P\{-c < \bar{x} - \theta < c\} = 1 - \alpha$$

o lo que es lo mismo, tipificando,

$$P\left\{|Z| < \frac{c}{\sigma/\sqrt{n}}\right\} = 1 - \alpha$$

con $Z \rightsquigarrow N(0, 1)$. Por tanto, deberá ser

$$c = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

en donde $z_{\alpha/2}$ es el valor de la abscisa de una $N(0, 1)$ que deja a su derecha —bajo la función de densidad— un área de probabilidad $\alpha/2$.

El intervalo buscado será, por tanto,

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

Como se ve, su longitud

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

depende de la probabilidad $1 - \alpha$ elegida en su construcción, a la que denominaremos *coeficiente de confianza*, y del tamaño muestral (a mayor tamaño muestral n , menor será la longitud del intervalo).

Para un tamaño muestral fijo, cuanto mayor sea el coeficiente de confianza, más grande será $z_{\alpha/2}$ y por tanto, mayor su longitud. Por tanto, antes de construir un intervalo de confianza, habrá que prefijar cuidadosamente el valor del coeficiente de confianza de manera que la *probabilidad con la que confiamos* el intervalo cubra al desconocido valor del parámetro sea alta, pero conservando inferencias válidas.

Así, de poco interés resultará concluir que hay probabilidad 0'999 de que el intervalo (en metros) $[\bar{x} - 2, \bar{x} + 2]$, cubra la estatura media de la población.

Los coeficientes de confianza que se suelen considerar son 0'90, 0'95 y 0'99, aunque esto dependerá del investigador, el cual deberá tener siempre en cuenta los comentarios anteriores. Por ejemplo, una varianza poblacional σ^2 pequeña o un tamaño muestral grande pueden permitir un mayor coeficiente de confianza sin un aumento excesivo de la longitud del intervalo.

Formalmente definimos el intervalo de confianza para un parámetro θ de la siguiente manera.

Definición

Supongamos que X es la variable aleatoria en estudio, cuya distribución depende de un parámetro desconocido θ , y X_1, \dots, X_n una muestra aleatoria simple de dicha variable.

Si $T_1(X_1, \dots, X_n)$ y $T_2(X_1, \dots, X_n)$ son dos estadísticos tales que

$$P\{T_1(X_1, \dots, X_n) \leq \theta \leq T_2(X_1, \dots, X_n)\} = 1 - \alpha$$

el intervalo

$$[T_1(x_1, \dots, x_n), T_2(x_1, \dots, x_n)]$$

recibe el nombre de *Intervalo de Confianza* para θ de *coeficiente de confianza* $1 - \alpha$.

Obsérvese que tiene sentido hablar de que, antes de tomar la muestra, el intervalo aleatorio

$$[T_1(X_1, \dots, X_n), T_2(X_1, \dots, X_n)]$$

cubra al verdadero y desconocido valor del parámetro θ con probabilidad $1 - \alpha$ pero, una vez elegida una muestra particular x_1, \dots, x_n , el intervalo no aleatorio

$$[T_1(x_1, \dots, x_n), T_2(x_1, \dots, x_n)]$$

cubrirá o no a θ , pero ya no tiene sentido hablar de la probabilidad con que lo cubre.

Es decir, podemos hacer afirmaciones del tipo de que en un $100(1 - \alpha)\%$ de las veces, el intervalo que obtengamos cubrirá al parámetro, pero nunca de que, por ejemplo, hay probabilidad $1 - \alpha$ de que el intervalo de confianza $[1'65, 1'83]$ cubra al parámetro, ya que los extremos de este último intervalo —y como siempre el parámetro— son números y no variables aleatorias.

Obsérvese también que el intervalo de confianza es un subconjunto de los posibles valores del parámetro precisamente por ser no aleatorio.

Así mismo mencionemos que cualquier par de estimadores T_1 y T_2 que cumplan la condición impuesta en la definición anterior darán lugar a un intervalo de confianza. Habitualmente éstos serán dos funciones del estimador natural obtenido para cada caso en el capítulo anterior. De hecho, en las siguientes secciones indicaremos cuál es el intervalo de confianza que razonablemente debe utilizarse en cada situación concreta, omitiendo su obtención en la mayoría de los casos, obtención que parte siempre de la correspondiente distribución en el muestreo obtenida en el Capítulo 5. La determinación del intervalo de confianza en las situaciones omitidas, constituye un buen ejercicio para el lector.

Un resumen de cada uno de estos intervalos aparece en ADD facilitando una rápida consulta.

Advertimos, finalmente, que las situaciones que aquí analizaremos se corresponden con las que aparecen de forma habitual en los problemas prácticos de inferencia, aunque, claro está, no son las únicas posibles.

Existen métodos generales de obtención de intervalos de confianza, como el *Método de Neyman* o el *Método de la Cantidad Pivotal*, que permitirían determinar intervalos de confianza en modelos poblacionales aquí no considerados.

No obstante, su análisis, especialmente para variables discretas, no es tan sencillo como el del método de la máxima verosimilitud estudiado en el capítulo anterior para la obtención de estimadores puntuales, por lo que serán omitidos. Una detallada descripción de los mismos puede verse, por ejemplo, en el libro de Vélez y García (1993).

Respecto a la notación que utilizaremos, tanto en los intervalos de confianza como en el resto del libro, digamos que denotaremos por z_p , $t_{n;p}$, $\chi^2_{n;p}$ y $F_{n_1, n_2; p}$, respectivamente, el valor de la abscisa de una distribución $N(0, 1)$, t_n de Student, χ^2_n de Pearson y F_{n_1, n_2} de Snedecor, que deja a su derecha —bajo la correspondiente función de densidad— un área de probabilidad p .

6.1.1. Cálculo de Intervalos de Confianza con R

En el capítulo siguiente veremos que el intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test bilateral. Por esta razón se utiliza la misma función de R para obtener intervalos de confianza y test de hipótesis sobre un parámetro.

En concreto, la función de R que nos va a proporcionar los intervalos (y los tests), es la función `t.test`. Con ella vamos a poder determinar los Intervalos de Confianza (y tests) para la media, para datos apareados y para la diferencia de medias, pero no para aquellos casos en los que la varianza, varianzas o medias poblacionales sean conocidas sino para cuando haya que estimarlas a partir de los datos. También queremos advertir que, para poder aplicar esta función, es necesario conocer los datos individualmente ya que no podremos

utilizarla cuando sólo conozcamos los valores de las medias o cuasivarianzas muestrales y no los datos de donde éstas proceden.

La función a utilizar en el caso de Intervalos de Confianza es

```
t.test(x, y = NULL, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Entrando a describir cada uno de sus argumentos, en primer lugar diremos que los valores que aparecen después del símbolo = son los que toma la función por defecto y que, por tanto, no será necesario especificar si son los valores que deseamos ejecutar. En **x** incorporamos los datos de la muestra, si se trata de inferencias para una sola muestra; si se trata de datos apareados o de dos muestras independientes, introduciremos los datos de la segunda muestra en el argumento **y**.

Si especificamos **paired=F** (lo cual no es necesario puesto que es la opción tomada por defecto), estamos es una situación de datos no apareados. Un caso de datos apareados debe especificarse con **paired=T**.

El argumento **var.equal** nos permite indicar qué tipo de situación tenemos en el caso de comparación de dos poblaciones independientes. Si es **var.equal=T** tendremos una situación en la que las varianzas de ambas poblaciones se suponen iguales, y el intervalo será el habitual basado en una *t* de Student. Si especificamos **var.equal=F** las varianzas de ambas poblaciones no se suponen iguales y, en ese caso, estamos requiriendo un intervalo basado en una *t* de Student pero en donde los grados de libertad se determina por la aproximación de Welch.

El último argumento permite especificar el coeficiente de confianza, tomándose por defecto el valor 0'95.

El intervalo de confianza para el cociente de varianzas poblacionales se obtiene con la función

```
var.test(x, y, conf.level = 0.95)
```

en donde incorporamos los datos en los argumentos **x** e **y**. De nuevo aquí necesitaremos conocer los datos concretos y no admite esta función la situación de ser las medias poblacionales conocidas.

Por último, en la obtención de intervalos de dos poblaciones binomiales, debemos utilizar la función de R **prop.test**,

```
prop.test(x, n, conf.level = 0.95, correct = TRUE)
```

Los argumentos de esta función son: **x**, vector de éxitos. **n**, vector de número de pruebas realizadas. El último argumento **correct** es utilizado para indicar

al ordenador que utilice una corrección de Yates, no considerada en este texto, por lo que deberemos indicar `correct = F` si queremos replicar los resultados obtenidos sin la ayuda de R.

6.2. Intervalo de confianza para la media de una población normal

Tanto en esta sección como en las siguientes, determinaremos intervalos de confianza de *colas iguales*. Es decir, aquellos tales que, si el coeficiente de confianza es $1 - \alpha$, dejan en cada uno de los extremos la mitad de la probabilidad, $\alpha/2$.

En esta sección suponemos que los n datos proceden de una población $N(\mu, \sigma)$, y lo que pretendemos determinar es el intervalo de confianza para la media μ .

Como vimos en la Sección 5.4, en esta situación, tanto si la varianza poblacional σ^2 es conocida como si no lo es, el estimador natural de μ es la media muestral \bar{x} , por lo que determinar un intervalo de confianza para μ significa buscar un número c tal que

$$P\{\bar{x} - c < \mu < \bar{x} + c\} = 1 - \alpha$$

es decir, tal que

$$P\{-c < \bar{x} - \mu < c\} = 1 - \alpha$$

o bien

$$P\{|\bar{x} - \mu| < c\} = 1 - \alpha.$$

σ conocida

La distribución en el muestreo de \bar{x} es, en este caso,

$$\bar{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

con lo que, tipificando, será

$$P\left\{|Z| < \frac{c}{\sigma/\sqrt{n}}\right\} = 1 - \alpha$$

es decir,

$$c = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}.$$

El intervalo buscado será, por tanto,

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

σ desconocida

En este caso la media muestral tiene por distribución

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

con lo que

$$P\{|\bar{x} - \mu| < c\} = 1 - \alpha$$

será equivalente a

$$P\left\{|t_{n-1}| < \frac{c}{S/\sqrt{n}}\right\} = 1 - \alpha$$

de donde se obtiene

$$c = \frac{t_{n-1;\alpha/2} \cdot S}{\sqrt{n}}.$$

Así pues, en el caso de que la varianza poblacional sea desconocida, el intervalo de confianza para la media resulta

$$\left[\bar{x} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right]$$

en donde S^2 es la cuasivarianza muestral.

Ejemplo 6.1

Un terapeuta desea estimar, con una confianza del 99%, la fuerza media de un músculo determinado en los individuos de una población. Admitiendo que las unidades de fuerza siguen una distribución normal de varianza 144, seleccionó una muestra aleatoria de 25 individuos de la población, para la que obtuvo una media muestral de $\bar{x} = 85$.

En estas condiciones, el intervalo de confianza será

$$\left[85 - z_{0'01/2} \frac{12}{\sqrt{25}}, 85 + z_{0'01/2} \frac{12}{\sqrt{25}} \right] = [78'82, 91'18].$$

Si, como es más razonable, el terapeuta no supone conocida la varianza poblacional, deberá estimarla con la cuasivarianza muestral de los 25 individuos seleccionados. Si ésta fue $S^2 = 139$, el intervalo de confianza será

$$\left[85 - t_{24;0'01/2} \sqrt{\frac{139}{25}}, 85 + t_{24;0'01/2} \sqrt{\frac{139}{25}} \right] = [78'4, 91'59].$$

6.3. Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes

Aquí consideraremos primero una situación general y luego supondremos —realmente como casos particulares— que la población es binomial y que es de Poisson.

Población no necesariamente normal

Si no suponemos modelo alguno para la variable aleatoria en estudio, excepto que tenga varianza σ^2 finita y que la muestra de tamaño n sea suficientemente grande, tenemos dos situaciones posibles dependiendo del conocimiento o no de la varianza poblacional.

Si σ es **conocida** el intervalo de confianza para μ de coeficiente de confianza $1 - \alpha$ será

$$I = \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

y si σ es **desconocida**

$$I = \left[\bar{x} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

siendo, como antes, S la cuasidesviación típica muestral.

Población binomial

Si suponemos que $X \sim B(1, p)$ y que la muestra es suficientemente grande, el intervalo de confianza para p de coeficiente de confianza $1 - \alpha$ es

$$I = \left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right]$$

en donde \hat{p} es la proporción muestral.

Población Poisson

Suponiendo que $X \sim \mathcal{P}(\lambda)$ y que la muestra es suficientemente grande, el intervalo de confianza para λ de coeficiente de confianza $1 - \alpha$ es

$$I = \left[\bar{x} - z_{\alpha/2} \sqrt{\bar{x}/n}, \bar{x} + z_{\alpha/2} \sqrt{\bar{x}/n} \right].$$

Ejemplo 6.2

Los siguientes datos son valores de actividad (en micromoles por minuto por gramo de tejido) de una cierta enzima observada en el tejido gástrico de 35 pacientes con carcinoma gástrico

0'360	1'185	0'524	0'870	0'356	2'567	0'566
1'789	0'578	0'578	0'892	0'345	0'256	0'987
0'355	0'989	0'412	0'453	1'987	0'544	0'798
0'634	0'355	0'455	0'445	0'755	0'423	0'754
0'452	0'452	0'450	0'511	1'234	0'543	1'501

El histograma de estos datos (Figura 6.1) muestra claramente una fuerte asimetría a la derecha, lo cual sugiere que los valores de actividad no siguen una distribución normal.

No obstante, al ser el tamaño muestral bastante grande —no todo lo deseado al no suponer la varianza poblacional conocida— la media muestral \bar{x} sí sigue una distribución normal. Es decir, si hiciéramos un histograma en el que representáramos los valores obtenidos por la media muestral en un gran número de muestras, éste tendría forma acampanada aunque, como ocurre en este caso, la variable poblacional no siga una distribución normal.

El intervalo de confianza a utilizar será

$$I = \left[\bar{x} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

el cual, para un coeficiente de confianza del 95 % es igual a

$$I = \left[0'753 - 1'96 \sqrt{\frac{0'2686}{35}}, 0'753 + 1'96 \sqrt{\frac{0'2686}{35}} \right] = [0'5813, 0'9247].$$

Si queremos resolver este ejemplo con R, primero introducimos los datos ejecutando (1), un histograma suyo, obtenido ejecutando (2) y que aparece en la Figura 6.1 muestra una fuerte asimetría a la derecha, lo cual sugiere que los valores de actividad no siguen una distribución normal.

```
> x<-c(0.360,1.185,0.524,0.870,0.356,2.567,0.566,
+ 1.789,0.578,0.578,0.892,0.345,0.256,0.987,
+ 0.355,0.989,0.412,0.453,1.987,0.544,0.798,
+ 0.634,0.355,0.455,0.445,0.755,0.423,0.754,
+ 0.452,0.452,0.450,0.511,1.234,0.543,1.501)
```

(1)

```
> hist(x,prob=T)
```

(2)

Si queremos determinar el intervalo de confianza para la media (de una población no necesariamente normal, muestras grandes), de coeficiente de confianza 0'95, ejecutaríamos (3), obteniendo el intervalo en (4).

```
> t.test(x)
```

(3)

One Sample t-test

```
data: x
t = 8.5953, df = 34, p-value = 4.842e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.5749635 0.9310365
sample estimates:
mean of x
 0.753
```

(4)

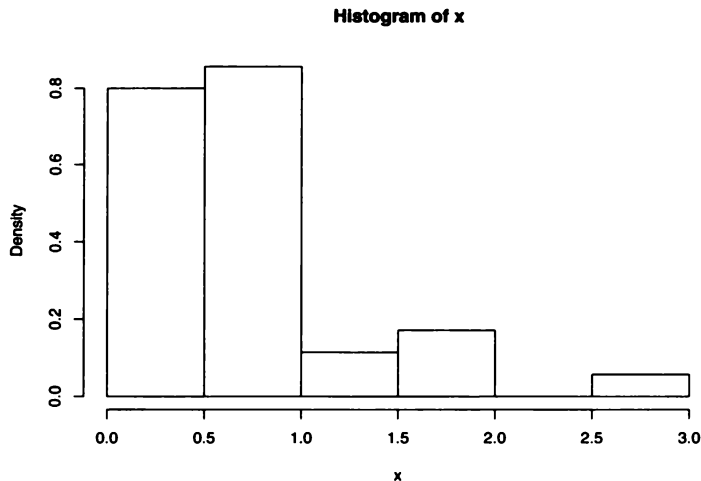


Figura 6.1 : Histograma del Ejemplo 6.2

El intervalo que obtenemos con R, $[0'5749, 0'9310]$ es algo diferente del que se obtuvo anteriormente debido a que antes se utilizaba la aproximación normal para la determinación de los cuantiles $z_{1-\alpha/2}$ y $z_{\alpha/2}$, mientras que aquí se utilizan los correspondientes de la distribución t de Student. Lo correcto sería lo que hicimos más arriba, pero a medida que n aumenta, apenas habrá diferencia entre ambos.

Ejemplo 6.3

Se quiere determinar un intervalo de confianza, con un coeficiente de confianza del 95 %, de la proporción de madrileños que han sido sometidos, alguna vez durante su vida, a algún tipo de tratamiento psiquiátrico.

Para ello se seleccionó, utilizando el padrón municipal, una muestra aleatoria simple de 200 residentes en la capital, con la que se estimó (utilizando técnicas de muestreo para preguntas comprometidas, las cuales están basadas en respuestas aleatorizadas, Warner, 1965) que el 15 % recibió este tipo de tratamiento.

El intervalo de confianza será

$$I = \left[0'15 \pm 1'96 \sqrt{\frac{0'15 \cdot 0'85}{200}} \right] = [0'1005, 0'1995].$$

Ejemplo 6.4

Admitiendo que el número de erratas por página de este libro sigue una distribución de Poisson, se quiere determinar un intervalo de confianza al 95 % del número medio de erratas por página que tiene.

Para ello se eligieron al azar y con reemplazamiento 100 páginas en las que se observó una media muestral de $\bar{x} = 0'04$ erratas por página.

El intervalo de confianza será,

$$I = \left[\bar{x} - z_{\alpha/2} \sqrt{\bar{x}/n}, \bar{x} + z_{\alpha/2} \sqrt{\bar{x}/n} \right]$$

es decir,

$$I = \left[0'04 - 1'96 \sqrt{0'04/100}, 0'04 + 1'96 \sqrt{0'04/100} \right] = [0'0008, 0'0792].$$

6.4. Intervalo de confianza para la varianza de una población normal

Dada una muestra aleatoria simple X_1, \dots, X_n de una población $N(\mu, \sigma)$, vamos a determinar el intervalo de confianza para σ^2 , distinguiendo dos casos según sea desconocida o no la media de la población μ .

μ desconocida

Como antes dijimos, queremos determinar el intervalo de colas iguales. Como es

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi_{n-1}^2$$

podemos encontrar en las tablas de la χ^2 dos abscisas tales que

$$P \left\{ \chi_{n-1;1-\alpha/2}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{n-1;\alpha/2}^2 \right\} = 1 - \alpha$$

de donde, despejando, se obtiene que

$$P \left\{ \frac{(n-1)S^2}{\chi_{n-1;\alpha/2}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1;1-\alpha/2}^2} \right\} = 1 - \alpha$$

es decir, el intervalo de confianza buscado será

$$I = \left[\frac{(n-1)S^2}{\chi_{n-1;\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1;1-\alpha/2}^2} \right]$$

con S^2 la cuasivarianza muestral.

μ conocida

En este caso, el intervalo de confianza será

$$I = \left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n;\alpha/2}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n;1-\alpha/2}^2} \right].$$

Ejemplo 6.1 (continuación)

Si el terapeuta del Ejemplo 6.1 quiere determinar un intervalo de confianza para la varianza de la variable en estudio, éste será

$$I = \left[\frac{(n-1)S^2}{\chi_{n-1;\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1;1-\alpha/2}^2} \right]$$

que para un coeficiente de confianza del 99 % proporciona los valores

$$I = \left[\frac{24 \cdot 139}{45'56}, \frac{24 \cdot 139}{9'886} \right] = [73'22, 337'45].$$

Obsérvese que para un tamaño muestral tan pequeño como el que tenemos, el intervalo de confianza al 99 % determinado resulta poco informativo, al tener éste una longitud muy grande.

El correspondiente al 90 %

$$I = \left[\frac{24 \cdot 139}{36'42}, \frac{24 \cdot 139}{13'85} \right] = [91'6, 240'9]$$

tampoco resulta mucho más informativo, perdiendo éste, además, parte del *grado de confianza* que el primero poseía. Una de las causas es que, habitualmente, estaremos interesados en estimar la desviación típica y no la varianza, puesto que ésta viene expresada en unidades al cuadrado lo que distorsiona en parte el resultado. El intervalo de confianza para la desviación típica será el de extremos la raíz cuadrada del correspondiente de la varianza. Así por ejemplo, el intervalo correspondiente al 90 % será

$$I = [\sqrt{91'6}, \sqrt{240'9}] = [9'57, 15'52].$$

6.5. Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes

Supondremos que X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} son dos muestras de tamaños n_1 y n_2 extraídas respectivamente de dos poblaciones independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

μ_1 y μ_2 conocidas

En este caso, el intervalo de colas iguales es

$$I = \left[\frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2 / \sum_{j=1}^{n_2} (Y_j - \mu_2)^2}{n_1 \cdot F_{n_1, n_2; \alpha/2}}, \frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2 / \sum_{j=1}^{n_2} (Y_j - \mu_2)^2}{n_1 \cdot F_{n_1, n_2; 1-\alpha/2}} \right].$$

μ_1 y μ_2 desconocidas

Si las medias poblacionales son desconocidas y las muestras proporcionan cuasivarianzas muestrales S_1^2 y S_2^2 respectivamente, el intervalo de confianza que se obtiene es

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; \alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 1-\alpha/2}} \right].$$

Ejemplo 6.5

Con objeto de estudiar la efectividad de un agente diurético, se eligieron al azar 11 pacientes, aplicando a 6 de ellos dicho fármaco y un placebo a los 5 restantes.

La variable observada en esta experiencia fue la concentración de sodio en la orina a las 24 horas, la cual dio los resultados siguientes:

Diurético :	20'4	62'5	61'3	44'2	11'1	23'7
Placebo :	1'2	6'9	38'7	20'4	17'2	

Supuesto que las concentraciones de sodio, tanto en la población a la que se aplicó el diurético $X_1 \sim N(\mu_1, \sigma_1)$ como a la que se aplicó el placebo $X_2 \sim N(\mu_2, \sigma_2)$, siguen distribuciones normales, en la determinación de un intervalo de confianza para la diferencia de medias poblacionales, veremos que, al ser las muestras pequeñas, necesitamos decidir si las varianzas poblacionales σ_1^2 y σ_2^2 pueden considerarse iguales o no.

Con este propósito se determina un intervalo de confianza para el cociente de dichas varianzas,

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; \alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 1-\alpha/2}} \right]$$

que resulta ser, para un coeficiente de confianza del 95%,

$$I = \left[\frac{483'12/208'52}{9'3645}, \frac{483'12 \cdot 7'3879}{208'52} \right] = [0'247, 17'117]$$

dado que

$$F_{n_1-1, n_2-1; \alpha/2} = F_{5, 4; 0'025} = 9'3645$$

y

$$F_{n_1-1, n_2-1; 1-\alpha/2} = \frac{1}{F_{n_2-1, n_1-1; \alpha/2}} = \frac{1}{F_{4, 5; 0'025}} = \frac{1}{7'3879}.$$

Si queremos resolver este ejemplo con R, primero incorporamos los datos en (1) y (2) y luego ejecutamos (3). El intervalo se obtiene en (4), lógicamente igual al acabado de calcular más arriba.

```

> x<-c(20.4,62.5,61.3,44.2,11.1,23.7) (1)
> y<-c(1.2,6.9,38.7,20.4,17.2) (2)
> var.test(x,y) (3)

```

F test to compare two variances

```

data: x and y
F = 2.3169, num df = 5, denom df = 4, p-value = 0.4359
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2474174 17.1172392 (4)
sample estimates:
ratio of variances
      2.316933

```

Este intervalo de confianza sugiere inferir que el cociente de ambas varianzas poblacionales es 1, es decir, que ambas son iguales, al pertenecer el 1 al intervalo de confianza calculado, razonamiento que justificaremos con detalle en el siguiente capítulo.

El que el 1 parezca estar muy cercano al extremo inferior del intervalo no debe confundirnos ya que la forma de la función de densidad de la F de Snedecor es asimétrica a la derecha por lo que tendrá, en consecuencia, más masa a la izquierda que a la derecha. De hecho, no es un mal ejercicio determinar intervalos de confianza para coeficientes de confianza menores, lo cual acortará la longitud del intervalo de confianza, aunque sensiblemente lo hará más por la derecha que por la izquierda, aunque se observará que éstos siguen conteniendo al 1.

6.6. Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes

Al igual que en la sección anterior suponemos que X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} son dos muestras de tamaños n_1 y n_2 respectivamente, extraídas de dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

σ_1 y σ_2 conocidas

En este caso sabemos, por la Sección 5.8, que es

$$\bar{x}_1 - \bar{x}_2 \rightsquigarrow N \left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

de donde el intervalo de confianza buscado será

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

En esta situación habrá que distinguir según sean

(a) $\sigma_1 = \sigma_2$

En cuyo caso, al ser

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \rightsquigarrow t_{n_1 + n_2 - 2}$$

obtendremos como intervalo de confianza

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{n_1 + n_2 - 2; \alpha/2} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

(b) $\sigma_1 \neq \sigma_2$

En este caso, la aproximación de Welch proporciona como intervalo de confianza

$$I = \left[\bar{x}_1 - \bar{x}_2 - t_{f; \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{f; \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

en donde S_1^2 y S_2^2 son las cuasivarianzas muestrales y f el entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{\left(\frac{S_1^2}{n_1} \right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2} \right)^2}{n_2 + 1}} - 2$$

Ejemplo 6.5 (continuación)

En la sección anterior concluimos infiriendo que las varianzas poblacionales podían considerarse iguales, admitiendo que las diferencias observadas entre sus estimadores, las cuasivarianzas muestrales, para la muestra concreta que allí se manejaba, era debida al azar y no a que existiera diferencia entre las varianzas poblacionales.

El intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$ será en consecuencia,

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{n_1+n_2-2; \alpha/2} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Utilizando la misma muestra que antes consideramos, práctica muy habitual pero algo más que discutible, obtendríamos el intervalo de confianza, para un coeficiente de confianza del 95 %,

$$I = \left[37'2 - 16'88 \mp 2'262 \sqrt{\frac{5 \cdot 483'12 + 4 \cdot 208'52}{9}} \sqrt{\frac{1}{6} + \frac{1}{5}} \right] = [-5'697, 46'347].$$

Para calcular este intervalo con R, ejecutamos (1) puesto que los datos los habíamos incorporado más arriba. El intervalo se obtiene en (2).

```
> t.test(x,y,var.equal=T) (1)
```

Two Sample t-test

```
data:  x and y
t = 1.766, df = 9, p-value = 0.1112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.708955 46.348955 (2)
sample estimates:
mean of x mean of y
 37.20    16.88
```

6.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

Si ahora X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} son dos muestras de tamaños n_1 y n_2 suficientemente grandes, extraídas de dos poblaciones independientes de medias μ_1 y μ_2 respectivamente, de las que sólo suponemos que tienen varianzas σ_1^2 y σ_2^2 finitas, tendremos que

Si σ_1 y σ_2 son conocidas

El intervalo de confianza para $\mu_1 - \mu_2$ con un coeficiente de confianza $1 - \alpha$ es

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

Si σ_1 y σ_2 son desconocidas

El intervalo de confianza se obtendrá sustituyendo las desconocidas varianzas por las cuasivarianzas muestrales, S_1^2 y S_2^2 , obteniéndose

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$$

Poblaciones binomiales

Por último, si admitimos, además de que los tamaños muestrales n_1 y n_2 son grandes, el que las poblaciones independientes son binomiales $X_1 \sim B(1, p_1)$ y $X_2 \sim B(1, p_2)$, la distribución aproximada en el muestreo de la diferencia de proporciones muestrales es

$$\hat{p}_1 - \hat{p}_2 \approx N \left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right)$$

conduce al siguiente intervalo de confianza para la diferencia de proporciones poblacionales $p_1 - p_2$

$$I = \left[\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right].$$

Ejemplo 6.6

En una granja experimental se intenta comparar la virulencia de dos organismos patógenos causantes de una determinada enfermedad en los pollos.

Para ello, se inoculó a 250 pollos con el organismo I, de los que 183 manifestaron signos durante las dos primeras semanas, y a 200 pollos con el organismo II, de los que 90 manifestaron signos de enfermedad en los 14 primeros días.

Supuesto que los pollos están aislados y que por tanto existe independencia entre ellos, el intervalo de confianza para la diferencia de proporciones con un coeficiente de confianza $100(1 - \alpha) \%$ será

$$I = \left[\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}} \right]$$

el cual, para un coeficiente de confianza del 95 %, es

$$I = \left[\frac{183}{250} - \frac{90}{200} \mp 1.96 \sqrt{\frac{0.732 \cdot 0.268}{250} + \frac{0.45 \cdot 0.55}{200}} \right] = [0.194, 0.370].$$

Para resolver este ejemplo con R, primero en (1) creamos un vector de éxitos y en (2) el vector de número de pruebas. El intervalo es obtenido en (4) ejecutando (3).

```
> x<-c(183,90) (1)
```

```
> n<-c(250,200) (2)
```

```
> prop.test(x,n,correct=F) (3)
```

```
2-sample test for equality of proportions without continuity
correction
```

```
data: x out of n
```

```
X-squared = 37.0292, df = 1, p-value = 1.164e-09
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
0.1938625 0.3701375 (4)
```

```
sample estimates:
```

```
prop 1 prop 2
```

```
0.732 0.450
```

6.8. Intervalos de confianza para datos apareados

Al igual que decíamos en la Sección 5.10, si la muestra que tenemos es de datos emparejados $(X_1, Y_1), \dots, (X_n, Y_n)$, en el sentido de proceder de una población bidimensional, la forma de actuar consiste en definir la variable unidimensional diferencia $D = X_i - Y_i$ y aplicar a sus parámetros los intervalos de confianza antes determinados.

Por ejemplo, si la variable bidimensional de donde proceden los datos es normal bivalente, la variable diferencia será

$$D \sim N\left(\mu_1 - \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}\right)$$

de donde si, por ejemplo, las muestras son pequeñas y la varianza poblacional $\sigma_d^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$ es desconocida, el que la media muestral —de las diferencias— tenga distribución en el muestreo

$$\frac{\bar{d} - \mu_d}{S_d/\sqrt{n}} \rightsquigarrow t_{n-1}$$

permite obtener como intervalo de confianza para $\mu_1 - \mu_2$ de coeficiente de confianza $1 - \alpha$, el siguiente

$$I = \left[\bar{d} - t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}}, \bar{d} + t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}} \right]$$

en donde es

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i) = \bar{x} - \bar{y} \quad y \quad S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - Y_i - \bar{d})^2.$$

Ejemplo 6.7

Con objeto de averiguar si la fuerza de la gravedad hace disminuir significativamente la estatura de las personas a lo largo del día, se seleccionaron al azar 10 individuos —mujeres de 25 años—, a las que se midió la estatura (en cm.) por la mañana al levantarse, X_i , y por la noche antes de acostarse, Y_i , obteniéndose los siguientes datos,

X_i	169'7	168'5	165'9	177'8	179'6	168'9	169'2	167'9	181'8	163'3
Y_i	168'2	165'5	164'4	175'7	176'6	166'1	167'1	166'3	179'7	161'5

Si queremos determinar un intervalo de confianza para la diferencia de estaturas medias poblacionales, en primer lugar deberemos calcular las diferencias $D_i = X_i - Y_i$

$$D_i: \quad 1'5 \quad 3 \quad 1'5 \quad 2'1 \quad 3 \quad 2'8 \quad 2'1 \quad 1'6 \quad 2'1 \quad 1'8$$

y como el tamaño muestral es pequeño, $n = 10$, y la varianza poblacional σ_d^2 desconocida, el intervalo de confianza será

$$I = \left[\bar{d} - t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}}, \bar{d} + t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}} \right]$$

que para un coeficiente de confianza de 0'95 resulta igual a

$$I = \left[2'15 - 2'262 \sqrt{\frac{0'349}{10}}, 2'15 + 2'262 \sqrt{\frac{0'349}{10}} \right] = [1'727, 2'573].$$

Si queremos resolver este ejemplo con R podemos, o bien calcular primero las diferencias $D_i = X_i - Y_i$ y luego ejecutar la función `t.test` a una muestra o, mejor, utilizarla para los pares de datos dados e indicarle que son datos apareados con el argumento `paired`. En concreto, incorporaremos primero los datos en (1) y (2); luego obtenemos un intervalo de confianza de coeficiente de confianza 0'95 ejecutando (3),

```
> x<-c(169.7,168.5,165.9,177.8,179.6,168.9,169.2,167.9,181.8,163.3) (1)
> y<-c(168.2,165.5,164.4,175.7,176.6,166.1,167.1,166.3,179.7,161.5) (2)
> t.test(x, y, paired = T) (3)
```

Paired t-test

data: x and y

t = 11.5014, df = 9, p-value = 1.104e-06

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

