Prácticas de Estadística con SPSS

CEU
Universidad
San Pablo

Prácticas de Estadística con SPSS

Santiago Angulo Díaz-Parreño, José Miguel Cárdenas Rebollo, Anselmo Romero Limón y Alfredo Sánchez Alberca.

Primera impresión, Abril 2013

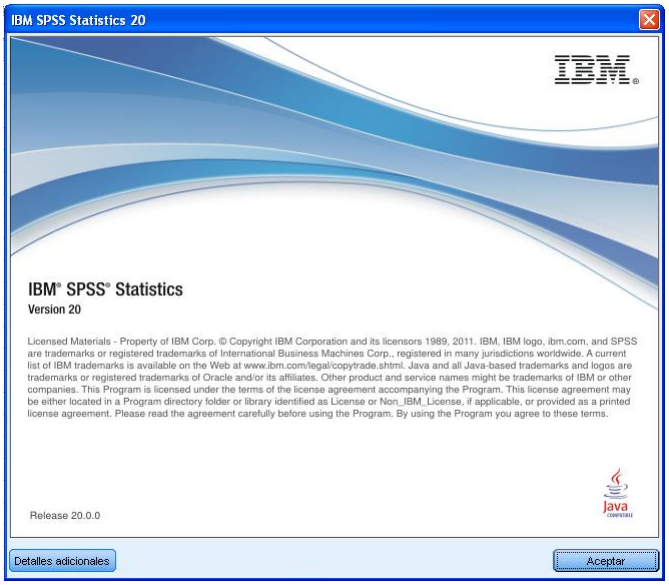
Índice general

1 — Introducción a SPSS

1.1 Introducción

La gran potencia de cálculo alcanzada por los ordenadores ha convertido a los mismos en poderosas herramientas al servicio de todas aquellas disciplinas que, como la estadística, requieren manejar un gran volumen de datos. Actualmente, prácticamente nadie se plantea hacer un estudio estadístico serio sin la ayuda de un buen programa de análisis estadístico.

SPSS®* es uno de los programas de análisis estadísticos más utilizados, sobre todo en el ámbito de las ciencias biosanitarias.



El objetivo de esta práctica es introducir al alumno en la utilización de este programa, enseñándole a realizar las operaciones básicas más habituales. A lo largo de la práctica, los alumnos aprenderán a crear variables, introducir datos de las muestras, transformar variables, filtrar datos y fundir e importar archivos de datos.

*Esta practica está basada en la versión 20.0 de SPSS® para Windows en español.

1.2 Funciones básicas

1.2.1 Arranque

Como cualquier otra aplicación de Windows, para arrancar el programa hay que hacer click sobre la opción correspondiente del menú *Inicio→Programas*, o bien sobre el icono de escritorio

Cuando el programa arranca, aparece la ventana del editor de datos (figura 1.1).

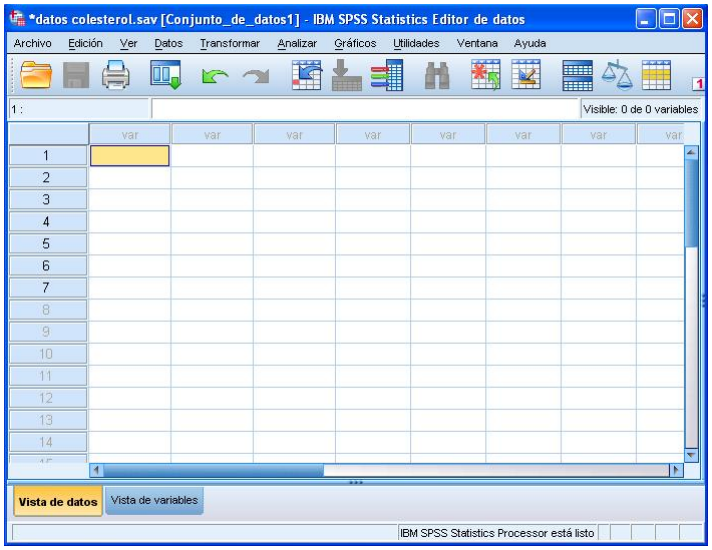


Figura 1.1: Ventana del editor de datos.

Como cualquier otra ventana de aplicación de Windows, la ventana del editor de datos tiene una barra de título, una barra de menús con las distintas funciones que puede hacer SPSS, entre ellas los análisis estadísticos de datos, una barra de botones que son atajos a las opciones más habituales de los menús, y una barra de estado en la parte inferior que nos indica lo que hace el programa en cada instante. Además, en la parte inferior aparecen dos pestañas que permiten pasar a la *Vista de datos* o a la *Vista de variables*.

1.2.2 Introducción de datos

Para realizar cualquier análisis, la ventana del editor de datos debe contener la matriz de datos a analizar. Una vez que el usuario obtiene los datos muestrales, estos deben introducirse en esta ventana. Para ello, lo primero es definir las variables que se han considerado en el estudio. Cada variable se corresponderá con una columna de la matriz de datos.

Para definir una variable debemos pasar a la *Vista de variables* haciendo click sobre la correspondiente pestaña (figura 1.2).

En esta otra ventana, debemos definir cada variable en una fila, rellenando los siguientes campos:

Nombre El nombre de la variable puede ser cualquier cadena de caracteres que comience por una letra y que no contenga espacios en blanco ni caracteres especiales como ?,¿,*, etc. Cada nombre de variable debe ser único y no se distingue entre mayúsculas y minúsculas.

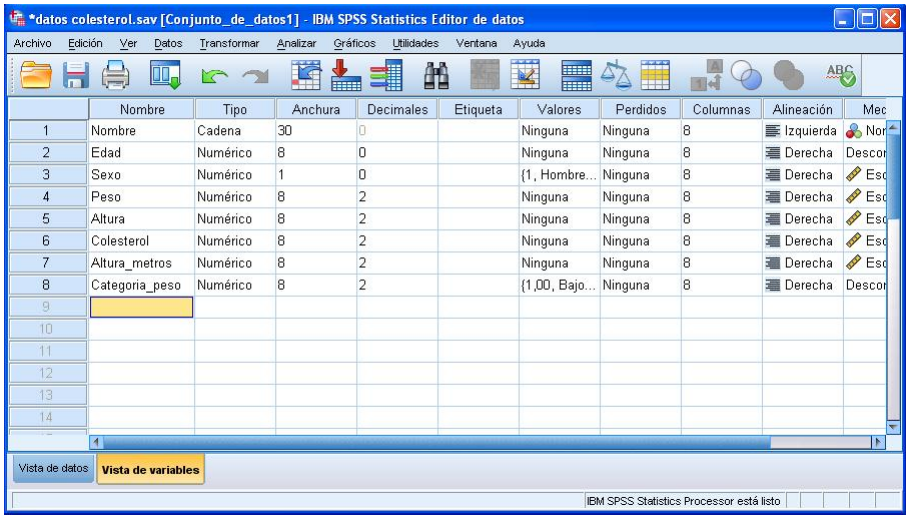


Figura 1.2: Vista de definición de variables.

Tipo Los tipos más comunes son Numérico (formato numérico estándar), Coma (con comas de separación cada tres cifras y punto para la parte decimal), Punto (con puntos de separación cada tres cifras y coma para la parte decimal), Notación Científica (utiliza la E para la exponenciación), Cadena (para datos alfanuméricos) y Fecha.

Anchura Es el número máximo de caracteres que pueden tener los valores de la variable.

Decimales Para las variables numéricas es el número de cifras decimales que podrán escribirse.

Etiqueta Es una descripción de la variable. Si el nombre de la variable es suficientemente descriptivo se puede omitir.

Valores Permite asignar etiquetas a los distintos valores que puede tomar la variable. No es obligatorio pero puede ser útil en algunos casos.

Al hacer click sobre la casilla aparece un cuadro de diálogo para asignar etiquetas a valores. Para ello basta con escribir un valor en el cuadro de texto *Valor* y la correspondiente etiqueta en el cuadro de texto *Etiqueta*. Después hay que hacer click sobre el botón *Añadir* y repetir los mismos pasos para todos los valores de la variable. Para finalizar hay que hacer click en el botón *Aceptar*.

Perdidos Permite definir qué valores se utilizarán para representar los datos perdidos por el usuario. Es útil para distinguir datos que se han perdido por distintas causas. Por ejemplo, puede ser interesante distinguir el dato perdido correspondiente a un entrevistado que se niega a responder, del dato perdido debido a que la pregunta no le afectaba al entrevistado. Los valores de datos especificados como perdidos por el usuario se excluyen de la mayoría de los cálculos.

Al hacer click sobre la casilla aparece un cuadro de diálogo donde deben indicarse los valores discretos que representarán valores perdidos (pueden introducirse hasta tres), o bien, el rango de valores que se representarán como valores perdidos.

Columnas Permite especificar el ancho de la columna en la que se introducirán los datos correspondientes a la variable.

Alineación Permite especificar la alineación de los datos correspondientes a la variable. Puede ser Izquierda, Derecha o Centrado.

Medida Permite especificar el tipo de escala utilizada para medir la variable. Puede ser *Escala* cuando la variable es numérica y la escala es de intervalo, *Ordinal* cuando los valores de la variable representan categorías con un cierto orden o *Nominal* cuando los valores representan categorías sin orden.

Rol Permite especificar la función que una variable tiene en el análisis. Puede ser *entrada*, cuando se trata de una variable independiente, *objetivo*, cuando es una variable dependiente, *ambos*, cuando la variable puede ser dependiente e independiente, *ninguna* si la variable no tiene ninguna función asignada, *particion*, cuando la variable se utilizará para dividir los datos en muestras separadas y *segmentar*, cuando se trata de una variable introducida para asegurar la compatibilidad en SPSS.

Una vez definidas las variables se procede a introducir los datos de la muestra. Para ello hay que volver a la *Ventana de datos* haciendo click en la correspondiente pestaña. Ahora aparecerán en las cabeceras de columna los nombres de las variables definidas. Cada individuo de la muestra se corresponde con una fila de la matriz de datos. Para introducir el valor de una variable en un individuo determinado, nos situamos en la celda de la fila de dicho individuo y de la columna de la variable, bien haciendo click sobre la misma, o bien desplazándonos por la matriz de datos con las flechas de movimiento del cursor del teclado, y se teclea el valor seguido de la tecla *Intro* (figura 1.3).