```
1.727125 2.572875
sample estimates:
mean of the differences
      2.15
```

(4)

Los resultados aparecen después. Se observa en (4) el intervalo de confianza buscado, idéntico al calculado anteriormente.

6.9. Ejercicios de Autoevaluación

Ejercicio 6.1

Se quiere estimar la eficacia de un tratamiento de fluoración del agua potable de una determinada ciudad. Para ello, antes de aplicar el tratamiento, se eligieron al azar 150 personas de la ciudad en cuestión y se observó que el 35 % de las mismas presentaba alguna caries dental. Pasado un año de tratamiento, se seleccionó otra muestra aleatoria simple de 150 habitantes de la misma ciudad, observándose un 30 % de personas con caries. En estas condiciones, determinar un intervalo de confianza, de coeficiente de confianza 0'95, para la diferencia de proporciones de personas con caries, antes y después de aplicar el tratamiento.

Ejercicio 6.2

El número de fraudes en compañías inmobiliarias de un determinado país sigue una distribución de Poisson de parámetro θ . Elegidas al azar cien compañías inmobiliarias del país en estudio, se obtuvo una media de 0'85 fraudes.

Determinar un intervalo de confianza, de coeficiente de confianza 0'95, para θ .

Ejercicio 6.3

Se quiere estimar, mediante un intervalo de confianza, la diferencia entre los pesos medios de los cerebros de vacas sanas y de vacas enfermas de una enfermedad degenerativa determinada, con objeto de ver si dicho intervalo contiene o no al cero.

Para ello se seleccionaron al azar $n_1 = 200$ vacas sanas, las cuales proporcionaron un peso medio cerebral de $\bar{x}_1 = 423$ gramos y una cuasidesviación típica muestral de 30 gramos e, independientemente de la muestra anterior, $n_2 = 100$ cerebros de vacas enfermas, que dieron un peso medio cerebral de $\bar{x}_2 = 410$ gramos y una cuasidesviación típica muestral de 50 gramos.

Determinar el intervalo de confianza deseado, con un coeficiente de confianza de 0'90.

Ejercicio 6.4

Se quiere estimar mediante un intervalo de confianza el porcentaje de suspensos de una determinada asignatura. Para ello se eligieron al azar 150 estudiantes de la asignatura en cuestión observándose, entre ellos, 60 repetidores. Determine el intervalo de confianza deseado, para un coeficiente de confianza de 0'95.

Ejercicio 6.5

Se sabe que el tiempo en días que transcurre desde la finalización de la segunda semana de Pruebas Presenciales de la UNED hasta la aparición de las notas es una variable aleatoria con distribución normal $N(\mu, \sigma)$.

En el curso académico pasado se tomaron al azar nueve asignaturas y se obtuvo un tiempo medio muestral $\bar{x} = 12$ días en la aparición de sus calificaciones y una cuasivarianza

muestral $S^2 = 25$.

Determinar un intervalo de confianza de coeficiente 0'95 para el tiempo medio μ .

6.10. Lecturas Recomendadas

Mood, A.M., Graybill, F.A. y Boes, D.C. (1974). *Introduction to the Theory of Statistics*.
Editorial McGraw-Hill.

Capítulo 7

Contraste de hipótesis

7.1. Introducción y conceptos fundamentales

Este capítulo es uno de los más importantes del libro ya que los *Contrastes de Hipótesis* son, sin duda alguna, los Métodos Estadísticos más utilizados.

Tanto es así, que el resto de los capítulos del libro son, básicamente, métodos estadísticos basados en contrastes de hipótesis.

Como ilustración de los conceptos que se irán definiendo, supongamos que estamos interesados en averiguar si el consumo habitual de un determinado producto modifica el nivel estándar de colesterol en las personas aparentemente sanas, el cual está fijado en 200 mg/dl. Actualmente parece concluirse que un nivel alto de colesterol es perjudicial en enfermedades cardiovasculares pero que, sin embargo, éste es necesario en la creación de defensas por parte del organismo, por lo que también se consideran perjudiciales niveles bajos de colesterol.

El primer punto a considerar en un contraste de hipótesis es precisamente ése: el establecer las *hipótesis* que se quieren contrastar, es decir, comparar.

Así, si en el ejemplo considerado representamos por μ el nivel medio de colesterol en la sangre de las personas que consumen habitualmente el producto en cuestión, el problema que tenemos planteado consiste en decidir si puede admitirse para μ un valor igual a 200 (el producto no modifica el nivel de colesterol) o un valor distinto de 200 (el producto modifica el contenido de colesterol).

Una de las dos hipótesis, generalmente la que corresponde a la situación estándar, recibe el nombre de *hipótesis nula* H_0 , mientras que la otra recibe el nombre de *hipótesis alternativa* H_1 , siendo el *contraste de hipótesis* el proceso de decisión basado en técnicas estadísticas mediante el cual decidimos —inferimos— cuál de las dos hipótesis creemos correcta, aceptándola y rechazando en consecuencia la otra. En este proceso medimos los dos posi-

bles errores que podemos cometer —aceptar H_0 cuando es falsa o rechazar H_0 cuando es cierta— en términos de probabilidades.

Por tanto, nuestro problema se puede plantear diciendo que lo que queremos es realizar el contraste de la hipótesis nula $H_0 : \mu = 200$, frente a la alternativa $H_1 : \mu \neq 200$.

Como todas las técnicas estadísticas, las utilizadas en el contraste de hipótesis se basan en la observación de una muestra, la cual aportará la información necesaria para poder decidir, es decir, para poder contrastar las hipótesis.

Si X representa la variable en observación: nivel de colesterol en la sangre, el contraste de hipótesis concluirá formulando una regla de actuación —denominada también contraste de hipótesis o por no ser excesivamente redundantes, *test de hipótesis* utilizando la terminología anglosajona— la cual estará basada en una muestra de X de tamaño n , X_1, \dots, X_n , o más en concreto en una función suya denominada *estadístico del contraste* $T(X_1, \dots, X_n)$, y que habitualmente será una función del estimador natural asociado al parámetro del que se quiere contrastar las hipótesis.

En la realización de un contraste de hipótesis suele ser habitual suponer un modelo probabilístico para la variable X , en cuyo caso hablaremos de *contrastos paramétricos*, en contraposición con los denominados *contrastos no paramétricos* —estudiados en el Capítulo 8—, en los que sólo serán necesarias suposiciones generales sobre el modelo probabilístico, tales como la simetría o continuidad de éste.

En todo caso, será imprescindible determinar la distribución en el muestreo del estadístico T del test, ya que la filosofía del contraste de hipótesis depende de su distribución en el muestreo, pudiendo formularse de la siguiente forma: si fuera cierta la hipótesis nula H_0 , la muestra, o mejor T , debería de comportarse de una determinada manera —tener una determinada distribución de probabilidad—. Si extraída una muestra al azar, acontece un suceso para T que tenía poca probabilidad de ocurrir si fuera cierta H_0 , —es decir, bajo H_0 —, puede haber ocurrido una de las dos cosas siguientes: o bien es que hemos tenido tan mala suerte de haber elegido una muestra *muy rara* o, lo que es más probable, que la hipótesis nula era falsa. La filosofía del contraste de hipótesis consiste en admitir la segunda posibilidad, rechazando en ese caso H_0 , aunque acotando la probabilidad de la primera posibilidad, mediante lo que más adelante denominaremos nivel de significación.

Así en nuestro ejemplo, parece razonable elegir al azar n personas aparentemente sanas a las que, tras haber consumido el producto en cuestión, midiéramos su nivel de colesterol en sangre, razonando de la siguiente forma: si la hipótesis nula $H_0 : \mu = 200$ fuera cierta, el estimador natural de μ , la media \bar{x} de la muestra obtenida tomaría un valor *cercano* a 200; si, tomada una muestra, este estimador está *lejos* de 200 deberemos rechazar H_0 .

No obstante, los términos *cercano* y *lejano* deben ser entendidos en el sen-

tido de algo con gran probabilidad de ocurrir o poca probabilidad de ocurrir, para lo cual necesitaremos conocer la distribución en el muestreo de T .

Además, estos términos dependen de la magnitud de los errores que estemos dispuestos a admitir, medidos éstos en términos de probabilidades. Puntualicemos estas ideas un poco más.

Errores de tipo I y de tipo II

Para determinar con precisión la regla de actuación en cada caso concreto, debemos considerar los dos errores posibles que podemos cometer al realizar un contraste de hipótesis, los cuales, como antes dijimos, son el de rechazar la hipótesis nula H_0 cuando es cierta, denominado *error de tipo I*, o el de aceptar H_0 cuando es falsa, denominado *error de tipo II*.

Ambos errores son de naturaleza bien distinta; así en el ejemplo considerado, si rechazamos H_0 cuando es cierta, tendremos un coste económico derivado de prohibir un producto no perjudicial, pero si aceptamos H_0 cuando es falsa y permitimos el consumo del producto, pueden producirse graves perjuicios en la salud de los consumidores.

La Estadística Matemática ha deducido tests de hipótesis, es decir reglas de actuación, siguiendo el criterio de fijar una cota superior para la probabilidad de error de tipo I, denominada *nivel de significación*, que maximizan $1 - P\{\text{error de tipo II}\}$, expresión ésta última denominada *potencia del contraste*.

Así, los tests que estudiemos en el libro serán tests de máxima potencia para un determinado nivel de significación α .

Región crítica y región de aceptación

Los tests de hipótesis, expresados siempre en función de un estadístico T adecuado al problema en cuestión, son de la forma

$$\begin{cases} \text{Aceptar } H_0 & \text{si } T \in C^* \\ \text{Rechazar } H_0 & \text{si } T \in C \end{cases}$$

en donde C y C^* son dos conjuntos disjuntos en los que se ha dividido el conjunto de valores posibles de T . C recibe el nombre de *región crítica* del test, y se corresponde con el conjunto de valores de T en donde se rechaza la hipótesis nula H_0 .

El conjunto complementario, C^* , se denomina *región de aceptación* y se corresponde, como su nombre indica, con el conjunto de valores del estadístico para los cuales se acepta H_0 .

Por completar la terminología propia de los contrastes de hipótesis, diremos que un test es *bilateral* cuando C esté formada por dos intervalos disjuntos y *unilateral* cuando la región crítica sea un intervalo.

Por último, se dice que una hipótesis —nula o alternativa— es *simple* cuando esté formada por un solo valor de parámetro. Si está formada por más de

uno, se denomina *compuesta*. Así, el ejemplo considerado se trata de un contraste de hipótesis nula simple —en H_0 está sólo el 200— frente a alternativa compuesta —en H_1 están todos los valores menos el 200.

Siguiendo con el mencionado ejemplo, y denotando $\mu_0 = 200$, hemos dicho que razonablemente deberemos aceptar H_0 cuando \bar{x} esté cerca de μ_0 , Figura 7.1, es decir, cuando sea

$$\mu_0 - c < \bar{x} < \mu_0 + c$$

para un c relativamente pequeño

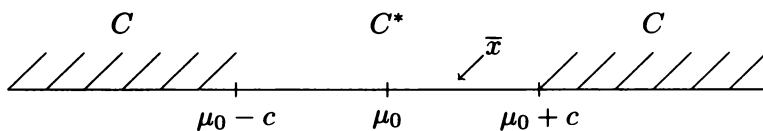


Figura 7.1

o bien, haciendo operaciones, cuando

$$|\bar{x} - \mu_0| < c.$$

Es decir, si $H_0 : \mu = \mu_0$ fuera cierta, cabría esperar que \bar{x} tomara un valor cercano a μ_0 ; en concreto del intervalo $[\mu_0 - c, \mu_0 + c]$, con gran probabilidad, $1 - \alpha$, dependiendo el valor de c de esta probabilidad.

Si observada una muestra concreta, \bar{x} no cae en el intervalo anterior, siguiendo la filosofía del contraste de hipótesis, rechazaremos H_0 , siendo, en consecuencia el mencionado intervalo, la región de aceptación del test.

Determinemos el valor de la constante c : si queremos que la probabilidad de cometer un error de tipo I, es decir, el nivel de significación sea α , deberá ser

$$P\{\bar{x} \in C\} = P\{|\bar{x} - \mu_0| > c\} = \alpha$$

es decir,

$$P\{|\bar{x} - \mu_0| < c\} = 1 - \alpha$$

cuando H_0 es cierta, es decir cuando $\mu = \mu_0$.

Ahora debemos distinguir diversas situaciones. Si por ejemplo admitimos un modelo poblacional normal, es decir que $X \sim N(\mu, \sigma)$, sabemos por la Sección 5.4 que, al no conocer la varianza poblacional, la distribución de \bar{x} es

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

con lo que, en la expresión anterior, c deberá ser tal que

$$P \left\{ |t_{n-1}| < \frac{c \sqrt{n}}{S} \right\} = 1 - \alpha$$

es decir,

$$c = t_{n-1; \alpha/2} \frac{S}{\sqrt{n}}$$

llevándonos, en definitiva, nuestros razonamientos intuitivos a considerar como test de hipótesis para contrastar a nivel α , $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ el siguiente,

$$\left\{ \begin{array}{ll} \text{Se acepta } H_0 \text{ si} & \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1; \alpha/2} \\ \text{Se rechaza } H_0 \text{ si} & \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} > t_{n-1; \alpha/2} \end{array} \right.$$

La Estadística Matemática nos dice que este test es óptimo en el sentido que mencionábamos más arriba.

En concreto, si elegida una muestra aleatoria simple de tamaño $n = 10$ se obtuvo una media muestral $\bar{x} = 202$ y una cuasivarianza muestral de $S^2 = 289$, el contraste $H_0 : \mu = 200$ frente a $H_1 : \mu \neq 200$ lleva a aceptar H_0 a nivel $\alpha = 0'05$ por ser

$$\frac{|202 - 200|}{\sqrt{289/10}} = 0'372 < 2'262 = t_{9; 0'025}$$

es decir, a concluir con la no existencia de diferencia significativa a ese nivel.

La deducción exacta de cada contraste óptimo depende de la situación concreta que se tenga: hipótesis de normalidad, muestras grandes, etc., ya que cada una de estas situaciones implica una distribución en el muestreo del estadístico a considerar.

De hecho, la determinación del estadístico a considerar en cada caso —es decir, la forma del contraste— es mucho más compleja que la determinación del estimador natural (método de la máxima verosimilitud) o la del intervalo de confianza correspondiente, ya omitida en el capítulo anterior, y depende de sofisticados métodos de la Estadística Matemática, los cuales están fuera del alcance de este libro. Si se quiere consultar más a fondo esta cuestión puede verse el texto de Vélez y García (1993).

No obstante, el lector no debe preocuparse por esta cuestión, de índole matemática, debiendo prestar atención a todo el proceso que un contraste de hipótesis conlleva. Una vez establecido con todo rigor el problema, la elección de la regla óptima será inmediata en los casos considerados en el libro, los cuales serán tratados en las siguientes secciones y en capítulos posteriores.

Relación entre intervalos de confianza y tests de hipótesis

En el ejemplo anterior, aceptábamos $H_0 : \mu = \mu_0$ cuando

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1; \alpha/2}$$

o bien, haciendo operaciones, cuando

$$\mu_0 \in \left[\bar{x} - t_{n-1; \alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1; \alpha/2} \frac{S}{\sqrt{n}} \right]$$

es decir, cuando la hipótesis nula pertenece al intervalo de confianza correspondiente.

Éste es un hecho bastante frecuente, aunque no una propiedad general, de los contrastes del tipo $H_0 : \theta = \theta_0$ frente a $H_0 : \theta \neq \theta_0$. El intervalo de confianza, de coeficiente de confianza uno menos el nivel de significación, constituye la región de aceptación del test.

Tests de hipótesis unilaterales

Supongamos en el ejemplo antes considerado, que el producto en cuestión es un *snack* elaborado con un determinado aceite. El interés estará entonces centrado en saber si este producto aumenta el nivel medio de colesterol o no. Es decir, en contrastar las hipótesis $H_0 : \mu \leq 200$ frente a $H_1 : \mu > 200$.

Ahora parece claro que la región crítica sea unilateral, Figura 7.2, del tipo $\mu_0 + c$.

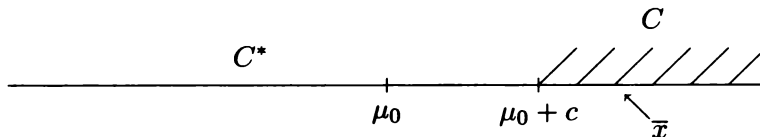


Figura 7.2

Si la probabilidad de error de tipo I es de nuevo α , deberá ser

$$P_{\mu=\mu_0} \{ \bar{x} > \mu_0 + c \} = \alpha.$$

Si admitimos la misma situación poblacional anterior, será de nuevo

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

con lo que en la expresión anterior, c deberá ser tal que

$$P \left\{ t_{n-1} > \frac{c \sqrt{n}}{S} \right\} = \alpha$$

es decir,

$$c = t_{n-1; \alpha} \frac{S}{\sqrt{n}}$$

con lo que se llegaría, en definitiva, a considerar como test de nivel α para contrastar $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ el siguiente,

$$\begin{cases} \text{Se acepta } H_0 \text{ si } & \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq t_{n-1; \alpha} \\ \text{Se rechaza } H_0 \text{ si } & \frac{\bar{x} - \mu_0}{S/\sqrt{n}} > t_{n-1; \alpha} \end{cases}$$

En el ejemplo considerado, al ser

$$\frac{202 - 200}{\sqrt{289/10}} = 0'372 < 1'833 = t_{9; 0'05}$$

se acepta $H_0 : \mu \leq 200$ al contrastarla frente a $H_1 : \mu > 200$, a nivel $\alpha = 0'05$.

P-valor

Una crítica que puede plantearse el lector respecto a la técnica de los tests de hipótesis, es la dependencia de nuestros resultados en el nivel de significación α elegido antes de efectuar el contraste.

Así surge de forma natural la pregunta: ¿Qué hubiera pasado en el ejemplo anterior si hubiéramos elegido otro α mucho mayor? ¿Se seguiría aceptando H_0 ?

La respuesta evidente es que depende de lo grande que sea α . Si para fijar ideas nos centramos en el contraste unilateral, al ser

$$\frac{\bar{x} - \mu_0}{S/\sqrt{10}} \rightsquigarrow t_9$$

y haber resultado un valor para el estadístico del contraste

$$\frac{\bar{x} - \mu_0}{S/\sqrt{10}} = \frac{202 - 200}{\sqrt{289/10}} = 0'372$$

si hubiéramos elegido por ejemplo $\alpha = 0'4$, hubiéramos rechazado H_0 , ya que $t_{9;0'4} = 0'261 < 0'372$, aunque obsérvese que en este caso la probabilidad de equivocarnos —rechazar H_0 siendo cierta— hubiera sido muy grande, $\alpha = 0'4$.

Parece razonable, por tanto, que independientemente del nivel de significación que hubiéramos elegido, debamos aceptar H_0 , puesto que el nivel de significación más pequeño que hubiéramos tenido que elegir para rechazar H_0 es demasiado grande como para admitir tal probabilidad de error de tipo I.

Este *nivel de significación observado* recibe el nombre de *p-valor* y se define con más precisión como el mínimo nivel de significación necesario para rechazar H_0 .

Obsérvese que al realizar un contraste de hipótesis debemos fijar un nivel de significación antes de tomar la muestra, que habitualmente suele ser $0'1$, $0'05$ ó $0'01$, y para ese nivel de significación elegido, aceptar o rechazar H_0 . Es decir, siempre se llega, por tanto, a una conclusión.

El cálculo del p-valor permite valorar la decisión ya tomada de rechazar o aceptar H_0 , de forma que un p-valor grande —digamos $0'2$ ó más— confirma una decisión de aceptación de H_0 . Tanto más nos lo confirma cuanto mayor sea el p-valor.

Por contra, un p-valor pequeño —digamos $0'01$ ó menos— confirma una decisión de rechazo de H_0 . Tanto más se nos confirmará esta decisión de rechazo cuanto menor sea el p-valor.

En situaciones intermedias, el p-valor no nos indica nada concreto salvo que quizás sería recomendable elegir otra muestra y volver a realizar el contraste.

Si una persona ha tomado una decisión que el p-valor contradice, confirmando éste precisamente la decisión contraria a la adoptada, el individuo lógicamente cambiará su decisión. Por esta razón, muchos de los usuarios de las técnicas estadísticas aplicadas no fijan ya el nivel de significación; simplemente hacen aparecer al final de sus trabajos el p-valor (el cual en muchos paquetes estadísticos se denomina *tail probability*), sacando conclusiones si éste se lo permite o simplemente indicándolo de forma que el lector las saque.

Esta postura, criticable en principio, no lo es más que la de otros investigadores que consideran —por definición— significativo un contraste para un p-valor menor que $0'05$, o la de aquellos otros que sólo contrastan hipótesis a una estrella, dos estrellas o tres estrellas, entendiendo estos niveles de significación, respectivamente como $0'1$, $0'05$ y $0'01$.

En nuestro ejemplo, el p-valor del contraste unilateral será

$$\text{p-valor} = P\{t_9 > 0'372\} = 0'35925$$

y en el bilateral

$$\text{p-valor} = P\{|t_9| > 0'372\} = 2 \cdot P\{t_9 > 0'372\} = 0'7185.$$

En su determinación hemos empleado unas tablas muy precisas de la t de Student. Por el significado del p -valor no es necesario llegar a tal precisión, ya que basta con dejarlo acotado, pudiendo haber concluido estos ejemplos, utilizando la *Tabla 5* de ADD, diciendo que en el primer caso es

$$0'3 < p\text{-valor} < 0'4$$

y en el segundo

$$0'6 < p\text{-valor} < 0'8$$

de interpretación suficientemente clara en ambos casos —aceptar H_0 —, especialmente en el segundo.

Contrastes óptimos

Ante una situación concreta que se nos plantee, la determinación del contraste óptimo, al igual que ocurría con los intervalos de confianza, dependerá fundamentalmente, de las suposiciones que se hagan en el modelo: normalidad, varianza (o varianzas) conocidas o desconocidas, muestras pequeñas o grandes, ..., y fundamentalmente de las hipótesis que se desee contrastar.

En las siguientes secciones de este capítulo, veremos los tests que la Estadística Matemática propone como óptimos en cada una de las situaciones que se consideran, para una y dos poblaciones.

Las que veremos son las que más frecuentemente suelen plantearse desde un punto de vista práctico, aunque no las únicas.

En el Capítulo 9 se aborda el problema de la comparación de más de dos poblaciones, dejando para el 8 los contrastes no paramétricos con los que se cubre gran parte de las situaciones restantes, aunque no todas.

Las reglas de actuación, es decir los tests, expuestos en éste y en los capítulos siguientes, son las consideradas como óptimas en las situaciones que se enuncian, pero su correcta utilización, o mejor dicho, el correcto planteamiento de la situación óptima ante un problema concreto que se nos presente, requerirá de mucha experiencia.

En las próximas secciones se estudian las situaciones referentes a los parámetros de una población (Secciones 7.2, 7.3 y 7.4) y luego a los de dos poblaciones independientes (Secciones 7.5, 7.6 y 7.7), finalizando, en la Sección 7.8, con el caso de datos apareados, el cual, como sabemos, se reduce adecuadamente al caso de una población.

En todas ellas se han incluido ejemplos de su aplicación y en ellas no deduciremos el test óptimo, sino que simplemente lo enunciaremos, apareciendo en ADD un resumen de todas las situaciones aquí consideradas.

Contraste de Hipótesis con R

Como hemos visto, el intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test de hipótesis bilateral. Por esta razón se utiliza una misma función de R para obtener intervalos de confianza y test de hipótesis sobre un parámetro. En concreto, la función de R que nos va a proporcionar los tests (y los intervalos) es la función `t.test` estudiada brevemente en el capítulo anterior y cuyos argumentos son

```
t.test(x, y = NULL, alternative = "two.sided", mu = 0, paired = FALSE,
       var.equal = FALSE, conf.level = 0.95)
```

Los argumentos `x` e `y` se utilizan para indicar el o los vectores de datos a utilizar en el contraste. El tercer argumento `alternative` presenta tres opciones: `two.sided`, que es la que se utiliza por defecto y que corresponde al caso de contrastes bilaterales; `greater`, correspondiente al caso de hipótesis nula *menor o igual* frente a hipótesis alternativa de *mayor*, y `less` para el caso de hipótesis nula de *mayor o igual* frente a alternativa de *menor*. Debemos especificar estas opciones entre comillas. Con el argumento `mu` indicamos el valor de la hipótesis nula.

De nuevo `paired` sirve para indicar una situación de datos apareados y `var.equal` si las varianzas poblacionales pueden considerarse o no iguales. El último argumento permite especificar el nivel de significación del test tomándose por defecto el valor 0'05.

En el caso de que queramos comparar dos muestras independientes, vamos a tener que suponer a las varianzas poblaciones como iguales o como distintas. Para contrastar cuál de las dos situaciones admitimos como válida, deberemos determinar el intervalo de confianza, o equivalentemente el test de hipótesis, de su cociente. Para ello podemos utilizar la función de R

```
var.test(x, y, ratio, alternative="two.sided", conf.level = 0.95)
```

en donde incorporamos los datos en los argumentos `x` e `y`. En `ratio` especificamos la hipótesis nula, que será `ratio = 1` si queremos contrastar la igualdad de las varianzas poblacionales. Con `alternative` indicamos el sentido de la hipótesis alternativa; como ocurría más arriba, `two.sided`, es la opción que se utiliza por defecto y que corresponde al caso de *igual* frente a *distinta*; `greater`, correspondiente al caso de hipótesis alternativa *mayor*, y `less` para el caso de hipótesis alternativa *menor*.

Advertimos que esta función da el intervalo de confianza expresado en la Sección 6.5 (para el caso de medias desconocidas) y no la región de aceptación con extremos los cuantiles de la F de Snedecor de la Sección 7.5. Es decir,

para contrastar, por ejemplo, la hipótesis nula de igualdad de ambas varianzas poblacionales, debemos analizar si el 1 pertenece o no a la región de aceptación (intervalo de confianza) dado por la función `var.test`; no si S_1^2/S_2^2 pertenece.

Por último, en la ejecución de tests de comparación de los parámetros de dos poblaciones binomiales, debemos utilizar la misma función de R `prop.test` que utilizamos para obtener intervalos de confianza para la diferencia de proporciones binomiales,