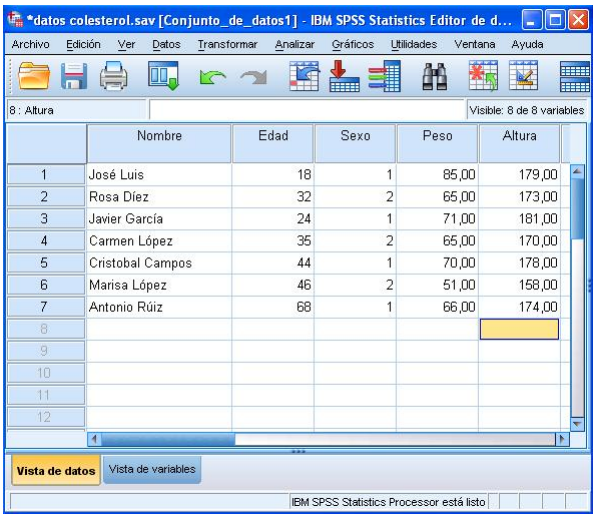


Figura 1.3: Introducción de datos en la matriz de datos. Cada columna corresponde a una variable y cada fila a un individuo de la muestra.

1.2.3 Guardar datos

Una vez introducidos los datos, conviene guardarlos en un fichero para no tener que volver a introducirlos en futuras sesiones. Para ello, se selecciona el menú *Archivo→Guardar*. Si el fichero ya existe, se actualizará su información, y si no, aparecerá un cuadro de diálogo en el que hay que introducir el nombre que queremos darle al fichero y la carpeta donde lo queremos ubicar. Los ficheros de datos de SPSS tienen por defecto extensión *.sav. Cuando los datos estén guardados en un fichero, el nombre del fichero aparecerá en el título de la ventana de datos (figura 1.3).

1.2.4 Recuperar datos

Si los datos con los que se pretende trabajar ya están guardados en un fichero, entonces tendremos que abrir dicho fichero. Para ello, se selecciona el menú *Archivo→Abrir→Datos* y se selecciona el fichero que se desea abrir. Automáticamente, los datos aparecerán en la vista de datos.

1.2.5 Modificación de datos

En ocasiones es necesario modificar los datos de la matriz de datos para corregir errores, añadir nuevos datos o eliminarlos. Para corregir un valor basta con seleccionar la celda que contiene el valor y teclear el nuevo. Otras operaciones habituales son:

- Insertar una variable nueva entre otras ya existentes. En la vista de variables se selecciona la fila que contiene la variable por encima de la cual queremos insertar la nueva, y se selecciona el menú *Edición→Insertar variable*.
- Eliminar una variable. En la vista de variables se selecciona la fila que contiene la variable a eliminar y se pulsa la tecla *Supr*.
- Insertar un individuo entre otros ya existentes. En la vista de datos se selecciona la fila que contiene los datos del individuo por encima del cual queremos insertar el nuevo, y se selecciona el menú *Edición→Insertar caso*.
- Eliminar un individuo. En la vista de datos se selecciona la fila que contiene los datos del individuo a eliminar y se se presiona la tecla *Supr*.

Cada vez que realicemos modificaciones en la matriz de datos, conviene volver a guardar los datos para que se actualice el fichero que los contiene.

¡Importante!: Cuando por equivocación realicemos una operación no deseada, podemos deshacerla mediante el menú *Edición→Deshacer*.

1.2.6 Transformación y generación de datos

En muchos análisis estadísticos se suelen transformar los datos de las variables originales en otros más convenientes para el análisis que se vaya a efectuar. Para generar una nueva variable mediante una transformación de otra ya existente o bien mediante funciones ya predefinidas se selecciona el menú *Transformar→Calcular Variable...* Entonces aparece la ventana de transformación de variables tal y como se muestra en la figura 1.4.

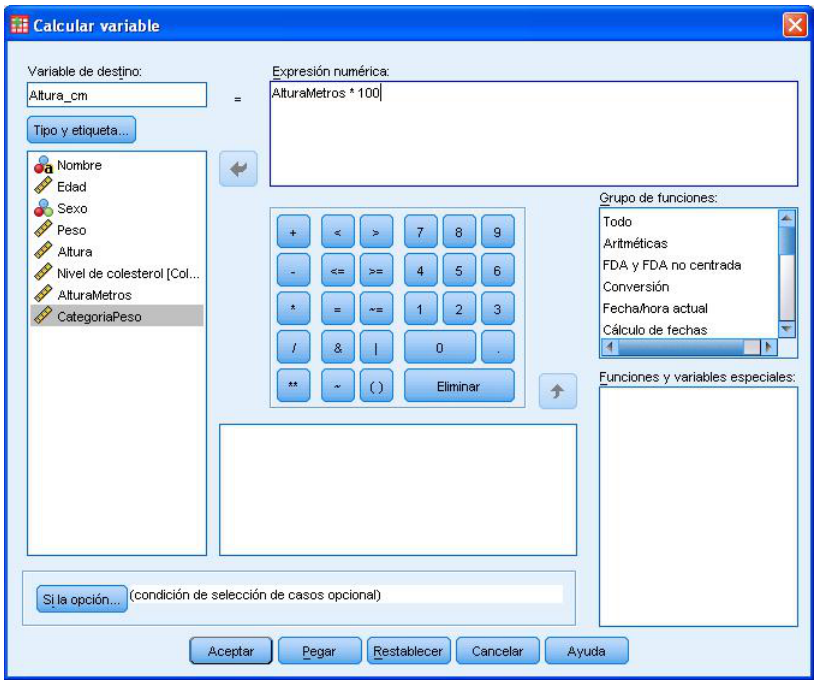


Figura 1.4: Ventana de transformación de variables. A la izquierda aparecen las variables ya definidas, a la derecha las funciones predefinidas que pueden utilizarse, y en el centro los operadores aritméticos y relacionales más comunes.

En esta ventana se debe introducir el nombre de la nueva variable en el cuadro *Variable de destino*, y la expresión cuyo resultado será el contenido de la nueva variable en el cuadro *Expresión numérica*. Para ello aparecen toda una serie de operadores y funciones para realizar la transformación, así como la lista de variables ya definidas que pueden utilizarse como argumentos de las distintas funciones de transformación.

Los operadores más habituales para construir expresiones son los aritméticos +, -, *, /, ** (potenciación), los relacionales =, <, >, ~=, <=, >= y los lógicos & (Y), | (O) y ~ (negación). Y algunas de las funciones más habituales son: ABS (valor absoluto), SQRT (raíz cuadrada), EXP (exponencial), LN (logaritmo neperiano), SIN (seno), COS (coseno), TAN (tangente), SUM (suma), MEAN (media aritmética), SD (desviación estándar), RND (redondeo al entero más cercano), TRUNC (parte entera de un número).

Haciendo click en el botón *Si la opción...* se pueden establecer condiciones de aplicación de la transformación. Para establecer una condición debemos activar la opción *Incluir si el caso satisface la condición* y después introducir una condición lógica como por ejemplo Sexo=1. De este modo, la transformación sólo se aplicará a los individuos que cumplan dicha condición.

Una vez definida la expresión hay que hacer click sobre el botón *Aceptar* y automáticamente aparecerá en la vista de datos una nueva columna con los datos transformados de la nueva variable.

1.2.7 Recodificación de datos

Otra forma de transformar una variable es crear otra cuyos valores sean una recodificación de los de la primera, por ejemplo agrupando en intervalos. Esta recodificación podemos hacerla tanto en la misma variable como en variables diferentes. Para ello se selecciona el menú *Transformar→Recodificar en distintas variables*. Automáticamente aparece la ventana de recodificación de variables tal y como se ve en la figura 1.5.

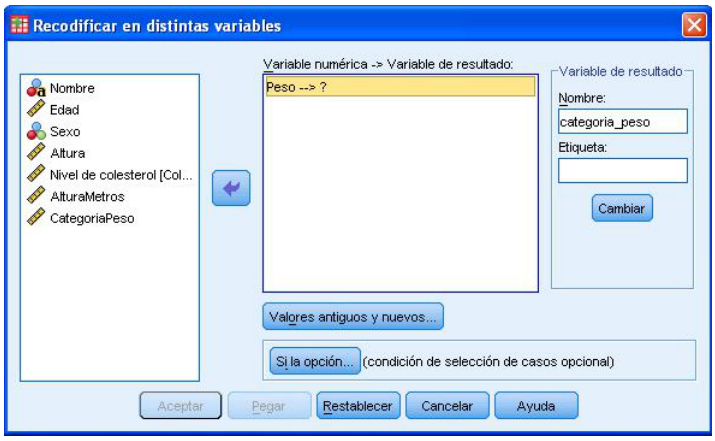


Figura 1.5: Ventana de recodificación de variables. A la izquierda aparecen las variables ya definidas, a la derecha deben especificarse las reglas de recodificación.

Para recodificar una variable en otra nueva, primero debemos seleccionar la variable que queremos recodificar y hacer click sobre el botón con una flecha que aparece al lado. Después hay que escribir el nombre de la nueva variable en el cuadro *Nombre* y hacer click sobre el botón *Cambiar*. A continuación hay que establecer las reglas de recodificación. Para ello hay que hacer click en el botón *Valores antiguos y nuevos* para que aparezca la ventana de definición de reglas (figura 1.6). Las reglas pueden establecer la conversión del valor de la variable original que introduzcamos en el cuadro *Valor antiguo* en el valor de la variable nueva que introduzcamos en el cuadro *Valor nuevo*, o bien la conversión de todo un intervalo de valores de la variable original en un valor de la variable nueva. Una vez definidos dichos valores hay que hacer click sobre el botón *Continuar*, y después sobre *Aceptar*.

1.2.8 Impresión

Para imprimir se utiliza el menú *Archivo→Imprimir*. Al instante aparece un cuadro de diálogo para la impresión donde debemos indicar si queremos imprimir todo o bien la selección que hayamos hecho. Tras esto se hace click sobre el botón *Aceptar* y la información se envía a la impresora.

Figura 1.6: Ventana de definición de reglas.

Antes de imprimir conviene hacer una previsualización de lo que se va a enviar a la impresora para estar seguros de que es eso lo que se quiere. Para ello se utiliza el menú *Archivo→Presentación preliminar*. Entonces aparece un visor donde se ve la página, tal y como se enviará a la impresora. Si todo parece correcto se puede hacer click sobre el botón *Imprimir* y aparecerá el cuadro de diálogo de impresión desde el que se puede enviar a la impresora definitivamente.

1.2.9 Salir del programa

Para terminar una sesión de trabajo se utiliza el menú *Archivo→Salir*, o bien se hace click sobre el aspa para cerrar la ventana del programa. Si quedan datos o resultados que no se han guardado, el programa nos preguntará antes de salir si deseamos guardarlos.

1.2.10 Ayuda

En esta práctica sólo hemos descrito las operaciones básicas en una sesión de trabajo. Pero quedan por describir todos los análisis estadísticos que pueden realizarse con los menús de la barra de menús. Aunque muchos de estos menús se explicarán en las siguientes prácticas, el programa dispone del menú de ayuda *Ayuda* en el que podemos encontrar una descripción de todos estos menús y al que podemos recurrir cada vez que tengamos dudas.

1.3 Ejercicios resueltos

1. Introducir en la matriz de datos los datos de la siguiente muestra y guardarlos en un fichero con el nombre datos_colesterol.sav.

Nombre	Sexo	Peso	Altura	Colesterol
José Luis Martínez Izquierdo	H	85	179	182
Rosa Díaz Díaz	M	65	173	232
Javier García Sánchez	H	71	181	191
Carmen López Pinzón	M	65	170	200
Marisa López Collado	M	51	158	148
Antonio Ruiz Cruz	H	66	174	249
Antonio Fernández Ocaña	H	62	172	276
Pilar Martín González	M	60	166	213
Pedro Gálvez Tenorio	H	90	194	241
Santiago Reillo Manzano	H	75	185	280
Macarena Álvarez Luna	M	55	162	262
José María de la Guía Sanz	H	78	187	198
Miguel Angel Cuadrado Gutiérrez	H	109	198	210
Carolina Rubio Moreno	M	61	177	194

i

(a) En la ventana de *Vista de variables*, crear las variables Nombre, Sexo, Peso, Altura y Colesterol e introducir los datos anteriores, siguiendo las indicaciones del apartado 1.2.2.

(b) Una vez introducidos los datos, se guardan en un fichero de nombre datos colesterol siguiendo lo indicado en el apartado 1.2.3.

2. Sobre la matriz de datos del ejercicio anterior realizar las siguientes operaciones:

   (a) Insertar detrás de la variable Nombre una nueva variable Edad con las edades de todos los individuos de la muestra.

Nombre	Edad
José Luis Martínez Izquierdo	18
Rosa Díaz Díaz	32
Javier García Sánchez	24
Carmen López Pinzón	35
Marisa López Collado	46
Antonio Ruiz Cruz	68
Antonio Fernández Ocaña	51
Pilar Martín González	22
Pedro Gálvez Tenorio	35
Santiago Reillo Manzano	46
Macarena Álvarez Luna	53
José María de la Guía Sanz	58
Miguel Angel Cuadrado Gutiérrez	27
Carolina Rubio Moreno	20

i

i. En la *Vista de variables* seleccionar la fila correspondiente a la variable Sexo haciendo click con el ratón sobre la cabecera de la misma y a continuación seleccionar el menú *Edición→Insertar variable*, con lo que aparece una nueva fila entre las variables Nombre y Sexo.

ii. En la nueva fila definir la variable Edad e introducir los datos anteriores.

iii. En la *Vista de datos* rellenar los datos de la columna correspondiente a la Edad.

   (b) Insertar entre los individuos 4º y 5º los datos correspondientes al siguiente individuo

Nombre: Cristóbal Campos Ruiz.
Edad: 44 años.
Sexo: Hombre.
Peso: 70 Kg.
Altura: 178 cm.
Colesterol: 220 mg/dl.

i. Seleccionar la fila correspondiente al 5º individuo, haciendo click con el ratón sobre la cabecera de la misma y a continuación seleccionar el menú *Edición→Insertar caso*, con lo que aparece una nueva fila entre las correspondientes a los individuos 4º y 5º.
ii. Introducir en la nueva fila los datos que se indican.

(c) Cambiar el valor de la variable Peso de Macarena Álvarez Luna por 58.

Hacer click con el ratón en la casilla cuyo contenido se desea modificar, escribir 58 y pulsar *Enter*.

(d) Transformar la variable Altura para que aparezca expresada en metros.

i. Seleccionar el menú *Transformar→Calcular variable*.
ii. En la ventana de transformación de datos introducir el nombre Altura_metros en el cuadro *Variable de destino*.
iii. Introducir la expresión Altura/100 en el cuadro *Expresión numérica*.
iv. Hacer click sobre el botón *Aceptar*.

(e) Recodificar la variable peso en las siguientes cuatro categorias, teniendo en cuenta el sexo:

Categoria	Hombres	Mujeres
Bajo	≤ 70	≤ 60
Medio	(70,85]	(60,70]
Alto	(85,100]	(70,80]
Muy Alto	> 100	> 80

i. Seleccionar el menú *Transformar→Recodificar en distintas variables*.
ii. En la ventana de recodificación de datos seleccionar la variable Peso y hacer click sobre el botón con una flecha que aparece al lado.
iii. Escribir el nombre de la variable recodificada Categoria_Peso en el cuadro *Nombre* de la *Variable de resultado* y hacer click sobre el botón *Cambiar*.
iv. Hacer click en el botón *Valores antiguos y nuevos* para abrir la ventana de definición de reglas de recodificación.
v. Para definir las reglas de recodificación de los hombres,
   A. Seleccionar la opción *Rango INFERIOR hasta valor* del cuadro *Valor antiguo* e introducir 70 en el cuadro correspondiente. Introducir 1 en el cuadro de la opción *Valor* del cuadro *Valor nuevo* y hacer click en el botón *Añadir*.
   B. Seleccionar la opción *Rango* del cuadro *Valor antiguo* e introducir 70 en el cuadro correspondiente y 85 en el cuadro *hasta*. Introducir 2 en el cuadro de la opción *Valor* del cuadro *Valor nuevo* y hacer click en el botón *Añadir*.
   C. Seleccionar la opción *Rango* del cuadro *Valor antiguo* e introducir 85 en el cuadro correspondiente y 100 en el cuadro *hasta*. Introducir 3 en el cuadro de la opción *Valor* del cuadro *Valor nuevo* y hacer click en el botón *Añadir*.

D. Seleccionar la opción *Rango valor hasta SUPERIOR* del cuadro *Valor antiguo* e introducir 100 en el cuadro correspondiente. Introducir 4 en el cuadro de la opción *Valor* del cuadro *Valor nuevo* y hacer click en el botón *Añadir*.

vi. Hacer click en el botón *Continuar* para cerrar la ventana.

vii. Hacer click en el botón *Si la opción...* para abrir la ventana de definición de concidiones.

viii. Seleccionar la opción *Incluir si el caso satisface la condición* e introducir la condición Sexo="H" en el cuadro correspondiente.

ix. Hacer click en el botón *Continuar* para cerrar la ventana.

x. Hacer click en el botón *Aceptar*.

xi. Repetir los mismos pasos para establecer las reglas de codificación de las mujeres.

xii. En *Vista de variables* hacer clik en *Valores de Categoria_Peso*, y en *Etiquetas de valor* ir asignando a los valores 1, 2, 3 y 4 las etiquetas Bajo, Medio, Alto y Muy alto respectivamente, haciendo click en el botón *Añadir* después de cada asignación, y una vez terminado hacer click en el *Aceptar*.

(f) Volver a guardar los cambios en el fichero anterior y salir del programa.

i

i. Seleccionar el menú *Archivo→Guardar*.

ii. Seleccionar el menú *Archivo→Salir*.

2 — Distribuciones de Frecuencias y Representaciones Gráficas

2.1 Fundamentos teóricos

Uno de los primeros pasos en cualquier estudio estadístico es el resumen y la descripción de la información contenida en una muestra. Para ello se van a aplicar algunos métodos de análisis descriptivo, que nos permitirán clasificar y estructurar la información al igual que representarla gráficamente.

Las características que estudiamos pueden ser o no susceptibles de medida; en este sentido definiremos una *variable* como un carácter susceptible de ser medido, es decir, cuantitativo y cuantificable mediante la observación, (por ejemplo el peso de las personas, la edad, etc...), y definiremos un *atributo* como un carácter no susceptible de ser medido, y en consecuencia observable tan sólo cualitativamente (por ejemplo el color de ojos, estado de un paciente, etc...). Se llaman modalidades a las posibles observaciones de un atributo.

Dentro de los atributos, podemos hablar de *atributos ordinales*, los que presentan algún tipo de orden entre las distintas modalidades, y de *atributos nominales*, en los que no existe ningún orden entre ellas.

Dentro de las variables podemos diferenciar entre *discretas*, si sus valores posibles son valores aislados, y *continuas*, si pueden tomar cualquier valor dentro de un intervalo.

En algunos textos no se emplea el término *atributo* y se denominan a todos los caracteres *variables*. En ese caso se distinguen *variables cuantitativas* para designar las que aquí hemos definido como *variables*, y *variables cualitativas* para las que aquí se han llamado *atributos*. En lo sucesivo se aplicará este criterio para simplificar la exposición.

2.1.1 Cálculo de frecuencias

Para estudiar cualquier característica, lo primero que deberemos hacer es un recuento de las observaciones, y el número de repeticiones de éstas. Para cada valor x_i de la muestra se define:

Frecuencia absoluta Es el número de veces que aparece cada uno de los valores x_i y se denota por n_i.

Frecuencia relativa Es el número de veces que aparece cada valor x_i dividido entre el tamaño muestral y se denota por f_i

$$f_i = \frac{n_i}{n}$$

Generalmente las frecuencias relativas se multiplican por 100 para que representen el tanto por ciento.

En el caso de que exista un orden entre los valores de la variable, a veces nos interesa no sólo conocer el número de veces que se repite un determinado valor, sino también el número de veces que aparece dicho valor y todos los anteriores. A este tipo de frecuencias se le denomina *frecuencias acumuladas*.

Frecuencia absoluta acumulada Es la suma de las frecuencias absolutas de los valores menores que x_i más la frecuencia absoluta de x_i, y se denota por N_i

$$N_i = n_1 + n_2 + \ldots + n_i$$

Frecuencia relativa acumulada Es la suma de las frecuencias relativas de los valores menores que x_i más la frecuencia relativa de x_i, y se denota por F_i

$$F_i = f_1 + f_2 + \ldots + f_i$$

Los resultados de las observaciones de los valores de una variable estadística en una muestra suelen representarse en forma de tabla. En la primera columna se representan los valores x_i de la variable colocados en orden creciente, y en la siguiente columna los valores de las frecuencias absolutas correspondientes n_i.

Podemos completar la tabla con otras columnas, correspondientes a las frecuencias relativas, f_i, y a las frecuencias acumuladas, N_i y F_i. Al conjunto de los valores de la variable observados en la muestra junto con sus frecuencias se le conoce como *distribución de frecuencias muestral*.