```
prop.test(x, n, alternative = "two.sided", conf.level = 0.95, correct = TRUE)
```

con una interpretación de sus argumentos ya comentada.

7.2. Contraste de hipótesis relativas a la media de una población normal

Supongamos que tenemos una muestra aleatoria simple X_1, \dots, X_n procedente de una población $N(\mu, \sigma)$ y que queremos contrastar hipótesis relativas a la media de la población, μ .

En primer lugar consideraremos el caso de *igual* frente a *distinta*, es decir, el caso en que queremos contrastar si puede admitirse para la media poblacional un determinado valor μ_0 o no.

$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$
--

En este caso, al igual que ocurre con casi todos los de *igual* frente a *distinta*, la región de aceptación se corresponde con el intervalo de confianza determinado en el capítulo anterior, aceptándose H_0 cuando y sólo cuando ésta pertenezca al intervalo de confianza.

Así, si suponemos σ **conocida**, fijado un nivel de significación α , aceptaremos $H_0 : \mu = \mu_0$ cuando y sólo cuando

$$\mu_0 \in \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

o equivalentemente, haciendo operaciones, cuando

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

con lo que podemos concluir diciendo que el test óptimo en esta situación es

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$

Ejemplo 7.1

Hace 10 años se realizó, en una determinada población, un estudio sobre su estatura cuyo histograma sugirió para dicha variable una distribución normal de media 1'68 m. y desviación típica 6'4 cm.

Ahora se quiere analizar si la estatura media de dicha población ha variado con el tiempo, para lo que se tomó una muestra de tamaño $n = 15$, la cual dio como resultado una media muestral de $\bar{x} = 1'73$ m.

Admitiendo que la distribución modelo sigue siendo normal y que la dispersión en la estatura de dicha población no ha variado en estos diez años, el averiguar si la estatura media de la población se mantiene en los niveles de hace una década o si ha variado significativamente, equivale a contrastar la hipótesis nula $H_0 : \mu = 1'68$ frente a la alternativa $H_1 : \mu \neq 1'68$, en donde μ representa la estatura media poblacional en la actualidad.

Si fijamos un nivel de significación $\alpha = 0'05$, al ser

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{|1'73 - 1'68|}{0'064/\sqrt{15}} = 3'026 > 1'96 = z_{0'05/2}$$

debemos rechazar la hipótesis nula H_0 de que la estatura media de la población no ha variado de forma significativa en estos 10 años.

Utilizando la *Tabla 3* de ADD se obtiene que el p-valor resultante es

$$P\{|Z| > 3'026\} = 2 \cdot P\{Z > 3'026\} \simeq 0'0025$$

el cual confirma la decisión adoptada.

Si se supone a σ **desconocida**, la distribución en el muestreo de la media muestral determinada en el Capítulo 5, o directamente el intervalo de confianza calculado en el 6, llevan a considerar que el test óptimo en este caso es

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1;\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} > t_{n-1;\alpha/2}$

a nivel de significación α .

Ejemplo 7.1 (continuación)

Si no se tiene certeza de que la varianza haya permanecido inalterable en los diez años, y la muestra obtenida hubiera dado una cuasivarianza muestral de $0'64 \text{ m}^2$ (la varianza se expresa en unidades al cuadrado), podíamos haber contrastado las hipótesis anteriores, $H_0 : \mu = 1'68$ frente a $H_1 : \mu \neq 1'68$, utilizando un test de la t de Student, que al mismo nivel hubiera aceptado también H_0 al ser

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} = \frac{|1'73 - 1'68|}{\sqrt{0'64/15}} = 0'242 < 2'145 = t_{14;0'05/2}.$$

De la *Tabla 5* de ADD se deduce que el p-valor es

$$\text{p-valor} = 2 \cdot P\{t_{14} > 0'242\} > 2 \cdot P\{t_{14} > 0'258\} = 2 \cdot 0'4 = 0'8$$

acotación suficiente para confirmar la aceptación de H_0 .

$$\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}$$

El estudio de los contrastes unilaterales es de suma importancia en el análisis de la efectividad de nuevos productos, donde el aumento de su efectividad ($H_1 : \mu > \mu_0$) o la disminución de alguna característica negativa asociada, como por ejemplo el tiempo que tarda en hacer efecto ($H_1 : \mu < \mu_0$) son las hipótesis de interés.

En estos casos, el objetivo es rechazar H_0 con un p-valor pequeño, lo que conduce a quedarnos con la hipótesis de interés H_1 , con un error pequeño en la inferencia, el error de rechazar H_0 siendo cierta, error suministrado por el p-valor.

La distribución en el muestreo de \bar{x} en los supuestos que se establecen, así como las consideraciones hechas al hablar de las hipótesis unilaterales, llevan a la Estadística Matemática a proponer como test óptimo para contrastar $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$,

Si σ es conocida

El test óptimo indica que

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

Si σ es desconocida

En este caso, el test óptimo indica que

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq t_{n-1;\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} > t_{n-1;\alpha}$

Ejemplo 7.2

Un laboratorio farmacéutico piensa que un nuevo medicamento fabricado por ellos prolonga significativamente la vida de los enfermos de SIDA, establecida en la actualidad en una media de dos años desde que la enfermedad se manifiesta.

Con objeto de validar su nuevo producto, y admitiendo que el tiempo de vida sigue una distribución normal de media μ , el laboratorio contrastó la hipótesis nula $H_0 : \mu \leq 2$ frente a la alternativa $H_1 : \mu > 2$, utilizando una muestra aleatoria de $n = 18$ pacientes, la cual le proporcionó una media de $\bar{x} = 2'8$ años y una cuasidesviación típica muestral de $S = 1'2$ años. Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{2'8 - 2}{1'2/\sqrt{18}} = 2'8284$$

el laboratorio rechazaría H_0 —validando en consecuencia su producto— con un p-valor suficientemente pequeño: aproximadamente igual a 0'005.

$$\begin{aligned} H_0 : \mu &\geq \mu_0 \\ H_1 : \mu &< \mu_0 \end{aligned}$$

Los mismos razonamientos anteriores llevan a proponer los siguientes tests para las hipótesis simétricas aquí consideradas.

Si σ es conocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$

Si σ es desconocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \geq t_{n-1;1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} < t_{n-1;1-\alpha}$

Ejemplo 7.3

La rapidez con la que un determinado medicamento actúa es esencial ante infartos agudos de miocardio. Los fármacos que se administran en la actualidad tardan en actuar una media de 30 segundos.

Un laboratorio afirma que el producto recién elaborado por ellos, actúa en menos tiempo. ¿Podemos recomendar su utilización?

El contraste de hipótesis que se plantea es $H_0 : \mu \geq 30$ frente a $H_1 : \mu < 30$. Si una muestra de $n = 10$ pacientes dio un tiempo medio de reacción de 28 segundos y una cuasivarianza de $S^2 = 16$ segundos al cuadrado, no podemos rechazar H_0 a nivel $\alpha = 0.05$ ya que

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{28 - 30}{4/\sqrt{10}} = -1.58 > -1.833 = t_{9;0.95}$$

con lo que podemos concluir afirmando que no existen evidencias claras de la efectividad del nuevo producto al nivel indicado.

7.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes

La obtención de tamaños muestrales suficientemente grandes —digamos mayores de 30— evita la obligación de suponer normalidad en la distribución modelo, alcanzándose, no obstante, resultados análogos a cuando se verifica tal suposición.

La normalidad en la distribución asintótica de \bar{x} , añade la peculiaridad de hacer que los puntos críticos sean ahora abscisas de normales estándar, tanto si la varianza poblacional es conocida como si no lo es.

Población no necesariamente normal

Supongamos que X_1, \dots, X_n es una muestra aleatoria simple de tamaño suficientemente grande como para poder admitir como distribución asintótica de \bar{x} la siguiente,

$$\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

En este caso, considerando los tres tipos de tests y distinguiendo, de nuevo, la situación en la que la varianza es conocida y la situación en la que es desconocida, tenemos los siguientes contrastes,

$$\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}$$

σ conocida

El test óptimo que se propone es la siguiente regla de actuación

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$

σ desconocida

Si σ es desconocida, entonces el test óptimo es

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} > z_{\alpha/2}$

Ejemplo 7.4

Un grupo de arqueólogos considera que la capacidad craneal es el factor determinante en la clasificación de restos humanos del paleolítico, variable que se admite sigue una distribución normal. En concreto, una capacidad craneal de 1500 cm^3 lleva a clasificar a un esqueleto como de *raza Neanderthal*.

Ante el hallazgo de 8 esqueletos en una necrópolis de la mencionada época, los arqueólogos calcularon una capacidad craneal media en dichos restos de 1450 cm^3 y una desviación típica muestral de 10 cm^3 .

En estas condiciones, la determinación de si los restos hallados pueden considerarse como de *raza Neanderthal* puede conseguirse contrastando la hipótesis nula $H_0 : \mu = 1500$ frente a

$H_1 : \mu \neq 1500$ en donde μ representa la capacidad craneal media de la población de restos encontrados. Como es

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} = \frac{|1450 - 1500|}{10'69/\sqrt{8}} = 13'23 > 3'49 = z_{0'0004/2}$$

la conclusión que puede sacarse es que claramente los restos no eran de *raza Neanderthal*.

$$H_0 : \mu \leq \mu_0$$

$$H_1 : \mu > \mu_0$$

Si σ es conocida

$$\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$$

$$\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$$

Si σ es desconocida

$$\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq z_\alpha$$

$$\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x} - \mu_0}{S/\sqrt{n}} > z_\alpha$$

Ejemplo 7.5

En una muestra de 49 adolescentes que sirvieron de sujetos en un estudio inmunológico, una variable de interés fue el diámetro de reacción en la piel ante un antígeno. La media y la desviación típica muestrales fueron 39 y 11 mm. respectivamente.

Si la reacción media habitual es de 30 mm. cabe preguntarse si la reacción observada fue mayor de lo esperado. Es decir, parece razonable contrastar la hipótesis nula $H_0 : \mu \leq 30$ frente a la alternativa $H_1 : \mu > 30$.

Obsérvese que no tiene sentido plantearse el contraste de las hipótesis complementarias $H_0 : \mu \geq 30$ frente $H_1 : \mu < 30$, ya que éste tiene como región crítica la cola de la izquierda y, al haberse observado una media muestral mayor que la hipótesis nula, siempre se aceptaría H_0 . Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{39 - 30}{11'114/\sqrt{49}} = 5'6685 > 1'645 = z_{0'05}$$

rechazaremos la hipótesis nula a nivel $\alpha = 0'05$. El p-valor

$$P\{Z > 5'6685\} < 0'0002$$

confirma, fuertemente, esta decisión.

$$H_0 : \mu \geq \mu_0$$

$$H_1 : \mu < \mu_0$$

Si σ es conocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$

Si σ es desconocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} < z_{1-\alpha}$

Ejemplo 7.6

Un grupo de historiadores norteamericanos está interesado en averiguar si la edad media de los soldados de la Unión en la época previa a la guerra civil americana de 1861 era menor de 30 años.

Con este propósito el grupo consideró *Fort Moultrie*, en Carolina del Sur, suficientemente representativo de los 75 fuertes con los que contaba Estados Unidos en 1850, eligiendo de allí una muestra de tamaño $n = 45$ para la que se obtuvo, según el Censo de Carolina del Sur de 1850, una media de $\bar{x} = 28'3$ años y una cuasidesviación típica $S = 5'96$.

Planteando el contraste de las hipótesis $H_0 : \mu \geq 30$ frente a $H_1 : \mu < 30$ y dado que el tamaño muestral es suficientemente grande, la suposición de normalidad para la variable edad no es requerida. Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{28'3 - 30}{5'96/\sqrt{45}} = -1'91 < -1'645 = z_{1-0'05}$$

podemos rechazar H_0 a nivel $\alpha = 0'05$, infiriendo, por tanto, una edad significativamente inferior a 30 años en los soldados, aunque con un p-valor, 0'0281, no concluyente.

Población binomial

La situación que se analiza en este apartado es la de una muestra aleatoria simple X_1, \dots, X_n de variables $B(1, p)$, es decir, variables que toman sólo los valores 1 (éxito), 0 (fracaso), siendo la probabilidad de éxito el parámetro p .

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Como el resto de los contrastes de *igual* frente a *distinta*, la región de aceptación del test óptimo se corresponde con el intervalo de confianza calculado en el capítulo anterior, aceptándose $H_0 : \mu = \mu_0$ cuando μ_0 pertenezca a dicho intervalo. Hacemos la observación de que allí, al ser la varianza de la proporción muestral, $p(1 - p)/n$, desconocida por depender del parámetro p , la estimamos con la proporción muestral mediante $\hat{p}(1 - \hat{p})/n$. Aquí, sin embargo, dado que los tests se realizan bajo H_0 , es decir, suponiendo que es cierta la hipótesis nula, ésta implica suponer como varianza de \hat{p} el valor $p_0(1 - p_0)/n$ (si es $p_0 \neq 0$). Ya adelantamos que esto mismo ocurrirá con la distribución de Poisson en el siguiente apartado.

En el caso que aquí nos ocupa,

- Se acepta H_0 si $\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1 - p_0)}{n}}} > z_{\alpha/2}$

Ejemplo 7.7

El dueño de un restaurante desea conocer si el porcentaje de clientes satisfechos con la relación calidad/precio de su comida ronda el 70 %.

Si éste fuera significativamente menor, incrementaría la calidad de sus comidas, mientras que si fuera claramente mayor, abarataría ligeramente los costes con objeto de obtener un mayor beneficio.

Para ello encuestó a la salida a 100 clientes, de los que obtuvo un porcentaje del 60 % de satisfechos.

Contrastadas las hipótesis $H_0 : p = 0'7$ frente a $H_1 : p \neq 0'7$, a nivel $\alpha = 0'05$ se obtuvo que

$$\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{|0'6 - 0'7|}{\sqrt{\frac{0'7 \cdot 0'3}{100}}} = 2'182 > 1'96 = z_{0'05/2}$$

con lo que se rechaza H_0 . Deberá mejorar la calidad de su comida o disminuir el precio de la misma.

$$H_0 : p \leq p_0$$

$$H_1 : p > p_0$$

En esta situación el test óptimo es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 & \text{ si } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_\alpha \\ \bullet \text{ Se rechaza } H_0 & \text{ si } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_\alpha \end{aligned}$$

Ejemplo 7.8

Antes de modificar el código de circulación y obligar a todos los conductores a usar cinturón de seguridad en las vías urbanas, se quiso averiguar si por propia iniciativa, la proporción de conductores que ya lo hacía era significativamente mayor del 60 %.

Con este propósito se contrastó a nivel 0'01 la hipótesis nula $H_0 : p \leq 0'6$ frente a la alternativa $H_1 : p > 0'6$, utilizando una muestra aleatoria de 35 conductores, lo que dio una proporción muestral del 70 %. Como

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0'7 - 0'6}{\sqrt{\frac{0'6 \cdot 0'4}{35}}} = 1'2076 < 2'325 = z_{0'01}$$

la Dirección General de Tráfico obligó a usar cinturón de seguridad en las vías urbanas.

Obsérvese que, en este último ejemplo, el nivel de significación fijado es

relativamente pequeño, ya que se quiere tener una cierta seguridad de que, si no se obliga a utilizar el cinturón de seguridad — porque se hubiese rechazado H_0 y en consecuencia aceptado H_1 — la probabilidad de error sea pequeña. Es decir, más vale obligar a utilizarlo, aunque la mayoría ya lo haga, que no obligarlo erróneamente. Se desea que la muestra dé una evidencia suficientemente clara en favor de H_1 , es decir una diferencia fuertemente significativa — mayor de 2'325 — respecto de la hipótesis nula.

$$H_0 : p \geq p_0$$

$$H_1 : p < p_0$$

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{1-\alpha} \end{aligned}$$

Ejemplo 7.9

Una Comunidad Autónoma española está interesada en averiguar si el índice de absentismo laboral es menor en dicha comunidad que la media europea, en donde se sitúa en el 9%.

Con este propósito, seleccionó al azar una muestra de 200 trabajadores de la comunidad, la cual proporcionó un porcentaje de absentismo del 5%.

Para averiguar si este porcentaje es significativamente menor de la media comunitaria, realizó el contraste de hipótesis $H_0 : p \geq 0'09$ frente a $H_1 : p < 0'09$, obteniendo, a nivel $\alpha = 0'025$,

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0'05 - 0'09}{\sqrt{\frac{0'09 \cdot 0'91}{200}}} = -1'9767 < -1'96 = z_{1-0'025}$$

es decir, que sí lo era.

Población Poisson

La situación que tenemos en este apartado es la de una muestra aleatoria simple X_1, \dots, X_n , también de tamaño suficientemente grande — n al menos 30 —, extraída de una población de Poisson de parámetro λ , $\mathcal{P}(\lambda)$, siendo los contrastes a considerar relativos a dicho parámetro.

Obsérvese que estos contrastes aparecen en la sección relativa a los contrastes de la media; de hecho λ es la media de la distribución.

Recuérdese, sin embargo, que λ también es su varianza, por lo que si queremos hacer contrastes sobre la varianza de una población y, después de analizada ésta, se considera en ella un modelo Poisson, los contrastes relativos a su varianza son los que aparecen a continuación.

$$H_0 : \lambda = \lambda_0$$

$$H_1 : \lambda \neq \lambda_0$$

En este caso, el test óptimo sugiere que

- Se acepta H_0 si $\frac{|\bar{x} - \lambda_0|}{\sqrt{\lambda_0/n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \lambda_0|}{\sqrt{\lambda_0/n}} > z_{\alpha/2}$

Ejemplo 7.10

Con objeto de valorar si un aeropuerto era utilizado de forma óptima, se observó el número de aterrizajes por minuto en 50 intervalos de un minuto elegidos al azar.

Si el número medio de dichos aterrizajes es 1, el aeropuerto se considera adecuadamente utilizado. En caso contrario serían necesarias reformas en el sentido, bien de ampliarlo o bien de reducir servicios.

Como para el número de aterrizajes por minuto puede admitirse una distribución de Poisson, un contraste de la forma $H_0 : \lambda = \lambda_0$ frente a $H_1 : \lambda \neq \lambda_0$, responderá a la cuestión planteada.

Fijado un nivel de significación $\alpha = 0.05$, y observados una media de 1.3 aterrizajes por minuto en los 50 elegidos al azar, se rechaza H_0 al ser

$$\frac{|\bar{x} - \lambda_0|}{\sqrt{\lambda_0/n}} = \frac{|1.3 - 1|}{\sqrt{1/50}} = 2.12 > 1.96 = z_{0.05/2}.$$

El p-valor es $P\{|Z| > 2.12\} = 0.034$, lo que indica que la decisión no es clara.

$$H_0 : \lambda \leq \lambda_0$$

$$H_1 : \lambda > \lambda_0$$

En esta situación, el contraste óptimo indica seguir la siguiente regla de actuación:

- Se acepta H_0 si $\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} \leq z_\alpha$
- Se rechaza H_0 si $\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} > z_\alpha$

Ejemplo 7.11

Un grupo de municipios rurales de una determinada región solicitó ayudas económicas al gobierno de la nación, argumentando que en dicha región el número medio de tormentas destructivas semanales era relativamente alto.

Con objeto de tomar una decisión al respecto, el gobierno encargó a sus técnicos que realizaran un contraste de hipótesis de la forma $H_0 : \lambda \leq \lambda_0$ frente a $H_1 : \lambda > \lambda_0$, ya que, en base a su experiencia, el gabinete considera que el número de tormentas semanales con esas características es una variable aleatoria con distribución de Poisson de parámetro λ , estando dispuesto a conceder la ayuda solicitada solamente en el caso de que ese número medio sea mayor que 2.

Con este propósito, los técnicos gubernamentales consultaron al servicio meteorológico nacional observando en 30 semanas de dicha zona una media de 2'5 tormentas destructivas a la semana.

Dado que ese gobierno es muy estricto en la concesión de determinadas ayudas, fijó como nivel de significación $\alpha = 0'01$, con objeto de conceder la ayuda solamente en el caso de que hubiera una gran evidencia en contra de una situación de normalidad, H_0 . Dado que

$$\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} = \frac{2'5 - 2}{\sqrt{2/30}} = 1'9365 < 2'325 = z_{0'01}$$

el gobierno no concedió la ayuda.

Solicitada la ayuda al gobierno autónomo, menos estricto en dicha concesión de fondos públicos, éste fijó como nivel de significación $\alpha = 0'1$, con lo que al ser $z_{0'1} = 1'285 < 1'9365$, el gobierno autónomo rechazó la hipótesis nula, concediendo la ayuda solicitada.

$$H_0 : \lambda \geq \lambda_0$$

$$H_1 : \lambda < \lambda_0$$

- Se acepta H_0 si $\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} < z_{1-\alpha}$

Ejemplo 7.12

Se admite como razonable el suponer una distribución de Poisson para el número de partículas radioactivas emitidas por un determinado elemento.

Con objeto de permitir o no la instalación de una fábrica en un polígono industrial cerca de un núcleo urbano, fábrica que usa ese elemento radioactivo en la elaboración de sus productos, se desea averiguar si la contaminación radioactiva que produce dicha fábrica es menor que 10.

Para ello se realizó el contraste de hipótesis $H_0 : \lambda \geq 10$ frente a $H_1 : \lambda < 10$, a nivel $\alpha = 0'01$, utilizando una muestra aleatoria de tamaño 30, consistente en 30 mediciones al azar en la zona, la cual proporcionó una media de $\bar{x} = 8'7$. Como

$$\frac{\bar{x} - \lambda_0}{\sqrt{\lambda_0/n}} = \frac{8'7 - 10}{\sqrt{10/30}} = -2'252 > -2'325 = z_{1-0'01}$$

se acepta la hipótesis nula, no permitiendo la instalación de la fábrica.

Obsérvese que esta conclusión ha sido adoptada porque hemos impuesto una probabilidad muy pequeña (0'01) de error, permitiendo la instalación de una fábrica que emite radiación elevada. Pero si hubiéramos aceptado como válida una probabilidad de error de 0'05, al ser el punto crítico $z_{0'95} = -1'645 > -2'252$ hubiéramos rechazado H_0 y admitido la instalación de la fábrica. De nuevo, ante la trascendencia de la decisión, se quiso estar muy seguro antes de rechazar erróneamente H_0 .

7.4. Contraste de hipótesis relativas a la varianza de una población normal

En toda la sección supondremos que tenemos una muestra X_1, \dots, X_n de una población normal $N(\mu, \sigma)$ y que estamos interesados en realizar contrastes sobre la varianza de dicha distribución.

Apuntemos, además, que las hipótesis referentes a la desviación típica se contrastarían utilizando las raíces cuadradas de los tests que aparecen a continuación.

$$\begin{aligned} H_0 : \sigma^2 &= \sigma_0^2 \\ H_1 : \sigma^2 &\neq \sigma_0^2 \end{aligned}$$

 μ conocida

Si la media es conocida, el test óptimo a utilizar de nivel de significación α , es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \in \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \notin \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \end{aligned}$$

 μ desconocida

En este caso la regla a utilizar será

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{(n-1)S^2}{\sigma_0^2} \in \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{(n-1)S^2}{\sigma_0^2} \notin \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \end{aligned}$$

Ejemplo 7.13

Se realizó un experimento con objeto de analizar la destreza de 18 estudiantes de enfermería, observando en ellos una medida de la destreza manual, la cual dio una cuasivarianza muestral de $S^2 = 1349$.

Supuesto que esta medida de la destreza sigue una distribución normal, ¿puede concluirse que la varianza poblacional es diferente de 2600, a nivel $\alpha = 0.05$?

Al no suponerse la media poblacional conocida, utilizaremos el segundo test. Como es

$$\left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] = \left[\chi_{17; 1-0.025}^2, \chi_{17; 0.025}^2 \right] = [7'564, 30'19]$$

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{17 \cdot 1349}{2600} = 8'82 \in [7'564, 30'19]$$

no podemos rechazar H_0 a ese nivel. Por interpolación lineal, de la *Tabla 4* de ADD obtenemos que el p-valor es

$$2 \cdot P\{\chi_{17}^2 < 8'82\} = 2 \cdot 0'05522 = 0'11044$$

bastante claro en la aceptación de la hipótesis nula.

$$\begin{array}{l} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}$$

μ conocida

En este caso el test óptimo es

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \leq \chi_{n;\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} > \chi_{n;\alpha}^2 \end{array}$$

μ desconocida

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1;\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} > \chi_{n-1;\alpha}^2 \end{array}$$

Ejemplo 7.14

Con objeto de estudiar la cantidad de proteínas contenidas en el líquido amniótico, se seleccionaron al azar 16 mujeres embarazadas, obteniéndose una cuasidesviación típica muestral

de $S = 0'7$ gramos por cada 100 ml. Admitiendo normalidad en dicha variable, contrastar, a nivel $0'05$, si la desviación típica poblacional puede considerarse mayor que $0'6$.

Como es $\chi^2_{15;0'05} = 25$ y

$$\frac{S \sqrt{n-1}}{\sigma_0} = \frac{0'7 \sqrt{15}}{0'6} = 4'518 < 5$$

se acepta $H_0 : \sigma \leq 0'6$. El p-valor

$$P \left\{ \sqrt{\chi^2_{15}} > 4'518 \right\} = P \{ \chi^2_{15} > 20'41 \}$$

será $0'1 < \text{p-valor} < 0'3$, bastante claro en la aceptación de H_0 .

$$\begin{array}{l} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{array}$$

μ conocida

En esta situación, el test óptimo indica que

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \geq \chi^2_{n;1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} < \chi^2_{n;1-\alpha} \end{array}$$

μ desconocida

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} \geq \chi^2_{n-1;1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1;1-\alpha} \end{array}$$

Ejemplo 7.15

Los pesos de 30 bebés recién nacidos que habían sido elegidos al azar, dieron una cuasidesviación típica muestral de 165 gramos. Admitiendo que los pesos en los recién nacidos siguen una distribución normal, contrastar las hipótesis $H_0 : \sigma^2 \geq 32000$ frente a $H_1 : \sigma^2 < 32000$, a nivel $\alpha = 0.05$.

Como es $\chi^2_{n-1;1-\alpha} = \chi^2_{29;0.95} = 17.71$ y

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{29 \cdot 165^2}{32000} = 24.67 > 17.71$$

se acepta H_0 . Además, el p-valor = $P\{\chi^2_{29} < 24.67\} \simeq 0.3$, confirma esta decisión.

7.5. Contraste de hipótesis relativas a las varianzas de dos poblaciones normales independientes

En esta sección se aborda el problema de la comparación de las varianzas de dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, utilizando muestras aleatorias de ambas poblaciones X_1, \dots, X_{n_1} , e Y_1, \dots, Y_{n_2} .

$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$
--

Los contrastes que aquí veremos son importantes por sí mismos y, especialmente, cuando comparemos las medias de dos poblaciones, ya que entonces habrá que distinguir si las varianzas poblacionales son o no iguales a la hora de aplicar el test adecuado.

μ_1 y μ_2 conocidas

Si las medias poblacionales son conocidas, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} \in \left[F_{n_1, n_2; 1 - \frac{\alpha}{2}}, F_{n_1, n_2; \frac{\alpha}{2}} \right] \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} \notin \left[F_{n_1, n_2; 1 - \frac{\alpha}{2}}, F_{n_1, n_2; \frac{\alpha}{2}} \right]
 \end{aligned}$$

μ_1 y μ_2 desconocidas

El test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{S_1^2}{S_2^2} \in \left[F_{n_1-1, n_2-1; 1 - \frac{\alpha}{2}}, F_{n_1-1, n_2-1; \frac{\alpha}{2}} \right] \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{S_1^2}{S_2^2} \notin \left[F_{n_1-1, n_2-1; 1 - \frac{\alpha}{2}}, F_{n_1-1, n_2-1; \frac{\alpha}{2}} \right]
 \end{aligned}$$

Ejemplo 7.16

Con objeto de averiguar si difieren significativamente los pesos medios de los individuos de dos poblaciones independientes, se eligieron al azar 13 individuos de la primera y 10 individuos de la segunda, obteniéndose los siguientes resultados

$$\begin{aligned}
 \bar{x}_1 &= 63'9 & S_1 &= 9'16 \\
 \bar{x}_2 &= 67'8 & S_2 &= 8'37
 \end{aligned}$$

Como veremos en la siguiente sección, para contrastar la igualdad de las medias poblacionales, necesitaremos averiguar primero si las varianzas poblacionales pueden suponerse iguales o no. Es decir, contrastar $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$. Si $\alpha = 0'1$, al ser

$$\begin{aligned}
 I &= \left[F_{n_1-1, n_2-1; 1 - \frac{\alpha}{2}}, F_{n_1-1, n_2-1; \frac{\alpha}{2}} \right] = [F_{12, 9; 0'95}, F_{12, 9; 0'05}] = \\
 &= [1/F_{9, 12; 0'05}, F_{12, 9; 0'05}] = [0'3576, 3'0729]
 \end{aligned}$$

y, al ser, $S_1^2/S_2^2 = 1'977 \in I$, aceptaremos H_0 .

$$\begin{array}{l} H_0 : \sigma_1^2 \leq \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{array}$$

μ_1 y μ_2 conocidas

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} \leq F_{n_1, n_2; \alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} > F_{n_1, n_2; \alpha} \end{array}$$

μ_1 y μ_2 desconocidas

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{S_1^2}{S_2^2} \leq F_{n_1-1, n_2-1; \alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{S_1^2}{S_2^2} > F_{n_1-1, n_2-1; \alpha} \end{array}$$

Ejemplo 7.17

Con objeto de medir el efecto de una experiencia traumática en los niveles de glucosa en la sangre, se formaron al azar dos poblaciones, la I constituida por ratones sometidos a la experiencia traumática y la II, formada por un grupo control. Los valores muestrales obtenidos fueron

$$\begin{array}{lll} n_1 = 12 & \bar{x}_1 = 70 & S_1 = 7'3 \\ n_2 = 14 & \bar{x}_2 = 81 & S_2 = 6'5 \end{array}$$

A nivel $\alpha = 0'05$, y admitiendo que los niveles de glucosa en la sangre siguen distribuciones normales, ¿puede considerarse mayor la varianza de la primera población?

Como es

$$\frac{S_1^2}{S_2^2} = \frac{53'29}{42'25} = 1'2613 < 2'63735 = F_{11,13;0'05}$$

se aceptará $H_0 : \sigma_1^2 \leq \sigma_2^2$, no pudiendo considerarse mayor σ_1^2 que σ_2^2 .

$$\begin{array}{l} H_0 : \sigma_1^2 \geq \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{array}$$

μ_1 y μ_2 conocidas

En esta situación, el test óptimo indica que

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} \geq F_{n_1, n_2; 1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^{n_1} (X_i - \mu_1)^2 / n_1}{\sum_{j=1}^{n_2} (Y_j - \mu_2)^2 / n_2} < F_{n_1, n_2; 1-\alpha} \end{array}$$

μ_1 y μ_2 desconocidas

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{S_1^2}{S_2^2} \geq F_{n_1-1, n_2-1; 1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{S_1^2}{S_2^2} < F_{n_1-1, n_2-1; 1-\alpha} \end{array}$$

Ejemplo 7.18

Se realizó un experimento para averiguar el efecto del humo de los cigarrillos en ratones de laboratorio. Con este propósito se expuso a 11 animales al humo de los cigarrillos y otros 11 fueron tomados como grupo control, observando en ambos grupos los latidos por minuto a 20 grados centígrados.

En el primer grupo se observó una cuasivarianza de $S_1^2 = 3400$, mientras que en el segundo la cuasivarianza observada fue de $S_2^2 = 4200$. Admitiendo normalidad, ¿puede considerarse mayor la varianza poblacional del segundo grupo a nivel $\alpha = 0'05$?

El contraste que se propone es $H_0 : \sigma_1^2 \geq \sigma_2^2$ frente a $H_1 : \sigma_1^2 < \sigma_2^2$. Como es $S_1^2/S_2^2 = 0'8095$ y $F_{n_1-1, n_2-1; 1-\alpha} = F_{10, 10; 0'95} = 1/F_{10, 10; 0'05} = 0'3358$, se aceptará H_0 .

Una cuestión interesante, tanto aquí como cada vez que se comparan parámetros de dos poblaciones, es lo referente a la denominación de población I y población II. Es igual contrastar, por ejemplo, $H_0 : \sigma_1^2 \leq \sigma_2^2$ frente a $H_1 : \sigma_1^2 > \sigma_2^2$, que, llamando población I a la anterior 2 y II a la anterior 1, contrastar $H_0 : \sigma_I^2 \geq \sigma_{II}^2$ frente a $H_1 : \sigma_I^2 < \sigma_{II}^2$, ya que en el primer caso rechazaremos H_0 cuando sea

$$\frac{S_1^2}{S_2^2} > F_{n_1-1, n_2-1; \alpha}$$

mientras que en el segundo rechazaremos H_0 —obteniendo la misma conclusión— cuando sea

$$\frac{S_I^2}{S_{II}^2} < F_{n_I-1, n_{II}-1; 1-\alpha} = 1/F_{n_{II}-1, n_I-1; \alpha}$$

es decir, haciendo operaciones y deshaciendo el cambio, cuando sea

$$\frac{S_1^2}{S_2^2} > F_{n_1-1, n_2-1; \alpha}.$$

La cuestión no radica en la denominación de las poblaciones, sino en, una vez denominadas éstas, cuál de las dos hipótesis parece razonable contrastar: *menor o igual frente a mayor*, ó, *mayor o igual frente a menor*.

Y la respuesta para esto es clara: Si es $S_1^2 > S_2^2$ debe contrastarse $H_0 : \sigma_1^2 \leq \sigma_2^2$ frente a $H_1 : \sigma_1^2 > \sigma_2^2$, ya que si contrastáramos las hipótesis simétricas, se aceptaría sin duda H_0 .

Análogamente, si es $S_1^2 < S_2^2$ debe contrastarse $H_0 : \sigma_1^2 \geq \sigma_2^2$ frente a $H_1 : \sigma_1^2 < \sigma_2^2$.

Una cuestión que queda abierta, no obstante, es qué ocurre cuando contrastamos $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$.

En este caso, por la asimetría de la distribución F de Snedecor, se recomienda utilizar como población I la de mayor cuasivarianza muestral.

7.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes

Al igual que en la sección anterior, la situación que tenemos aquí planteada es la de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, de las que hemos extraído sendas muestras aleatorias independientes de tamaños n_1 y n_2 respectivamente, X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} , representando, como siempre, por \bar{x}_1 , S_1^2 y por \bar{x}_2 , S_2^2 la media y cuasivarianza de la primera y segunda muestra respectivamente.

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \\ H_1 : \mu_1 &\neq \mu_2 \end{aligned}$$

σ_1 y σ_2 conocidas

En este caso el test óptimo es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \end{aligned}$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

Como ocurría en los Capítulos 5 y 6, si las muestras no son suficientemente grandes, con objeto de determinar el test óptimo, no sólo en el contraste bilateral, sino también en los unilaterales, habrá que distinguir los casos en que las varianzas poblacionales puedan considerarse iguales y aquellos en los que no puedan ser consideradas iguales.

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales se pueden considerar iguales, entonces el test óptimo es

- Se acepta H_0 si
$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{n_1+n_2-2; \alpha/2}$$
- Se rechaza H_0 si
$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{n_1+n_2-2; \alpha/2}$$

(b) $\sigma_1 \neq \sigma_2$

En el caso de que las varianzas poblacionales no puedan considerarse iguales, el test óptimo es

- Se acepta H_0 si
$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq t_{f; \alpha/2}$$
- Se rechaza H_0 si
$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > t_{f; \alpha/2}$$

Ejemplo 7.16 (continuación)

En la sección anterior llegamos a la conclusión de que las varianzas poblacionales podían considerarse iguales. Por tanto, a nivel $\alpha = 0'05$, al ser $t_{21; 0'025} = 2'08$ y

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{|63'9 - 67'8|}{\sqrt{\frac{12 \cdot 9'16^2 + 9 \cdot 8'37^2}{21}} \sqrt{\frac{1}{13} + \frac{1}{10}}} = 1'05 < 2'08$$

se aceptará H_0 . Por interpolación, de la *Tabla 5* de ADD se obtiene que

$$p\text{-valor} = 2 \cdot P\{t_{21} > 1'05\} = 2 \cdot 0'1588 = 0'3176.$$

$$\begin{aligned} H_0 : \mu_1 &\leq \mu_2 \\ H_1 : \mu_1 &> \mu_2 \end{aligned}$$

Como en el apartado anterior, habrá que distinguir si las varianzas poblacionales pueden considerarse conocidas o no, y en ese caso, si pueden admitirse como iguales.

σ_1 y σ_2 conocidas

La regla óptima es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_\alpha \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha \end{aligned}$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

De nuevo hay que distinguir si las varianzas pueden ser consideradas iguales o no.

(a) $\sigma_1 = \sigma_2$

En este caso la regla óptima es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq t_{n_1 + n_2 - 2; \alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} > t_{n_1 + n_2 - 2; \alpha} \end{aligned}$$

(b) $\sigma_1 \neq \sigma_2$

El test óptimo es

<ul style="list-style-type: none"> • Se acepta H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq t_{f;\alpha}$ • Se rechaza H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > t_{f;\alpha}$

Ejemplo 7.19

Se quiere comparar el coeficiente intelectual de los estudiantes de dos universidades. Para ello se extrajeron muestras aleatorias de ambas instituciones, obteniéndose los siguientes resultados

$$\begin{array}{lll} n_1 = 13 & \bar{x}_1 = 100 & S_1^2 = 100 \\ n_2 = 10 & \bar{x}_2 = 90 & S_2^2 = 25 \end{array}$$

A nivel $\alpha = 0'05$, ¿puede admitirse el coeficiente intelectual de la segunda universidad es menor que el de la primera?

Lo primero que tenemos que comprobar es si las varianzas pueden admitirse como iguales o no. Es decir, contrastar $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$. Al ser

$$\begin{aligned} I &= [F_{n_1-1, n_2-1; 1-\alpha/2}, F_{n_1-1, n_2-1; \alpha/2}] = [F_{12, 9; 0'975}, F_{12, 9; 0'025}] = \\ &= [1/F_{9, 12; 0'025}, F_{12, 9; 0'025}] = [0'291, 3'8682] \end{aligned}$$

y ser $S_1^2/S_2^2 = 4 \notin I$ rechazaremos la hipótesis nula de igualdad de las varianzas.

Por tanto, el estadístico del test óptimo a utilizar, para contrastar la igualdad de medias, será

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \rightsquigarrow t_f$$

con f el número entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 + \left(\frac{S_2^2}{n_2}\right)^2} - 2 = \frac{\left(\frac{100}{13} + \frac{25}{10}\right)^2}{\left(\frac{100}{13}\right)^2 + \left(\frac{25}{10}\right)^2} - 2 = 19'666$$

Como es

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{100 - 90}{\sqrt{\frac{100}{13} + \frac{25}{10}}} = 3'1323 > 1'725 = t_{20;0'05}$$

rechazaremos H_0 .

El p-valor, igual a $P\{t_{20} > 3'1323\} \simeq 0'0025$, confirma claramente esta decisión, la cual hasta cierto punto podría sorprender por haber sólo una diferencia de 10 puntos. No obstante, ésta ha resultado ser una diferencia significativa.

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Realmente este apartado es innecesario, ya que se corresponde exactamente con el anterior, intercambiando los papeles de las dos poblaciones, debido a la simetría de las distribuciones normal y t de Student.

Podríamos, por tanto, prescindir de él llamando siempre población 1 a la de mayor media muestral, siendo entonces el contraste de interés el de la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a la alternativa $H_1 : \mu_1 > \mu_2$.

No obstante, por completar el esquema, la hemos mantenido, aunque debamos tener siempre claro cuál es el contraste de interés.

El simétrico no es que no tenga interés, es que su resultado es obvio: Si $\bar{x}_1 > \bar{x}_2$, la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ se aceptará siempre frente a $H_1 : \mu_1 < \mu_2$, con tal que sea $\alpha < 0'5$ —cota que, desde todo punto de vista, no se debe rebasar—, ya que la región crítica para el contraste de esas hipótesis es la cola izquierda de la distribución —incluida en el semieje de los números negativos por ser $\alpha < 0'5$ —, mientras que al ser $\bar{x}_1 > \bar{x}_2$, el estadístico del contraste será siempre positivo.

σ_1 y σ_2 conocidas

En este caso el test óptimo es

<ul style="list-style-type: none"> • Se acepta H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{1-\alpha}$ • Se rechaza H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha}$

σ_1 y σ_2 desconocidas. Muestras pequeñas

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales pueden suponerse iguales y las muestras no tienen ambas, tamaños suficientemente grandes, el test óptimo es

<ul style="list-style-type: none"> • Se acepta H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \geq t_{n_1 + n_2 - 2; 1-\alpha}$ • Se rechaza H_0 si $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < t_{n_1 + n_2 - 2; 1-\alpha}$

(b) $\sigma_1 \neq \sigma_2$

Si las varianzas poblacionales son distintas, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \geq t_{f;1-\alpha} \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < t_{f;1-\alpha}
 \end{aligned}$$

Ejemplo 7.20

Un grupo de científicos de una estación antártica, estuvo de acuerdo en participar en un estudio nutricional el cual se proponía analizar los niveles de vitamina C en personas que viven en un clima extremadamente frío.

Con este objetivo, las personas de la estación fueron divididas al azar en dos grupos. Al Grupo 1 le fue administrado un suplemento de vitamina C y el Grupo 2 fue utilizado como grupo control.

Los datos de los niveles, en $\mu\text{g}/10^8$ células, de ácido ascórbico en sangre fueron (Fuente: Dr. P. Gormley, Antarctic Division, Australian Department of Science and Technology)

Grupo 1	18'3	9'3	12'6	15'7	14'2	13'1	14'3	16'2	18'1	19'4	15'5	11'7
Grupo 2	24'9	16	26'3	25'5	19'3	16'8	15'7	24'6	19'9	9'4	17'4	

los cuales suministraron unos valores de

$$\begin{aligned}
 n_1 &= 12 & \bar{x}_1 &= 14'87 & S_1^2 &= 8'7 \\
 n_2 &= 11 & \bar{x}_2 &= 19'62 & S_2^2 &= 27'79
 \end{aligned}$$

Las hipótesis a contrastar serán $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$.

Admitiendo que los niveles de ácido ascórbico siguen distribuciones normales en ambas poblaciones, lo primero que deberemos contrastar es la igualdad de las varianzas. Al ser, para $\alpha = 0'05$,

$$\begin{aligned}
 [F_{n_1-1, n_2-1; 1-\alpha/2}, F_{n_1-1, n_2-1; \alpha/2}] &= [F_{11, 10; 0'975}, F_{11, 10; 0'025}] = \\
 &= [1/F_{10, 11; 0'025}, F_{11, 10; 0'025}] = [0'28363, 3'66885]
 \end{aligned}$$

y $S_1^2/S_2^2 = 0'31$, aceptaremos la igualdad de las varianzas.

Ahora, al ser

$$\begin{aligned}
 &\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \\
 &= \frac{14'87 - 19'62}{\sqrt{\frac{11 \cdot 8'7 + 10 \cdot 27'79}{21}} \sqrt{\frac{1}{12} + \frac{1}{11}}} = -2'698 < -1'721 = t_{21; 0'95}
 \end{aligned}$$

rechazaremos H_0 a nivel $\alpha = 0.05$.

El p-valor, entre 0.01 y 0.005, es suficientemente pequeño como para confirmar el rechazo de H_0 e inferir, en base a estos datos, que la administración de vitamina C en ambientes muy fríos disminuye los niveles de ácido ascórbico en la sangre.

Si queremos resolver este ejemplo con R, ejecutaremos (3), después de incluir los datos en (1) y (2)

```
> x<-c(18.3,9.3,12.6,15.7,14.2,13.1,14.3,16.2,18.1,19.4,15.5,11.7) (1)
```

```
> y<-c(24.9,16,26.3,25.5,19.3,16.8,15.7,24.6,19.9,9.4,17.4) (2)
```

```
> t.test(x,y,alternative="less",var.equal=T) (3)
```

Two Sample t-test

```
data: x and y
```

```
t = -2.6989, df = 21, p-value = 0.006722 (4)
```

```
alternative hypothesis: true difference in means is less than 0
```

```
95 percent confidence interval:
```

```
NA -1.722055
```

```
sample estimates:
```

```
mean of x mean of y
```

```
14.86667 19.61818
```

El p-valor del test ahora se puede determinar con exactitud; viene dado en (4).

Respecto a la observación que antes hacíamos, queríamos decir que hubiera dado lo mismo cambiar los nombres de los grupos en el ejemplo anterior y llamar grupo 1 al grupo control, contrastando $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$, puesto que, en ese caso, también habríamos rechazado H_0 —obteniendo la misma conclusión: el grupo control tiene mayores niveles de ácido ascórbico— al ser, con esta nueva denominación

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 2.698 > 1.721 = t_{21;0.05}.$$

De hecho, esto habría sido más correcto, ya que ahora sería $S_1^2 = 27.79 > 8.7 = S_2^2$, teniendo una asignación más adecuada a la hora de contrastar la igualdad de las varianzas.

Como conclusión, ya que en el texto damos las dos versiones del contraste unilateral, aconsejamos denominar población 1 a la de mayor cuasivarianza muestral, ya que frecuentemente deberemos contrastar la igualdad de las varianzas previamente, y luego elegir el contraste unilateral de interés según las observaciones realizadas más arriba.

7.7. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

La situación que se estudia en esta sección es la de dos muestras aleatorias independientes X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} , de tamaños similares y suficientemente grandes —digamos $n_1 \approx n_2$ y $n_1 + n_2 > 30$.

Precisamente por esta razón no se requiere normalidad en las distribuciones modelo.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

σ_1 y σ_2 conocidas

En este caso el test óptimo es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \end{aligned}$$

σ_1 y σ_2 desconocidas

Si las varianzas poblacionales no se suponen conocidas —situación por otro lado habitual—, el test óptimo es

- Se acepta H_0 si $\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq z_{\alpha/2}$
 - Se rechaza H_0 si $\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > z_{\alpha/2}$

Ejemplo 7.21

Se realizó un estudio a lo largo de 12 meses, en el cual se recogieron datos sobre las mujeres que daban a luz en hospitales de Tasmania, sobre del uso de *Syntocinon*, un medicamento utilizado para provocar el parto.

El grupo 1 fue un grupo control formado por mujeres que no usaron el medicamento, y el grupo 2 el formado por mujeres que lo usaron dentro de un periodo de dos horas desde que rompieron aguas.

Los datos, en horas, desde que rompieron aguas hasta el momento del parto fueron (Fuente: Profess. J. Correy, Depart. of Obstets., University of Tasmania)

$$\begin{aligned} n_1 &= 315 & \bar{x}_1 &= 9'43 & S_1^2 &= 32'4616 \\ n_2 &= 301 & \bar{x}_2 &= 9'14 & S_2^2 &= 26'2455 \end{aligned}$$

A nivel $\alpha = 0'05$, ¿puede inferirse una diferencia significativa entre ambos grupos?

Como es

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|9'43 - 9'14|}{\sqrt{\frac{32'4616}{315} + \frac{26'2455}{301}}} = 0'6649 < 1'96 = z_{0'025}$$

se acepta la no existencia de diferencias significativas entre ambos grupos, es decir, se acepta la hipótesis $H_0 : \mu_1 = \mu_2$.

$$\begin{aligned} H_0 : \mu_1 &\leq \mu_2 \\ H_1 : \mu_1 &> \mu_2 \end{aligned}$$

σ_1 y σ_2 conocidas

Si las varianzas de las poblaciones son, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_\alpha \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha
 \end{aligned}$$

σ_1 y σ_2 desconocidas

Caso de que se desconozcan las varianzas de las poblaciones, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq z_\alpha \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > z_\alpha
 \end{aligned}$$

Ejemplo 7.22

Se cree que un nuevo fertilizante puede aumentar los contenidos de ácido ascórbico en los guisantes. Con objeto de comprobar esta teoría, se consideraron dos grupos: El Grupo 1, en el que se usó el fertilizante, y el Grupo 2, utilizado como grupo control.

Los contenidos de ácido ascórbico medidos en miligramos/100 g., dieron los siguientes resultados:

$$\begin{array}{llll}
 \text{Grupo 1:} & n_1 = 63 & \bar{x}_1 = 39 & S_1^2 = 123 \\
 \text{Grupo 2:} & n_2 = 80 & \bar{x}_2 = 35 & S_2^2 = 80
 \end{array}$$

A nivel $\alpha = 0.05$, ¿puede aceptarse la eficacia de este nuevo fertilizante?

Llamando respectivamente μ_1 y μ_2 a los niveles medios de ácido ascórbico en los guisantes sometidos al fertilizante y en los guisantes no fertilizados, el contraste que se propone es $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$. Como es

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{39 - 35}{\sqrt{\frac{123}{63} + \frac{80}{80}}} = 2'33 > 1'645 = z_{0'05}$$

se rechaza H_0 . El p-valor = $P\{Z > 2'33\} < 0'0099$ confirma la decisión tomada.

$$H_0 : \mu_1 \geq \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

Al igual que ocurría en la sección anterior, esta situación se incluye por completar el esquema, ya que, como sabemos, puede reducirse a la del apartado anterior.

σ_1 y σ_2 conocidas

Si las varianzas poblacionales son conocidas, el test óptimo es

$$\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{1-\alpha}$$

$$\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha}$$

σ_1 y σ_2 desconocidas

Si son desconocidas, el test a utilizar es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \geq z_{1-\alpha} \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < z_{1-\alpha}
 \end{aligned}$$

Poblaciones binomiales

En este apartado las observaciones X_1, \dots, X_{n_1} e Y_1, \dots, Y_{n_2} se suponen procedentes de dos poblaciones independientes $B(1, p_1)$ y $B(1, p_2)$ respectivamente, siendo los tamaños muestrales similares y grandes —digamos $n_1 \approx n_2$ y $n_1 + n_2 > 100$.

$$\begin{aligned}
 H_0 : p_1 &= p_2 \\
 H_1 : p_1 &\neq p_2
 \end{aligned}$$

En este caso, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \leq z_{\alpha/2} \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} > z_{\alpha/2}
 \end{aligned}$$

siendo $\hat{p}_1 = x_1/n_1$ y $\hat{p}_2 = x_2/n_2$, respectivamente, las proporciones de la primera y segunda muestras, y $\bar{p} = (x_1 + x_2)/(n_1 + n_2)$.

Ejemplo 7.23

Se quiere investigar si la subida de la acetona en una población infantil es semejante en los niños y en las niñas.

Con este propósito se tomaron al azar muestras de orina de 50 niños y de 50 niñas, obteniéndose para los primeros una proporción muestral de subidas de la acetona de 7/50, mientras que para los segundos la subida fue de 9/50.

A nivel $\alpha = 0'05$, ¿puede aceptarse que no influye significativamente el sexo en la subida de la acetona infantil?

Como es

$$\frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{|7/50 - 9/50|}{\sqrt{\frac{0'16 \cdot 0'84}{50} + \frac{0'16 \cdot 0'84}{50}}} = 0'546 < 1'96 = z_{0'025}$$

se concluye considerando no significativo el sexo en cuanto a la subida de la acetona infantil. El p-valor, $P\{|Z| > 0'546\} = 0'5858$, confirma esta decisión.

Si queremos resolver este ejemplo con R ejecutaríamos la siguiente secuencia de sentencias. Primero creamos el vector de frecuencias absolutas en (1) y el de frecuencias totales en (2), obteniendo en (4) el p-valor del test, con hipótesis nula la igualdad de las proporciones muestrales, al ejecutar (3).