Ejemplo 1. *En una encuesta a 25 matrimonios, sobre el número de hijos que tienen, se obtienen los siguientes datos:*

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.

Los valores distintos de la variable son: 0, 1, 2, 3 y 4. Así frecuencia absoluta sería:

x_i	*Recuento*	n_i
0	*II*	2
1	*IIIIII*	6
2	*IIIIIIIIIIII*	14
3	*II*	2
4	*I*	1

Y la tabla de distribución de las frecuencias sería:

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
Suma	25	1		

Cuando el tamaño de la muestra es grande en el caso de variables discretas con muchos valores distintos de la variable, y en cualquier caso si se trata de variables continuas, se agrupan las observaciones en *clases*, que son intervalos contiguos, preferiblemente de la misma amplitud.

Para decidir el número de clases a considerar, una regla frecuentemente utilizada es tomar el entero más próximo a $\sqrt{n}$ donde n es el número de observaciones en la muestra. Pero conviene probar con distintos números de clases y escoger el que proporcione una descripción más clara. Así se prefijan los intervalos $(a_{i-1}, a_i], i = 1, 2, \ldots, l$ siendo $a = a_0 < a_1 < .... < a_l = b$ de tal modo que todos los valores observados estén dentro del intervalo $(a, b]$, y sin que exista ambigüedad a la hora de decidir a qué intervalo pertenece cada dato.

Llamaremos *marca de clase* al punto medio de cada intervalo. Así la *marca de la clase* $(a_{i-1}, a_i]$ es el punto medio x_i de dicha clase, es decir

$$x_i = \frac{a_{i-1} + a_i}{2}$$

En el tratamiento estadístico de los datos agrupados, todos los valores que están en una misma clase se consideran iguales a la marca de la clase. De esta manera si en la clase $(a_{i-1}, a_i]$ hay n_i valores observados, se puede asociar la marca de la clase x_i con esta frecuencia n_i.

2.1.2 Representaciones gráficas

Hemos visto que la tabla estadística resume los datos de una muestra, de forma que ésta se puede analizar de una manera más sistemática y resumida. Para conseguir una percepción visual de las características de la población resulta muy útil el uso de gráficas y diagramas. Dependiendo del tipo de variable y de si trabajamos con datos agrupados o no, se utilizarán distintos tipos.

Diagrama de barras y polígono de frecuencias

Consiste en representar sobre el eje de abscisas de un sistema de ejes coordenados los distintos valores de la variable X, y levantar sobre cada uno de esos puntos una barra cuya altura sea igual a la frecuencia absoluta o relativa correspondiente a ese valor, tal y como se muestra en la figura 2.1(a). Esta representación se utiliza para distribuciones de frecuencias con pocos valores distintos de la variable, tanto cuantitativas como cualitativas, y en este ultimo caso se suele representar con rectángulos de altura igual a la frecuencia de cada modalidad.

En el caso de variables cuantitativas se puede representar también el diagrama de barras de las frecuencias acumuladas, tal y como se muestra en la figura 2.1(b).

Otra representación habitual es el *polígono de frecuencias* que consiste en la línea poligonal cuyos vertices son los puntos (x_i, n_i), tal y como se ve en la figura 2.1(c), y si en vez de considerar las frecuencias absolutas o relativas se consideran las absolutas o relativas acumuladas, se obtiene el *polígono de frecuencias acumuladas*, como se ve en la figura 2.1(d).

Histogramas

Este tipo de representaciones se utiliza en variables continuas y en variables discretas en que se ha realizado una agrupación de las observaciones en clases. Un *histograma* es un conjunto de rectángulos, cuyas bases son los intervalos de clase $(a_{i-1}, a_i]$ sobre el eje OX y su altura la correspondiente frecuencia absoluta , relativa, absoluta acumulada, o relativa acumulada, tal y como se muestra en la figuras 2.2(a) y 2.2(b).

Si unimos los puntos medios de las bases superiores de los rectángulos del histograma, se obtiene el *polígono de frecuencias* correspondiente a datos agrupados (figura 2.2(c)).

El polígono de frecuencias también se puede utilizar para representar las frecuencias acumuladas, tanto absolutas como relativas. En este caso la línea poligonal se traza uniendo los extremos derechos de las bases superiores de los rectángulos del histograma de frecuencias acumuladas, en lugar de los puntos centrales (figura 2.2(d)).

Para variables cualitativas y cuantitativas discretas también se pueden usar las superficies representativas; de éstas, las más empleadas son los *sectores circulares*.

Sectores circulares o diagrama de sectores

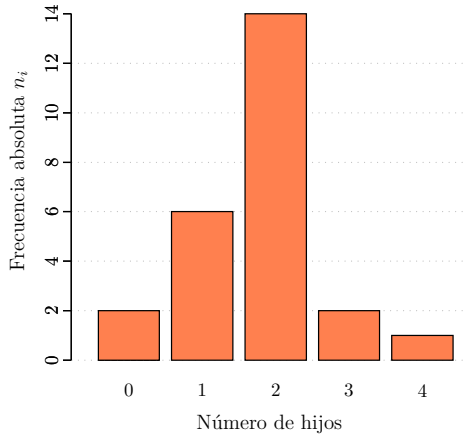
Es una representación en la que un círculo se divide en sectores, de forma que los ángulos, y por tanto las áreas respectivas, sean proporcionales a la frecuencia.

Ejemplo 2. *Se está haciendo un estudio en una población el grupo sanguíneo de sus ciudadanos. Para ello disponemos de una muestra de 30 personas, con los siguientes resultados: 5 personas con grupo 0, 14 con grupo A, 8 con grupo B y 3 con grupo AB.*

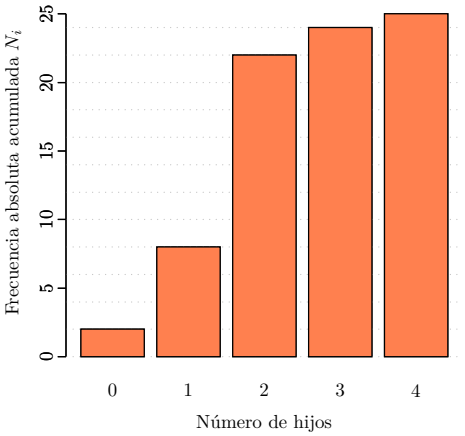
El diagrama de sectores de frecuencias relativas correspondiente aparece en la figura 2.3.

Diagrama de cajas y datos atípicos

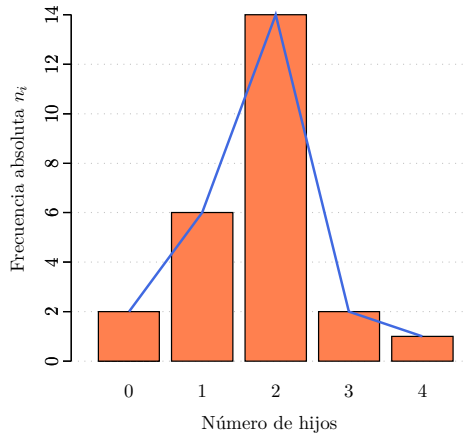
Los datos extremadamente altos o bajos, en comparación con los del resto de la muestra, reciben el nombre de datos influyentes o *datos atípicos*. Tales datos que, como su propio nombre indica, pueden modificar las conclusiones de un estudio, deben ser considerados atentamente antes de aceptarlos, pues no pocas veces podrán ser, simplemente, datos erróneos. La representación gráfica más apropiada para detectar estos datos es el *diagrama de cajas*. Este diagrama está formado por una caja que contiene el 50% de los datos centrales de la distribución, y unos



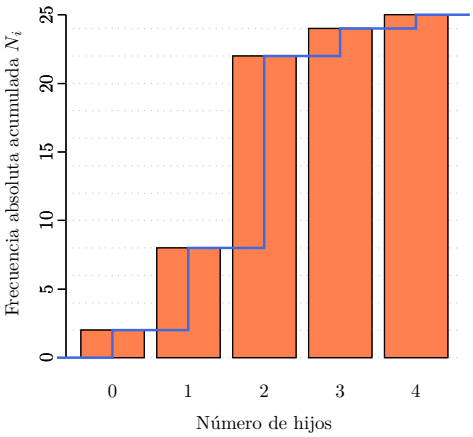
(a) Diagrama de barras de frecuencias absolutas.



(b) Diagrama de barras de frecuencias absolutas acumuladas.



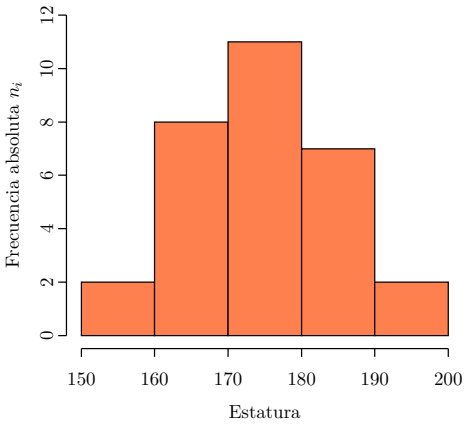
(c) Polígono de frecuencias absolutas.



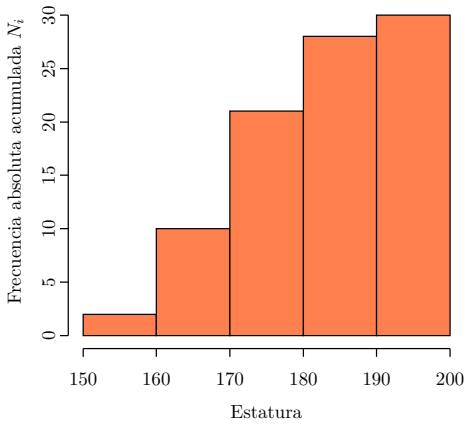
(d) Polígono de frecuencias absolutas acumuladas

Figura 2.1: Diagramas de barras y polígonos asociados para datos no agrupados.

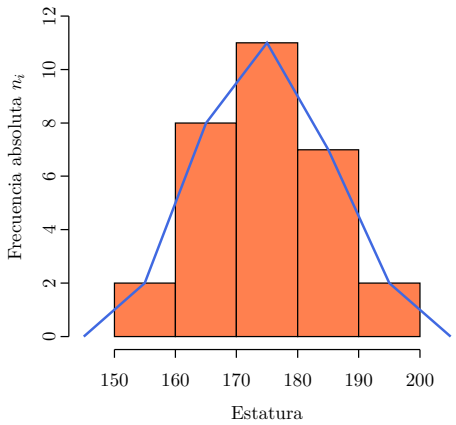
segmentos que salen de la caja, que indican los límites a partir de los cuales los datos se consideran atípicos. En la figura 2.4 se puede observar un ejemplo en el que aparecen dos datos atípicos.