```
> x<-c(7,9) (1)
```

```
> n<-c(50,50) (2)
```

```
> prop.test(x,n,correct=F) (3)
```

```
2-sample test for equality of proportions without continuity
correction
```

```
data: x out of n
```

```
X-squared = 0.2976, df = 1, p-value = 0.5854
```

```
(4)
```

```
alternative hypothesis: two.sided
```

```
95 percent confidence interval:
```

```
-0.1834929 0.1034929
```

```
sample estimates:
```

```
prop 1 prop 2
```

```
0.14 0.18
```

$$H_0 : p_1 \leq p_2$$

$$H_1 : p_1 > p_2$$

El test óptimo es

- Se acepta H_0 si $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \leq z_\alpha$
- Se rechaza H_0 si $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} > z_\alpha$

Ejemplo 7.24

Se quiere averiguar si las mujeres embarazadas que conviven con gatos —población 1— presentan un índice mayor de toxoplasmosis que las que no tienen felinos en el hogar.

Para ello se eligieron al azar mujeres embarazadas en el octavo mes de gestación y se analizó la presencia o ausencia de anticuerpos del toxoplasma, clasificándolas según tuvieran o no gatos. Los resultados obtenidos fueron

$$\begin{array}{llll} n_1 = 40 & \hat{p}_1 = 0'375 & = 15/40 \\ n_2 = 60 & \hat{p}_2 = 0'1 & = 6/60 \end{array}$$

Las hipótesis que se contrastaron fueron $H_0 : p_1 \leq p_2$ frente a $H_1 : p_1 > p_2$. Como es

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{0'375 - 0'1}{\sqrt{\frac{0'21 \cdot 0'79}{40} + \frac{0'21 \cdot 0'79}{60}}} = 3'3076$$

y $P\{Z > 3'31\} = 0'0005$, se rechaza H_0 con gran seguridad, concluyendo que la convivencia con gatos conlleva un aumento significativo de la presencia de anticuerpos al toxoplasma.

Si se quiere resolver este ejemplo con R, dado que las hipótesis a contrastar son $H_0 : p_1 \leq p_2$ frente a $H_1 : p_1 > p_2$, la expresión a ejecutar es (1) obteniendo el p-valor del test en (2).

```
> x<-c(15,6)
> n<-c(40,60)
> prop.test(x,n,alternative="greater",correct=F)                                     (1)
```

2-sample test for equality of proportions without continuity correction

data: x out of n

X-squared = 10.9403, df = 1, p-value = 0.0004705
(2)

alternative hypothesis: greater

95 percent confidence interval:

0.1338933 1.0000000

sample estimates:

prop 1 prop 2

0.375 0.100

$$\begin{aligned} H_0 : p_1 &\geq p_2 \\ H_1 : p_1 &< p_2 \end{aligned}$$

De nuevo esta situación se expone por completar el esquema aunque puede reducirse a la anterior. Como ahora ya no necesitamos comparar las varianzas, no tenemos restricciones acerca de cuál debería ser la población 1.

Para realizar contrastes de interés se aconseja tomar como población 1 la de mayor proporción muestral, y considerar el contraste del apartado anterior.

Si se considera el de este apartado, el test óptimo que debe utilizarse es

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \geq z_{1-\alpha} \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} < z_{1-\alpha} \end{aligned}$$

7.8. Contrastes de hipótesis para datos apareados

Como siempre que tratamos esta situación, supondremos que tenemos n parejas de datos $(X_1, Y_1), \dots, (X_n, Y_n)$ en donde las variables X_i e Y_i no pueden calificarse de independientes.

El tratamiento que se hace aquí para este tipo de datos es similar al que se hizo en las Secciones 5.10 y 6.8, el cual consistía en definir la variable unidimensional diferencia, $D_i = X_i - Y_i$, y trasladar los posibles contrastes sobre los parámetros de X e Y a los correspondientes de la variable unidimensional D , utilizando los tests óptimos ya estudiados en las Secciones 7.2, 7.3 y 7.4.

Ejemplo 7.25

Un grupo de investigadores canadienses afirma haber descubierto un tipo de alimentación para las gallinas a base de lino, trigo y soja, bajo la cual éstas producen huevos que no aumentan el colesterol en las personas que los consumen.

Para comprobar dicha teoría, se seleccionaron al azar 35 personas a las que se les midió su nivel de colesterol habitual X_i , observando en ellos de nuevo dicho nivel, Y_i , después de una dieta a base de los huevos en estudio.

Dado que se obtuvo un número suficientemente grande de datos, no resultó necesario comprobar la normalidad en la distribución modelo de la variable $D_i = X_i - Y_i$, con lo que, para contrastar $H_0 : \mu_d = 0$ —los huevos no modifican el colesterol— frente a $H_1 : \mu_d \neq 0$, con $\mu_d = \mu_1 - \mu_2$, deberemos utilizar el estadístico

$$\frac{|\bar{d} - 0|}{S_d/\sqrt{n}} \approx N(0, 1)$$

en donde es

$$\bar{d} = \bar{x}_1 - \bar{x}_2 \quad \text{y} \quad S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - Y_i - \bar{d})^2$$

Los resultados obtenidos fueron

$$\bar{x}_1 = 200 \quad , \quad \bar{x}_2 = 203 \quad , \quad \frac{1}{34} \sum_{i=1}^{35} (X_i - Y_i - \bar{d})^2 = 196$$

con lo que será

$$\frac{|\bar{d} - 0|}{S_d/\sqrt{n}} = \frac{3}{14/\sqrt{35}} = 1'268 < 1'96 = z_{0'025}$$

aceptándose que no existe diferencia significativa, a nivel $\alpha = 0'05$, en los niveles de colesterol antes y después de una alimentación a base de los huevos en cuestión.

7.9. Ejercicios de Autoevaluación

Ejercicio 7.1

Una famosa pizzería afirma que el tiempo en minutos que tarda el cliente en recibir su pedido es una variable aleatoria con distribución normal de media μ . Además asegura que μ nunca es mayor que 12.

No obstante, un cliente se ha quejado de que en los 9 últimos pedidos, efectuados en días elegidos al azar, el tiempo medio por él calculado ha sido de 17'792 minutos y la cuasivarianza muestral igual a 36.

Contrastar la hipótesis nula de que efectivamente se puede asegurar que es $\mu \leq 12$, para un nivel de significación de 0'05.

Ejercicio 7.2

Se esté estudiando el tiempo de vida entre los pacientes a una determinada enfermedad. A tal fin se eligieron al azar 100 fichas de pacientes fallecidos por la enfermedad en estudio, obteniéndose una media muestral de 740 días y una cuasidesviación típica muestral de 32 días.

¿Puede admitirse para los pacientes de la enfermedad en cuestión un tiempo medio de vida superior a 730 días?

Ejercicio 7.3

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 trató a 10 pacientes esquizofrénicos hospitalizados con una medicación antipsicótica y se les clasificó, después del tratamiento, en dos grupos: el de *psicóticos* (es decir, el de los que seguían padeciendo la enfermedad después del tratamiento) y el de *no psicóticos*. Se les extrajo una muestra de fluido cerebro-espinal a cada paciente 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>	0'0105	0'0145	0'0160	0'0130	0'0156	0'0104
<i>Psicóticos</i>	0'0222	0'0245	0'0320	0'0150		

Suponiendo que los datos anteriores proceden de dos distribuciones normales independientes, una para cada uno de los dos grupos de pacientes, ¿difiere la actividad DBH entre estos dos grupos, a nivel $\alpha = 0'05$?

Ejercicio 7.4

Se ha demostrado que el número de ordenadores que compra un estudiante a lo largo de su carrera es una variable aleatoria con distribución de Poisson. Se quiere averiguar si puede admitirse la hipótesis nula de que la media de dicha variable es igual a 1, o alternativamente la hipótesis de que es igual a 2. Para ello se seleccionaron al azar 100 estudiantes, obteniéndose una media muestral de $\bar{x} = 1'3$ ordenadores. Se pide: Contrastar ambas hipótesis a nivel $\alpha = 0'05$. Calcular el p-valor e interpretarlo. Calcular la potencia del test e interpretarla.

Ejercicio 7.5

Un matemático está convencido que puede afirmarse que el tiempo medio μ de espera de su autobús es de más de 15 minutos. Para ello, anota tiempos de espera elegidos al azar de dicho autobús, obteniendo los siguientes valores en minutos:

14 , 19 , 20 , 14 , 17 , 24 , 14 , 20 , 20

Si admite que dicho tiempo de espera sigue una distribución normal, contrastar a nivel $\alpha = 0'05$ la hipótesis del matemático. Acotar el p-valor.

7.10. Lecturas Recomendadas

Mood, A.M., Graybill, F.A. y Boes, D.C. (1974). *Introduction to the Theory of Statistics*. Editorial McGraw-Hill.

Capítulo 8

Contrastes no paramétricos

8.1. Introducción

Los contrastes estudiados en el capítulo anterior requerían suponer un modelo poblacional para los datos observados, a menos que los tamaños muestrales fueran grandes. Además, las hipótesis a contrastar hacían referencia a los parámetros de ese modelo.

En este capítulo estudiaremos algunos tests, denominados *no paramétricos*, que por un lado no requieren especificar un modelo para la variable en estudio y, por otro, contrastan hipótesis que no se refieren a los valores de la media, ni de la varianza.

En la Sección 8.2 estudiaremos los denominados *tests* χ^2 , que se utilizan cuando los datos que tengamos sean recuentos de observaciones y que presenten la característica común de que el estadístico de contraste sigue, aproximadamente, una distribución χ^2 . En las Secciones 8.3 y 8.4 analizaremos algunos tests sobre la mediana poblacional respectivamente de una o dos poblaciones, alternativos a los contrastes sobre las medias estudiados en el capítulo anterior.

8.2. Pruebas χ^2

En esta sección estudiaremos tres tests que tienen la peculiaridad de estar definidos en base a recuentos o frecuencias de k posibles clases o grupos en los que se clasifican los datos y no en valores concretos de las variables en análisis. Se trata del contraste de *bondad del ajuste*, mediante el cual analizamos si puede admitirse una determinada distribución como modelo probabilístico de nuestros datos; del contraste de *homogeneidad de varias muestras*, con el que analizamos si puede admitirse la igualdad de las poblaciones de donde se extrajeron los datos y, por último, del contraste de *independencia de caracteres*, con el que contrastamos la hipótesis nula de independencia de dos variables.

Además, los tres tienen un estadístico de contraste con distribución (aproximada) χ^2 . Veamos a continuación un ejemplo de cada uno.

Ejemplo 8.1

Se quiere comprobar si un dado está *equilibrado* o si por el contrario está *cargado*. A tal fin se lanzó 600 veces, obteniéndose los resultados de la siguiente tabla

	1	2	3	4	5	6	Total
n_i	103	98	89	109	100	101	600

Tabla 8.1

Ejemplo 8.2

Se ha realizado un estudio sobre caries dental en niños de seis ciudades con diferentes cantidades de flúor en el suministro de agua, con objeto de analizar si existen diferencias significativas entre las seis ciudades. Seleccionada una muestra aleatoria de 125 niños de cada ciudad, los resultados obtenidos fueron los siguientes,

Comunidad	nº de niños sin caries	nº de niños con caries	Total
A	38	87	125
B	8	117	125
C	30	95	125
D	44	81	125
E	64	61	125
F	32	93	125
	216	534	750

Tabla 8.2

Ejemplo 8.3

Se quiere analizar si existe independencia entre el Peso y la Talla de los individuos de una población. Con tal propósito se seleccionó una muestra de 100 individuos de la mencionada población, obteniéndose los siguientes resultados

Talla	1'55 – 1'65	1'65 – 1'75	1'75 – 1'85	1'85 – 1'95	
Peso					
50-60	10	8	2	1	
60-70	6	14	6	2	
70-80	2	8	18	5	
80-90	0	4	6	8	
					100

Tabla 8.3

Como se ve en estos tres ejemplos, los datos aparecen divididos en una serie de clases E_1, E_2, \dots, E_k ; en el Ejemplo 8.1, las clases eran 1, 2, ..., 6; en

el Ejemplo 8.2 eran *niños sin caries* y *niños con caries* en cada una de las seis comunidades, y en el Ejemplo 8.3 las clases eran las 16 combinaciones de modalidades de Pesos y Tallas. Además, en los tres ejemplos, los datos son frecuencias absolutas, es decir, recuentos, n_1, n_2, \dots, n_k , de las mencionadas clases; es decir, número de individuos de la muestra de tamaño $n = \sum_{i=1}^k n_i$ que pertenecen a cada una de las clases.

Como dijimos más arriba, los tres contrastes que veremos en esta sección están basados en el denominado *estadístico λ de Pearson*, el cual mide, para cada clase E_i , las discrepancias entre las frecuencias (relativas) observadas n_i/n y las esperadas de ser cierta la hipótesis nula H_0 de que las clases E_i tienen probabilidades p_i , determinadas éstas por la hipótesis nula de la distribución modelo supuesta, o por la hipótesis nula de homogeneidad de las muestras, o por la de independencia de las dos variables en análisis. Este estadístico,

$$\lambda = \sum_{i=1}^k \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2$$

mide las discrepancias normalizadas. Si toma un valor grande, entonces deberemos rechazar la hipótesis nula.

En la formalización del contraste necesitaremos determinar la distribución en el muestreo, bajo H_0 , del estadístico del contraste λ . Si el tamaño muestral n es suficientemente grande —digamos $n > 30$ —, el estadístico λ de Pearson se distribuye, aproximadamente, según una χ^2 con $k - 1$ grados de libertad. Ello permitirá determinar los puntos críticos de los contrastes que veremos en los siguientes apartados, los cuales, por este hecho común, se denominan pruebas χ^2 .

El estadístico λ de Pearson, después de simplificar, puede calcularse como

$$\lambda = \sum_{i=1}^k \frac{(n_i - n p_i)^2}{n p_i} = \sum_{i=1}^k \left(\frac{n_i^2}{n p_i} \right) - n.$$

8.2.1. Pruebas χ^2 con R

Para ejecutar con R los tres tipos de contrastes χ^2 estudiados la función a utilizar será

```
chisq.test(x, correct=TRUE, p)
```

en donde incluiremos en el primer argumento x el vector de observaciones (frecuencias absolutas) en el test de bondad del ajuste, o la tabla de doble entrada en el caso de los otros dos tests.

El segundo argumento es opcional y permite utilizar la *corrección de Yates* aunque sólo en el caso de tablas de contingencia 2×2 . Esta opción es la

que se toma por defecto; para no utilizarla deberemos ejecutar `correct=F`. Observamos que siempre que tengamos tablas 2×2 calculará la corrección de Yates (a menos que le indiquemos que no lo haga) aunque las frecuencias esperadas no sean menores que 5. Esta corrección la definiremos más adelante.

El tercer argumento también es opcional y es utilizado, únicamente, en los tests de bondad del ajuste para indicar el vector de probabilidades teóricas que comparamos con las observadas, es decir, las que establecemos en la hipótesis nula si es que el modelo teórico es distinto del *uniforme* que asigna igual probabilidad a todas las clases consideradas ya que, en ese caso, podemos prescindir de esta opción al ser la que se toma por defecto.

8.2.2. Contraste de bondad del ajuste

Este contraste tiene por objeto averiguar si puede admitirse o no la hipótesis nula H_0 de seguir la variable aleatoria en observación una determinada distribución modelo F_0 , expresando dicha hipótesis nula de la forma $H_0 : F(x) = F_0(x) \forall x$, o más brevemente, $H_0 : X \sim F_0$.

Dado que la mayoría de las técnicas inferenciales estudiadas en los capítulos anteriores requerían suponer un modelo para la variable aleatoria en estudio, dependiendo los resultados obtenidos de la distribución supuesta, la comprobación de dicho modelo, resulta, cuanto menos, de gran interés, si no imprescindible, en la mayoría de las inferencias que realicemos.

Ejemplo 8.4

Supongamos que queremos comprobar si la teoría de Mendel sobre la herencia genética es correcta, observando guisantes de una determinada variedad, los cuales pueden ser clasificados, atendiendo a su color y su forma, en cuatro clases, $E_1 = \text{lisos y amarillos}$, $E_2 = \text{lisos y verdes}$, $E_3 = \text{arrugados y amarillos}$ y $E_4 = \text{arrugados y verdes}$.

Según dicha teoría, las proporciones esperadas para cada una de las clases son 9/3/3/1 respectivamente.

Seleccionados al azar 556 guisantes de la variedad en estudio, se obtuvieron las siguientes cantidades de cada clase: 315 de E_1 , 108 de E_2 , 101 de E_3 y 32 de E_4 .

Si la teoría de Mendel fuera correcta, es decir si la hipótesis nula de que los datos siguen una distribución con probabilidades asociadas 9/16, 3/16, 3/16, 1/16 fuese correcta, cabría esperar que las frecuencias relativas de la muestra no difiriesen mucho de esas probabilidades. Es decir, si la hipótesis nula fuese correcta, cabría esperar que el estadístico λ de Pearson fuera pequeño. En caso contrario deberemos rechazar H_0 . Estas ideas se formalizan en el siguiente apartado.

Contraste de hipótesis

Sea $(\Omega, \mathcal{A}, P_0)$ un espacio probabilístico (con función de distribución asociada $F_0(x)$), en el que se considera una partición de Ω formada por k sucesos de \mathcal{A} , E_1, E_2, \dots, E_k , tales que $p_i = P_0(E_i) > 0$ $i = 1, \dots, k$ y $\sum_{i=1}^k p_i = 1$.

Si realizado un experimento aleatorio se obtuvieron las frecuencias absolutas n_1, \dots, n_k para las clases E_1, \dots, E_k , y por

$$\lambda = \sum_{i=1}^k \frac{(n_i - n p_i)^2}{n p_i} = \sum_{i=1}^k \left(\frac{n_i^2}{n p_i} \right) - n$$

representamos el estadístico de Pearson, para contrastar a nivel α la hipótesis nula $H_0 : X \sim F_0$, frente a la alternativa de que los datos no se ajustan al modelo F_0 , el test óptimo a utilizar es

- Se acepta H_0 si $\lambda < \chi_{k-1; \alpha}^2$
- Se rechaza H_0 si $\lambda \geq \chi_{k-1; \alpha}^2$

teniendo perfecto sentido, tanto en este como en todos los contrastes que veremos, el cálculo e interpretación de su p-valor.

Ejemplo 8.4 (continuación)

Las frecuencias observadas y esperadas son

	f. observadas n_i	f. esperadas $n \cdot p_i$	
E_1	315	312'75	$(= 556 \cdot 9/16)$
E_2	108	104'25	
E_3	101	104'25	
E_4	32	34'75	
	556	556	

Para contrastar, a nivel $\alpha = 0'05$, la hipótesis nula H_0 : *los datos se ajustan a la distribución teórica 9/3/3/1*, frente a la alternativa H_1 : *los datos no se ajustan a dicha distribución teórica*, debemos calcular el estadístico λ de Pearson, el cual toma el valor

$$\lambda = \sum_{i=1}^k \frac{(n_i - n p_i)^2}{n p_i} = 0'47$$

y compararlo con el punto crítico proporcionado por una χ_3^2 . Como es $\lambda = 0'47 < 7'815 = \chi_{3; 0'05}^2$, aceptaremos H_0 , concluyendo que dicha variedad de guisantes respeta la teoría mendeliana. El p-valor, entre 0'9 y 0'95, confirma esta decisión.

Si queremos resolver este ejemplo con R, deberemos ejecutar la siguiente secuencia de instrucciones. El p-valor, dado en (1), confirma la aceptación de la hipótesis nula, es decir, podemos estar tranquilos, Mendel estaba en lo cierto.

```
> x<-c(315,108,101,32)
> p1<-c(9/16, 3/16, 3/16, 1/16)
> chisq.test(x,p=p1)
```

Chi-squared test for given probabilities

data: x

X-squared = 0.47, df = 3, p-value = 0.9254

(1)

Ejemplo 8.1 (continuación)

En este caso, las frecuencias observadas y esperadas son

	n_i	$n \cdot p_i$	
1	103	100	(= 600 · 1/6)
2	98	100	
3	89	100	
4	109	100	
5	100	100	
6	101	100	
	600	600	

Al contrastar, a nivel $\alpha = 0.05$, la hipótesis nula H_0 : *los datos se ajustan a la distribución teórica 1/6, ..., 1/6*, frente a la alternativa de que no se ajustan, debemos calcular el estadístico λ de Pearson, el cual toma el valor

$$\lambda = \sum_{i=1}^6 \frac{(n_i - n p_i)^2}{n p_i} = 2.16.$$

Como $\lambda = 2.16 < 11.07 = \chi_{5,0.05}^2$ concluiremos con que el dado es *equilibrado*.

Para resolver este ejemplo con R, observemos primero que, si el dado está equilibrado, los datos deben ajustarse a la distribución teórica que asigna probabilidad 1/6 a cada una las seis caras. Ése será el vector de probabilidades a utilizar en el test.

Así, primero incorporamos los datos observados en (1) y el vector de probabilidades teóricas en (2). Finalmente ejecutamos (3) obteniendo el valor del estadístico en (4) y el p-valor del test en (5). Éste es lo suficientemente grande como para concluir con la aceptación de la hipótesis nula de seguir los datos el modelos teórico, es decir, de estar el dado equilibrado.

> x<-c(103,98,89,109,100,101) (1)

> p1<-c(1/6,1/6,1/6,1/6,1/6,1/6) (2)

> chisq.test(x,p=p1) (3)

Chi-squared test for given probabilities

data: x

X-squared = 2.16, df = 5, p-value = 0.8266

(4)

(5)

Como dijimos más arriba, al analizar la bondad del ajuste a un modelo *uniforme*, podemos prescindir del último argumento ya que este modelo es el que la función toma por defecto, obteniendo los mismos resultados si ejecutamos

> chisq.test(x)

Como veremos más abajo, la corrección de Yates (no suministrada por R en el caso de la Bondad del Ajuste) es una modificación del estadístico de Pearson, utilizada cuando las frecuencias esperadas son pequeñas (tradicionalmente cuando son menores que 5) para que la aproximación por una distribución χ^2 del estadístico de Pearson sea buena. El lector

se podría preguntar cómo saber si las frecuencias esperadas son o no son pequeñas; pues bien, se pueden obtener tanto éstas como las observadas ejecutando al final de la sentencia, respectivamente, `$expected` y `$observed`. Lógicamente las observadas coinciden con el vector `x`.

```
> chisq.test(x,p=p1)$expected
[1] 100 100 100 100 100 100
> chisq.test(x,p=p1)$observed
[1] 103 98 89 109 100 101
```

Ejemplo 8.5

Encuestados 100 matrimonios acerca del número X de hijos varones que tenían, se obtuvo la siguiente distribución de frecuencias

X	n_i
0	22
1	42
2	28
3	8
	100

¿Puede aceptarse como modelo para X una distribución binomial $B(3, 0'5)$?

Las probabilidades y las frecuencias esperadas que la distribución $B(3, 0'5)$ asigna a cada una de las clases E_i —en este caso los valores 0, 1, 2 y 3— vienen dadas en la siguiente tabla,

X	p_i	$n \cdot p_i$
0	0'1250	12'5
1	0'3750	37'5
2	0'3750	37'5
3	0'1250	12'5
	1	100

Como el estadístico de Pearson toma el valor $\lambda = 11'787 > 7'815 = \chi_{3,0'05}^2$, se rechaza la hipótesis nula de que X sigue una distribución $B(3, 0'5)$, a nivel $\alpha = 0'05$.

Para resolver este ejemplo con R, observemos en primer lugar que el vector de probabilidades teóricas debe obtenerse a partir de una distribución de probabilidad, en concreto, de una $B(3, 0'5)$, por lo que obtenemos el vector de probabilidades en (2) utilizando la función de masa de la binomial, después de incluir las frecuencias observadas en (1). La ejecución del test en (3) proporciona el p-valor en (4), valor suficientemente pequeño como para rechazar la hipótesis nula de que los datos observados se ajustan a esa binomial.

```
> x<-c(22,42,28,8) (1)
> p1<-c(dbinom(0,3,0.5),dbinom(1,3,0.5),dbinom(2,3,0.5),dbinom(3,3,0.5)) (2)
> chisq.test(x,p=p1) (3)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 11.7867, df = 3, p-value = 0.008151
(4)
```

Es interesante destacar como los valores de la binomial que definen las clases son irrelevantes. Es decir, lo que hace el estadístico de Pearson en este test es comparar frecuencias observadas con frecuencias esperadas si fuera cierta la hipótesis nula de que los datos se ajustan bien al modelo propuesto, sin considerar para nada qué formó las clases o qué representan; lo importante son ambos vectores de frecuencias.

Las frecuencias esperadas, dadas a continuación, coinciden con las obtenidas anteriormente sin utilizar R.

```
> chisq.test(x,p=p1)$expected
[1] 12.5 37.5 37.5 12.5
```

Observación 1

Tanto en la definición del contraste anterior, como en los ejemplos que le seguían, suponíamos que la distribución modelo a contrastar estaba completamente especificada.

En muchas ocasiones, no obstante, el que admitamos como razonable un determinado tipo de distribución para los datos, no implica necesariamente conocer unos valores para los parámetros de esa distribución.

Así por ejemplo, si para realizar un contraste de la t de Student es necesario que la distribución de la variable en estudio X sea normal, necesitaremos contrastar la normalidad de los datos, pero de entre todas las normales a contrastar, ¿cuál debemos elegir?

La respuesta a esta pregunta es: aquella que tenga como parámetros sus estimaciones de máxima verosimilitud dadas por los datos.

Pero ocurre que, si utilizamos los datos de la muestra en consideración para realizar w estimaciones de parámetros, la distribución χ^2 del estadístico de Pearson tendrá ahora w grados de libertad menos, es decir, $k - 1 - w$ grados de libertad, si se consideran k clases.

Ejemplo 8.6

Las alas de 100 moscas domésticas elegidas al azar, suministraron la siguiente distribución de frecuencias absolutas para sus longitudes X , en $mm \times 10^{-1}$

I	n_i
36 – 38	6
38 – 40	9
40 – 42	33
42 – 44	48
44 – 46	54
46 – 48	57
48 – 50	45
50 – 52	30
52 – 54	12
54 – 56	6
	300

Tabla 8.4

El histograma de esta distribución de frecuencias, dado por la Figura 8.1, sugiere una distribución normal para X . De entre todas las distribuciones normales elegiremos aquella que tenga como parámetros las estimaciones de máxima verosimilitud. En el Ejemplo 5.4 vimos que éstas eran $\hat{\mu} = \bar{x}$ y $\hat{\sigma} = \sqrt{s^2}$.

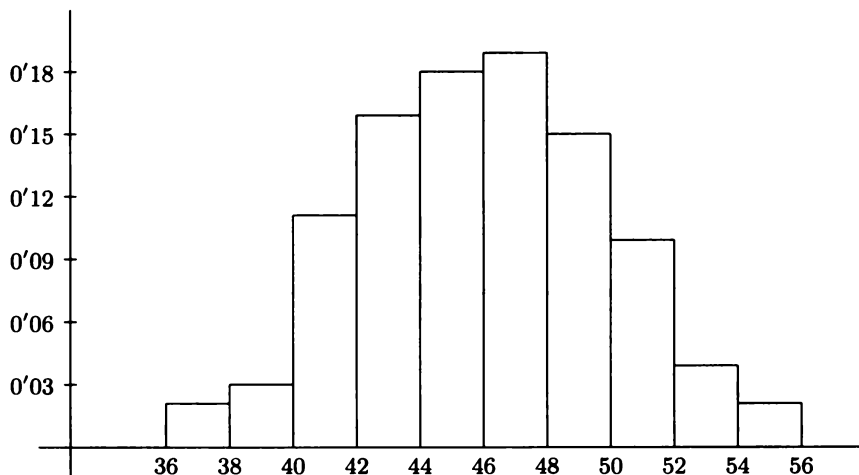


Figura 8.1

De la Tabla 8.4, obtenemos $\bar{x} = 46$ y $\hat{\sigma} = 3'88$, con lo que contrastaremos $H_0 : X \sim N(46, 3'88)$ frente a la alternativa de que los datos no provienen de esta distribución normal. Obsérvese que como los estimadores máximo-verosímiles son los que proporcionan, en casi la totalidad de los casos, las distribuciones *más cercanas*, si los datos no se ajustan a esta normal, cualquier otra distribución normal se ajustará peor, con lo que no merece la pena intentar la bondad del ajuste con otra normal.

Sucede algo parecido a lo que ocurría con la recta de mínimos cuadrados. Si ésta no se ajusta bien a los datos, debemos intentar con otra función —una parábola, por ejemplo—, pero es inútil probar con otra recta, ya que, la *más cercana*, está demasiado lejos.

Debido a que la suma de las frecuencias esperadas tiene que ser n , y la distribución normal toma valores en toda la recta real, debemos considerar como primera clase *Longitudes* < 38 y como última clase *Longitudes* ≥ 54, en lugar de los intervalos de la Tabla 8.4.

Las probabilidades que la distribución $N(46, 3'88)$ asigna a cada una de las clases consideradas aparece en la Tabla 8.5 junto con las frecuencias esperadas para cada clase. Así por ejemplo, es

$$P\{X < 38\} = P\{Z < (38 - 46)/3'88\} = P\{Z < -2'06\} = 0'0197$$

$$P\{38 \leq X < 40\} = P\{-2'06 \leq Z < -1'55\} = 0'0409 = P\{52 \leq X < 54\}$$

El estadístico λ de Pearson toma el valor $\lambda = 3'18$. Al haber estimado 2 parámetros con la muestra en consideración, la χ^2 en la que buscar el punto crítico tendrá $10 - 2 - 1 = 7$ grados de libertad. Como es $\lambda = 3'18 < 14'07 = \chi^2_{7;0'05}$ aceptaremos H_0 a nivel $\alpha = 0'05$. El p-valor, entre 0'7 y 0'9, confirma esta decisión.

I	p_i	$n \cdot p_i$
< 38	0'0197	5'91
38 – 40	0'0409	12'27
40 – 42	0'0909	27'27
42 – 44	0'1500	45
44 – 46	0'1985	59'55
46 – 48	0'1985	59'55
48 – 50	0'1500	45
50 – 52	0'0909	27'27
52 – 54	0'0409	12'27
≥ 54	0'0197	5'91
	1	300

Tabla 8.5

Observación 2

Para que la aproximación χ^2 a la distribución del estadístico λ sea aceptable, no sólo es necesario que el tamaño muestral sea grande, sino que además las frecuencias esperadas no sean demasiado pequeñas —digamos $n \cdot p_i \geq 5$.

Si esto no es así, deberemos agrupar clases E_i contiguas hasta que se tenga esta acotación, reduciendo en igual medida los grados de libertad de la distribución límite χ^2 . Es decir, si por ejemplo una vez agrupadas las clases contiguas nos quedamos con 5 clases E_i , la χ^2 de la que obtendríamos los puntos críticos, debería tener 4 grados de libertad.

Y esto supuesto que no estimemos parámetros de la distribución modelo a partir de la muestra, ya que de ser así, aún deberíamos reducir más los grados de libertad, de acuerdo con la observación anterior.

Ejemplo 8.7

Se cree que el número de descendientes X de los extintos Periquitos de Carolina (*conuropsis carolinensis*) seguía una distribución de Poisson de parámetro 2, por ser éste el número medio de descendientes que tenían cuando fueron estudiados hace 200 años.

Con objeto de comprobar si puede admitirse hoy en día tal modelo para la variable X en los periquitos actuales, se eligieron al azar 100 parejas de estas aves, en las que se obtuvieron los datos dados por la siguiente distribución de frecuencias

X	n_i
0	25
1	30
2	24
3	14
4	5
5	1
6	1
	100

La hipótesis nula a contrastar será, por tanto, $H_0 : X \rightsquigarrow \mathcal{P}(2)$ frente a la alternativa H_1 : los datos no se distribuyen según una $\mathcal{P}(2)$.

Las probabilidades que esta distribución asigna a cada una de las clases E_i —en este caso los valores 0, 1, 2, 3, 4 y 5— vienen dadas en la siguiente tabla, en la que aparecen también las frecuencias esperadas

X	p_i	$n \cdot p_i$
0	0'1353	13'53
1	0'2707	27'07
2	0'2707	27'07
3	0'1804	18'04
4	0'0902	9'02
5	0'0361	3'61
6	0'0166	1'66
	1	100

Obsérvese que, al igual que ocurría en el ejemplo anterior, las clases E_i tiene que ser tales que la suma de sus probabilidades sea 1, consiguiendo así que la suma de las frecuencias esperadas sea n . Para ello, alguna clase (la última y/o la primera) deberá reunir las probabilidad cola, alterando ligeramente su definición. Por eso en este ejemplo la última clase ya no es en realidad $X = 6$, sino $X \geq 6$.

Al observarse que las dos últimas frecuencias esperadas son pequeñas, deberemos agrupar las dos últimas clases, quedándonos con la tabla

X	n_i	p_i	$n \cdot p_i$
0	25	0'1353	13'53
1	30	0'2707	27'07
2	24	0'2707	27'07
3	14	0'1804	18'04
4	5	0'0902	9'02
≥ 5	2	0'0527	5'27
	100	1	100

de la que se obtiene un valor para el estadístico de Pearson $\lambda = 15'114 > 11'07 = \chi_{5;0'05}^2$ por lo que se rechaza la distribución propuesta para X .

Obsérvese que rechazar H_0 no quiere decir que los datos no se ajusten a una distribución de Poisson; solamente que no se ajustan a la Poisson de parámetro 2.

Si hubiéramos estimado el parámetro de la distribución de Poisson a partir de los datos (recuérdese que el estimador máximo verosímil era la media muestral), hubiéramos probado con la mejor distribución de Poisson que se puede ajustar, en este caso, la $\mathcal{P}(1'5)$, al ser $\bar{x} \simeq 1'5$.

En este caso, la tabla resultante hubiera sido

X	n_i	p_i	$n \cdot p_i$
0	25	0'2231	22'31
1	30	0'3347	33'47
2	24	0'2510	25'10
3	14	0'1255	12'55
≥ 4	7	0'0657	6'57
	100	100	100

en la que de nuevo se han agrupado las últimas clases, y para la que se obtiene un valor de $\lambda = 0'928$.

Como se ha estimado un parámetro, la χ^2 en la que hay que buscar el punto crítico sería una $\chi^2_{k-w-1} = \chi^2_3$ de la que se obtiene, para un nivel de significación $\alpha = 0.05$, un valor $\chi^2_{3,0.05} = 7.815 > 0.928 = \lambda$, pudiendo concluir con que los datos sí se ajustan a una distribución de Poisson $\mathcal{P}(1.5)$.

La secuencia de comandos de R con la que contrastar las hipótesis antes planteadas sería la siguiente

```
> x<-c(25,30,24,14,5,1,1)
> p1<-c(dpois(0,2),dpois(1,2),dpois(2,2),dpois(3,2),dpois(4,2),dpois(5,2),dpois(6,2))
> chisq.test(x,p=p1)
Error en chisq.test(x, p = p1) : probabilities must sum to 1.
```

pero al ejecutar `chisq.test(x,p=p1)` vemos como el programa nos da un error, debido a que el soporte del modelo que estamos ajustando, la $\mathcal{P}(2)$, no termina en 6, quedando, por consiguiente, probabilidad (para valores mayores que 6) sin utilizar y, en consecuencia, el vector de probabilidades a analizar en el ajuste, `p1`, no suma 1. Para que esto no ocurra, debemos incluir como última probabilidad el resto no incluido, es decir, $P\{X \geq 6\} = 1 - P\{X < 6\} = 1 - P\{X \leq 5\} = 1 - F(5)$, es decir, $1 -$ el valor de la función de distribución de la $\mathcal{P}(2)$ en 5; es decir, ejecutaremos

```
> p1<-c(dpois(0,2),dpois(1,2),dpois(2,2),dpois(3,2),dpois(4,2),dpois(5,2),
+ 1-ppois(5,2))
> chisq.test(x,p=p1)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 15.2265, df = 6, p-value = 0.01857
(1)
```

Warning message:

```
In chisq.test(x, p = p1) : Chi-squared approximation may be incorrect
```

cuyo p-valor, dado en (1), indica rechazar que los datos se ajusten bien a una $\mathcal{P}(2)$. Si, como sugerimos antes, analizamos la bondad del ajuste a una $\mathcal{P}(1.5)$, deberíamos ejecutar

```
> p1<-c(dpois(0,1.5),dpois(1,1.5),dpois(2,1.5),dpois(3,1.5),dpois(4,1.5),
+ dpois(5,1.5),1-ppois(5,1.5))
> chisq.test(x,p=p1)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 1.7272, df = 6, p-value = 0.943
```

Warning message:

```
In chisq.test(x, p = p1, correct = T) :
Chi-squared approximation may be incorrect
```

que nos indica un buen ajuste a una $\mathcal{P}(1.5)$.

Si las frecuencias esperadas son pequeñas y no queremos agrupar clases

contiguas, podemos corregir el estadístico de Pearson mediante la denominada *corrección de Yates*, y utilizar como estadístico de contraste

$$\lambda_c = \sum_{i=1}^k \frac{(|n_i - n p_i| - 0'5)^2}{n p_i}$$

el cual seguirá teniendo una distribución χ_{k-1}^2 grados de libertad (menos los parámetros que estimemos a partir de la muestra).

Así, en el Ejemplo 8.7, para la $\mathcal{P}(1'5)$, la tabla a utilizar sería

X	n_i	$n \cdot p_i$
0	25	22'31
1	30	33'47
2	24	25'10
3	14	12'55
4	5	4'71
5	1	1'41
6	1	0'45
	100	100

de la que se obtiene un valor para la corrección de Yates de $\lambda_c = 0'58544$. Como es $\chi_{5;0'05}^2 = 11'07$ también aceptaríamos H_0 .

El utilizar la corrección de Yates generalmente conduce a tests más conservadores, es decir, tendentes a aceptar en muchos más casos la hipótesis nula de los que se hubiera aceptado de haber utilizado el estadístico λ de Pearson. Este hecho hace desaconsejar su uso, salvo en aquellas situaciones en las que una reducción del número de clases condujese a una χ^2 sin grados de libertad.

Observación 3

Digamos también que, siempre que sea posible, debemos elegir los sucesos E_k de forma que sea $p_i = 1/k$, consiguiendo de esta manera una mejor aproximación de la distribución de λ a la χ^2 .

8.2.3. Contraste de homogeneidad de varias muestras

Este contraste tiene por objeto averiguar si existen o no diferencias significativas entre r poblaciones, de las que se han extraído sendas muestras aleatorias simples.

Es decir, es un contraste semejante —en cuanto a sus propósitos— a los contrastes de análisis de la varianza que estudiaremos en el Capítulo 9, aunque con la diferencia de que ahora los datos son frecuencias o recuentos del número

de individuos pertenecientes a cada una de las clases en las que se han dividido las poblaciones, y no valores de una variable observable.

Ejemplo 8.8

Con objeto de averiguar si existen o no diferencias significativas entre los hábitos fumadores de tres comunidades, se seleccionó una muestra aleatoria simple de 100 individuos de cada una de las tres comunidades, obteniéndose los siguientes resultados,

Comunidad	fumadores	no fumadores	Total
A	13	87	100
B	17	83	100
C	18	82	100
	48	252	300

Tabla 8.6

¿Pueden considerarse homogéneas las tres poblaciones en cuanto a sus hábitos fumadores?

En general, tendremos s clases en las que se han dividido las r poblaciones, estando clasificadas las r muestras aleatorias extraídas (una de cada población) en una tabla de frecuencias absolutas de la forma

Muestras	Clases				Totales
	C_1	C_2	\cdots	C_s	
M_1	n_{11}	n_{12}	\cdots	n_{1s}	n_1
M_2	n_{21}	n_{22}	\cdots	n_{2s}	n_2
\vdots	\cdots	\cdots	\cdots	\cdots	\vdots
M_r	n_{r1}	n_{r2}	\cdots	n_{rs}	n_r
Totales	m_1	m_2	\cdots	m_s	n

Tabla 8.7

en donde n_{ij} es el número de individuos de la muestra i -ésima que pertenecen a la clase j -ésima, $n_i = \sum_{j=1}^s n_{ij}$ el tamaño de la muestra i -ésima, $m_j = \sum_{i=1}^r n_{ij}$ la frecuencia absoluta marginal de la clase C_j y $n = \sum_{i=1}^r n_i = \sum_{j=1}^s m_j = \sum_{i=1}^r \sum_{j=1}^s n_{ij}$ el tamaño muestral.

El propósito de este test es contrastar la hipótesis nula H_0 : las r poblaciones son homogéneas, frente a la alternativa de no serlo.

Denominando p_j a la probabilidad teórica de la clase C_j , $j = 1, \dots, s$, podemos aplicar a la muestra M_i el estadístico λ de Pearson, obteniendo que

$$\sum_{j=1}^s \frac{(n_{ij} - n_i p_j)^2}{n_i p_j} \approx \chi_{s-1}^2$$

Si es cierta la hipótesis nula de igualdad de las r poblaciones, la probabilidad de la clase C_j seguirá siendo p_j en cada una de las r muestras y, como además éstas son independientes, la suma de los estadísticos de Pearson utilizados en cada una de las muestras tendrá también una distribución χ^2 (aproximadamente), de grados de libertad la suma de los grados de libertad de cada una de las χ^2 . Por tanto, será

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_i p_j)^2}{n_i p_j} \approx \chi_{r(s-1)}^2$$

De forma casi generalizada, las probabilidades p_j serán desconocidas, por lo que será necesario estimarlas mediante los estimadores de máxima verosimilitud, $\hat{p}_j = m_j/n$, teniendo que restar $s - 1$ grados de libertad a la $\chi_{r(s-1)}^2$ (una de las p_j no hay que estimarla puesto que la suma de todas ellas debe ser 1) quedando en definitiva que

$$\sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n} \approx \chi_{(r-1)(s-1)}^2$$

al ser $r(s - 1) - (s - 1) = (s - 1)(r - 1)$.

Contraste de hipótesis

Supongamos n datos como los de la Tabla 8.7. Para contrastar, a nivel α , la hipótesis nula H_0 : *son homogéneas las r poblaciones*, de las que se extraen las muestras M_1, \dots, M_r , frente a la alternativa de *no homogeneidad de las r poblaciones*, y si es

$$\lambda = \sum_{i=1}^r \sum_{j=1}^s \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n}$$

entonces el contraste óptimo a utilizar consiste en

- Aceptar H_0 si $\lambda < \chi_{(r-1)(s-1);\alpha}^2$
- Rechazar H_0 si $\lambda \geq \chi_{(r-1)(s-1);\alpha}^2$

Observación 4

De forma análoga a como ocurría en la sección anterior, con objeto de obtener una rápida convergencia hacia la χ^2 , las frecuencias esperadas deberán ser suficientemente grandes —digamos $n_i m_j/n \geq 5$.

Si esto no se cumple, deberemos agrupar clases contiguas —reduciendo adecuadamente los grados de libertad—, o de forma alternativa utilizar el estadístico corregido

$$\lambda_c = \sum_{i=1}^r \sum_{j=1}^s \frac{(|n_{ij} - n_i m_j / n| - 0.5)^2}{n_i m_j / n}$$

manteniendo, en este caso, los mismos grados de libertad.

Como ocurría antes, la corrección de Yates conducirá, en general, a tests más conservadores.

Ejemplo 8.8 (continuación)

Como el estadístico de Pearson toma el valor

$$\lambda = 1.042 < 5.991 = \chi^2_{2;0.05}$$

aceptamos la hipótesis nula de homogeneidad de las tres poblaciones en cuanto a sus hábitos fumadores.

Para resolver este ejercicio con R, primero incorporamos los datos en (1) creando la matriz de datos. En (2) y (3) asignamos nombres a las clases que presentan las variables en estudio. Finalmente, en (4) ejecutamos la función `chisq.test` que nos dará la información necesaria sobre el test de homogeneidad de las tres poblaciones.

```
> fuma<-matrix(c(13,17,18,87,83,82),ncol=2) (1)
> colnames(fuma)<-c("fumadores","no fumadores") (2)
> rownames(fuma)<-c("A","B","C") (3)
> chisq.test(fuma) (4)
      Pearson's Chi-squared test
data:  fuma
X-squared = 1.0417, df = 2, p-value = 0.594 (5)
> chisq.test(fuma)$expected (6)
```

En concreto, en (5) obtenemos el valor del estadístico de Pearson, $\lambda = 1.0417$ y del p-valor, 0.594, suficientemente grande como para concluir con la aceptación de la hipótesis nula de homogeneidad de las tres poblaciones, es decir, con que no existen diferencias significativas entre las tres comunidades.

Observamos que, como la tabla de datos no es 2×2 , la función `chisq.test` no calcula la corrección de Yates, por lo que es interesante analizar si las frecuencias esperadas son o no menores que 5 para calcularla si es necesario (es decir, si son menores que 5) mediante su fórmula general; no con la función `chisq.test` ya que ésta no la calcula si no son tablas 2×2 . Ejecutando (6) observamos que las frecuencias esperadas son lo suficientemente grandes como para no requerir corrección de Yates.

```
> chisq.test(fuma)$expected (6)
      fumadores no fumadores
A           16           84
B           16           84
C           16           84
```

Ejemplo 8.2 (continuación)

En este caso, el estadístico de Pearson es

$$\lambda = 65'63 > 11'07 = \chi^2_{5;0'05}$$

con lo que puede concluirse con la afirmación de que existe diferencia significativa entre las seis comunidades en cuanto a la existencia de caries dental.

Para resolver este problema con R, primero creamos la matriz de datos en (1) y, aunque ello es irrelevante en la ejecución del test, a continuación asignamos nombres a las filas y columnas. Finalmente, en (2), ejecutamos el test de homogeneidad de varias muestras, cuyos resultados, que aparecen en (3), indican rechazar claramente la hipótesis nula de homogeneidad de las tres poblaciones en cuanto a la presencia o no de caries.

```
> caries<-matrix(c(38,8,30,44,64,32,87,117,95,81,61,93),ncol=2) (1)
```

```
> colnames(caries)<-c("niños sin caries","niños con caries")
```

```
> rownames(caries)<-c("A","B","C","D","E","F")
```

```
> chisq.test(caries) (2)
```

Pearson's Chi-squared test

```
data: caries
```

```
X-squared = 65.8552, df = 5, p-value = 7.448e-13 (3)
```

8.2.4. Contraste de independencia de caracteres

El último contraste de la χ^2 que vamos a estudiar analiza la posible independencia entre dos caracteres observados en los individuos de una población.

Ejemplo 8.9

Se desea investigar una posible dependencia entre los síntomas de *deterioro psicogenético del pensamiento y depresión* en una determinada población.

Con tal fin se seleccionó una muestra aleatoria simple de 100 individuos de la población en cuestión, la cual dio los siguientes resultados

<i>Deterioro</i>	<i>Depresión</i>		
	SI	NO	
SI	38	9	
NO	31	22	
			100

En general tendremos dos caracteres, *A* con *a* modalidades y *B* con *b* modalidades, estando los *n* individuos de la muestra clasificados en una *tabla de doble entrada* o de *contingencia* de la forma

B	1	2	...	b	
A					
1	n_{11}	n_{12}	...	n_{1b}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2b}	$n_{2.}$
\vdots					\vdots
a	n_{a1}	n_{a2}	...	n_{ab}	$n_{a.}$
	$n_{.1}$	$n_{.2}$...	$n_{.b}$	n

en donde n_{ij} es el número de individuos de la muestra de tamaño n que presentan a la vez la modalidad i -ésima del carácter A y la j -ésima del carácter B .

Las hipótesis a contrastar son, H_0 : *los caracteres A y B son independientes*, frente a la alternativa, H_1 : *A y B no son independientes*.

Llamando p_i a la probabilidad —marginal— de obtener un individuo de la población que presente la modalidad i -ésima del carácter A , y q_j la probabilidad —marginal— de obtener un individuo de la población que presente la modalidad j -ésima del carácter B , si la hipótesis nula fuese correcta, la probabilidad p_{ij} de obtener un individuo de la población que presente a la vez la modalidad i -ésima del carácter A y j -ésima del carácter B , sería $p_i \cdot q_j$, con lo que, en la muestra de tamaño n , cabría esperar que $n \cdot p_i \cdot q_j$ presenten a la vez ambas modalidades.

La comparación de las frecuencias observadas, n_{ij} , con las esperadas, $n \cdot p_i \cdot q_j$, para cada una de las $k = a \cdot b$ clases se hará a través del estadístico λ de Pearson.

En efecto, por lo estudiado en la introducción de la sección, el estadístico

$$\lambda = \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij} - n p_i q_j)^2}{n p_i q_j}$$

seguirá, aproximadamente, una distribución χ^2 con $ab - 1$ grados de libertad.

Como ocurría en la sección anterior, las probabilidades p_i y q_j serán habitualmente desconocidas, por lo que deberemos estimarlas utilizando sus estimadores máximo-verosímiles, las frecuencias relativas, $\hat{p}_i = n_{i.}/n$ y $\hat{q}_j = n_{.j}/n$, quedando el estadístico de Pearson a utilizar habitualmente de la forma

$$\lambda = \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n}$$

el cual seguirá —aproximadamente— una distribución χ^2 con $ab - 1 - (a - 1) - (b - 1) = (a - 1)(b - 1)$ grados de libertad.

Contraste de hipótesis

Así pues, para contrastar, a nivel α , la hipótesis nula H_0 : *los caracteres A y B son independientes*, frente a la alternativa H_1 : *los caracteres A y B no son independientes*, el contraste a utilizar es

- Se acepta H_0 si $\lambda < \chi^2_{(a-1)(b-1);\alpha}$
- Se rechaza H_0 si $\lambda \geq \chi^2_{(a-1)(b-1);\alpha}$

siendo

$$\lambda = \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij} - n_{i.} n_{.j} / n)^2}{n_{i.} n_{.j} / n}$$

Observación 5

De nuevo las frecuencias esperadas no deben ser muy pequeñas (digamos $n_{i.} n_{.j} / n \geq 5$), debiendo agruparse las clases contiguas en caso contrario, o utilizar la corrección de Yates.

Por último queremos hacer notar que, matemáticamente, el contraste de homogeneidad de varias muestras y el de independencia de caracteres han resultado técnicamente idénticos.

Es decir, que al expresarse los datos de ambos en una tabla de contingencia, el estadístico de Pearson *se calcula* de la misma manera.

Existe, no obstante, una diferencia fundamental entre ambos: en el contraste de homogeneidad, los totales marginales $n_{i.}$ son los que fija el investigador, el cual decide, por tanto, cuántos individuos deben elegirse de cada población. Por otro lado, en el contraste de independencia, lo que fija el experimentador es n , quedando los totales marginales $n_{i.}$ fuera del control del investigador.

Ejemplo 8.9 (continuación)

En este caso, el estadístico de Pearson es tal que

$$\lambda = 5'823 > 3'841 = \chi^2_{1;0'05}$$

por lo que deberemos rechazar la independencia entre la depresión y el deterioro psicogenético, a nivel $\alpha = 0'05$.

Para resolver este problema con R primero incorporamos los datos como sigue,

```
> deterioro<-matrix(c(38,31,9,22),ncol=2)
> colnames(deterioro)<-c("SI","NO")
> rownames(deterioro)<-c("SI","NO")
```

Luego, observamos que, al ser la tabla de contingencia de dimensión 2×2 , podemos o no calcular la corrección de Yates; como por defecto la función `chisq.test` la calcula, ejecutando (1) obtenemos el valor del p-valor en (2) con dicha corrección (que sugiere rechazar la hipótesis nula de independencia aunque no con mucha claridad). Si no calculamos el test con la corrección de Yates ejecutando (3), el p-valor, dado en (4), es un poco más contundente en cuanto al rechazo de la hipótesis nula (como de hecho ocurre siempre con la corrección de Yates; es un poco *conservadora*). Parece, por tanto razonable analizar si la mencionada corrección es necesaria o no; para ello ejecutamos (5) obteniendo frecuencias esperadas no menores que 5, por lo que la corrección de Yates es un tanto engañosa y no deberíamos utilizarla, concluyendo con el rechazo de H_0 .

```
> chisq.test(deterioro) (1)
Pearson's Chi-squared test with Yates' continuity correction
```

```
data: deterioro
X-squared = 4.8243, df = 1, p-value = 0.02806 (2)
```

```
> chisq.test(deterioro,correct=F) (3)
Pearson's Chi-squared test
```

```
data: deterioro
X-squared = 5.8227, df = 1, p-value = 0.01582 (4)
```

```
> chisq.test(deterioro)$expected (5)
      SI      NO
SI 32.43 14.57
NO 36.57 16.43
```

Ejemplo 8.3 (continuación)

En este ejemplo, el estadístico de Pearson es tal que

$$\lambda = 41'83 > 16'92 = \chi_{9,0'05}^2$$

por lo que deberemos rechazar la independencia entre el peso y la talla de los individuos de la población, a nivel $\alpha = 0'05$.

No obstante, las frecuencias esperadas de varias clases (entre paréntesis en la Tabla 8.8) son algo menores que 5. Aunque el valor del estadístico, el tamaño de la tabla y el de la muestra son suficientemente grandes como para no hacer necesario usar la corrección de Yates, calcularemos su valor, por razones pedagógicas.

Talla	1'55 – 1'65	1'65 – 1'75	1'75 – 1'85	1'85 – 1'95	
Peso					
50-60	10 (3'78)	8 (7'14)	2 (6'72)	1 (3'36)	21
60-70	6 (5'04)	14 (9'52)	6 (8'96)	2 (4'48)	28
70-80	2 (5'94)	8 (11'22)	18 (10'56)	5 (5'28)	33
80-90	0 (3'24)	4 (6'12)	6 (5'76)	8 (2'88)	18
	18	34	32	16	100

Tabla 8.8

Su valor es

$$\lambda_c = \sum_{i=1}^a \sum_{j=1}^b \frac{(|n_{ij} - n_{i.} n_{.j}/n| - 0'5)^2}{n_{i.} n_{.j}/n} = 33.$$

Como se ve, algo menor —más conservador— que el estadístico de Pearson, aunque sigue rechazándose la hipótesis nula.

La resolución de este ejercicio con R comienza con la incorporación de los datos en la primera sentencia que sigue. Como la tabla de contingencia de los datos no es 2×2 , la función `chisq.test`, ejecutada en (1), no va a calcular la corrección de Yates aunque fuera necesaria. El p-valor del test, dado en (2), indica rechazar la hipótesis nula de independencia claramente, pero el programa ya nos avisa de que la aproximación puede ser incorrecta porque, si calculamos las frecuencias esperadas con (3), vemos que hay varias menores que 5 siendo necesario el uso de la corrección de Yates, como hicimos más arriba, cuyo valor era 33, confirmando (a pesar de ser un poco conservadora, $33 < 44'834$), el rechazo de la hipótesis H_0 .

```
> pesotalla<-matrix(c(10,6,2,0,8,14,8,4,2,6,18,6,1,2,5,8),ncol=4)
> chisq.test(pesotalla) (1)
```

Pearson's Chi-squared test

```
data: pesotalla
X-squared = 41.834, df = 9, p-value = 3.524e-06
(2)
```

Warning message:

```
In chisq.test(pesotalla) : Chi-squared approximation may be incorrect
```

```
> chisq.test(pesotalla)$expected (3)
  [,1] [,2] [,3] [,4]
[1,] 3.78 7.14 6.72 3.36
[2,] 5.04 9.52 8.96 4.48
[3,] 5.94 11.22 10.56 5.28
[4,] 3.24 6.12 5.76 2.88
```

Warning message:

```
In chisq.test(pesotalla) : Chi-squared approximation may be incorrect
```

Una última cuestión a destacar sobre los tests de independencia de caracteres. Si se acepta la hipótesis nula de independencia de las dos variables, ya hemos terminado, pero si se acepta la hipótesis alternativa de no independencia, ¿es posible averiguar qué valores de las dos variables no independientes están más relacionados? La respuesta a esta pregunta es el Método Estadístico del *Análisis de Correspondencias*, el cual se estudia en el Capítulo 3 del texto TA.

8.3. Tests relativos a una muestra y datos apareados

La hipótesis nula que aquí contrastaremos hará referencia a la mediana de la población de donde se extrajeron los datos o, si son datos apareados, a la mediana de la diferencia de las variables (que puede ser distinta de la diferencia de las medianas). Es decir, la hipótesis nula será $H_0 : M = M_0$ que se contrastará frente a la hipótesis alternativa $H_1 : M \neq M_0$. También se podrán contrastar, lógicamente, las correspondientes hipótesis unilaterales.

8.3.1. El contraste de los signos

El *test de los signos*, al igual que el resto de los estudiados en el capítulo, no necesita suponer una distribución modelo específica para la variable aleatoria en estudio; sólo se exige que la distribución modelo sea de tipo continuo, al menos en un entorno de la mediana poblacional M . Pero además, este test es tan genérico que no requiere ni de los valores de las observaciones, sólo de sus *rangos*, es decir, para ser aplicado sólo necesita de las *ordenaciones* de las observaciones, y no los valores numéricos de éstas.

Para su definición deberemos distinguir los casos en los que las hipótesis a contrastar sean bilaterales o sean unilaterales.

$H_0 : M = M_0$ $H_1 : M \neq M_0$

Dada una muestra aleatoria simple de la población, X_1, \dots, X_n , si la hipótesis nula $H_0 : M = M_0$ es cierta, aproximadamente la mitad de las observaciones serán menores que M_0 y la otra mitad mayores, ya que la mediana poblacional se define como aquel valor M tal que

$$P\{X < M\} = P\{X > M\} = 0'5.$$

Por tanto, si consideramos como estadístico del contraste el *número T de observaciones mayores que M_0* , o equivalentemente, el *número de signos positivos* de entre todas las diferencias $X_i - M_0$, $i = 1, \dots, n$, el observar un valor de T muy grande o muy pequeño tenderá a desacreditar la hipótesis nula en favor de la alternativa.

A pesar de ser éste un contraste no paramétrico, con objeto de determinar los puntos críticos del contraste que, basado en los anteriores razonamientos definiremos más adelante, necesitamos conocer la distribución en el muestreo del estadístico del test, T , con objeto de poder precisar lo que se entiende por *muy grande* o *muy pequeño*.

Afortunadamente, la distribución de T bajo la hipótesis nula no depende del modelo de X . Por esta razón, a este tipo de contrastes se les suele deno-

minar de *distribución libre*, además de no paramétricos. Si llamamos *éxito* al suceso en el que $X_i > M_0$ y *fracaso* al suceso $X_i < M_0$, T será el número de éxitos en n pruebas de Bernoulli, con lo que su distribución, si es cierta la hipótesis nula, será (Sección 4.4.1) binomial $B(n, 0'5)$.

Contraste de hipótesis

Como antes dijimos, si el valor de T es muy grande o muy pequeño, rechazaremos la hipótesis nula $H_0 : M = M_0$, aceptando en consecuencia la alternativa $H_1 : M \neq M_0$. En concreto, fijado un nivel de significación α

- Se acepta H_0 si $t_{1-\alpha/2} < T < t_{\alpha/2}$
- Se rechaza H_0 si $T \leq t_{1-\alpha/2}$ ó $T \geq t_{\alpha/2}$

en donde t_β es el valor de una binomial $B(n, 0'5)$ que deja a la derecha una área de probabilidad β , es decir, tal que $P\{W \geq t_\beta\} = \beta$ con $W \sim B(n, 0'5)$.

Como por las propiedades de la distribución binomial, es $t_{1-\alpha/2} = n - t_{\alpha/2}$, podemos expresar todo el test en función del punto crítico $t_{\alpha/2}$ en la forma

- Se acepta H_0 si $n - t_{\alpha/2} < T < t_{\alpha/2}$
- Se rechaza H_0 si $T \leq n - t_{\alpha/2}$ ó $T \geq t_{\alpha/2}$

en donde $t_{\alpha/2}$ es tal que

$$\sum_{t=t_{\alpha/2}}^n \binom{n}{t} (0'5)^n = \frac{\alpha}{2}.$$

Como la distribución binomial es de tipo discreto, es posible que no exista ningún valor $t_{\alpha/2}$ que cumpla la relación anterior. Por tanto, el $t_{\alpha/2}$ que se toma es el menor número entero tal que

$$\sum_{t=t_{\alpha/2}}^n \binom{n}{t} (0'5)^n \leq \frac{\alpha}{2}.$$

Ejemplo 8.10

Se realizó un estudio con objeto de averiguar si el retraso de los trabajadores de una empresa es, en promedio, de 5 minutos, o si por el contrario es significativamente mayor o menor. Para ello se seleccionó al azar una muestra de 11 empleados de la firma y se midió el tiempo en minutos que llegaron tarde el día que fueron seleccionados. Los resultados obtenidos fueron

Empleado	1	2	3	4	5	6	7	8	9	10	11
Retraso	2	0'1	7	1'8	4	2'3	5'6	7'4	5'1	6'1	6

Las hipótesis que se contrastaron fueron $H_0 : M = 5$ frente a $H_1 : M \neq 5$. Fijado como nivel de significación $\alpha = 0'1$, al ser

$$\sum_{t=9}^{11} \binom{11}{t} (0'5)^{11} = 0'0328 < 0'05$$

y

$$\sum_{t=8}^{11} \binom{11}{t} (0'5)^{11} = 0'1134 > 0'05$$

será $t_{\alpha/2} = 9$. Como es $2 < T = 6 < 9$, aceptaremos $H_0 : M = 5$.

Para resolver este ejemplo con R, observemos primero que, al ser un test bilateral, deberemos dejar a cada lado de la región de aceptación una probabilidad de $\alpha/2 = 0'05$, con lo que podemos obtener con R los dos extremos del intervalo de aceptación ejecutando (1) y (2)

```
> qbinom(0.05,11,0.5) (1)
```

```
[1] 3
```

```
> qbinom(1-0.05,11,0.5) (2)
```

```
[1] 8
```

Por tanto, la región de aceptación es $[3, 8]$. Como $T = 6$ está en la región de aceptación, aceptaremos H_0 .

$$H_0 : M \leq M_0$$

$$H_1 : M > M_0$$

En este caso, si el número de *signos positivos* es grande, rechazaremos H_0 en favor de H_1 . En concreto, fijado un nivel de significación α

• Se acepta H_0 si $T < t_\alpha$

• Se rechaza H_0 si $T \geq t_\alpha$

en donde t_α es el valor de una binomial $B(n, 0'5)$ que deja a la derecha una área de probabilidad α , es decir, tal que $P\{W \geq t_\alpha\} = \alpha$ con $W \sim B(n, 0'5)$.

De nuevo puede ocurrir que α no sea accesible, por lo que t_α se toma como el menor número entero tal que

$$\sum_{t=t_\alpha}^n \binom{n}{t} (0'5)^n \leq \alpha.$$

$$H_0 : M \geq M_0$$

$$H_1 : M < M_0$$

Ahora, un número pequeño de *signos positivos*, es decir un valor de T pequeño, desacreditará la hipótesis nula en favor de la alternativa. Por tanto, fijado un nivel de significación α

- Se acepta H_0 si $T > n - t_\alpha$
- Se rechaza H_0 si $T \leq n - t_\alpha$

siendo de nuevo t_α el menor número entero tal que

$$\sum_{t=t_\alpha}^n \binom{n}{t} (0'5)^n \leq \alpha.$$

Muestras grandes

Vimos en la Sección 4.7 que cuando el tamaño muestral es suficientemente grande, la distribución binomial se puede aproximar por una distribución normal. Esta aproximación es especialmente buena cuando el parámetro p de dicha distribución es 0'5, como aquí sucede.

Por tanto, cuando n es grande —digamos $n \geq 12$ —, podremos determinar los puntos críticos del test de los signos por la distribución normal, habiéndose añadido un factor de corrección por estar aproximando una distribución discreta por una de tipo continuo como es la normal.

En concreto, si fijado un nivel de significación α , es como siempre z_α el valor de la abscisa de una normal $N(0, 1)$ que deja a la derecha una área de probabilidad α , el punto crítico t_α en los contrastes anteriores es

$$t_\alpha = 0'5 (z_\alpha \sqrt{n} + n + 1).$$

Obviamente, para el contraste bilateral deberá cambiarse α por $\alpha/2$ en la fórmula anterior.

El problema de los empates

Aunque teóricamente no deberían observarse valores iguales a la mediana a contrastar M_0 , al haberse supuesto la distribución continua en la vecindad de la mediana, de hecho se producen.

Existen varias alternativas para solucionar este problema. La primera y más razonable, es medir con mayor precisión cerca de M_0 de forma que podamos discriminar si el dato es menor o mayor que M_0 , para poder decidir el signo aportado por el valor muestral.

Si los datos ya vienen dados, lo más aconsejable es ignorar las diferencias cero, disminuyendo, en consecuencia, el tamaño de la muestra.

Ejemplo 8.11

Se realizó un estudio con objeto de averiguar si el número de linfocitos en los animales de laboratorio era mayor de 2500 por milímetro cúbico.

Para ello se seleccionaron al azar 15 de dichos animales para los que se obtuvieron los siguientes datos sobre su número de linfocitos, expresados en miles por milímetro cúbico

Animal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Linfo.	2'3	2'9	1'6	2	4'2	3'1	2'3	2'5	2	1'6	3'3	4'1	4	3	2'8

Tabla 8.9

Las hipótesis a contrastar son $H_0 : M \leq 2'5$ frente a $H_1 : M > 2'5$.

Al haberse observado un valor igual a 2'5, lo ignoraremos, considerando como tamaño muestral $n = 14$, el cual es suficientemente grande como para utilizar la aproximación normal.

Si fijamos un nivel de significación $\alpha = 0'05$, es $z_\alpha = 1'645$, con lo que el punto crítico será

$$t_\alpha = 0'5 (1'645 \sqrt{14} + 15) = 10'577513.$$

Al ser $T = 8 < 10'58 = t_\alpha$, aceptaremos H_0 . El p-valor $= P\{T \geq 8\} = \sum_{i=8}^{14} \binom{14}{i} (0'5)^{14} = 0'395264$, confirma la decisión tomada.

La forma más rápida de ejecutar el test solicitado es calcular el p-valor de dicho test; es decir, calcular $P\{T \geq 8\} = 1 - P\{T \leq 7\}$ siendo $T \sim B(14, 0'5)$. Ahora no tenemos que estar preocupados de hacer aproximaciones normales; solamente de calcular la probabilidad cola de una binomial. Para ello, ejecutando con R la sentencia (1), obtenemos el valor 0'3952637, muy similar al que calculamos más arriba con la aproximación normal, 0'395264. Estos valores del p-valor conducen a aceptar la hipótesis nula H_0 y a no poder concluir con que los animales de laboratorio, como pensábamos, tengan un número *mediano* de linfocitos mayor de 2'5.

```
> 1-pbinom(7,14,0.5)
```

(1)

```
[1] 0.3952637
```

Datos apareados

Los resultados vistos hasta ahora pueden aplicarse al caso de datos apareados $(X_1, Y_1), \dots, (X_n, Y_n)$ definiendo la variable diferencia $D = X - Y$.

Es necesaria, no obstante, mucha cautela en las conclusiones, ya que los contrastes que se hagan harán referencia a la mediana de la variable diferencia, pero no necesariamente a la diferencia de las medianas. Ambas cantidades coincidirán cuando las poblaciones X e Y sean simétricas con el mismo centro de simetría y además la población diferencia D también sea simétrica.

8.3.2. El contraste de los rangos signados de Wilcoxon

El contraste de los signos acabado de estudiar tiene la gran ventaja de su sencillez, pero el inconveniente de utilizar solamente el signo suministrado por cada observación, es decir, si es $X_i - M_0 > 0$, o si es $X_i - M_0 < 0$, sin considerar la magnitud de dicha diferencia, y esta información que proporciona la muestra debe ser utilizada cuando exista, con objeto de mejorar la potencia del contraste.

El *contraste de rangos signados de Wilcoxon*, que aquí estudiaremos, recoge esta información, aunque requiere a cambio, que la distribución modelo sea continua y simétrica.

Aunque no haremos referencia explícita de ello, el test de los rangos signados de Wilcoxon se puede utilizar también en el caso de datos apareados, análogamente a como ocurría con el test de los signos.

$H_0 : M = M_0$ $H_1 : M \neq M_0$

Sea X_1, \dots, X_n una muestra aleatoria de la variable en observación X y $D_i = X_i - M_0$ las diferencias de la muestra con la mediana a contrastar M_0 .

Si ordenamos sus valores absolutos $|D_1|, \dots, |D_n|$ asignando a cada uno su rango $r(|D_i|)$, es decir, al menor $|D_i|$ el valor 1 y así hasta el último al que asignamos el valor n , el test de Wilcoxon utiliza como estadístico de contraste, T^+ , la *suma de los rangos de las diferencias positivas*, es decir, los *rangos signados*. Analíticamente,

$$T^+ = \sum_{i=1}^n z_i r(|D_i|)$$

con

$$z_i = \begin{cases} 1 & \text{si } D_i > 0 \\ 0 & \text{si } D_i < 0. \end{cases}$$

Con objeto de determinar los puntos críticos del test deberemos conocer la distribución en el muestreo de T^+ .

Los valores extremos de T^+ son 0 (todas las diferencias son negativas) y $n(n+1)/2$ (todas las diferencias son positivas).

La determinación de la función de masa —es una variable discreta— de T^+ resulta complicada por lo que en la *Tabla 13* de ADD aparecen los puntos críticos para tamaños muestrales pequeños y, en el caso de que el tamaño muestral sea suficientemente grande, se puede aproximar, como veremos más adelante, por una distribución normal.

Contraste de hipótesis

Valores muy grandes o muy pequeños de T^+ desacreditarán la hipótesis nula $H_0 : M = M_0$ en favor de la alternativa $H_1 : M \neq M_0$, con lo que fijado un nivel de significación α ,

- Se acepta H_0 si $\frac{n(n+1)}{2} - t_{\alpha/2} < T^+ < t_{\alpha/2}$
- Se rechaza H_0 si $T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2}$ ó $T^+ \geq t_{\alpha/2}$

en donde $t_{\alpha/2}$ es tal que $P\{T^+ \geq t_{\alpha/2}\} = \alpha/2$.

Como la distribución binomial es discreta, es posible que no exista ningún valor $t_{\alpha/2}$ que cumpla la relación anterior. Por tanto, el $t_{\alpha/2}$ que se toma es el menor número entero tal que

$$P\{T^+ \geq t_{\alpha/2}\} \leq \frac{\alpha}{2}.$$

Contraste de los rangos signados de Wilcoxon con R

El test de los rangos signados de Wilcoxon se ejecuta con la función `wilcox.test`, que será la misma que utilizaremos para el contraste de Wilcoxon-Mann-Whitney más adelante,

```
wilcox.test(x, alternative="two.sided", mu=0, exact=T, correct=T)
```

en donde incluiremos en el primer argumento `x` el vector de observaciones. Con el argumento `alternative` podemos elegir el tipo de test que vamos a ejecutar, bilateral (que es el que se utiliza por defecto), `less` o `greater` si la hipótesis alternativa que queremos contrastar es, respectivamente, menor o mayor. Con `mu` podemos señalar el valor de la hipótesis a contrastar, eligiendo la función el valor 0 por defecto. Con `exact` indicamos si queremos que R calcule el valor exacto de la distribución del estadístico T^+ de Wilcoxon (opción tomada por defecto) o que calcule el valor aproximado del p-valor para muestras grandes cuya expresión daremos más abajo, opción que se elige

ejecutando `exact=F`. Finalmente, con `correct` indicamos si queremos utilizar la corrección de continuidad.

Recordamos que, al igual que pasaba con el test de los signos, debemos eliminar las observaciones iguales a la hipótesis a contrastar reduciendo el tamaño muestral. El ordenador nos avisará si aparecen empates entre los valores absolutos de las diferencias a ordenar por rangos aunque no las elimina sino que las promedia como indicaremos más adelante.

Ejemplo 8.12

Se está llevando a cabo un experimento con objeto de medir los efectos que produce la inhalación prolongada de óxido de cadmio.

Los niveles de hemoglobina, en gramos, de cuatro ratones elegidos al azar de un laboratorio en donde existe la contaminación en estudio fueron 14'4, 15'9, 13'8, 15'3.

¿Puede admitirse a nivel $\alpha = 0'3$ la hipótesis nula de un promedio poblacional de 15 gramos? Las hipótesis que contrastaremos son $H_0 : M = 15$ frente a $H_1 : M \neq 15$. Como el tamaño de la muestra es $n = 4$, podemos utilizar la distribución de T^+ antes determinada.

Al ser

$$P\{T^+ \geq 9\} = 0'125 \quad \text{y} \quad P\{T^+ \geq 8\} = 0'1875$$

será $t_{\alpha/2} = t_{0'15} = 9$ y como es

$$\begin{array}{lll} D_1 = 14'4 - 15 = -0'6 & |D_1| = 0'6 & r(|D_1|) = 2 \\ D_2 = 15'9 - 15 = 0'9 & |D_2| = 0'9 & r(|D_2|) = 3 \\ D_3 = 13'8 - 15 = -1'2 & |D_3| = 1'2 & r(|D_3|) = 4 \\ D_4 = 15'3 - 15 = 0'3 & |D_4| = 0'3 & r(|D_4|) = 1 \end{array}$$

será $T^+ = 3 + 1 = 4$. Dado que se verifica la relación

$$\frac{n(n+1)}{2} - t_{\alpha/2} = 1 < T^+ = 4 < 9 = t_{\alpha/2}$$

aceptaremos H_0 .

Para resolver este ejercicio con R, después de incorporar los datos en (1), ejecutamos (2) para obtener en (3) el valor del estadístico $T^+ = 4$ y el p-valor, 0'875, suficientemente grande como para aceptar la hipótesis nula.