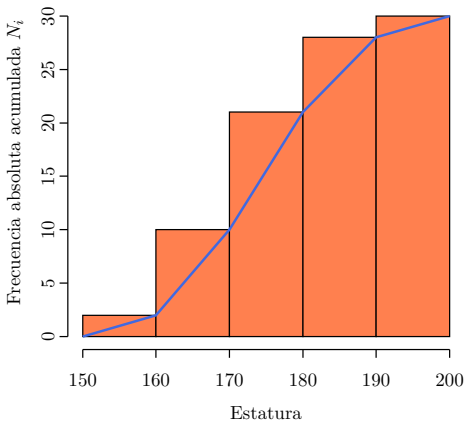
(a) Histograma de frecuencias absolutas.



(b) Histograma de frecuencias absolutas acumuladas.



(c) Polígono de frecuencias absolutas.



(d) Polígono de frecuencias absolutas acumuladas

Figura 2.2: Histograma y polígonos asociados para datos agrupados.

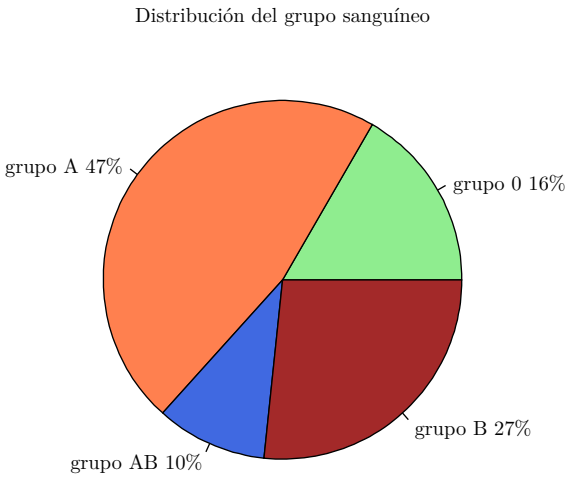


Figura 2.3: Diagrama de sectores de frecuencias relativas del grupo sanguineo.

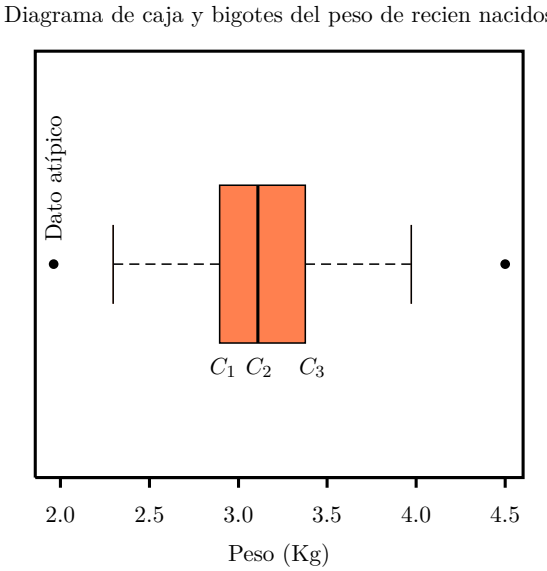


Figura 2.4: Diagrama de cajas para una muestra de recién nacidos. Existen dos niños con pesos atípicos, uno con peso extremadamente bajo 1.9 kg, y otro con peso extremadamente alto 4.5 kg.

2.2 Ejercicios resueltos

1. Se realizó una encuesta a 40 personas de más de 70 años sobre el número de medicamentos distintos que tomaban habitualmente. El resultado de dicha encuesta fue el siguiente:

$$3-1-2-2-0-1-4-2-3-5-1-3-2-3-1-4-2-4-3-2$$
$$3-5-0-1-2-0-2-3-0-1-1-5-3-4-2-3-0-1-2-3$$

Se pide:

(a) Crear la variable medicamentos e introducir los datos.

(b) Construir la tabla de frecuencias.

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
ii. Seleccionar la variable medicamentos en el campo *Variables* del cuadro de diálogo.
iii. Activar la opción *Mostrar tabla de frecuencias* y hacer click en el botón *Aceptar*.

(c) Dibujar el diagrama de barras de las frecuencias absolutas.

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Barras→Definir*.
ii. Seleccionar la variable medicamentos en el campo *Eje de categorías* del cuadro de diálogo y seleccionar la opción *N^o de casos*.

(d) Dibujar el polígono de frecuencias absolutas.

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Líneas→Definir*.
ii. Seleccionar la variable medicamentos en el campo *Eje de categorías* del cuadro de diálogo y seleccionar la opción *N^o de casos*.

(e) Dibujar el diagrama de barras de las frecuencias relativas acumuladas.

Repetir los mismos pasos del apartado c) pero seleccionando esta vez la opción *% acum*.

(f) Dibujar el diagrama de sectores.

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Sectores→Definir*.
ii. Seleccionar la variable medicamentos en el campo *Definir sectores por* del cuadro de diálogo.

2. En un hospital se realizó un estudio sobre el número de personas que ingresaron en urgencias en el mes de noviembre. Los datos observados fueron:

$$15-23-12-10-28-7-12-17-20-21-18-13-11-12-26$$
$$29-6-16-39-22-14-17-21-28-9-16-13-11-16-20$$

Se pide:

(a) Crear la variable urgencias e introducir los datos.

(b) Dibujar el histograma de las frecuencias absolutas agrupando en 5 clases desde el 0 hasta el 40.

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Histograma*.
ii. Seleccionar la variable urgencias en el campo *Variable* del cuadro de diálogo y hacer click en el botón *Aceptar*.

iii. Editar el histograma haciendo doble click sobre él.
iv. En el editor de gráficos hacer click con el botón derecho del ratón en la zona del histograma, el cual quedará rodeado de una línea amarilla y hacer click en la ventana emergente sobre *Ventana Propiedades*.
v. Seleccionar la opción *Clases de punto*, en *Eje X* elegir *Personalizado* y en *Número de Intervalos* poner 5.
vi. Hacer click sobre el botón *Aplicar*, después sobre el botón *Cerrar* de la ventana de propiedades y cerrar el editor de gráficos.

(c) Dibujar el diagrama de cajas. ¿Existe algún dato atípico?

i

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Diagramas de caja*.
ii. Seleccionar la opción *Resúmenes para distintas variables* y hacer click en el botón *Definir*.
iii. Seleccionar la variable urgencias en el campo *Las cajas representan* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.

(d) En el caso de que exista algún dato atípico, eliminarlo y dibujar el histograma de frecuencias absolutas, de forma que aparezcan clases de amplitud 5, comenzando en el 5 y terminando en el 30.

i

i. Identificar el caso que corresponde al dato atípico y eliminarlo en el editor de datos.
ii. Repetir los pasos del apartado b) para dibujar el histograma.

2.3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

$$0-1-2-1-3-0-1-0-1-2-0-1$$
$$1-1-2-0-1-3-2-1-2-1-0-1$$

Se pide:

(a) Crear la variable lesiones e introducir los datos.
(b) Construir la tabla de frecuencias.
(c) Dibujar el diagrama de barras de las frecuencias relativas acumuladas.
(d) Dibujar el polígono de frecuencias de las frecuencias absolutas acumuladas.
(e) Dibujar el diagrama de sectores.

2. Para realizar un estudio sobre la estatura de los estudiantes universitarios, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

Se pide:

(a) Crear la variable estatura e introducir los datos.
(b) Dibujar el histograma de las frecuencias absolutas agrupando desde 150 a 200 en clases de amplitud 10.
(c) Dibujar el diagrama de cajas. ¿Existe algún dato atípico?.

3 — Estadísticos Muestrales

3.1 Fundamentos teóricos

Hemos visto cómo podemos presentar la información que obtenemos de la muestra, a través de tablas o bien a través de gráficas. La tabla de frecuencias contiene toda la información de la muestra pero resulta difícil sacar conclusiones sobre determinados aspectos de la distribución con sólo mirarla. Ahora veremos cómo a partir de esos mismos valores observados de la variable estadística, se calculan ciertos números que resumen la información muestral. Estos números, llamados *Estadísticos*, se utilizan para poner de manifiesto ciertos aspectos de la distribución, tales como la dispersión o concentración de los datos, la forma de su distribución, etc. Según sea la característica que pretenden reflejar se pueden clasificar en Medidas de posición, Medidas de dispersión y Medidas de forma.

3.1.1 Medidas de posición

Son valores que indican cómo se sitúan los datos. Los más importantes son la Media aritmética, la Mediana y la Moda.

Media aritmética $\overline{x}$

Se llama *media aritmética* de una variable estadística X, y se representa por $\overline{x}$, a la suma de todos los resultados observados, dividida por el tamaño muestral. Es decir, la media de la variable estadística X, cuya distribución de frecuencias (x_i, n_i), viene dada por

$$\overline{x} = \frac{x_1 + \ldots + x_1 + \ldots + x_k + \ldots + x_k}{n_1 + \ldots + n_k} = \frac{x_1 n_1 + \ldots + x_k n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

La media aritmética sólo tiene sentido en variables cuantitativas.

Mediana *Me*

Se llama *mediana* y lo denotamos por Me, a aquel valor de la muestra que, una vez ordenados todos los valores de la misma en orden creciente, tiene tantos términos inferiores a él como superiores. En consecuencia, divide la distribución en dos partes iguales.

La mediana sólo tiene sentido en atributos ordinales y en variables cuantitativas.

Moda *Mo*

La *moda* es el valor de la variable que presenta una mayor frecuencia en la muestra. Cuando haya más de un valor con frecuencia máxima diremos que hay más de una moda. En variables continuas o discretas agrupadas llamaremos clase modal a la que tenga la máxima frecuencia. Se puede calcular la moda tanto en variables cuantitativas como cualitativas.

Cuantiles

Si el conjunto total de valores observados se divide en r partes que contengan cada una $\frac{n}{r}$ observaciones, los puntos de separación de las mismas reciben el nombre genérico de *cuantiles*.

Según esto la mediana también es un cuantil con $r = 2$. Algunos cuantiles reciben determinados nombres como:

Cuartiles. Son los puntos que dividen la distribución en 4 partes, con igual número de observaciones en cada una de ellas y se designan por C_1, C_2, C_3. Es claro que $C_2 = Me$.

Deciles. Son los puntos que dividen la distribución en 10 partes, con igual número de observaciones en cada una de ellas y se designan por $D_1, D_2, \ldots, D_9$.

Percentiles. Son los puntos que dividen la distribución en 100 partes, con igual número de observaciones en cada una de ellas y se designan por $P_1, P_2, \ldots, P_{99}$.

3.1.2 Medidas de dispersión

Miden la separación existente entre los valores de la muestra. Las más importantes son el Rango o Recorrido, el Rango Intercuartílico, la Varianza, la Desviación Típica y el Coeficiente de Variación.

Rango o Recorrido *Re*

La medida de dispersión más inmediata es el rango. Llamamos *recorrido* o *rango* y lo designaremos por *Re* a la diferencia entre los valores máximo y mínimo que toma la variable en la muestra. Es decir

$$Re = max\{x_i, i = 1, 2, \ldots, n\} - min\{x_i, i = 1, 2, \ldots, n\}$$

Este estadístico sirve para medir el campo de variación de la variable, aunque es la medida de dispersión que menos información proporciona sobre la mayor o menor agrupación de los valores de la variable alrededor de las medidas de tendencia central. Además tiene el inconveniente de que se ve muy afectado por los datos atípicos.

Rango intercuartílico *RI*

El *rango intercuartílico RI* es la diferencia entre el tercer y el primer cuartil, y mide, por tanto, el campo de variación del 50% de los datos centrales de la distribución. Por consiguiente

$$RI = C_3 - C_1$$

La ventaja del rango intercuartílico frente al recorrido es que no se ve tan afectado por los datos atípicos.

Varianza s_x^2

Llamamos *varianza* de una variable estadística X, y la designaremos por s_x^2, a la media de los cuadrados de las desviaciones de los valores observados respecto de la media de la muestra. Así

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i$$

Desviación típica s_x

La raíz cuadrada positiva de la varianza se conoce como *desviación típica* de la variable X, y se representa por s

$$s = +\sqrt{s_x^2}$$

Coeficiente de variación de Pearson Cv_x

Al cociente entre la desviación típica y el valor absoluto de la media se le conoce como *coeficiente de variación de Pearson* o simplemente *coeficiente de variación*:

$$Cv_x = \frac{s_x}{|\overline{x}|}$$

El coeficiente de variación es adimensional, y por tanto permite hacer comparaciones entre variables expresadas en distintas unidades. Cuanto más próximo esté a 0, menor será la dispersión de la muestra en relación con la media, y más representativa será ésta última del conjunto de observaciones.

3.1.3 Medidas de forma

Indican la forma que tiene la distribución de valores en la muestra. Se pueden clasificar en dos grupos: Medidas de *asimetría* y medidas de *apuntamiento o curtosis*.

Coeficiente de asimetría de Fisher g_1

El *coeficiente de asimetría de Fisher*, que se representa por g_1, se define como

$$g_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 f_i}{s_x^3}$$

Dependiendo del valor que tome tendremos:

- $g_1 = 0$. Distribución simétrica.
- $g_1 < 0$. Distribución asimétrica hacia la izquierda.
- $g_1 > 0$. Distribución asimétrica hacia la derecha.

Coeficiente de apuntamiento o curtosis g_2

El grado de apuntamiento de las observaciones de la muestra, se caracteriza por el *coeficiente de apuntamiento o curtosis* y se representa por g_2

$$g_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 f_i}{s_x^4} - 3$$

Dependiendo del valor que tome tendremos:

- $g_2 = 0$. La distribución tiene un apuntamiento igual que el de la distribución normal de la misma media y desviación típica. Se dice que es una distribución *mesocúrtica*.
- $g_2 < 0$. La distribución es menos apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *platicúrtica*.
- $g_2 > 0$. La distribución es más apuntada que la distribución normal de la misma media y desviación típica. Se dice que es una distribución *leptocúrtica*.

Tanto g_1 como g_2 suelen utilizarse para comprobar si los datos muestrales provienen de una población no normal. Cuando g_1 está fuera del intervalo [-2,2] se dice que la distribución es demasiado asimétrica como para que los datos provengan de una población normal. Del mismo modo, cuando g_2 está fuera del intervalo [-2,2] se dice que la distribución es, o demasiado apuntada, o demasiado plana, como para que los datos provengan de una población normal.

3.1.4 Estadísticos de variables en las que se definen grupos

Ya sabemos cómo resumir la información contenida en una muestra utilizando una serie de estadísticos. Pero hasta ahora sólo hemos estudiado ejemplos con un único carácter objeto de estudio.

En la mayoría de las investigaciones no estudiaremos un único carácter, sino un conjunto de caracteres, y muchas veces será conveniente obtener información de un determinado carácter, en función de los grupos creados por otro de los caracteres estudiados en la investigación. A estas variables que se utilizan para formar grupos se les conoce como *variables clasificadoras* o *discriminantes*.

Por ejemplo, si se realiza un estudio sobre un conjunto de niños recién nacidos, podemos estudiar su peso. Pero si además sabemos si la madre de cada niño es fumadora o no, podremos hacer un estudio del peso de los niños de las madres fumadoras por un lado y los de las no fumadoras por otro, para ver si existen diferencias entre ambos grupos.

3.2 Ejercicios resueltos

1. Se realizó una encuesta a 40 personas de más de 70 años sobre el número de medicamentos distintos que tomaban habitualmente. El resultado de dicha encuesta fue el siguiente:

$$3-1-2-2-0-1-4-2-3-5-1-3-2-3-1-4-2-4-3-2$$
$$3-5-0-1-2-0-2-3-0-1-1-5-3-4-2-3-0-1-2-3$$

Se pide:

(a) Crear la variable medicamentos e introducir los datos. Si ya se tienen los datos, simplemente recuperarlos.

(b) Calcular la media aritmética, mediana, moda, varianza y desviación típica de dicha variable. Interpretar los estadísticos.

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
ii. Seleccionar la variable medicamentos en el campo *Variables* del cuadro de diálogo.
iii. Hacer click sobre el botón *Estadísticos*. Para seleccionar únicamente los estadísticos que nos piden, marcar las casillas correspondientes a dichos estadísticos y hacer click sobre los botones *Continuar* y *Aceptar*.

(c) Calcular el coeficiente de asimetría y el de curtosis e interpretar los resultados

i Seguir los mismos pasos del apartado anterior, seleccionando ahora los estadísticos que se piden.

(d) Calcular los cuartiles.

i Seguir los mismos pasos de los apartados anteriores, activando la opción *Cuartiles*.

2. En un grupo de 20 alumnos, las calificaciones obtenidas en Matemáticas fueron:

SS - AP - SS - AP - AP - NT - NT - AP - SB - SS
SB - SS - AP - AP - NT - AP - SS - NT - SS - NT

Se pide:

(a) Crear la variable calificaciones e introducir los datos.

(b) Recodificar esta variable, asignando 2.5 al SS, 5.5 al AP, 7.5 al NT y 9.5 al SB.

i

i. Seleccionar el menú *Transformar→Recodificar en distintas variables*.
ii. Seleccionar la variable calificaciones y hacer click sobre el botón con la flecha del cuadro de diálogo para llevarla a Variable de entrada.
iii. Introducir el nombre de la nueva variable en el campo *Nombre* del cuadro de diálogo y hacer click en el botón *Cambiar*.
iv. Hacer click en el botón *Valores antiguos y nuevos* e introducir las reglas de recodificación y hacer click sobre los botones *Continuar* y *Aceptar*.

(c) Calcular la moda y la mediana.

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
ii. Seleccionar la variable recodificada en el campo *Variables* del cuadro de diálogo.
iii. Hacer click sobre el botón *Estadísticos*, seleccionar los estadísticos que se piden y hacer click sobre los botones *Continuar* y *Aceptar*.

3. Para realizar un estudio sobre la estatura de los estudiantes universitarios, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

   179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
   162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
   175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

   Se pide:

   (a) Crear la variable estatura e introducir los datos.

   (b) Obtener un resumen de estadísticos en el que se muestren la media aritmética, mediana, moda, varianza, desviación típica y cuartiles. Interpretar los estadísticos.

   i

   i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
   ii. Seleccionar la variable estatura en el campo *Variables* del cuadro de diálogo.
   iii. Hacer click sobre el botón *Estadísticos*, seleccionar los estadísticos que se piden y hacer click sobre los botones *Continuar* y *Aceptar*.

   (c) Calcular el tercer decil e interpretarlo.

   i Seguir los mismos pasos de los apartados anteriores, activando la opción *Percentiles* e introduciendo el percentil deseado en el correspondiente cuadro de texto.

   (d) Con los datos obtenidos en apartados anteriores, calcular el coeficiente de variación de Pearson y el rango intercuartílico, e interpretar los resultados.

4. Para realizar un estudio sobre la estatura de los estudiantes universitarios, seleccionamos, mediante un proceso de muestreo aleatorio, una muestra de 30 estudiantes, obteniendo los siguientes resultados (medidos en centímetros):

x_i	Marca	n_i	f_i	N_i	F_i
[150,160)	155	2	0,07	2	0,07
[160,170)	165	7	0,23	9	0,3
[170,180)	175	12	0,4	21	0,7
[180,190)	185	7	0,23	28	0,93
[190,200)	195	2	0,07	30	1

   Se pide:

   (a) Crear la variable estatura, en la que vamos a introducir las marcas de la clase y crear la variable frecuencias, en la que se introducirán las frecuencias absolutas.

   (b) Ponderar los casos de la variable estatura con las frecuencias de la variable frecuencias

   i

   i. Seleccionar el menú *Datos→Ponderar casos*.
   ii. Activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencias y hacer click sobre el botón *Aceptar*.

   (c) Obtener un resumen de estadísticos en el que se muestren la media aritmética, mediana, moda, varianza, desviación típica y cuartiles.

   i

   i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
   ii. Seleccionar la variable estatura en el campo *Variables* del cuadro de diálogo.
   iii. Hacer click sobre el botón *Estadísticos*, seleccionar los estadísticos que se piden

> y hacer click sobre los botones *Continuar* y *Aceptar*.

¿Existen diferencias entre estos estadísticos y los del ejercicio anterior? ¿A qué se deben?

(d) Calcular el tercer decil.

> *i* Seguir los mismos pasos de los apartados anteriores, activando la opción *Percentiles* e introduciendo el percentil correspondiente en el cuadro de texto.

(e) Calcular el percentil 62.

> *i* Seguir los pasos de los apartados anteriores seleccionando el estadístico deseado.

(f) Con los datos obtenidos en apartados anteriores, calcular el coeficiente de variación de Pearson y el rango intercuartílico, e interpretar los resultados.