```
> x<-c(14.4,15.9,13.8,15.3) (1)
```

```
> wilcox.test(x,mu=15,correct=F) (2)
```

Wilcoxon signed rank test

```
data: x
```

```
V = 4, p-value = 0.875 (3)
```

```
alternative hypothesis: true location is not equal to 15
```

$$H_0 : M \leq M_0$$

$$H_1 : M > M_0$$

En este caso, fijado un nivel de significación α

- Se acepta H_0 si $T^+ < t_\alpha$
- Se rechaza H_0 si $T^+ \geq t_\alpha$

en donde de nuevo t_α es el menor número entero tal que

$$P\{T^+ \geq t_\alpha\} \leq \alpha.$$

$$\begin{aligned} H_0 : M &\geq M_0 \\ H_1 : M &< M_0 \end{aligned}$$

Para este último contraste unilateral, fijado un nivel de significación α

- Se acepta H_0 si $T^+ > \frac{n(n+1)}{2} - t_\alpha$
- Se rechaza H_0 si $T^+ \leq \frac{n(n+1)}{2} - t_\alpha$

siendo de nuevo t_α el menor número entero tal que

$$P\{T^+ \geq t_\alpha\} \leq \alpha.$$

Muestras grandes

La distribución del estadístico T^+ se puede aproximar por una $N(0, 1)$ aplicando el teorema central del límite, cuando el tamaño muestral es suficientemente grande (digamos $n > 15$). En ese caso es

$$\frac{4T^+ - n(n+1)}{\sqrt{2n(n+1)(2n+1)/3}} \approx N(0, 1)$$

con lo que, despejando, el punto crítico (para el contraste unilateral) quedaría

$$t_\alpha = \frac{n(n+1)}{4} + \frac{1}{4} z_\alpha \sqrt{\frac{2n(n+1)(2n+1)}{3}}.$$

Si la muestra no es muy grande podría utilizarse una corrección de continuidad, restando 0'5 al valor absoluto del numerador de la distribución de T^+ .

El problema de los empates y el de las diferencias iguales

Aunque de nuevo, teóricamente, no deberían observarse ni valores iguales a la mediana a contrastar M_0 , ni diferencias D_i iguales, cuestión esta última que produce problemas al asignar los rangos, de hecho se obtendrán.

Respecto a los empates, la solución que se propone es la misma que en el test de los signos: ignorarlos disminuyendo el tamaño de la muestra.

Por otro lado, si dos o más diferencias absolutas son iguales, $|D_i| = |D_j|$ para al menos un $i \neq j$, se propone tomar como rango común a todas las diferencias iguales, la media aritmética de los rangos que tendrían si fueran distinguibles, aunque conservando cada D_i su signo.

Ejemplo 8.11 (continuación)

A partir de los datos de la Tabla 8.9 obtenemos la siguiente tabla de diferencias D_i para cada uno de los individuos

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	-0'2	0'4	-0'9	-0'5	1'7	0'6	-0'2	0	-0'5	-0'9	0'8	1'6	1'5	0'5	0'3

y en valor absoluto, una vez eliminado el empate de la octava observación

Animal	1	2	3	4	5	6	7	9	10	11	12	13	14	15
$ D_i $	0'2	0'4	0'9	0'5	1'7	0'6	0'2	0'5	0'9	0'8	1'6	1'5	0'5	0'3

de rangos y signos

Animal	1	2	3	4	5	6	7	9	10	11	12	13	14	15
$r(D_i)$	1'5	4	10'5	6	14	8	1'5	6	10'5	9	13	12	6	3
Signos	-	+	-	-	+	+	-	-	-	+	+	+	+	+

Por tanto, el estadístico del test, suma de los rangos de las diferencias positivas, tomará el valor

$$T^+ = 4 + 14 + 8 + 9 + 13 + 12 + 6 + 3 = 69.$$

Para contrastar, a nivel $\alpha = 0'05$, la hipótesis $H_0 : M \leq 2'5$ frente a $H_1 : M > 2'5$, podemos determinar el punto crítico utilizando la aproximación normal, para la que se obtiene el valor

$$t_{0'05} = \frac{n(n+1)}{4} + \frac{1}{4} z_\alpha \sqrt{\frac{2n(n+1)(2n+1)}{3}} = \frac{14 \cdot 15}{4} + \frac{1}{4} 1'645 \sqrt{\frac{2 \cdot 14 \cdot 15 \cdot 29}{3}} = 78'704.$$

Como es $T^+ = 69 < 78'704$, aceptaremos H_0 . El p-valor es ahora,

$$P \left\{ Z > \frac{4 \cdot 69 - 14 \cdot 15}{\sqrt{2 \cdot 14 \cdot 15 \cdot 29/3}} \right\} = P\{Z > 1'04\} = 0'1492.$$

suficientemente grande como para confirmar la aceptación de H_0 .

Para resolver este ejemplo con R, primero incorporamos los datos en (1), puesto que no los habíamos incluido antes al ejecutar el test de los signos. Recordemos que este test no tiene en cuenta el valor de las observaciones; sólo si son mayores o menores que la hipótesis a contrastar. En (2) ejecutamos el test de Wilcoxon, calculando el valor aproximado del p-valor y sin corrección de continuidad.

```
> x<-c(2.3,2.9,1.6,2,4.2,3.1,2.3,2,1.6,3.3,4.1,4,3,2.8) (1)
```

```
> wilcox.test(x,alternative="greater",mu=2.5,exact=F,correct=F) (2)
```

Wilcoxon signed rank test

```
data: x
```

```
V = 69, p-value = 0.1498 (3)
```

```
alternative hypothesis: true location is greater than 2.5
```

En (3) obtenemos el valor del estadístico del test, $V=69$, y el valor aproximado del p-valor, 0'1498, suficientemente grande como para aceptar la hipótesis nula.

8.4. Tests relativos a dos muestras independientes

En esta sección estudiaremos dos contrastes no paramétricos para contrastar la hipótesis nula de igualdad de dos poblaciones independientes, expresada ésta mediante la igualdad de sus medianas poblacionales, $H_0 : M_X = M_Y$.

8.4.1. El contraste de Wilcoxon-Mann-Whitney

Este contraste, introducido primero por Wilcoxon en 1945, resultó ser equivalente al propuesto por Mann y Whitney dos años más tarde, de ahí que se le conozca con el nombre de los tres estadísticos que lo diseñaron.

Este test requiere que las distribuciones poblacionales F y G sean continuas.

$$H_0 : M_X = M_Y$$

$$H_1 : M_X \neq M_Y$$

Sea X_1, \dots, X_m una muestra aleatoria simple de tamaño m de la primera población e Y_1, \dots, Y_n una de tamaño n de la segunda.

La idea del contraste consiste en medir las magnitudes de los valores Y_i en relación con los X_i , es decir, las posiciones de los Y_i en la muestra conjunta de las X_i e Y_i . Si observamos que la mayoría de los Y_i están hacia el principio o hacia el final de la muestra conjunta, deberemos rechazar la hipótesis nula de igualdad de ambas poblaciones.

En concreto, si llamamos

$$D_{ij} = \begin{cases} 1 & Y_j < X_i \\ 0 & Y_j \geq X_i \end{cases}$$

$\forall i = 1, \dots, m$ y $j = 1, \dots, n$, el estadístico U en el que está basado el contraste es

$$U = \sum_{i=1}^m \sum_{j=1}^n D_{ij}$$

es decir, el número de Y_j que preceden estrictamente a cada X_i .

La distribución exacta en el muestreo de U es complicada, apareciendo en la *Tabla 14* de ADD los puntos críticos en el caso de tamaños muestrales pequeños. Además, veremos más adelante que cuando m y n sean mayores que 5, la aproximación normal es adecuada. Apuntemos, no obstante, que los valores de U están entre 0 y $m \cdot n$, así como que la distribución de U es simétrica respecto a su media $m \cdot n/2$.

Contraste de hipótesis

Valores muy grandes o muy pequeños de U desacreditarán la hipótesis nula de igualdad de ambas poblaciones. Así pues, fijado un nivel de significación α ,

- Se acepta H_0 si $m \cdot n - u_{m,n;\alpha/2} < U < u_{m,n;\alpha/2}$
- Se rechaza H_0 si $U \leq m \cdot n - u_{m,n;\alpha/2}$ ó $U \geq u_{m,n;\alpha/2}$

en donde $u_{m,n;\alpha/2}$ es el menor número entero tal que

$$P\{U \geq u_{m,n;\alpha/2}\} \leq \frac{\alpha}{2}.$$

$$\begin{aligned} H_0 : M_X &\leq M_Y \\ H_1 : M_X &> M_Y \end{aligned}$$

En este caso, la existencia de muchas Y_i que preceden a X_i , como puede verse en la *Figura 8.2*, hará que U tome valores altos, situación que parece confirmar la hipótesis alternativa.

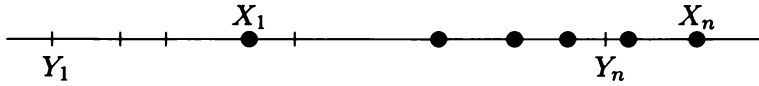


Figura 8.2

Por tanto, fijado un nivel de significación α

- Se acepta H_0 si $U < u_{m,n;\alpha}$
- Se rechaza H_0 si $U \geq u_{m,n;\alpha}$

en donde de nuevo $u_{m,n;\alpha}$ es el menor número entero tal que

$$P\{U \geq u_{m,n;\alpha}\} \leq \alpha.$$

Las hipótesis H_0 y H_1 las hemos expresado en función de las medianas poblacionales, queriendo destacar con ello el hecho de que si se acepta, por ejemplo, la hipótesis alternativa, $H_1 : M_X > M_Y$, se concluye con que la variable en observación tiende a tomar valores significativamente mayores en la población denominada X que en la población denominada Y .

No obstante, el contraste de Wilcoxon, Mann y Whitney está diseñado para hipótesis nulas y alternativas que hacen referencia a las funciones de distribución poblacionales. Así por ejemplo, sería más adecuado hablar en este contraste unilateral de que las hipótesis nula y alternativa son $H_0 : F(x) = G(x) \quad \forall x$ y $H_1 : F(x) > G(x) \quad \forall x$.

$$\begin{aligned} H_0 : M_X &\geq M_Y \\ H_1 : M_X &< M_Y \end{aligned}$$

En este último contraste unilateral rechazaremos H_0 cuando U tome valores pequeños. Es decir, fijado un nivel de significación α

- Se acepta H_0 si $U > m \cdot n - u_{m,n;\alpha}$
- Se rechaza H_0 si $U \leq m \cdot n - u_{m,n;\alpha}$

siendo de nuevo $u_{m,n;\alpha}$ el menor número entero tal que

$$P\{U \geq u_{m,n;\alpha}\} \leq \alpha.$$

Muestras grandes

Como antes dijimos, la distribución de U se aproximará por una normal en cuanto el tamaño de las muestras sea parecido y lo suficientemente grande ($m \approx n$, y además, $m, n > 5$), pudiendo usarse, como de costumbre, una corrección de continuidad (restar 0'5 al valor absoluto del numerador), si los tamaños muestrales son cercanos a 5.

En concreto se tiene que

$$\frac{U - \frac{mn}{2}}{\sqrt{mn(m+n+1)/12}} \approx N(0,1)$$

con lo que el punto crítico $u_{m,n;\alpha}$ será

$$u_{m,n;\alpha} = \frac{mn}{2} + z_{\alpha} \sqrt{\frac{mn(n+m+1)}{12}}.$$

Ejemplo 8.13

Se realizó un estudio con objeto de averiguar si el número de pulsaciones por minuto puede considerarse igual entre los hombres y mujeres de una determinada población.

Para ello se eligieron al azar 12 hombres y 12 mujeres de la mencionada población obteniéndose los siguientes datos

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
Hombres	74	77	71	76	79	74	83	79	83	72	79	77
Mujeres	81	84	80	73	78	80	82	84	80	84	75	82

Si representamos por X la pulsación en la población de hombres y por Y la pulsación en la de mujeres, las hipótesis que se quieren contrastar son $H_0 : M_X = M_Y$, frente a $H_1 : M_X \neq M_Y$.

Ordenando la muestra combinada y subrayando las observaciones Y_i tendremos

71, 72, 73, 74, 74, 75, 76, 77, 77, 78, 79, 79, 79, 80, 80, 80, 81, 82, 82, 83, 83, 84, 84, 84

Contando ahora el número de Y_i que preceden a cada X_j fijo, se obtendrá, al ir variando los X_j , un valor de U

$$U = \sum_{i=1}^{12} \sum_{j=1}^{12} D_{ij} = (0+0) + (1+1) + (2+2+2) + (3+3+3) + (9+9) = 35.$$

Fijado como nivel de significación $\alpha = 0'05$, será $z_{\alpha/2} = 1'96$ y por tanto,

$$u_{m,n;\alpha/2} = \frac{mn}{2} + z_{\alpha/2} \sqrt{\frac{mn(n+m+1)}{12}} = \frac{12 \cdot 12}{2} + 1'96 \sqrt{\frac{12 \cdot 12 \cdot 25}{12}} = 105'95.$$

Como es $U = 35 \notin (38'05, 105'95)$ se rechaza la hipótesis nula de igualdad de ambas poblaciones, a ese nivel de significación.

Para resolver este ejemplo con R, utilizaremos de nuevo la función antes introducida,

```
wilcox.test(x,y,alternative="two.sided",mu=0,exact=T,correct=T)
```

en donde incluiremos en el primer argumento x el vector de observaciones de una de las dos poblaciones a comparar y en el segundo, y , los datos de la otra población. El resto de los argumentos son los anteriormente explicados.

Para este ejemplo, incorporamos los datos en (1) y (2) y ejecutamos la función en (3). No hemos incluido los argumentos `alternative` ni `mu` porque vamos a ejecutar los que toma por defecto, respectivamente, la igualdad de las medianas de ambas poblaciones y que su diferencia es 0. Como no queremos que calcule la distribución exacta del estadístico de Wilcoxon-Mann-Whitney sino la aproximación normal (por obtener un resultado análogo al antes conseguido), le decimos que no ejecute la opción `exact=T` que es la que toma por defecto, sino la opción `exact=F`. Por último, como tampoco queremos que utilice la corrección de continuidad, ejecutamos la función con la opción `correct=F`.

```
> x<-c(74,77,71,76,79,74,83,79,83,72,79,77) (1)
```

```
> y<-c(81,84,80,73,78,80,82,84,80,84,75,82) (2)
```

```
> wilcox.test(x,y,exact=F,correct=F) (3)
```

Wilcoxon rank sum test

data: x and y

$W = 35$, $p\text{-value} = 0.03206$ (4)

alternative hypothesis: true location shift is not equal to 0

Los resultados del estadístico de contraste, 35, y de su p -valor, 0'03206, aparecen en (4). Este p -valor no es concluyente, pero indica rechazar la hipótesis nula de igualdad entre las medianas de ambas poblaciones a un nivel de significación $\alpha = 0'05$ por ser este valor, mayor que el p -valor lo que indica que el estadístico toma un valor perteneciente a la región crítica del test.

8.4.2. El contraste de la Mediana

El *contraste de la Mediana* es un contraste en el que, de nuevo, las hipótesis hacen referencia a las medianas poblacionales, M_X y M_Y .

$$H_0 : M_X = M_Y$$

$$H_1 : M_X \neq M_Y$$

Sean X_1, \dots, X_m e Y_1, \dots, Y_n muestras aleatorias de las dos poblaciones en consideración. Si la hipótesis nula es cierta, entonces ambas muestras procederán de poblaciones con la misma mediana, por lo que, en la muestra combinada, de tamaño $m + n$ y de mediana muestral M_s , cabría esperar que la mitad de las observaciones fueran menores que M_s y la otra mitad mayores.

Por tanto, si consideramos como estadístico del test, $A =$ número de observaciones x_i menores o iguales que M_s , valores muy grandes o muy pequeños suyos desacreditarán la hipótesis nula.

Desgraciadamente la distribución de A resulta complicada y como en cuanto los tamaños muestrales sean moderadamente grandes (digamos $m > 10$ y $n > 10$) se puede utilizar una distribución χ^2 en la determinación de los puntos críticos, omitiremos el caso de muestras pequeñas.

Así pues, supongamos que los tamaños muestrales son suficientemente grandes. En ese caso, si expresamos los datos como en la Tabla 8.10

	Valores menores o iguales que M_s	Valores mayores que M_s	Total muestral
X_1, \dots, X_m	a	$m - a$	m
Y_1, \dots, Y_n	b	$n - b$	n
	$a + b$	$m + n - a - b$	$m + n$

Tabla 8.10

podemos aplicar la técnica de la χ^2 estudiada en la Sección 8.2.3 y utilizar como estadístico del contraste el λ de Pearson, que para el caso particular de la Tabla 8.10 queda, después de simplificar,

$$\lambda = \frac{(m+n)(an - bm)^2}{mn(a+b)(m+n-a-b)}.$$

Contraste de hipótesis

Por tanto, fijado un nivel de significación α ,

- Se acepta H_0 si $\lambda < \chi_{1;\alpha}^2$
- Se rechaza H_0 si $\lambda \geq \chi_{1;\alpha}^2$

en donde por $\chi_{1;\alpha}^2$ representamos, como de costumbre, el valor de una abscisa de una χ_1^2 que deja a la derecha un área de probabilidad α .

Ejemplo 8.13 (continuación)

La muestra combinada con las observaciones y_j subrayadas era

71, 72, 73, 74, 74, 75, 76, 77, 77, 78, 79, 79, 79, 80, 80, 80, 81, 82, 82, 83, 83, 84, 84, 84

de donde se deduce que es $M_s = 79$.

La tabla de contingencia 8.10, aplicada a los datos del ejemplo, queda en la forma

	Valores menores o iguales que 79	Valores mayores que 79	Total muestral
X_1, \dots, X_m	10	2	12
Y_1, \dots, Y_n	3	9	12
	13	11	24

con lo que el estadístico es

$$\lambda = \frac{(m+n)(an-bm)^2}{mn(a+b)(m+n-a-b)} = \frac{24 \cdot (10 \cdot 12 - 3 \cdot 12)^2}{12 \cdot 12 \cdot 13 \cdot 11} = 8'224.$$

Fijado como nivel de significación $\alpha = 0'05$, al ser $\chi^2_{1;0'05} = 3'841 < 8'224 = \lambda$, rechazaremos la hipótesis nula de igualdad de ambas poblaciones, concluyendo que existen diferencias significativas entre ellas.

Como ejecutar este test es en realidad hacer un test de homogeneidad de dos muestras, para resolverlo con R utilizaremos las sentencias utilizadas con este test. En concreto, primero creamos la tabla en (1) (prescindiendo en este ejemplo de la asignación de nombres a las filas y columnas, cuestión que es irrelevante a la hora de resolver el problema). En (2) ejecutamos el test de homogeneidad de la χ^2 mediante la función `chisq.test` que, al serlo sobre una tabla 2×2 , aplicará por defecto la corrección de Yates. Si ejecutamos (3), vemos que las frecuencias esperadas no son menores que 5 por lo que ejecutaremos el test en (4) sin corrección de Yates. Se observa en (5) el mismo valor del estadístico de contraste, $8'2238$, y un p-valor suficientemente pequeño como para rechazar la *homogeneidad* de las dos poblaciones. Se aprecia de nuevo como la corrección de Yates, obtenida en (6) es más conservadora; es decir, tiende a aceptar más la hipótesis nula, es decir, el valor del estadístico de contraste con corrección de Yates suele tomar valores menores que el estadístico sin corrección de Yates.

```
> pulsaciones<-matrix(c(10,3,2,9),ncol=2) (1)
```

```
> chisq.test(pulsaciones) (2)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: pulsaciones
```

```
X-squared = 6.042, df = 1, p-value = 0.01397 (6)
```

```
> chisq.test(pulsaciones)$expected (3)
```

```
 [,1] [,2]
```

```
[1,] 6.5 5.5
```

```
[2,] 6.5 5.5
```

```
> chisq.test(pulsaciones,correct=F) (4)
```

Pearson's Chi-squared test

```
data: pulsaciones
```

```
X-squared = 8.2238, df = 1, p-value = 0.004135 (5)
```

8.5. Ejercicios de Autoevaluación

Ejercicio 8.1

Con objeto de probar la eficacia de tres nuevos tratamientos antivirales en aves infectadas del virus H5N1 se sometió a 150 aves infectadas elegidas al azar al tratamiento T_1 , a otras 200 al tratamiento T_2 y a 180 al tratamiento T_3 , obteniéndose los siguientes resultados de número de aves supervivientes o no al cabo de 20 días de iniciado el tratamiento,

	Supervivientes	No Supervivientes
T_1	122	28
T_2	145	55
T_3	110	70

¿Existen diferencias significativas entre los tres tratamientos?

Ejercicio 8.2

Se llevó a cabo un estudio en unas regiones de la antigua Mesopotamia en donde se realizaron intercambios locales, analizándose un tipo específico de cerámica. Un examen inicial de éstas sugiere que la anchura de las líneas representadas en ellas es distinta según las zonas de producción, clasificando las 90 cerámicas analizadas, según la siguiente tabla de doble entrada,

	Línea Gruesa	Línea Fina
Zona Oriental	42	17
Zona Occidental	10	21

¿Confirman estos datos las diferencias de los grosores de líneas en ambas zonas?

Ejercicio 8.3

La confianza de los consumidores es un índice muy valorado en los mercados financieros. Con objeto de analizar si ha disminuido esa confianza entre dos meses determinados, se pidió a 10 personas elegidas al azar que dieran una valoración de su confianza en la economía del país, de 0 (pésima) a 10 (óptima), en los dos meses (las mismas 10 personas en ambos meses). Los resultados obtenidos fueron los siguientes:

	Índice de confianza									
Mes 1	7	7	6	5	6	8	5	3	5	4
Mes 2	7	5	5	4	6	7	5	4	4	3

Utilizando estos datos y el test de los signos, se pide:

a) Analizar con este test si ha habido una disminución significativa del índice de confianza a nivel $\alpha = 0.05$. Calcular también el p-valor del test.

b) Determinar la potencia del test anterior para una disminución de 2 unidades en el índice de confianza.

Ejercicio 8.4

Los griegos utilizaron en muchas de sus construcciones (por ejemplo en El Partenón) el denominado *rectángulo de oro*, definido como aquel en el que la proporción entre el lado menor y el lado mayor es igual a $1/(0.5(\sqrt{5} + 1)) = 0.618$.

En un reciente yacimiento arqueológico de indios americanos de la tribu *shoshoni* se encontraron prendas de cuero decoradas con rectángulos de proporciones parecidas. En concreto, en diez de tales rectángulos elegidos al azar las proporciones obtenidas fueron las siguientes:

0'662 , 0'690 , 0'628 , 0'606 , 0'570 , 0'654 , 0'672 , 0'576 , 0'611 , 0'670

A la vista de estos datos, ¿puede aceptarse, a nivel $\alpha = 0'1$, la hipótesis de que dicha tribu utilizó también el mencionado *rectángulo de oro* en sus decoraciones, usando el test de los rangos signados de Wilcoxon?

Ejercicio 8.5

El origen de la civilización etrusca sigue siendo todavía un misterio para los antropólogos. En concreto, una cuestión que se plantea es la de si fueron originarios de la península italiana o si inmigraron a ella procedentes de algún otro lugar. Se pensó que una forma de contestar a esta pregunta sería comparar a los actuales italianos con los restos arqueológicos etruscos mediante un estudio antropométrico. Para ello, se midió, en milímetros, la máxima anchura, X , de 8 cráneos de restos de varones etruscos y la máxima anchura, Y , de la cabeza de 10 varones italianos, todos ellos elegidos al azar. Los resultados obtenidos fueron los siguientes:

Etruscos	141	132	154	142	141	150	134	140			
Italianos	133	138	136	125	135	130	127	131	116	128	

En base a los datos obtenidos y utilizando un contraste de Wilcoxon-Mann-Whitney, ¿se puede concluir con la existencia de diferencias significativas entre las dos poblaciones a nivel $\alpha = 0'05$?

8.6. Lecturas Recomendadas

Gibbons, J.D. y Chakraborti S. (2003). *Nonparametric Statistical Inference*. Editorial Marcel Dekker.

Greenwood, P.E. y Nikulin M.S. (1996). *A Guide to Chi-Squared Testing*. Editorial Wiley.

Capítulo 9

Análisis de la Varianza

9.1. Introducción

En los dos capítulos anteriores estudiamos la manera de contrastar la *igualdad* de las medias o medianas de dos poblaciones. En este capítulo expondremos técnicas que las generalizan, denominadas de *Análisis de la Varianza*, las cuales permiten comparar (las medias de) más de dos poblaciones. Las suposiciones que estas técnicas requieren son, básicamente, la normalidad de las poblaciones a comparar, el que tengan la misma varianza (suposición de *homocedasticidad*) y el que sean independientes. La tercera suposición es fácilmente alcanzable al realizar el experimento, pero si fallan las otras dos, deberemos transformar los datos, o utilizar Métodos no paramétricos o, alternativamente, utilizar Métodos Robustos, según se estudia en el Capítulo 5 del texto MR.

La técnica del Análisis de la Varianza se basa en dividir la variabilidad total existente en un conjunto de datos, en diversas *fuentes de variación*, analizando, mediante un contraste de hipótesis, si la aportación relativa de cada una de estas fuentes de variación a la variación total, es significativa o no.

La técnica del Análisis de la Varianza, introducida en los años cincuenta del pasado siglo XX por Ronald Fisher, es utilizada, fundamentalmente, en el análisis de datos procedentes de experimentos, los cuales, por otra parte, son diseñados teniendo en cuenta el futuro Análisis de la Varianza que se hará de sus resultados: el investigador, antes de realizar su experimento, identifica aquellas fuentes de variación que considera importantes y elige un *diseño* que le permita medir la importancia de la contribución de cada una de estas fuentes a la variación total, eliminando de esta variación total aquellas otras fuentes de variación debidas a causas perturbadoras y sin interés en dicha comparación. Por esta razón, la continuación natural de este capítulo es lo que se conoce como *Diseño de Experimentos*.

Un ejemplo que consideraremos más adelante, consiste en averiguar si tres

dietas, A , B y C presentan diferencias significativas en cuanto a sus efectos sobre el aumento de peso en ratones.

Formalmente, el problema se puede plantear diciendo que lo que queremos es comparar tres poblaciones, o, con más precisión, contrastar la hipótesis nula de igualdad de los efectos medios de las tres poblaciones, $H_0 : \mu_A = \mu_B = \mu_C$.

Utilizando la terminología del Diseño de Experimentos hablaremos de un *factor* en estudio, la dieta, el cual se presenta a tres *niveles*, A , B y C , o también, que estamos interesados en estudiar la igualdad de tres *tratamientos*, A , B y C .

En este capítulo estudiaremos solamente el Análisis de la Varianza para un factor en un Diseño Completamente Aleatorizado.

9.2. Análisis de la Varianza para un Factor: Diseño Completamente Aleatorizado

Como hemos dicho más arriba, en este capítulo analizaremos situaciones como la del ejemplo anterior en las que hay *un factor* en estudio, el cual actúa a r *niveles*.

En estos casos en los que sólo se considera un factor, a los niveles se les suele llamar *tratamientos*.

Si denotamos por μ_1, \dots, μ_r los efectos medios de los tratamientos, el interés del investigador se centra en contrastar la hipótesis nula de igualdad de dichos efectos medios, $H_0 : \mu_1 = \mu_2 = \dots = \mu_r$ frente a la alternativa de no ser iguales todos estos efectos medios, $H_1 : \text{no todos son iguales}$, en base a observar los valores de cada uno de los tratamientos en individuos elegidos al azar.

Concretamente, si designamos por X_{ij} el valor o respuesta observado en el individuo j -ésimo, $j = 1, \dots, n_i$ sometido al tratamiento i -ésimo, $i = 1, \dots, r$, los $n = \sum_{i=1}^r n_i$ datos correspondientes a las r muestras aleatorias, de tamaños n_1, \dots, n_r pueden representarse en la forma

Tratamiento	Observaciones				Totales	Medias muestrales
1	x_{11}	x_{12}	\cdots	x_{1n_1}	T_1	\bar{x}_1
2	x_{21}	x_{22}	\cdots	x_{2n_2}	T_2	\bar{x}_2
\vdots	\vdots	\vdots	\dots	\vdots	\vdots	\vdots
r	x_{r1}	x_{r2}	\cdots	x_{rn_r}	T_r	\bar{x}_r
					T	

Tabla 9.1

en donde es $T_i = \sum_{j=1}^{n_i} x_{ij}$, $T = \sum_{i=1}^r T_i$ y $\bar{x}_i = T_i/n_i$

Para contrastar las hipótesis $\begin{cases} H_0 : \mu_1 = \dots = \mu_r \\ H_1 : \text{alguna distinta} \end{cases}$ serán necesarias las siguientes suposiciones:

(a) La i -ésima población o tratamiento X_i se distribuye según una $N(\mu_i, \sigma)$ $i = 1, \dots, r$.

(b) Las r poblaciones o tratamientos son independientes entre sí.

(c) La muestra de tamaño (prefijado) n_i de la población i -ésima es aleatoria simple.

Obsérvese que la suposición (a) lleva implícita no sólo la normalidad sino también la *homocedasticidad*, es decir, el que todos los tratamientos tengan igual varianza.

Insistimos en que el cumplimiento de estas condiciones es un requisito necesario para poder utilizar las técnicas desarrolladas a continuación.

Modelo del Diseño

De forma trivial puede escribirse que

$$x_{ij} = \mu + (\mu_i - \mu) + (x_{ij} - \mu_i)$$

y llamando $\alpha_i = \mu_i - \mu$ y $e_{ij} = x_{ij} - \mu_i$, la expresión anterior puede escribirse en la forma

$$\boxed{x_{ij} = \mu + \alpha_i + e_{ij}} \quad [9.1]$$

Las expresiones anteriores son válidas para cualquier constante μ , pero aquí tomaremos

$$\mu = \frac{\sum_{i=1}^r n_i \mu_i}{n}$$

es decir, una media ponderada de los efectos medios de los tratamientos.

Si es cierta H_0 , el efecto medio común será precisamente μ . Por esta razón, α_i puede interpretarse como el *efecto del tratamiento i -ésimo*. Cuanto más se distancie μ_i del efecto medio común μ , mayor será α_i .

Con esta notación, las hipótesis a contrastar se expresan de la forma $H_0 : \alpha_i = 0 \forall i = 1, \dots, r$ frente a $H_1 : \text{no todas las } \alpha_i = 0$.

Como x_{ij} es un valor muestral obtenido por la variable X_i , la cual tiene media μ_i , la diferencia e_{ij} puede interpretarse como el *error*, debido al azar, que se produce en todo muestreo, el cual hace que no todas las observaciones muestrales sean iguales a su media.

Por tanto, la ecuación [9.1] puede interpretarse diciendo que cada dato observado x_{ij} es el resultado de un efecto común, μ , más el efecto propio del tratamiento i -ésimo de donde procede el dato, α_i , más un término de error, e_{ij} , fruto del muestreo aleatorio efectuado dentro de la población i -ésima.

Fuentes de variación

Si llamamos $\bar{\bar{x}}$ a la media de la muestra global,

$$\bar{\bar{x}} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \bar{x}_i = \frac{T}{n}$$

mediante sencillas operaciones algebraicas puede comprobarse la igualdad

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{x}_i - \bar{\bar{x}})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2. \quad [9.2]$$

El miembro de la izquierda se denomina *suma total de cuadrados*, se representa por SST y se calcula por la expresión

$$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$$

La interpretación de SST se deduce de su denominación y es clara. Si todos los datos x_{ij} fueran iguales, $\bar{\bar{x}}$ también sería igual a este valor común, siendo la variación total existente en los datos igual a cero. En ese caso, SST también sería cero.

En otros casos, SST recoge la dispersión existente en los datos de la Tabla 9.1. Cuanto mayor sea la dispersión en los datos, mayor será SST .

SST no tiene en cuenta de dónde procede la dispersión existente en los datos, si de las filas o de las columnas, y este punto es muy importante, ya que si las filas de la Tabla 9.1 están formadas por números idénticos, Tabla 9.2(a), es decir, es $x_{ij} = c_i$, será $\bar{x}_i = c_i$ no existiendo variación dentro del tratamiento i -ésimo, procediendo toda la dispersión de las diferencias entre filas. Pero si, aunque las filas no sean constantes, sus efectos medios muestrales lo son, Tabla 9.2(b), será $\bar{x}_i = k = \bar{\bar{x}} \forall i$, y el primer miembro de la derecha en la igualdad [9.2], denominado *suma de cuadrados debida a los tratamientos*, SST_i , será cero.

Éste se calcula por la expresión

$$SST_i = \sum_{i=1}^r \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

Trat.	Observ.	Medias
1	3 3 3	3
2	5 5 5	5
3	1 1 1	1

Trat.	Observ.	Medias
1	1 5 3	3
2	5 1 3	3
3	3 1 5	3

Tabla 9.2(a) . $SST = 24$ Tabla 9.2(b) . $SST = 24$

Por último, el tercer miembro de [9.2] se denomina *suma residual de cuadrados*, se representa por SSE y se corresponde con aquella parte de la variación total *no explicada* por los tratamientos.

Se calcula por diferencia de las otras dos sumas de cuadrados,

$$SSE = SST - SST_i.$$

La razón de haber descompuesto la suma total de cuadrados de la manera anterior está motivada porque si H_0 es cierta, las medias muestrales \bar{x}_i tenderán a ser iguales y, por tanto, iguales a $\bar{\bar{x}}$, siendo la suma de cuadrados SST_i cercana a cero o, con más precisión, pequeña en relación a SSE .

Por tanto, si H_0 es cierta, el cociente SST_i/SSE tenderá a ser pequeño, mientras que valores grandes de este cociente tenderán a desacreditar la hipótesis nula. Ése será, salvos constantes, nuestro estadístico de contraste.

Para formalizar el test óptimo y poder calcular los puntos críticos, necesitamos determinar su distribución en el muestreo. Para ello es necesario el siguiente resultado.

Teorema 9.1

(i) $SSE/\sigma^2 \rightsquigarrow \chi_{n-r}^2$.

(ii) Si H_0 es cierta, entonces $SST_i/\sigma^2 \rightsquigarrow \chi_{r-1}^2$.

(iii) SSE y SST_i son independientes.

Como conclusión se tiene que, si H_0 es cierta, el estadístico

$$F = \frac{\frac{SST_i}{\sigma^2} \frac{1}{r-1}}{\frac{SSE}{\sigma^2} \frac{1}{n-r}} = \frac{SST_i/(r-1)}{SSE/(n-r)}$$

seguirá una distribución F de Snedecor con $(r-1, n-r)$ grados de libertad por ser el cociente de dos χ^2 independientes divididas por sus grados de libertad.

Contraste de hipótesis

Como antes dijimos, si H_0 es falsa, el estadístico F tenderá a ser grande por lo que, en ese caso, deberemos rechazar la hipótesis nula.

En concreto, la Estadística Matemática propone como test óptimo de nivel α para contrastar $\begin{cases} H_0: \mu_1 = \dots = \mu_r \\ H_1: \text{alguna distinta} \end{cases}$ cuando se verifican las suposiciones (a), (b) y (c), el siguiente

- Se acepta H_0 si $F < F_{(r-1, n-r); \alpha}$
- Se rechaza H_0 si $F \geq F_{(r-1, n-r); \alpha}$

Teniendo perfecto sentido, al ser éste un contraste de hipótesis, el cálculo e interpretación del p-valor del test.

Tabla de Análisis de la Varianza

Los resultados anteriores se resumen en una tabla la cual suele denominarse ANOVA, apareciendo una reproducción de la misma en ADD.

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = \sum_{i=1}^r \frac{T_i^2}{n_i} - \frac{T^2}{n}$	$r - 1$	$\frac{SST_i}{r - 1}$	$\frac{SST_i / (r - 1)}{SSE / (n - r)}$
Residual	$SSE = SST - SST_i$	$n - r$	$\frac{SSE}{n - r}$	
Total	$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$	$n - 1$		

Estimador de la Varianza

Una consecuencia directa que también se obtiene del Análisis de la Varianza es la estimación de la varianza poblacional común σ^2 . En concreto, por la propiedades de la distribución χ^2 de Pearson, un estimador de σ^2 con buenas propiedades estadísticas es

$$\widehat{\sigma^2} = S^2 = \frac{SSE}{n - r}.$$

9.3. Análisis de la Varianza con R

La función de R que vamos a utilizar para ejecutar el Análisis de la Varianza será

`aov(modelo,datos)`

incluyendo en el argumento `modelo` el “Modelo Lineal” mediante el cual expresamos la variable dependiente cuantitativa observada, en función del factor que define las poblaciones a comparar. En `datos` incluiremos las observaciones que tendrán que venir expresadas en formato *data frame*.

Ejemplo 9.1

Con objeto de analizar si existen diferencias en el aumento de peso entre tres dietas, se decidió someter a 5 ratones a cada una de ellas, obteniéndose los siguientes aumentos de peso

Dieta	Aumento de peso					T_i	\bar{x}_i
A	32	37	34	33	30	166	33'2
B	36	38	37	30	34	175	35
C	35	30	36	29	31	161	32'2
						502	

Para contrastar $\begin{cases} H_0 : \mu_A = \mu_B = \mu_C \\ H_1 : \text{alguna distinta} \end{cases}$ la tabla de Análisis de la Varianza que se obtiene es

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = 16820'4 - 16800'3 = 20'1$	2	10'05	$F = 1'142$
Residual	$SSE = SST - SST_i = 105'63$	12	8'8025	
Total	$SST = 16926 - 16800'27 = 125'73$	14		

Si fijamos un nivel de significación $\alpha = 0'1$, al ser $F = 1'142 < 2'8068 = F_{(2,12);0'1}$, se acepta H_0 . El p-valor es mayor que 0'1.

Para resolver este problema con R, primero creamos los datos, los cuales tendrán que venir en formato *data frame* para que los entienda R, mediante la secuencia (1), (2) y (3),

```
> peso<-c(32,37,34,33,30,36,38,37,30,34,35,30,36,29,31)      (1)
> dieta<-c("A","A","A","A","A","B","B","B","B","B","C","C","C","C","C") (2)
> ejem8_1<-data.frame(dieta,peso)                             (3)
```

Para obtener la tabla de Análisis de la Varianza ejecutamos (4),

```
> summary(aov(peso~dieta,ejem8_1)) (4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dieta	2	20.133	10.067	1.1439	0.351
Residuals	12	105.600	8.800		

(5)

El p-valor del test, que aparece en (5), así como el resto de valores obtenidos, por supuesto coinciden con los obtenidos anteriormente, indicando el p-valor, claramente, la aceptación de la hipótesis nula de igualdad de los efectos medias de las tres dietas.