5. En un hospital se ha tomado nota de la concentración de anticuerpos de inmunoglobulina M en el suero sanguíneo de personas sanas, y han resultado los siguientes datos por litro. Entre paréntesis figura el sexo de la persona (H para hombre y M para mujer).

(H) 1.071	(H) 0.955	(H) 0.73	(M) 0.908	(M) 0.859
(H) 0.927	(M) 0.962	(M) 1.543	(H) 1.094	(M) 0.847
(H) 1.214	(M) 1.456	(M) 1.516	(M) 1.002	(M) 0.799
(M) 0.881	(M) 1.096	(M) 0.964	(H) 0.973	(H) 1.222
(H) 0.887	(H) 1.022	(M) 0.881	(M) 1.42	(M) 1.205

Se pide

(a) Crear las variables sexo e inmunoglobulina e introducir los datos.

(b) Dividir el archivo, usando como variables de segmentación la variable sexo

> *i*
> i. Seleccionar el menú *Datos→Dividir archivo...*
> ii. Seleccionar la opción *comparar los grupos* u *Organizar los resultados por grupos* (Se diferencian en la forma de presentar los resultados).
> iii. Seleccionar la variable sexo en el campo *Grupos basados en* del cuadro de diálogo y hacer clic sobre el botón *Aceptar*.

(c) Calcular la media aritmética, la moda y la mediana de la inmunoglobolina, tanto en hombres como en mujeres.

> *i*
> i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
> ii. Seleccionar la variable inmunoglobulina en el campo *Variables* del cuadro de diálogo.
> iii. Hacer click sobre el botón *Estadísticos*, seleccionar los estadísticos que se piden, hacer click sobre el botón *Continuar* y finalmente hacer click sobre el botón *Aceptar*.

(d) Calcular la varianza y la desviación típica tanto en hombres como en mujeres.

> *i* Seguir los mismos pasos del apartado anterior.

(e) ¿En qué población es más representativa la media, en la de hombres o en la de mujeres?

> *i* Para responder a la pregunta será necesario calcular el coeficiente de variación.

6. Se reciben dos lotes de un determinado fármaco, fabricados con dos modelos de maquinaria diferentes, además uno proviene de Madrid y el otro de Valencia, se toma una muestra de 10 cajas, cinco de cada lote y se mide la concentración del principio activo, obteniendo los siguientes resultados:

Modelo Maquinaria	A	B	A	B	A
Procedencia	Madrid	Madrid	Valencia	Madrid	Valencia
Concentración (mg/mm^3)	17,6	19,2	21,3	15,1	17,6

Modelo Maquinaria	B	A	B	B	A
Procedencia	Valencia	Madrid	Valencia	Madrid	Valencia
Concentración (mg/mm^3)	18,9	16,2	18,3	19	16,4

Se pide :

(a) Crear las variables procedencia, maquinaria y concentracion e introducir los datos.

(b) Calcular la media aritmética, desviación típica, coeficiente de asimetría y curtosis de la concentración según el lugar de procedencia.

i

i. Seleccionar el menú *Datos→Dividir archivo...*
ii. Seleccionar la opción *comparar los grupos* u *Organizar los resultados por grupos.*
iii. Seleccionar la variable procedencia en el campo *Grupos basados en* del cuadro de diálogo y hacer clic sobre el botón *Aceptar.*
iv. Seguir los mismos pasos del ejercicio anterior para seleccionar los estadísticos.

(c) Dibujar el diagrama de cajas de la concentración de principio activo, según la maquinaria de fabricación.

i

i. Seleccionar el menú *Datos→Dividir archivo...*
ii. Seleccionar la opción *comparar los grupos* u *Organizar los resultados por grupos.*
iii. Seleccionar la variable maquinaria en el campo *Grupos basados en* del cuadro de diálogo y hacer clic sobre el botón *Aceptar.*
iv. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Diagramas de caja....*
v. Seleccionar la opción *Resúmenes para distintas variables* y hacer click en el botón *Definir.*
vi. Seleccionar la variable concentracion en el campo *Las cajas representan* del cuadro de diálogo y hacer click sobre el botón *Aceptar.*

3.3 Ejercicios propuestos

1. El número de lesiones padecidas durante una temporada por cada jugador de un equipo de fútbol fue el siguiente:

$$0-1-2-1-3-0-1-0-1-2-0-1$$
$$1-1-2-0-1-3-2-1-2-1-0-1$$

Se pide:

(a) Crear la variable lesiones e introducir los datos. Si ya se tienen los datos, simplemente recuperarlos.

(b) Calcular: media aritmética, mediana, moda, varianza y desviación típica.

(c) Calcular los coeficientes de asimetría y curtosis e interpretar los resultados.

(d) Calcular el cuarto y el octavo decil.

2. En una encuesta sobre la intención de voto en unas elecciones en las que se presentaban tres partidos A, B y C, se preguntó a 30 personas y se obtuvieron las siguientes respuestas:

$$A-B-VB-A-C-A-VB-C-A-A-B-B-A-B-B$$
$$B-A-A-C-B-B-B-A-VB-A-B-VB-A-B-B$$

Se pide:

(a) Crear la variable voto e introducir los datos.

(b) Calcular aquellos estadísticos que sea posible para este atributo

3. La siguiente tabla expresa la distribución de las puntuaciones obtenidas por un grupo de alumnos.

0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
7	8	13	6	7	6	6	5	6	2

Se pide:

(a) Calcular la media aritmética, la mediana y la moda.

(b) Calcular el percentil 92.

(c) Calcular la desviación típica.

(d) Calcular el coeficiente de asimetría.

(e) Calcular del coeficiente de curtosis.

4. En un estudio de población se tomó una muestra de 27 personas, y se les preguntó por su edad y estado civil, obteniendo los siguientes resultados:

Estado Civil	Casado	Soltero	Soltero	Viudo	Casado	Casado	Divorciado	Soltero	Soltero
Edad	62	31	45	100	39	62	31	45	100

Estado Civil	Soltero	Viudo	Casado	Soltero	Divorciado	Viudo	Divorciado	Soltero	Viudo
Edad	21	38	59	62	65	38	59	62	65

Estado Civil	Casado	Viudo	Casado	Divorciado	Divorciado	Viudo	Viudo	Soltero	Viudo
Edad	21	31	62	59	65	38	59	31	65

Se pide:

(a) Crear las variables adecuadas e introducir los datos.

(b) Calcular la media y la desviación típica de la edad según el estado civil.

(c) Dibujar el diagrama de barras para las frecuencias absolutas de la edad según el estado civil.

5. En un estudio se ha medido la tensión arterial de 25 individuos. Además se les ha preguntado si fuman y beben:

Fumador	si	no	si	si	si	no	no	si	no	si	no	si	no
Bebedor	no	no	si	si	no	no	si	si	no	si	no	si	si
Tensión arterial	80	92	75	56	89	93	101	67	89	63	98	58	91

Fumador	si	no	no	si	no	no	no	si	no	si	no	si
Bebedor	si	no	si	si	no	no	si	si	si	no	si	no
Tensión arterial	71	52	98	104	57	89	70	93	69	82	70	49

Se pide :

(a) Crear las variables correspondientes e introducir los datos.

(b) Calcular la media aritmética, desviación típica, coeficiente de asimetría y curtosis de la tensión arterial por grupos dependiendo de si beben y/o fuman.

(c) Dibujar el histograma para las frecuencias absolutas de la tensión arterial según lo grupos hechos anteriormente.

4 — Regresión Lineal Simple y Correlación

4.1 Fundamentos teóricos

4.1.1 Regresión

La *regresión* es la parte de la estadística que trata de determinar la posible relación entre una variable numérica Y, que suele llamarse *variable dependiente*, y otro conjunto de variables numéricas, $X_1, X_2, \ldots, X_n$, conocidas como *variables independientes*, de una misma población. Dicha relación se refleja mediante un modelo funcional $y = f(x_1, \ldots, x_n)$.

El caso más sencillo se da cuando sólo hay una variable independiente X, y entonces se habla de *regresión simple*. En este caso el modelo que explica la relación entre X e Y es una función de una variable $y = f(x)$.

Dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Familia de curvas	Ecuación genérica
Lineal	$y = b_0 + b_1 x$
Cuadrática	$y = b_0 + b_1 x + b_2 x^2$
Cúbica	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Potencia	$y = b_0 \cdot x^{b_1}$
Exponencial	$y = b_0 \cdot e^{b_1 x}$
Logarítmica	$y = b_0 + b_1 \ln x$
Inversa	$y = b_0 + \frac{b_1}{x}$
Compuesto	$y = b_0 b_1^x$
Crecimiento	$y = e^{b_0 + b_1 x}$
G (Curva-S)	$y = e^{b_0 + \frac{b_1}{x}}$

Para elegir un tipo de modelo u otro, se suele representar el *diagrama de dispersión*, que consiste en dibujar sobre unos ejes cartesianos correspondientes a las variables X e Y, los pares de valores (x_i, y_j) observados en cada individuo de la muestra.

Ejemplo 3. *En la figura 4.1 aparece el diagrama de dispersión correspondiente a una muestra de 30 individuos en los que se ha medido la estatura en cm (X) y el peso en kg (Y). En este caso la forma de la nube de puntos refleja una relación lineal entre la estatura y el peso.*

Según la forma de la nube de puntos del diagrama, se elige el modelo más apropiado (figura 4.2), y se determinan los parámetros de dicho modelo para que la función resultante se ajuste lo mejor posible a la nube de puntos.