Alternativamente podíamos haber dado nombre al resultado obtenido con `aov` ejecutando (6) y (7) porque así facilitaremos el análisis que veremos en la siguiente sección.

```
> resul<-aov(peso~dieta,ejem8_1) (6)
```

```
> summary(resul) (7)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dieta	2	20.133	10.067	1.1439	0.351
Residuals	12	105.600	8.800		

9.4. Análisis de las condiciones

Como dijimos más arriba, las poblaciones a comparar deben seguir un modelo normal y además debe verificarse la suposición de homocedasticidad, es decir, que todas ellas deben tener la misma varianza.

El Análisis de la Normalidad de unos datos se puede efectuar gráficamente con ayuda del denominado *Gráfico de normalidad* que consiste en representar en el eje de abscisas los cuantiles de la normal estándar y en el eje de ordenadas los cuantiles de la muestra; si estos pares de puntos están más o menos en la diagonal del gráfico, se tendrá que los cuantiles muestrales serán similares a los de la $N(0,1)$ y podremos concluir con la normalidad de los datos. Este gráfico se puede obtener fácilmente con R gracias a la función `qqnorm`

Obtendremos también el diagrama de hojas y ramas, que vimos en el Capítulo 2 que se podría conseguir con la función `stem` para completar el Análisis de Normalidad

El Análisis de la homocedasticidad se puede hacer gráficamente mediante un *Gráfico de cajas*, obtenido con la función `boxplot`, como vimos en el Capítulo 2, y también con un test que incluimos por completar esta cuestión aunque no analizamos con detalle (puede verlo en García Pérez, 2008b), denominado test de Barlett y que contrasta la hipótesis nula de igualdad de las varianzas; se ejecuta con la función de R, `bartlett.test`.

Ejemplo 9.1 (continuación)

Para analizar la normalidad de los datos del ejemplo anterior, después de *abrir* una ventana

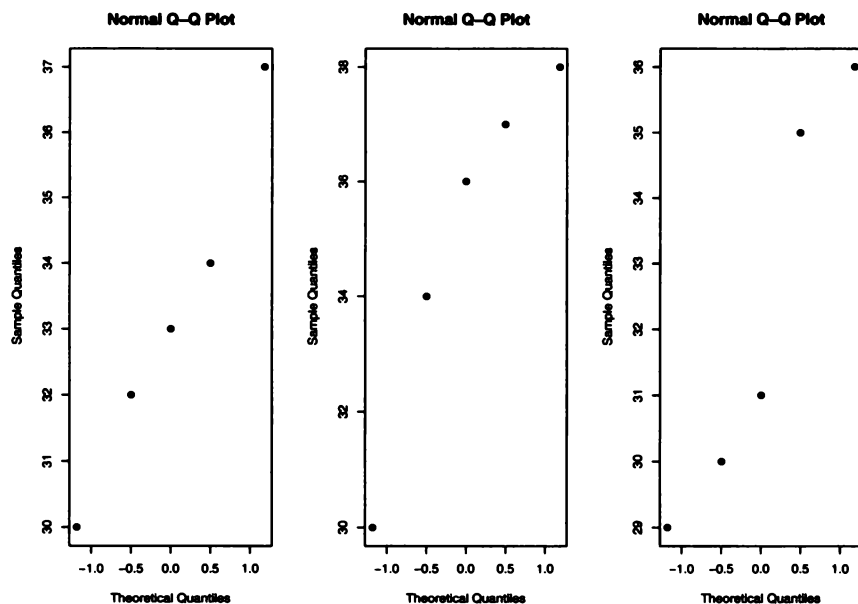


Figura 9.1 : Gráficos de normalidad del Ejemplo 9.1

de tres gráficos en (1), hacemos los gráficos de normalidad para las tres variables, cuya representación obtenemos en la Figura 9.1. Se ve que los datos se sitúan aproximadamente en la diagonal del gráfico, lo que sugiere que se puede admitir para ellos una distribución normal. El gráfico de hojas y ramas para la primera población, obtenido tras ejecutar (2), lo confirma aunque desmerece al considerar sólo cinco datos

```
> par(mfrow=c(1,3))
> qqnorm(peso[1:5],pch=16)
> qqnorm(peso[6:10],pch=16)
> qqnorm(peso[11:15],pch=16)
```

(1)

```
> stem(peso[1:5])
```

(2)

The decimal point is at the |

```
30 | 0
32 | 00
34 | 0
36 | 0
```

El Análisis de la homocedastidad se puede realizar gráficamente con el gráfico de cajas ejecutando (3), gráfico que aparece en la Figura 9.2. En él se ve que las cajas no son muy distintas aunque, como pasa siempre con los análisis gráficos, éstos son opinables. Por esta razón, ejecutamos el test de Barlett en (4), cuyo p-valor, dado en (5), es bastante concluyente en la aceptación de la hipótesis nula de igualdad de las varianzas.

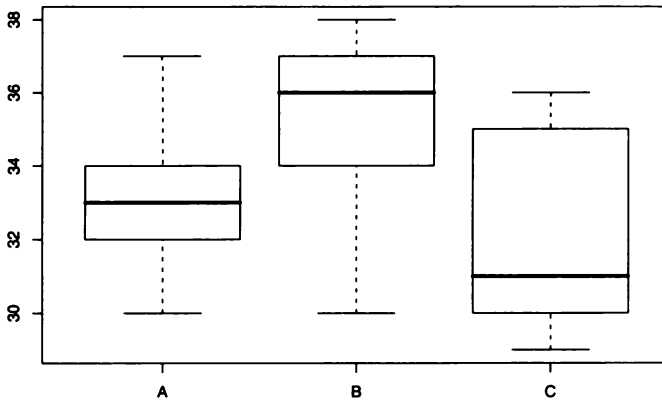


Figura 9.2 : Gráfico de cajas del Ejemplo 9.1

```
> boxplot(peso~dieta) (3)
```

```
> bartlett.test(peso~dieta) (4)
```

Bartlett test of homogeneity of variances

data: peso by dieta

Bartlett's K-squared = 0.1708, df = 2, p-value = 0.9182

(5)

R tiene una función que puede utilizarse para cuando no puede admitirse la igualdad de la varianzas, la cual ejecuta un test similar a la aproximación de Welch en la comparación de dos poblaciones independientes. Se trata de la función `oneway.test`. Así, si para los datos del ejemplo anterior no se hubiera podido aceptar la igualdad de las varianzas o ésta fuera dudosa, ejecutaríamos (6) obteniendo en (7) un p-valor, de nuevo lo suficientemente alto como para aceptar la hipótesis nula de igualdad de los efectos medios de las tres dietas.

```
> oneway.test(peso~dieta) (6)
```

One-way analysis of means (not assuming equal variances)

data: peso and dieta

F = 0.9462, num df = 2.000, denom df = 7.927, p-value = 0.4280

(7)

9.5. Comparaciones Múltiples

En el ejemplo anterior hemos aceptado la hipótesis nula de igualdad de los efectos medios de las poblaciones a comparar pero, en muchas ocasiones, rechazaremos esta hipótesis, pudiendo hacer *Comparaciones Múltiples* entre los diversos tratamientos sobre los que hemos rechazado la igualdad común de todos ellos, con la idea de formar grupos de tratamientos equivalentes.

La primera idea que se le ocurrirá al lector es la de hacer tests de comparación de dos poblaciones, de nivel α , formando grupos de dos tratamientos. Este método es erróneo porque, en ese caso, el nivel de significación global ya no sería α . En este apartado expondremos dos tests que sí tienen en cuenta este problema, tests que se denominan de *comparaciones múltiples*.

Estos tests sólo son válidos para el caso que aquí nos ocupa de un Análisis de la Varianza para un factor y un diseño completamente aleatorizado.

Ejemplo 9.2

En un estudio sobre el efecto de la glucosa en la eliminación de insulina, fueron tratados especímenes de tejidos pancreáticos de animales experimentales con cinco estimulantes diferentes. Más tarde fue determinada la cantidad de insulina eliminada obteniéndose los siguientes resultados:

Estimulante	Observaciones							
1	1'53	1'61	3'75	2'89	3'26	2'83	2'86	2'59
2	3'15	3'96	3'59	1'89	1'45	3'49	1'56	2'44
3	3'89	4'80	3'68	5'70	5'62	5'79	4'75	5'33
4	8'18	5'64	7'36	5'33	8'82	5'26	8'75	7'10
5	5'86	5'46	5'69	6'49	7'81	9'03	7'49	8'98

Se quiere saber si existe diferencia entre los estimulantes en relación con la cantidad de insulina eliminada. Es decir, se trata de contrastar la hipótesis $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ frente a H_1 : alguna distinta, utilizando un diseño completamente aleatorizado.

La tabla de Análisis de la Varianza es

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = 154'9199$	4	38'73	$F = 29'76$
Residual	$SSE = 45'5574$	35	1'3016	
Total	$SST = 200'4773$	39		

Para un nivel de significación de $\alpha = 0'05$ se obtiene un punto crítico $F_{(4,35);0'05} = 2'6478$ que conduce a rechazar la hipótesis nula; de hecho, con gran seguridad ya que p-valor es muy pequeño, $P\{F_{(4,35)} > 29'76\} < 0'005$.

En los dos tests que exponemos a continuación se requiere que el tamaño muestral de cada tratamiento sea el mismo, es decir que sea n_i constante.

Contraste de la mínima diferencia significativa (LSD):

Este contraste propone calcular la mínima diferencia significativa, definida como

$$LSD = t_{n-r;\alpha/2} \sqrt{\frac{2SSE/(n-r)}{n/r}}$$

y concluir diciendo que existe diferencia significativa, a nivel α , entre dos medias poblacionales μ_i y μ_j cuando y sólo cuando sea $|\bar{x}_i - \bar{x}_j| \geq LSD$. Es decir,

- Se acepta $\mu_i = \mu_j$ si $|\bar{x}_i - \bar{x}_j| < LSD$
- Se acepta $\mu_i \neq \mu_j$ si $|\bar{x}_i - \bar{x}_j| \geq LSD$

Ejemplo 9.2 (continuación)

Tanto para este contraste como para el *HSD* que veremos a continuación es útil construir la tabla de la diferencia de medias (en valor absoluto) que, para este problema es igual a

$ \bar{x}_i - \bar{x}_j $	\bar{x}_1	\bar{x}_2	\bar{x}_3	\bar{x}_4	\bar{x}_5
\bar{x}_1	--	0'02625	2'28	4'39	4'43625
\bar{x}_2	--	--	2'25375	4'36375	4'41
\bar{x}_3	--	--	--	2'11	2'15625
\bar{x}_4	--	--	--	--	0'04625
\bar{x}_5	--	--	--	--	--

El valor *LSD* es, en este caso,

$$LSD = t_{n-r;\alpha/2} \sqrt{\frac{2S^2}{n/r}} = t_{40-5;0'05/2} \sqrt{\frac{2 \cdot 1'3016}{8}} = 2'03 \cdot 0'5704384 = 1'158.$$

Observando la tabla de diferencias de medias anterior, se obtienen tres grupos de tratamientos "equivalentes", el $\{I, II\}$, el $\{III\}$ y el $\{IV, V\}$.

Hemos puesto entre comillas lo de equivalentes, porque las clasificaciones proporcionadas por los tests de comparaciones múltiples no tiene porqué ser disjuntas. Es decir, puede darse el caso de no existir diferencias significativas entre, por ejemplo, el primer y segundo tratamiento, no existir diferencias significativas entre el segundo y el tercero, y sí existir diferencias significativas entre el primero y el tercero.

Contraste de Tukey para una diferencia francamente significativa (HSD):

Este contraste se basa en calcular el valor HSD , definido por

$$HSD = q_{r,n-r;\alpha} \sqrt{\frac{SSE/(n-r)}{n/r}}$$

y declarar significativa cualquier diferencia que exceda dicho valor.

El valor del punto crítico $q_{r,n-r;\alpha}$ se obtiene en unas tablas del *Recorrido Studentizado*, como la *Tabla 7* de ADD.

Ejemplo 9.2 (continuación)

En nuestro ejemplo, el valor del HSD es

$$HSD = q_{r,n-r;\alpha} \sqrt{\frac{S^2}{n/r}} = q_{5,40-5;0'05} \sqrt{\frac{1'3016}{8}} = 4'07 \cdot 0'4034 = 1'642$$

con lo que, observando de nuevo la tabla de diferencias de medias, se obtienen los mismos grupos que antes $\{I, II\}$, $\{III\}$ y $\{IV, V\}$.

9.6. Comparaciones Múltiples con R

Con R sólo haremos comparaciones múltiples utilizando el *Contraste de Tukey HSD* mediante la función

`TukeyHSD(x, conf.level=0.95)`

cuyo primer argumento x debe ser un objeto creado con la función `aov`. El segundo es el 1 – el nivel de significación (coeficiente de confianza del intervalo de confianza/región de aceptación) de los tests donde la hipótesis nula es la igualdad de las medias de las poblaciones comparadas.

Ejemplo 9.2 (continuación)

Para resolver esta problema con R, primero incorporamos los datos a partir de (1), ejecutamos el Análisis de la Varianza en (2) obteniendo la tabla ANOVA con (3). En (4) se observa un p-valor casi cero lo que lleva a rechazar la igualdad de los efectos medios de los cinco estimulantes. El contraste HSD de Tukey, a nivel 0'05, se obtiene ahora ejecutando (5)

```
> insulina<-c(1.53,1.61,3.75,2.89,3.26,2.83,2.86,2.59,                (1)
+ 3.15,3.96,3.59,1.89,1.45,3.49,1.56,2.44,3.89,4.8,3.68,5.7,5.62,5.79,4.75,5.33,
+ 8.18,5.64,7.36,5.33,8.82,5.26,8.75,7.1,5.86,5.46,5.69,6.49,7.81,9.03,7.49,8.98)
> estimula<-factor(rep(LETTERS[1:5],c(8,8,8,8,8)))
> ejem8_2<-data.frame(estimula,insulina)

> resul<-aov(insulina~estimula)                                (2)
```

```

> summary(resul)
      Df Sum Sq Mean Sq F value    Pr(>F)
estimula  4 154.920   38.730   29.755 7.956e-11 ***
      (4)
Residuals 35  45.557    1.302

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> TukeyHSD(resul)
      Tukey multiple comparisons of means
      95% family-wise confidence level

Fit: aov(formula = insulina ~ estimula)

$estimula
      diff      lwr      upr      p adj
B-A 0.02625 -1.6138197 1.666320 0.9999989
C-A 2.28000  0.6399303 3.920070 0.0027393
D-A 4.39000  2.7499303 6.030070 0.0000000
E-A 4.43625  2.7961803 6.076320 0.0000000
C-B 2.25375  0.6136803 3.893820 0.0031151
D-B 4.36375  2.7236803 6.003820 0.0000001
E-B 4.41000  2.7699303 6.050070 0.0000000
D-C 2.11000  0.4699303 3.750070 0.0062262
E-C 2.15625  0.5161803 3.796320 0.0049938
E-D 0.04625 -1.5938197 1.686320 0.9999897

```

Los intervalos (regiones de aceptación) obtenidos a partir de (6), cuyo extremo inferior está encabezado con *lwr* y el superior con *upr*, que contengan al cero implicarán la igualdad de los efectos medios cuyas letras aparecen al comienzo de la línea. Así, por ejemplo, el primer intervalo de aceptación es $[-1'61, 1'66]$ el cual, al contener al cero, implica la igualdad de los efectos medios de los tratamiento B-A. De esta manera vemos que podemos considerar tres clases de tratamientos equivalentes: el $\{A, B\}$, $\{C\}$, $\{D, E\}$. La última columna nos da los *p*-valores de los tests, los cuales confirman la clasificación anterior: sólo pueden considerarse iguales los de la primera línea y los de la última,

9.7. Ejercicios de Autoevaluación

Ejercicio 9.1

La tabla del final del enunciado da el contenido de nitrógeno, en miligramos, de plantas de trébol rojo inoculadas con cultivos de *Rhizobium trifolii* más cinco cepas diferentes de *Rhizobium meliloti*, además de un compuesto de trébol rojo que se utilizó como control. Se pide:

a) Contrastar si todas las plantas de trébol rojo pueden considerarse homogéneas desde el punto de vista de la cantidad de nitrógeno que contienen.

b) Caso de que no se pueda aceptar esa hipótesis, contrastar, a nivel $\alpha = 0'05$, qué tratamientos pueden considerarse homogéneos, utilizando un contraste HSD.

						T_i
<i>3Dok1</i>	19'4	32'6	27	32'1	33	144'1
<i>3Dok5</i>	17'7	24'8	27'9	25'2	24'3	119'9
<i>3Dok4</i>	17	19'4	9'1	11'9	15'8	73'2
<i>3Dok7</i>	20'7	21	20'5	18'8	18'6	99'6
<i>3Dok13</i>	14'3	14'4	11'8	11'6	14'2	66'3
<i>Compuesto</i>	17'3	19'4	19'1	16'9	20'8	93'5

Ejercicio 9.2

Se quiere averiguar si tres tipos de gasolina presentan diferencias significativas en cuanto a sus efectos contaminantes. Para ello se seleccionaron al azar doce vehículos en los que se aplicaron aleatoriamente los tres tipos de gasolina, obteniéndose los siguientes datos respecto a reducción de óxido de nitrógeno:

<i>Gasolina I</i>	23	26	25	25
<i>Gasolina II</i>	28	29	27	25
<i>Gasolina III</i>	22	25	26	27

Con estos datos, ¿pueden inferirse diferencias significativas entre los tres tipos de gasolina, con un nivel de significación $\alpha = 0'05$?

Ejercicio 9.3

Se quiere averiguar si el porcentaje medio de proteínas en la leche, de vacas alimentadas con piensos elaborados con productos transgénicos, es diferente de las alimentadas con piensos tradicionales. Para ello se anotó la producción de 5 vacas elegidas al azar alimentadas con piensos elaborados con soja A5403 (*Población I*), 5 vacas elegidas al azar alimentadas con piensos elaborados con maíz T25 (*Población II*) y 5 vacas elegidas al azar con alimentación tradicional (*Población III*).

Los resultados obtenidos fueron los siguientes:

<i>Población I</i>	82'1	84'6	83'1	82'3	84'1
<i>Población II</i>	86'5	83'2	84'2	83'2	83'3
<i>Población III</i>	90'3	91'5	89'3	86'5	90'2

a) Contrastar si puede aceptarse la hipótesis nula de igualdad en el porcentaje de proteínas de la leche de las tres poblaciones.

b) Caso de que no se pueda aceptar esa hipótesis, contrastar, a nivel $\alpha = 0'05$, qué poblaciones pueden considerarse homogéneas, utilizando un contraste HSD.

Ejercicio 9.4

Con objeto de comparar la duración de tres procesos industriales *Proceso1*, *Proceso2* y *Proceso3*, se eligieron al azar 3 elementos elaborados con el *Proceso1*, 4 elaborados con el *Proceso2* y 3 elaborados con el *Proceso3*, anotándose el tiempo (en minutos) empleado en la fabricación de cada elemento. Los resultados obtenidos fueron los siguientes:

Proceso1	24	26	31	
Proceso2	26	29	23	27
Proceso3	24	23	28	

Respecto a la duración del proceso de producción, ¿pueden considerarse homogéneos los tres procesos industriales a nivel $\alpha = 0'05$?

Ejercicio 9.5

Se están investigando cuatro métodos diferentes de preparación del compuesto superconductor $PbMo_6S_8$. Los investigadores afirman que la presencia de oxígeno durante el proceso de preparación, afecta a la temperatura de transición T_c de la superconductividad. Los métodos de preparación $M1$ y $M2$ utilizan técnicas que están diseñadas para eliminar la presencia de oxígeno, mientras que los métodos $M3$ y $M4$ permiten la presencia del oxígeno. Se midió la temperatura de transición (en grados Kelvin) de cinco unidades experimentales producidas con cada uno de los cuatro métodos en comparación, obteniéndose los siguientes datos:

Método	Temperaturas de Transición				
$M1$	14'8	14'8	14'7	14'8	14'9
$M2$	14'6	15'0	14'9	14'8	14'7
$M3$	12'7	11'6	12'4	12'7	12'1
$M4$	14'2	14'4	14'4	12'2	11'7

a) ¿Puede inferirse que existen diferencias significativas entre los cuatro métodos de preparación?

b) Si es así, a nivel $\alpha = 0'05$, ¿pueden considerarse algunos métodos equivalentes utilizando un contraste HSD?

9.8. Lecturas Recomendadas

Box, G.E.P., Hunter, W.G. y Hunter J.S. (1978). *Statistics for Experimenters*. Editorial Wiley.

García Pérez, A. (2008b). *Estadística Aplicada: Conceptos Básicos*. Segunda Edición. Editorial UNED. Colección Educación Permanente.

Capítulo 10

Regresión Lineal y Correlación

10.1. Introducción

A principios del siglo XX, los astrónomos E. Hertzsprung y H.N. Russell, observaron una cierta relación entre el tipo espectral de las estrellas y su luminosidad, de tal forma que las estrellas azules parecían ser las más brillantes.

Como el color de las estrellas es una expresión de su temperatura superficial —las estrellas azules son las más calientes y las rojas las más frías— pensaron en la existencia de una relación entre la temperatura de una estrella y su luminosidad.

Con objeto de estudiar esta posible relación, representaron en un diagrama de dispersión la temperatura superficial en el eje de abscisas y la luminosidad en el de ordenadas, obteniendo un gráfico como el de la Figura 10.1.

En él se observa que, en efecto, parece existir una relación entre luminosidad y temperatura. Además se puede deducir la forma de esta relación: lineal creciente para un gran número de estrellas, las incluidas en lo que se denominó “secuencia principal”, y lineal, casi constante, para el grupo superior de estrellas rojas.

El propósito del análisis de regresión y correlación es el estudio de la relación existente entre dos variables aleatorias, una denominada *independiente* o *covariable*, bajo el control del experimentador, habitualmente representada por X y con valores en el eje de abscisas, y otra denominada *dependiente*, habitualmente representada por Y y con valores en el eje de ordenadas. En el ejemplo anterior, la Luminosidad sería la variable dependiente y la Temperatura Superficial la covariable independiente.

El *Análisis de la Regresión* se ocupa de estudiar la *forma* de la relación existente entre dos o más variables aleatorias, mientras que el *Análisis de la*

Correlación investiga el grado o fuerza de dicha relación.

Esta relación lineal existente entre dos variables aleatorias se denomina *Regresión Lineal Simple*, mientras que cuando se consideran más de dos covariables se hablará de *Regresión Lineal Múltiple*.

El poder inferir la existencia de una fuerte relación entre dos variables aleatorias permite obtener interesantes aplicaciones. Así, en el ejemplo anterior, al conocerse que la luminosidad de una estrella es directamente proporcional a su superficie y a la cuarta potencia de su temperatura superficial, a igualdad de temperaturas superficiales, las estrellas situadas en la parte superior del diagrama de Hertzsprung-Russell debían tener un radio mayor por lo que se las denominó estrellas *gigantes*, mientras que las situadas en la parte inferior recibieron el calificativo de *enanas*.

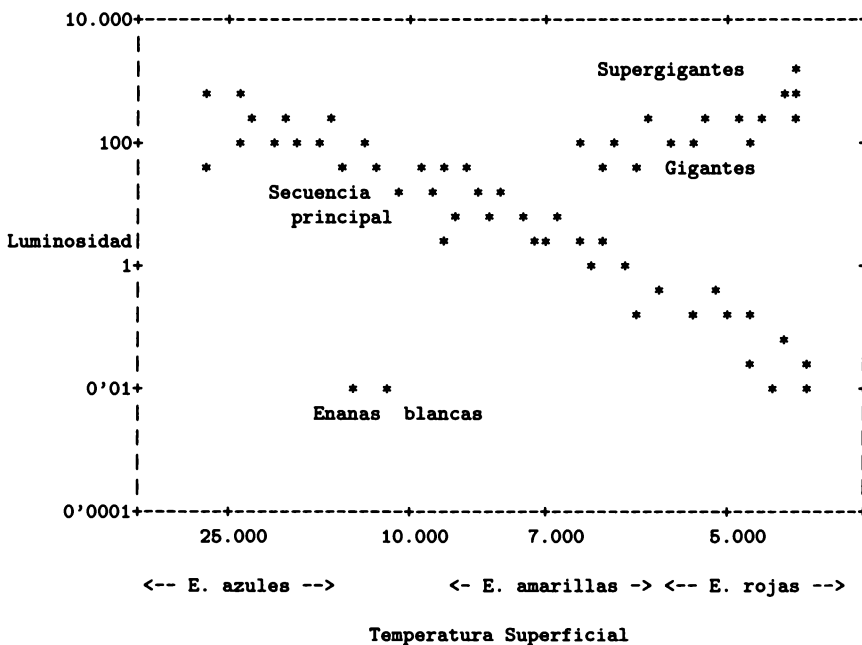


Figura 10.1

Las estrellas de la franja casi horizontal, de gran radio, por tener gran luminosidad, y de baja temperatura, se denominaron *gigantes rojas*.

Por el contrario, las que aparecían abajo y a la izquierda, muy calientes pero de escasa luminosidad —y por tanto de radio pequeño—, se denominaron *enanas blancas*.

El admitir la relación lineal propuesta por el diagrama de Hertzsprung-Russell implica también el que —para una misma temperatura— las estrellas

gigantes tengan una densidad muy baja en comparación con las enanas cuya densidad media puede medirse en toneladas por centímetro cúbico.

La confirmación estadística de una relación de tipo lineal entre dos variables aleatorias, permite obtener, como hemos visto, conclusiones que no habrían podido ser obtenidas de no existir dicha relación.

10.2. Modelo de la Regresión Lineal Simple

La situación general que se plantea para la regresión lineal simple es la de dos variables aleatorias, X e Y , estando interesados en inferir la existencia o no de una relación lineal entre ambas, de la forma

$$Y = \beta_0 + \beta_1 X + e$$

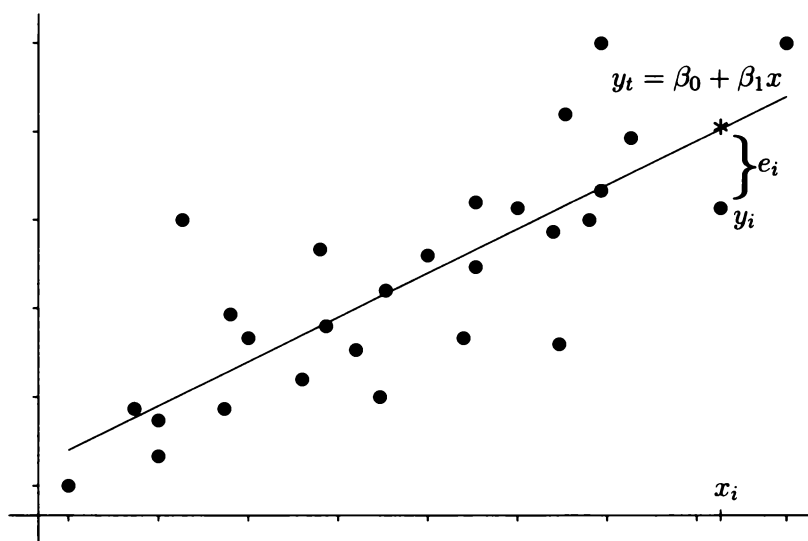


Figura 10.2

interpretada ésta en el sentido de que, fijados unos valores x_i de la variable X , obtendremos valores

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

de la variable Y , los cuales no llegan a estar sobre la recta $y_t = \beta_0 + \beta_1 x$ debido al error de muestreo e_i , (véase la Figura 10.2). Los parámetros β_0 y β_1 se denominan *coeficientes de regresión*.

El modelo de regresión lineal supone que los errores e_i son independientes y con distribución $N(0, \sigma)$; es decir, que dado un valor x de la variable

aleatoria X , la distribución condicionada Y/x es normal $N(\mu_{y/x}, \sigma)$, con $\mu_{y/x} = E[Y/x] = \beta_0 + \beta_1 x$, siendo σ^2 la varianza común a todas las distribuciones condicionadas (hipótesis de homocedasticidad), y que las distribuciones condicionadas por distintos x son independientes entre sí.

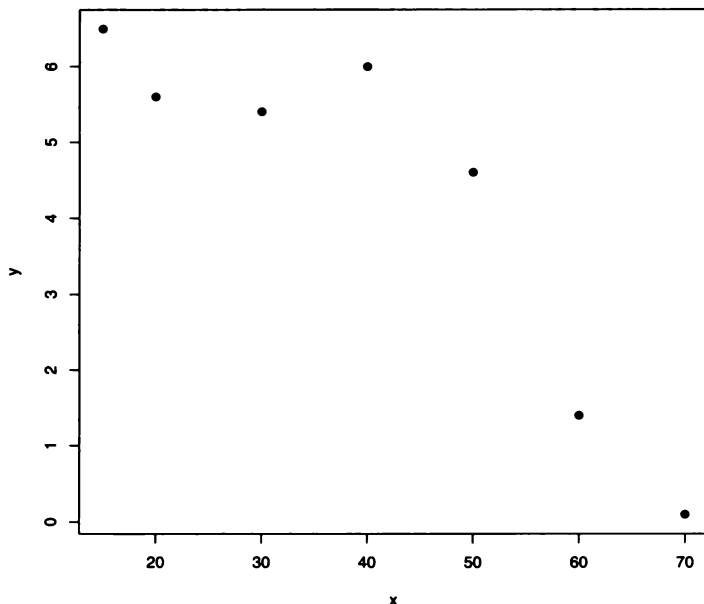


Figura 10.3

Ejemplo 10.1

Se midió el contenido de oxígeno, variable Y , a diversas profundidades, variable X , en el lago Worther de Australia, obteniéndose los siguientes datos, en miligramos por litro

X	15	20	30	40	50	60	70
Y	6'5	5'6	5'4	6	4'6	1'4	0'1

La representación de estos datos en un diagrama de dispersión es la Figura 10.3. De ella parece desprenderse que existe una relación lineal inversa entre Profundidad y Cantidad de Oxígeno: a mayor profundidad, menor cantidad de oxígeno.

En situaciones como la del ejemplo anterior, en las que parece existir una relación lineal entre X e Y , la recta que las relaciona debería ser la más próxima posible a la nube de puntos en el sentido de mínimos cuadrados (veremos en el texto MR lo inadecuado de este planteamiento), es decir, la recta que minimiza

la suma de los residuos r_i al cuadrado, entendidos éstos como las diferencias entre los valores y_i observados y los y_{ti} proporcionados por la recta ajustada,

$$\sum_{i=1}^n r_i^2 = \sum_{i=1}^n (y_i - y_{ti})^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

La recta así caracterizada, vimos en la Sección 2.4.2 que era la Recta de Mínimos Cuadrados, también denominada Recta de Regresión (con más precisión, de Y sobre X)

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x$$

en donde $\hat{\beta}_0$ y $\hat{\beta}_1$ se determinan por las ecuaciones

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

y

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n}$$

En el Ejemplo 10.1,

$$y_t = 8'631 - 0'108 x.$$

10.2.1. Interpretación de los coeficientes de regresión

Como vimos más arriba, en un modelo de regresión lineal se supone que la media de Y/x es de la forma

$$\mu_{y/x} = E[Y/x] = \beta_0 + \beta_1 x$$

por lo que el estimador $\hat{\beta}_1$ de la pendiente de la recta de regresión ajustada

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x$$

se interpreta como el cambio en promedio de la variable Y por el incremento en una unidad de la variable X . La estimación $\hat{\beta}_0$, es decir, la estimación de la ordenada en el origen, se interpreta como el valor promedio cuando la covariable es igual a cero. De hecho, esta interpretación es la evidente ya que

la función anterior y_t es una función (una recta) de la variable x , digamos $h(x)$ que, por las observaciones del comienzo de este apartado, podemos denominar función valor promedio de Y . Si $x = 0$, entonces es $h(0) = \hat{\beta}_0$ por lo que se interpreta $\hat{\beta}_0$ como el valor promedio cuando la covariable X es igual a cero. De la misma manera, la interpretación de $\hat{\beta}_1$ es la habitual para la derivada de una función ya que, en este caso, la derivada de la función valor promedio de Y , es $h'(x) = \hat{\beta}_1$.

En el Ejemplo 10.1 anterior, en donde la variable Y era el Contenido de Oxígeno y la variable X la Profundidad, obtuvimos una recta de regresión de la forma

$$y_t = 8'631 - 0'108 x$$

interpretándose como que el promedio en el contenido de oxígeno disminuye en 0'108 por unidad de medida en la variable X ; como en este ejemplo la variable X se medía en metros, se diría que por metro que bajamos en profundidad.

Respecto al otro coeficiente de regresión, se interpreta diciendo que 8'631 es el valor promedio de contenido de oxígeno para una profundidad igual a cero.

Si en otro ejemplo hubiéramos ajustado una recta de regresión de la forma

$$y_t = 98 + 0'44 x$$

siendo Y la covariable Presión Sanguínea Sistólica medida en mmHg. y X la Edad en años de los pacientes, la interpretación de los coeficientes ajustados sería la de que, el promedio de presión sanguínea se incrementa 0'44 mmHg. por año que se incrementa la Edad y, respecto a $\hat{\beta}_0$, que 98 es el valor promedio de presión sanguínea cuando la Edad es cero.

10.3. Contraste de la Regresión Lineal Simple

Más arriba hemos explicado cómo obtener la recta de regresión lineal, pero esta recta siempre se puede determinar y en unos casos explicará bien a la variable dependiente en función de la independiente y en otros casos, no lo hará. Es decir, en unos casos la recta de regresión podrá ser utilizada para, por ejemplo, hacer predicciones de Y dados unos x concretos y en otros casos no podrá ser utilizada para este propósito porque las predicciones serían desastrosas.

Será la Inferencia Estadística la que deberá ahora validar o no la recta de regresión obtenida, mediante un test de hipótesis basado en un Análisis de la Varianza semejante a los efectuados en el capítulo anterior o, equivalentemente, mediante un test de la hipótesis nula de igualdad a cero del coeficiente β_1 .

10.3.1. Análisis de la variación explicada frente a la no explicada por la recta de regresión

En la Sección 2.4.3, dijimos que entre dos funciones que ajustáramos por mínimos cuadrados a una nube de puntos, deberíamos elegir aquella para la cual se obtuviera una menor varianza residual. Si sólo consideramos una función, dijimos que el ajuste por ésta se podía considerar bueno, si el coeficiente de determinación (relacionado con la varianza residual) era cercano a 1.

En esta sección vamos a precisar estas ideas mediante tests de hipótesis para contrastar la regresión lineal simple.

Si llamamos *suma total de cuadrados SST* a

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}$$

la cual representa la dispersión de los datos y_i entorno a su media muestral \bar{y} , se puede demostrar fácilmente que es

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_{t_i} - \bar{y})^2 + \sum_{i=1}^n (y_i - y_{t_i})^2.$$

El primer miembro de la derecha, denominado *suma de cuadrados debida a la regresión lineal*, $SSEX$ representa la parte de la suma total de cuadrados explicada por la recta de mínimos cuadrados $y_t = \beta_0 + \beta_1 x$, suma de cuadrados que se calcula por la expresión

$$SSEX = \hat{\beta}_1^2 \left(\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n} \right).$$

La variación restante, es decir, la *suma de cuadrados no explicada* por la recta de mínimos cuadrados, será

$$SSNEX = \sum_{i=1}^n (y_i - y_{t_i})^2 = \sum_{i=1}^n r_i^2 = SST - SSEX$$

igual, por otra parte, a n veces la varianza residual.

Si la hipótesis nula es $H_0 : X$ e Y no están relacionadas linealmente, (es decir, la recta de regresión no sirve para explicar a la variable dependiente en función de la independiente), y la alternativa $H_1 : X$ e Y están relacionadas

linealmente, (es decir, la recta de regresión es útil), el contraste a construir parece claro:

Si la variación explicada por la recta de mínimos cuadrados $SSEX$ es grande con respecto a la variación residual $SSNEX$, deberemos rechazar H_0 ; en otro caso aceptarla. Salvo constantes necesarias para obtener una distribución en el muestreo conocida, el estadístico del test a considerar será por tanto, $SSEX/SSNEX$.

Teorema 10.1

En las condiciones de normalidad antes especificadas, se tiene que

$$(i) \ SSNEX/\sigma^2 \rightsquigarrow \chi_{n-2}^2.$$

$$(ii) \text{ Si } H_0 \text{ es cierta, entonces } SSEX/\sigma^2 \rightsquigarrow \chi_1^2.$$

$$(iii) \ SSEX \text{ y } SSNEX \text{ son independientes.}$$

Como conclusión se tiene que, si H_0 es cierta, el estadístico

$$F = \frac{\frac{SSEX}{\sigma^2} \frac{1}{1}}{\frac{SSNEX}{\sigma^2} \frac{1}{n-2}} = \frac{SSEX}{SSNEX/(n-2)}$$

seguirá una distribución F de Snedecor con $(1, n-2)$ grados de libertad por ser el cociente de dos χ^2 independientes divididas por sus grados de libertad.

Contraste de hipótesis

Por lo que antes dijimos, si H_0 es falsa, el estadístico F tenderá a tomar valores grandes, rechazando en ese caso H_0 . Por tanto, el test óptimo de nivel α para contrastar $\begin{cases} H_0 : X \text{ e } Y \text{ no están relacionadas linealmente} \\ H_1 : X \text{ e } Y \text{ están relacionadas linealmente} \end{cases}$ es el siguiente

- Se acepta H_0 si $F < F_{1,n-2;\alpha}$
- Se rechaza H_0 si $F \geq F_{1,n-2;\alpha}$

teniendo perfecto sentido el cálculo e interpretación del p-valor del test.

Tabla de Análisis de la Varianza

Los resultados anteriores se resumen en una tabla denominada de Análisis de la Varianza (ANOVA) para la regresión lineal simple, una reproducción de la cual aparece en ADD.

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal simple	$SSEX = \widehat{\beta}_1^2 \left(\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i \right)^2}{n} \right)$	1	$SSEX$	$\frac{\frac{SSEX}{SSNEX}}{n-2}$
Residual	$SSNEX = SST - SSEX$	$n-2$	$\frac{SSNEX}{n-2}$	
Total	$SST = \sum_{i=1}^n y_i^2 - \frac{\left(\sum_{i=1}^n y_i \right)^2}{n}$	$n-1$		

Estimación de la varianza común σ^2

Como vimos en la Sección 5.3.1, la media de una distribución χ^2 es igual a sus grados de libertad, por lo que del apartado (i) del Teorema 10.1 anterior, será $E[SSNEX/\sigma^2] = n-2$, o bien, $E[SSNEX/(n-2)] = \sigma^2$, con lo que

$$\widehat{\sigma^2} = \frac{SSNEX}{n-2}$$

será un estimador insesgado, es decir, un buen estimador, de la varianza común σ^2 . Además, observemos que este valor lo obtenemos como cuadrado medio de la suma de cuadrados residual en la tabla ANOVA anterior.

Ejemplo 10.1 (continuación)

Para contrastar si existe regresión lineal significativa entre el Contenido de oxígeno Y y la Profundidad X, la tabla de análisis de la varianza para la regresión lineal simple sería,

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Regresión lineal	$SSEX = 29'481$	1	29'481	$F = 20'32$
Residual	$SSNEX = 7'2533$	5	1'4507	
Total	$SST = 36'7343$	6		

Como para un nivel de significación $\alpha = 0'05$ es $F_{1,5;0'05} = 6'6079 < 20'32 = F$, rechazamos H_0 , con un p-valor entre $0'005$ y $0'01$, concluyendo que sí existe relación lineal entre ambas variables, es decir, que la recta de regresión determinada, es útil para explicar a la variable dependiente en función de la independiente.

Además, una buena estimación de la varianza común es $\widehat{\sigma}^2 = 1'4507$.

10.3.2. Contraste de hipótesis para β_1

Una forma alternativa al Análisis de la Varianza anterior, para analizar si puede considerarse válida la recta de regresión determinada, es contrastar si se puede aceptar que es cero o no el parámetro β_1 de la ecuación de regresión lineal entre ambas variables.

Si se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ y se acepta la alternativa $H_1 : \beta_1 \neq 0$ la regresión lineal dada por la recta de regresión será aceptable, o en terminología de tests de hipótesis, existe una relación lineal *significativa*, ya que de hecho, el test ha resultado significativo.

El contraste que veremos, se basa en que la distribución en el muestreo de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ es normal de parámetros

$$\hat{\beta}_0 \rightsquigarrow N\left(\beta_0, \sqrt{\frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}}\right) \quad \hat{\beta}_1 \rightsquigarrow N\left(\beta_1, \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}\right)$$

siendo σ^2 la varianza de la distribución condicionada Y/x .

Al conocer las distribuciones en el muestreo de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, podremos determinar, de forma similar a como lo hicimos en los Capítulos 6 y 7, intervalos de confianza y tests de hipótesis para β_0 y, especialmente, para β_1 .

Si denominamos

$$S_b^2 = \frac{\widehat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{SSNEX/(n-2)}{\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2 / n} = \frac{SSNEX/(n-2)}{SSEX/\hat{\beta}_1^2}$$

al ser independientes

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \rightsquigarrow N(0, 1) \quad \text{y} \quad \frac{SSNEX}{\sigma^2} \rightsquigarrow \chi_{n-2}^2$$

el estadístico

$$\frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}{\sqrt{\frac{SSNEX}{(n-2)\sigma^2}}} = \frac{\hat{\beta}_1 - \beta_1}{S_b}$$

seguirá una distribución t de Student con $n - 2$ grados de libertad, por lo que si queremos contrastar $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$,

- Se acepta H_0 si $|t| < t_{n-2;\alpha/2}$
- Se rechaza H_0 si $|t| \geq t_{n-2;\alpha/2}$

siendo el estadístico del contraste

$$t = \frac{\hat{\beta}_1}{S_b} = \sqrt{\frac{SSEX(n-2)}{SSNEX}}.$$

En el Ejemplo 10.1, al ser $\hat{\sigma}^2 = 1'4507$, será

$$S_b^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\hat{\sigma}^2}{SSEX/\hat{\beta}_1^2} = \frac{1'4507}{29'481} 0'108^2 = 0'000574$$

y, por tanto,

$$t = \frac{-0'108}{\sqrt{0'000574}} = -4'508$$

con lo que se rechaza claramente H_0 , siendo el p -valor menor que 0'01. Es decir, la recta de regresión es válida para explicar la variable dependiente Y en función de la independiente X .

Obsérvese que el estadístico de contraste ha tomado aquí el valor $t = -4'508$. Por el otro método alternativo de contrastar la validez de la recta, obtuvimos el valor $F = 20'32$, que es igual a $(-4'508)^2$. Este hecho no debe sorprendernos ya que el cuadrado de una t de Student con m grados de libertad es una F de Snedecor con $F_{(1,m)}$ grados de libertad. Por tanto, esta circunstancia se deberá producir siempre.

10.4. Regresión Lineal con R

Básicamente, los objetivos que perseguimos con R van a ser, primero determinar la recta en el caso de regresión lineal simple (o, como veremos más adelante, el hiperplano en el caso de regresión lineal múltiple), mediante la función `lm`. Después, analizar cuáles de las covariables X_i son significativas a la hora de predecir la variable dependiente Y , mediante la función `summary`.

Ejemplo 10.1 (continuación)

Para resolver este ejemplo con R, primero incorporaremos los datos en (1) y (2), obteniendo la recta, que aquí denominamos `ajus`, al ejecutar (3).

Podemos obtener los estimadores de los coeficientes de regresión ejecutando el objeto creado mediante (4). La recta de regresión ajustada ha sido la que tiene por coeficientes los dados en (5) y que es

$$y = 8'6310 - 0'1081 x$$

Ahora tenemos que analizar si la covariable X explica suficientemente bien a la variable dependiente Y ; es decir, si puede aceptarse o no la hipótesis nula de ser cero el coeficiente de regresión de X , es decir, $H_0 : \beta_1 = 0$. Para ello ejecutamos (6) obteniendo en (7) el p-valor de dicho test, 0'00635, suficientemente pequeño como para rechazar esta hipótesis nula y concluir con que β_1 es significativamente distinto de cero, es decir, que la covariable independiente X es significativa para explicar a la variable dependiente Y mediante la ecuación de la recta de regresión determinada.

```
> x<-c(15,20,30,40,50,60,70) (1)
```

```
> y<-c(6.5,5.6,5.4,6,4.6,1.4,0.1) (2)
```

```
> ajus<-lm(y~x) (3)
```

```
> ajus (4)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

```
(Intercept)          x
      8.6310       -0.1081 (5)
```

```
> summary(ajus) (6)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

```
      1      2      3      4      5      6      7
-0.50907 -0.86841  0.01289  1.69419  1.37550 -0.74320 -0.96190
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.63102    1.07747   8.010  0.00049 ***
```



```

x          -0.10813    0.02399  -4.508  0.00635 **
                                (7)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.204 on 5 degrees of freedom
Multiple R-Squared:  0.8025,    Adjusted R-squared:  0.7631
F-statistic: 20.32 on 1 and 5 degrees of freedom,    p-value:    0.006352

```

Hemos obtenido más arriba una tabla ANOVA para analizar la regresión lineal. Esta tabla, no obstante, sólo nos permitirá contrastar la hipótesis nula de que todo el modelo lineal es adecuado frente a la hipótesis alternativa de no ser todo el modelo lineal ajustado adecuado para explicar los datos que, en el caso de una regresión lineal simple, coincidirá con el test sobre el coeficiente de regresión. Es posible obtener esta tabla ANOVA aplicando la función `anova` al objeto creado con la función `lm`.

No cabe duda de que es más interesante la vía recién estudiada mediante la cual contrastamos la significación de cada covariable que el análisis de todas a la vez.

Ejemplo 10.1 (continuación)

Si queremos obtener la tabla de Análisis de la Varianza para la regresión lineal, ejecutamos (1) obteniendo dicha tabla a continuación que coincide, lógicamente, con la calculada más arriba. Se observa también que coinciden los resultados con el resumen acabado de obtener con R y con el p-valor así obtenido, 0'006352; es decir, con un estadístico de distribución F de Snedecor (de valor del estadístico 20'322) es el mismo que el obtenido con el estadístico t de Student (de valor $-4'508$). La razón es que se puede demostrar que una F de Snedecor con $(1, n)$ grados de libertad es el cuadrado de una t de Student con n grados de libertad (es decir, $20'322 = -4'508^2$, salvo los decimales que se pierden).

```
> anova(ajus) (1)
```

Analysis of Variance Table

```

Response: y
      Df Sum Sq Mean Sq F value    Pr(>F)
x       1  29.4810  29.4810   20.322 0.006352 **
Residuals  5   7.2533   1.4507
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Una vez determinada la recta de mínimos cuadrados, es posible predecir un valor de Y para un $X = x_i$ dado, simplemente sustituyendo dicho x_i en la ecuación de la recta de ajuste y determinando y_{ti} . Esta sería una estimación

por punto aunque también se podría determinar un intervalo de confianza para una predicción (García Pérez, 2008b).

10.5. Correlación Lineal

Situaciones como las analizadas en las secciones anteriores, en las que parecía existir una relación entre dos variables aleatorias, como la luminosidad y la temperatura de las estrellas, o la profundidad y la cantidad de oxígeno del lago Worther, son las que vamos a estudiar aquí.

Hasta ahora hemos visto si el tipo de relación existente entre ellas era o no lineal. Ahora contrastaremos el grado o fuerza de esta relación.

Con más precisión, si los datos $(x_1, y_1), \dots, (x_n, y_n)$ son valores tomados por una variable aleatoria bidimensional (X, Y) , la cual suponemos con distribución normal bivalente —véase Sección 4.6.1— de medias $\mu_1 = E[X]$, $\mu_2 = E[Y]$, varianzas $\sigma_1^2 = V(X)$, $\sigma_2^2 = V(Y)$ y coeficiente de correlación ρ , el propósito de esta sección es el hacer inferencias acerca del coeficiente de correlación poblacional ρ .

Ejemplo 10.2

Las calificaciones obtenidas, en dos asignaturas, por 17 alumnos de un centro escolar fueron las siguientes

X	3	4	6	7	5	8	7	3	5	4	8	5	5	8	8	8	5
Y	5	5	8	7	7	9	10	4	7	4	10	5	7	9	10	5	7

¿Qué se puede decir acerca del coeficiente de correlación poblacional entre ambas variables?

10.5.1. Estimación por punto de ρ

Vimos en el Capítulo 5 que un buen estimador puntual del coeficiente de correlación poblacional ρ es el coeficiente de correlación muestral, r , definido en la Sección 2.4.3 como coeficiente de correlación lineal de Pearson de la forma

$$r = \frac{n \sum_{i=1}^n X_i Y_i - \left(\sum_{i=1}^n X_i \right) \left(\sum_{i=1}^n Y_i \right)}{\sqrt{n \sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i \right)^2} \sqrt{n \sum_{i=1}^n Y_i^2 - \left(\sum_{i=1}^n Y_i \right)^2}}.$$

Pues bien, si se determinó la recta de mínimos cuadrados y, además, la tabla de análisis de la varianza para la regresión lineal, es más sencillo calcular r como la raíz cuadrada de

$$r^2 = \frac{SSEX}{SST} = \frac{\hat{\beta}_1^2 \left(\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2 / n \right)}{\sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2 / n}.$$

aunque hay que analizar por separado el signo de r . En el Ejemplo 10.2, r toma el valor

$$r = \frac{17 \cdot 739 - 99 \cdot 119}{\sqrt{17 \cdot 629 - (99)^2} \sqrt{17 \cdot 903 - (119)^2}} = 0'759.$$

La Estadística Matemática nos sugiere también otro estimador para ρ^2 , muy parecido al anterior y que usaremos menos en la estimación por punto, que, al igual que r^2 , se puede obtener de la tabla ANOVA; se trata de

$$\hat{\rho}^2 = 1 - \frac{SSNEX/(n-2)}{SST/(n-1)}.$$

10.5.2. Contraste de hipótesis sobre ρ

Como es habitual, suele ser más interesante determinar intervalos de confianza y tests de hipótesis para un parámetro poblacional que calcular estimaciones puntuales, ya que aquellos son más informativos al venir acompañados del coeficiente de confianza y del nivel de significación, o p-valor.

En este apartado vamos a explicar cómo ejecutar el test sobre la hipótesis nula $\rho = 0$. Otros tests e intervalos de confianza pueden estudiarse en García Pérez (2008b).

Contraste de $H_0 : \rho = 0$ frente a $H_1 : \rho \neq 0$

Se puede demostrar que cuando H_0 es cierta, el estadístico

$$t = r \sqrt{\frac{n-2}{1-r^2}}$$

sigue una distribución t de Student con $n-2$ grados de libertad, con lo que, fijado un nivel de significación α , se define el siguiente test

- | |
|---|
| <ul style="list-style-type: none"> • Se acepta H_0 si $t < t_{n-2; \alpha/2}$ • Se rechaza H_0 si $t \geq t_{n-2; \alpha/2}$ |
|---|

Ejemplo 10.2 (continuación)

Si queremos contrastar las hipótesis $H_0 : \rho = 0$ frente a $H_1 : \rho \neq 0$, es decir, no existe relación entre las dos notas de los alumnos del centro escolar frente a que sí existe una relación —o mejor dicho, correlación— significativa entre ambas variables, calcularíamos el valor del estadístico t , que en este caso toma el valor

$$t = 0'759 \sqrt{\frac{15}{1 - 0'5761}} = 4'515.$$

Como para un nivel de significación $\alpha = 0'05$ es $t > 2'131 = t_{15;0'025}$, rechazaremos H_0 , cosa que ya podíamos prever por el valor de r , el cual, recuérdese, es una buena estimación de ρ . Con R podemos calcular el valor del coeficiente de correlación lineal con la función `cor` y el contraste de la hipótesis nula de ser cero con la función `cor.test`. Así, los datos de este ejemplo los incorporamos en (1) y (2), obteniendo el valor del coeficiente de correlación ejecutando (3). El test de $H_0 : \rho = 0$ se ejecuta mediante (4), con el que se obtiene el p-valor en (5), que indica rechazar H_0 claramente.

```
> x<-c(3,4,6,7,5,8,7,3,5,4,8,5,5,8,8,5) (1)
```

```
> y<-c(5,5,8,7,7,9,10,4,7,4,10,5,7,9,10,5,7) (2)
```

```
> cor(x,y) (3)
```

```
[1] 0.7590159
```

```
> cor.test(x,y) (4)
```

Pearson's product-moment correlation

```
data: x and y
```

```
t = 4.5151, df = 15, p-value = 0.0004108
```

```
(5)
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
0.4382534 0.9082981
```

```
sample estimates:
```

```
cor
```

```
0.7590159
```

10.6. Modelo de la Regresión Lineal Múltiple

El Modelo de Regresión Lineal Múltiple supone una relación del tipo

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + e$$

entre la variable dependiente Y y las k independientes X_1, \dots, X_k .

Al igual que hacíamos en el con la regresión lineal simple, estaremos los denominados *coeficientes de regresión* $\beta_0, \beta_1, \dots, \beta_k$, con objeto de determinar el mejor *hiperplano de regresión muestral* de entre todos los de la forma

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k.$$

Para ello, deberemos tomar una muestra de tamaño n , la cual consistirá en una matriz de datos de la forma

$$\begin{array}{ccc|c} x_{11} & & x_{k1} & y_1 \\ x_{12} & \cdots & x_{k2} & y_2 \\ \cdots & & \cdots & \cdots \\ x_{1n} & \cdots & x_{kn} & y_n \end{array}$$

suponiendo que las distribuciones condicionadas por distintos (x_1, \dots, x_k) de $Y/X_1 = x_1, \dots, X_k = x_k$ son normales de varianza constante σ^2 , e independientes entre sí unas de otras.

Como hacíamos en el caso de la regresión lineal simple, los estimadores $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, de los coeficientes de regresión serán los de mínimos cuadrados, es decir, aquellos que hagan mínima la suma de cuadrados

$$\sum_{j=1}^n e_j^2 = \sum_{j=1}^n \left(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{1j} - \hat{\beta}_2 x_{2j} - \dots - \hat{\beta}_k x_{kj} \right)^2$$

los cuales resultan ser las soluciones en $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$, de un sistema de ecuaciones denominado *sistema de ecuaciones normales* (García Pérez, 2008b). No obstante, en su cálculo recomendamos utilizar siempre la función `lm` de R ya estudiada en la regresión lineal simple, aunque ahora expresaremos en el modelo más de una covariable.

Ejemplo 10.3

Se consideró que el Número de admisiones previas del paciente, X_1 , y su Edad, X_2 , podrían servir para predecir la Estancia en días, Y , que pasaban en un determinado hospital ciertos enfermos crónicos.

Con dicho propósito se tomó una muestra aleatoria simple de 15 pacientes la cual suministró los siguientes datos

X_1	0	0	0	1	1	1	1	2	2	2	3	3	4	4	5
X_2	21	18	22	24	25	25	26	34	25	38	44	51	39	54	55
Y	15	15	21	28	30	35	40	35	30	45	50	60	45	60	50

Se quiere analizar si alguna o ambas variables independientes X_1, X_2 , pueden servir para explicar a la variable dependiente Y , estimado previamente los coeficientes de regresión de las variables significativas.

El análisis de los coeficientes de regresión lo haremos más adelante, pero ya podemos determinar su estimación con R. Primero incorporamos los datos y, a continuación, se ejecuta (1), obteniendo las estimaciones en (2),

```
> x1<-c(0,0,0,1,1,1,1,2,2,2,3,3,4,4,5)
> x2<-c(21,18,22,24,25,25,26,34,25,38,44,51,39,54,55)
> y<-c(15,15,21,28,30,35,40,35,30,45,50,60,45,60,50)
> hiper<-lm(y~x1+x2)
> hiper
```

(1)

Call:

```
lm(formula = y ~ x1 + x2)
```

Coefficients:

(Intercept)	x1	x2	
2.08572	0.05699	1.05002	(2)

Es decir, el hiperplano de regresión muestral es

$$y_i = 2.0857 + 0.057 x_1 + 1.05 x_2.$$

10.6.1. Contraste de la Regresión Lineal Múltiple

Como en el caso de la regresión lineal simple, se puede contrastar la adecuación global del modelo o la igualdad a cero de los coeficientes de regresión. Antes vimos que estos dos tests eran equivalentes pues sólo había una covariable independiente; aquí no y, desde luego, son mucho más interesantes los segundos porque permitirán decidir cuáles covariables X_i son significativas y cuales no, en la explicación de la variable dependiente Y , de manera que se puedan descartar algunas de estas covariables independientes no significativas, antes de determinar la ecuación del hiperplano de regresión a utilizar en las predicciones. Haremos este análisis con la función `summary`.

Ejemplo 10.3 (continuación)

Para analizar si ambas covariables son o no significativas ejecutamos (1), observando en (2) los p-valores de los tests sobre los coeficientes de regresión, los cuales indican que puede aceptarse la hipótesis nula de ser cero el coeficiente de regresión de X_1 , debiendo eliminar esta variable del modelo.

Se puede observar en (3) que el p-valor del test sobre el modelo global indica aceptarle.

```
> summary(hiper) (1)
```

Call:

```
lm(formula = Y ~ x1 + x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.122	-3.543	1.542	2.317	10.557

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	2.08572	6.73931	0.309	0.76226
x1	0.05699	2.61310	0.022	0.98296
x2	1.05002	0.32621	3.219	0.00737 **

(2)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.059 on 12 degrees of freedom

Multiple R-Squared: 0.8503, Adjusted R-squared: 0.8254

F-statistic: 34.08 on 2 and 12 DF, p-value: 1.125e-05 (3)

Con objeto de completar el ejemplo, ejecutamos (4) y (5), obteniendo en (6) la recta de regresión lineal ajustada, cuyo p-valor asociado, (7), confirma que la Edad del paciente, X_2 , es significativa (ahora aún más) para explicar a la variable dependiente, Estancia en días en el hospital.

> hiper2<-lm(Y ~ x2) (4)

> summary(hiper2) (5)

Call:

lm(formula = Y ~ x2)

Residuals:

Min	1Q	Median	3Q	Max
-10.089	-3.561	1.534	2.345	10.552

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.977	4.373	0.452	0.659
x2	1.057	0.123	8.593	1.01e-06 ***

(6)

(7)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.821 on 13 degrees of freedom

Multiple R-Squared: 0.8503, Adjusted R-squared: 0.8388

F-statistic: 73.84 on 1 and 13 DF, p-value: 1.014e-06

La recta de regresión finalmente ajustada,

$$y_t = 1'977 + 1'057 x_2$$

permite predecir que, por ejemplo, un paciente de 60 años que ingrese en el hospital en estudio, es muy probable que esté en él,

$$y_t = 1'977 + 1'057 \cdot 60 = 65'397$$

días.

Una última observación sobre una pregunta que, seguramente se estará haciendo el lector: ¿por qué se aceptó más arriba que era válida una regresión lineal múltiple con las dos covariables explicativas si una de ellas no lo era? La respuesta es que el modelo global

$$y_t = 2'0857 + 0'057 x_1 + 1'05 x_2$$

es válido si sólo nos preguntamos si es válido o si no lo es (es decir, si todos los coeficientes de regresión son cero o no todos son cero), pero si realizamos un análisis más preciso, mediante

los tests individuales para los coeficientes de regresión, vemos que aún hay otro modelo (el de regresión lineal simple) que es mejor.

Observemos además que, si hubiéramos considerado este modelo más amplio, la estimación del coeficiente de regresión de la variable no significativa fue 0'057, lo que quiere decir que, en ese caso, apenas si ponderaría valores de X_1 .

10.7. Ejercicios de Autoevaluación

Ejercicio 10.1

Se cree que existe una relación de tipo lineal entre el nivel de ingresos de una familia y la cantidad de basura producida por ésta.

Con objeto de averiguar si puede confirmarse tal hipótesis, se eligieron al azar seis comunidades de vecinos para las que se anotó su renta X y el peso de la basura producida Y . Los resultados obtenidos fueron los siguientes:

X	18	20	16	10	9	15
Y	2	3	1'2	0'7	0'5	1'8

Determinar la recta de regresión y analizar si es significativa.

Ejercicio 10.2

Se cree que existe una relación de tipo lineal entre el punto de ebullición del agua y la presión atmosférica del lugar en el que ésta se pone a hervir. Para analizar esta hipótesis, se obtuvieron seis mediciones en Los Alpes a seis alturas diferentes en la que se observó una determinada presión atmosférica (en pulgadas de mercurio) X , anotándose la temperatura Y a la que comenzaba a hervir el agua (en grados Fahrenheit) en cada una de esas seis alturas. Los resultados obtenidos fueron los siguientes:

X	20'79	22'40	23'89	24'02	25'14	29'04
Y	194'5	197'9	200'9	201'4	203'6	210'7

Determinar la recta de regresión y analizar si es significativa.

Ejercicio 10.3

Se cree que el tamaño X de los asentamientos prehistóricos puede servir para predecir el tamaño de la población Y del lugar donde se aquellos se produjeron. Por ello se quiere determinar la recta de regresión basándose en datos actuales y, con ella, hacer estimaciones de tiempos pasados. Con este propósito se obtuvieron los siguientes datos de Tamaño de Asentamientos en hectáreas, X y Número de habitantes, Y , de los pueblos actuales del área en estudio:

X	0'6	1'0	1'1	1'2	1'6	1'9	2'3	3'0	3'1	3'3
Y	20	70	100	130	120	170	195	190	210	360

X	3'7	4'0	4'5	5'4	5'9	6'1	6'4	8'9	10'0	12'0
Y	300	250	500	270	190	630	650	310	730	850

Determinar la recta de regresión y analizar si es significativa.

Ejercicio 10.4

La Obsidiana es un mineral de origen volcánico al que los pobladores de Mesoamérica, en la época prehispánica, atribuían propiedades mitológicas (a causa de la leyenda del guerrero Obsid) y era muy utilizado en la fabricación de elementos de caza y defensa (tales como puntas de flecha, de lanza, raspadores, cuchillos) así como objetos rituales. En Arqueología se cree que los dos factores que influyen en la Densidad en gramos, Y , de hallazgos de este mineral en los yacimientos arqueológicos son, la Distancia en kilómetros, X_1 , a la cual se hallaba la fuente de donde se extraía el mineral y el Tamaño en hectáreas, X_2 , del asentamiento. Examinados cinco asentamientos, se obtuvieron los siguientes datos:

Y	40	35	30	20	25
X_1	100	90	80	75	70
X_2	35	32	28	20	30

Realizar un análisis de regresión lineal múltiple.

Ejercicio 10.5

En un estudio de restos de fauna en varias cuevas del Pleistoceno, se cree que el número de fragmentos de huesos de lobo, X_1 , y de huesos de bóvido, X_2 , son significativos para predecir el total de fragmentos de la cueva, Y . Los datos de que se dispone son los siguientes:

X_1	1	111	278	63	48	161	24	0	0	18
X_2	31	0	1622	150	13	12	0	33	58	107
Y	1211	618	4260	820	137	2916	249	128	505	998

Realizar un análisis de regresión lineal múltiple.

10.8. Lecturas Recomendadas

García Pérez, A. (2008b). *Estadística Aplicada: Conceptos Básicos*. Segunda Edición. Editorial UNED. Colección Educación Permanente.

Bibliografía General

- Bartlett, M.S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc.*, (Serie A), **160**, 268-282.
- Bermúdez de Castro, J.M. (2010). *La Evolución del Talento*. Editorial Debate.
- Bortkiewicz, L. von (1898). *Das Gesetz der kleinen Zahlen*, B. G. Teubner, Leipzig.
- Box, G.E.P., Hunter, W.G. y Hunter J.S. (1978). *Statistics for Experimenters*. Editorial Wiley.
- Braun W.J. y Murdoch, D.J. (2007). *A First Course in Statistical Programming*. Editorial Cambridge.
- Dalgaard, P. (2002). *Introductory Statistics with R*. Editorial Springer.
- De Moivre, A. (1733). Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi. Opúsculo en Latín del 12 de Noviembre de 1733.
- Feller W. (1975). *Introducción a la Teoría de Probabilidades y sus Aplicaciones*. Volumen I. Editorial Limusa.
- García Pérez, A. (1993a). *Estadística Aplicada con BMDP*. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (1993b). *Estadística Aplicada con SAS*. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (1998). *Fórmulas y Tablas Estadísticas*. Editorial UNED. Colección Adendas.
- García Pérez, A. (1998). *Problemas Resueltos de Estadística Básica*. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (2005a). *Métodos Avanzados de Estadística Aplicada. Técnicas Avanzadas*. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (2005b). *Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo*. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (2008a). *Ejercicios de Estadística Aplicada*. Editorial UNED. Colección Cuadernos de la UNED.
- García Pérez, A. (2008b). *Estadística Aplicada: Conceptos Básicos*. Segunda Edición. Editorial UNED. Colección Educación Permanente.
- García Pérez, A. (2008c). *Estadística Aplicada con R*. Editorial UNED. Colección Varia.
- García Pérez, A. y Cabrero Ortega, M.Y. (2009a). *Estadística Aplicada con SPSS*. Editorial UNED.
- García Pérez, A. y Cabrero Ortega, M.Y. (2009b). *Cuadernos de Estadística Aplicada: Arqueología y Paleontología*. Editorial UNED.

- García Pérez, A. (2010). *Cuadernos de Estadística Aplicada: Área de la Salud*. Editorial UNED.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionis conicis solem ambientum*, Hamburgo.
- Gibbons, J.D. y Chakraborti S. (2003). *Nonparametric Statistical Inference*. Editorial Marcel Dekker.
- Greenwood, P.E. y Nikulin M.S. (1996). *A Guide to Chi-Squared Testing*. Editorial Wiley.
- Haigh, J. (2003). *Matemáticas y Juegos de Azar: Jugar con la Probabilidad*. Editorial Tusquets.
- Johnson, N.L., Kemp, A.W. y Kotz, S. (2005). *Univariate Discrete Distributions*. Tercera Edición. Editorial Wiley.
- Johnson, N.L., Kotz, S. y Balakrishnan, N. (1994). *Continuous Univariate Distributions*. Volumen I. Segunda Edición. Editorial Wiley.
- Johnson, N.L., Kotz, S. y Balakrishnan, N. (1995). *Continuous Univariate Distributions*. Volumen II. Segunda Edición. Editorial Wiley.
- Laplace, P-S de (1814). *Essai Philosophique sur les probabilités*. (Existe traducción: *Ensayo filosófico sobre las probabilidades*, Alianza.)
- Mood, A.M., Graybill, F.A. y Boes, D.C. (1974). *Introduction to the Theory of Statistics*. Editorial McGraw-Hill.
- Quesada Paloma, V. y García Pérez, A. (1988). *Lecciones de Cálculo de Probabilidades*. Editorial Díaz de Santos.
- Rutherford, E. y Geiger, H. (1910). The probability variations in the distribution of α particles. *Philosophical Magazine*, Sixth Ser., **20**, 698-704.
- Sánchez-Crespo Rodríguez, J.L. y García España, E. (1961). *Estadística Descriptiva*. Editorial Instituto Nacional de Estadística.
- Sturges, H.A. (1926). The choice of a class interval. *Journal of the Americal Statistical Association*, **21**, 65-66.
- Student (1908). The probable error of a mean. *Biometrika*, **6**, 1-25.
- Tukey, J.W. (1977). *Exploratory Data Analysis*. Editorial Addison-Wesley.
- Vélez Ibarrola, R. y García Pérez, A. (1993). *Principios de Inferencia Estadística*. Editorial UNED.
- Warner, S.L.(1965): Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the Americal Statistical Association*, **60**, 63-69.

Soluciones a los Ejercicios de Autoevaluación

Se han incluido las soluciones de casi todos los ejercicios de autoevaluación. Su resolución detallada está en los textos *Problemas Resueltos de Estadística Básica y Ejercicios de Estadística Aplicada*.

Ejercicio 2.2

$$a = 0'125, M_e = 0'115, M_d = 0'06, p_{1/4} = 0'06, p_{3/4} = 0'185, R = 0'16 \\ s^2 = 0'0042, S^2 = 0'0048, V_p = 51'848, A_p = 1'00293, A_f = 0'159$$

Ejercicio 2.3

$$a = 30'93, M_e = 16'72, M_d = 9'78, p_{1/4} = 10'716, p_{6/10} = 19'72 \\ R = 746, s^2 = 3192'417, S^2 = 3195'988, V_p = 182'67, A_p = 0'374$$

Ejercicio 2.4

$$y_t = -3'865 + 0'027x, \quad r = 0'769$$

Ejercicio 2.5

$$y_t = -300 + 4'118x, \quad r = 0'976$$

Ejercicio 3.1

$$5/9$$

Ejercicio 3.2

$$P(A \cup B) = 13/40, \quad P(A \cup C) = 11/40$$

Ejercicio 3.3

$$0'204$$

Ejercicio 3.4

$$a) \quad 0'01, \quad b) \quad 1/30$$

Ejercicio 3.5

0'6346

Ejercicio 4.1

$$E[X] = -0'1 \quad , \quad D(X) = 0'6633$$

Ejercicio 4.2

0'3953

Ejercicio 4.3

$$a) \quad 0'8185 \quad , \quad b) \quad 0'6826$$

Ejercicio 4.4

$$a) \quad 1 - e^{-0'001x} \text{ si } x > 0 \quad , \quad b) \quad D(X) = 1000 \quad , \quad c) \quad 0'2231 \quad , \quad d) \quad 0'4065$$

Ejercicio 4.5

$$a) \quad 0'9415 \quad , \quad b) \quad 0'0014 \quad , \quad c) \quad 10 \quad , \quad d) \quad 0'0559$$

Ejercicio 5.1

$$a) \quad 0'0739 \quad , \quad b) \quad 0'5292$$

Ejercicio 5.2

$$a) \quad \text{Poisson}(2'3) \quad , \quad b) \quad 2'3 \quad , \quad c) \quad 16$$

Ejercicio 5.3

2977

Ejercicio 5.4

404

Ejercicio 5.5

0'08

Ejercicio 6.1

$$[-0'0558, 0'1558]$$

Ejercicio 6.2

$$[0'6693, 1'0307]$$

Ejercicio 6.3

$$[4'0654, 21'9346]$$

Ejercicio 6.4

[0'3216 , 0'4784]

Ejercicio 6.5

[8'157 , 15'843]

Ejercicio 7.1

p-valor= 0'01

Ejercicio 7.2

p-valor= 0'0009

Ejercicio 7.3

p-valor= 0'05033

Ejercicio 7.4

a) p-valor= 0'0013 , b) Potencia \approx 1

Ejercicio 7.5

0'01 < p-valor < 0'025

Ejercicio 8.1

p-valor < 0'005

Ejercicio 8.2

p-valor < 0'005

Ejercicio 8.3

a) p-valor= 0'0625 , b) Potencia= 0'2266

Ejercicio 8.4

p-valor > 0'1

Ejercicio 8.5

p-valor < 0'05

Ejercicio 9.1

a) p-valor < 0'01 , b) $\{I, II\}, \{II, IV, VI\}$ y $\{III, IV, V, VI\}$

Ejercicio 9.2

p-valor > 0'05

Ejercicio 9.3

a) $p\text{-valor} < 0'005$, b) $\{I\}$ y $\{II, III\}$

Ejercicio 9.4

$p\text{-valor} > 0'05$

Ejercicio 9.5

a) $p\text{-valor} < 0'005$, b) $\{I, II\}$ y $\{III, IV\}$

Ejercicio 10.1

$y_t = -1'3538 + 0'19685 x$, Regresión significativa

Ejercicio 10.2

$y_t = 154'03 + 1'9605 x$, Regresión significativa

Ejercicio 10.3

$y_t = 39'79 + 63'36 x$, Regresión significativa

Ejercicio 10.4

$y_t = -21'85 + 0'38 x_1 + 0'7 x_2$, Regresión dudosa

Ejercicio 10.5

$y_t = 320'9 + 10'33 x_1 + 0'67 x_2$, Restos de lobo significativos

Obtención de R

Debe ir a dirección de Internet

`http:// lib.stat.cmu.edu/R/CRAN`

o equivalentemente a

`http:// cran.r-project.org`

luego, en principio (según el sistema operativo que tenga) ir a

Windows

allí seleccionar

base

y *bajarse* el ejecutable último, algo así como

R-2.7.2-win32.exe

si la versión de R es la 2.7.2, versión que ha sido la utilizada para escribir este texto.

Suele ser también muy interesante *bajarse* el Manual (en pdf) de utilización de R. Para ello, en la columna de la izquierda de la pantalla de

`http:// lib.stat.cmu.edu/R/CRAN`

en la última sección de **Documentation** aparece la opción **Manuals**. Elija allí, **The R Reference Index** y seleccione con el botón de la derecha del ratón la opción **Guardar destino como...** para *bajarlo* a su ordenador.

Instalación y Manejo Básico de R

Una vez que se haya *bajado* el ejecutable más arriba comentado, debe descomprimirlo en su ordenador. Para ello lo más simple es que seleccione

Siguiente en todos los pasos, con lo que obtendrá incluso un icono de acceso directo al programa en el Escritorio de su ordenador.

Una vez que lo hayamos abierto, aparecerá el *prompt* `>` después de las siguientes observaciones

```
R version 2.7.2 (2008-08-25)
```

```
Copyright (C) 2008 The R Foundation for Statistical Computing
```

```
ISBN 3-900051-07-0
```

```
R es un software libre y viene sin GARANTIA ALGUNA.
```

```
Usted puede redistribuirlo bajo ciertas circunstancias.
```

```
Escriba 'license()' o 'licence()' para detalles de distribucion.
```

```
R es un proyecto colaborativo con muchos contribuyentes.
```

```
Escriba 'contributors()' para obtener más información y
```

```
'citation()' para saber cómo citar R o paquetes de R en publicaciones.
```

```
Escriba 'demo()' para demostraciones, 'help()' para el sistema on-line de ayuda,  
o 'help.start()' para abrir el sistema de ayuda HTML con su navegador.
```

```
Escriba 'q()' para salir de R.
```

```
[Previously saved workspace restored]
```

```
>
```

La línea donde aparece el *prompt* `>` es la *línea de comandos*. Allí es donde deberemos ejecutar nuestras *instrucciones*. Una de dichas instrucciones, con la que finalizar una sesión de R es

```
> q()
```

(Advertimos que, en los ejemplos de los textos, siempre incluiremos el *prompt* `>` el cual, evidentemente, no debe ser *teclado* si se quieren repetir aquellos.)

Al terminar una de las sesiones de R, el ordenador nos preguntará si queremos conservar los cálculos que hayamos realizado en la sesión, mediante la pregunta *Guardar imagen de área de trabajo?* (*Save workspace image?* si eligió el idioma inglés en la instalación). Si respondemos *Sí*, al comenzar la sesión siguiente podremos volver a utilizar los resultados de la sesión recién finalizada.