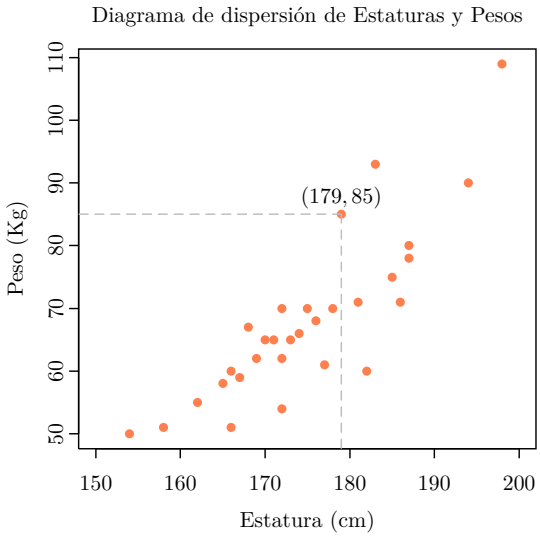
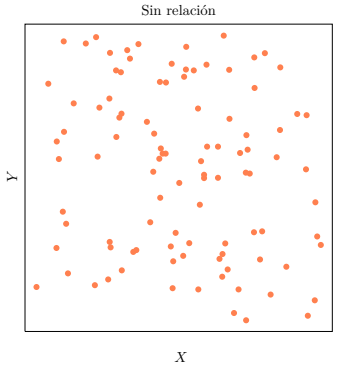
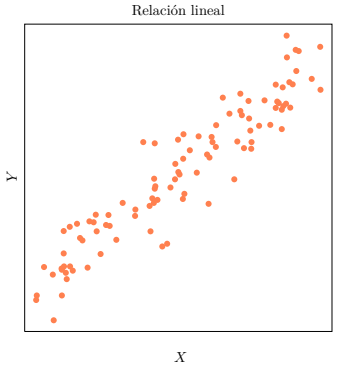


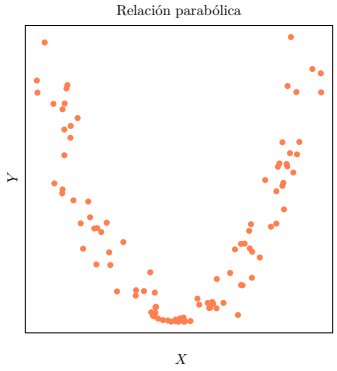
Figura 4.1: Diagrama de dispersión. El punto (179,85) indicado corresponde a un individuo de la muestra que mide 179 cm y pesa 85 Kg.



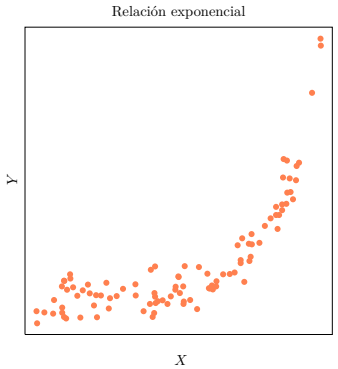
(a) Sin relación.



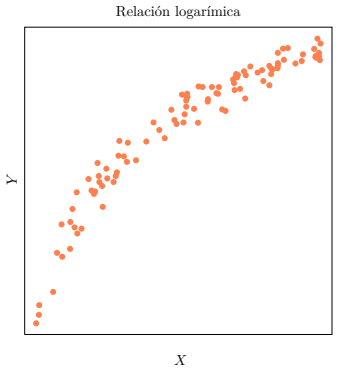
(b) Relación lineal.



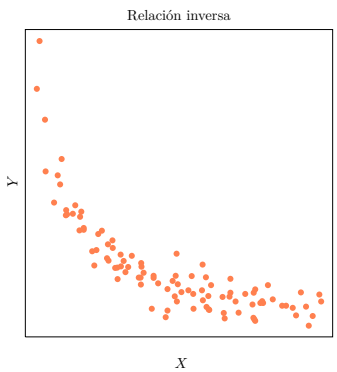
(c) Relación polinómica.



(d) Relación exponencial.



(e) Relación logarítmica.



(f) Relación inversa.

Figura 4.2: Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

El criterio que suele utilizarse para obtener la función óptima, es que la distancia de cada punto a la curva, medida en el eje Y, sea lo menor posible. A estas distancias se les llama *residuos* o *errores* en Y (figura 4.3). La función que mejor se ajusta a la nube de puntos será, por tanto, aquella que hace mínima la suma de los cuadrados de los residuos.*

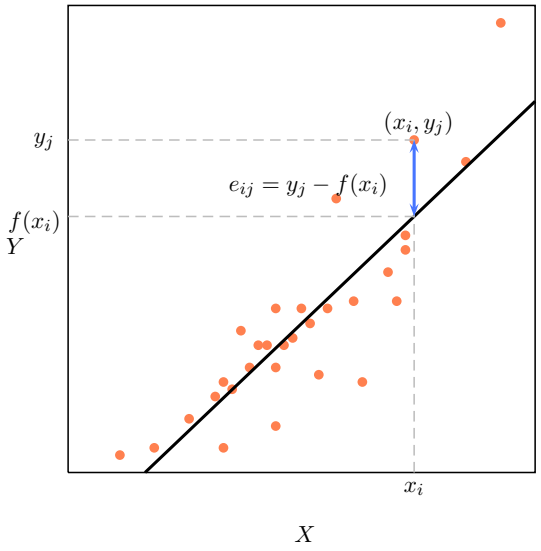


Figura 4.3: Residuos o errores en Y. El residuo correspondiente a un punto (x_i, y_j) es la diferencia entre el valor y_j observado en la muestra, y el valor teórico del modelo $f(x_i)$, es decir, $e_{ij} = y_j - f(x_i)$.

Rectas de regresión

En el caso de que la nube de puntos tenga forma lineal y optemos por explicar la relación entre X e Y mediante una recta $y = a + bx$, los parámetros a determinar son a (punto de corte con el eje de ordenadas) y b (pendiente de la recta). Los valores de estos parámetros que hacen mínima la suma de residuos al cuadrado, determinan la recta óptima. Esta recta se conoce como *recta de regresión de Y sobre X* y explica la variable Y en función de la variable X. Su ecuación es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}),$$

donde s_{xy} es un estadístico llamado *covarianza* que mide el grado de relación lineal, y cuya fórmula es

$$s_{xy} = \frac{1}{n}\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}.$$

Ejemplo 4. *En la figura 4.4 aparecen las rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura del ejemplo anterior.*

La pendiente de la recta de regresión de Y sobre X se conoce como *coeficiente de regresión de Y sobre X*, y mide el incremento que sufrirá la variable Y por cada unidad que se incremente la variable X, según la recta.

Cuanto más pequeños sean los residuos, en valor absoluto, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la recta pasa por todos los puntos de la nube, y la relación es perfecta. En este caso ambas rectas, la de Y sobre X y la de X sobre Y coinciden (figura 4.5(a)).

*Se elevan al cuadrado para evitar que en la suma se compensen los residuos positivos con los negativos.

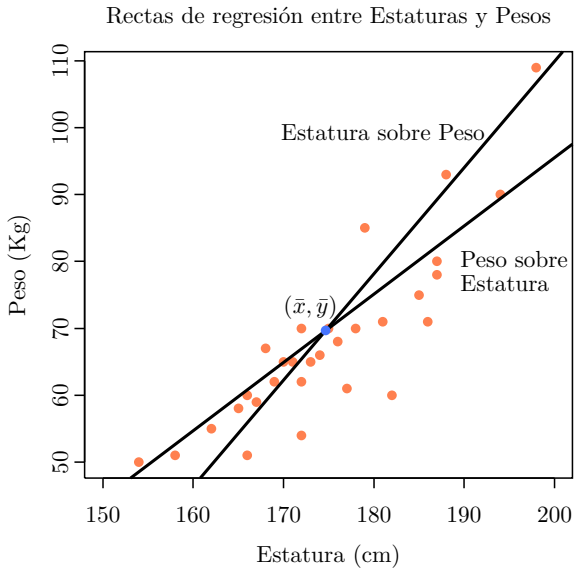


Figura 4.4: Rectas de regresión de Estatura sobre Peso y de Peso sobre Estatura. Las rectas de regresión siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$

Por contra, cuando no existe relación lineal entre las variables, la recta de regresión de Y sobre X tiene pendiente nula, y por tanto la ecuación es $y = \bar{y}$, en la que, efectivamente no aparece x, o $x = \bar{x}$ en el caso de la recta de regresión X sobre Y, de manera que ambas rectas se cortan perpendicularmente (figura 4.5(b)).

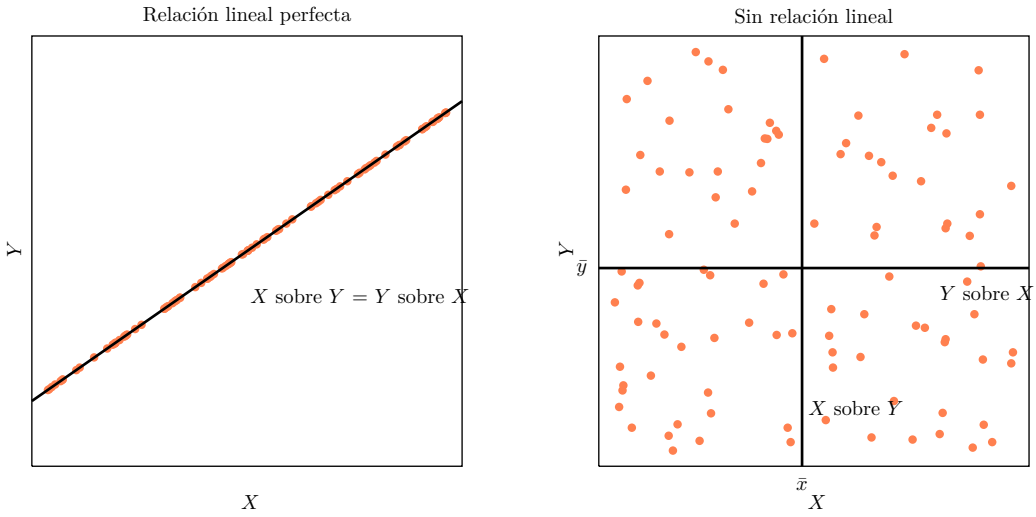
4.1.2 Correlación

El principal objetivo de la regresión simple es construir un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables X (variable independiente) e Y (variable dependiente) medidas en una misma muestra. Generalmente, el modelo construido se utiliza para realizar inferencias predictivas de Y en función de X en el resto de la población. Pero aunque la regresión garantiza que el modelo construido es el mejor posible, dentro del tipo de modelo elegido (lineal, polinómico, exponencial, logarítmico, etc.), puede que aún así, no sea un buen modelo para hacer predicciones, precisamente porque no haya relación de ese tipo entre X e Y. Así pues, con el fin de validar un modelo para realizar predicciones fiables, se necesitan medidas que nos hablen del grado de dependencia entre X e Y, con respecto a un modelo de regresión construido. Estas medidas se conocen como medidas de *correlación*.

Dependiendo del tipo de modelo ajustado, habrá distintos tipos de medidas de correlación. Así, si el modelo de regresión construido es una recta, hablaremos de correlación lineal; si es un polinomio, hablaremos de correlación polinómica; si es una función exponencial, hablaremos de correlación exponencial, etc. En cualquier caso, estas medidas nos hablarán de lo bueno que es el modelo construido, y como consecuencia, de si podemos fiarnos de las predicciones realizadas con dicho modelo.

La mayoría de las medidas de correlación surgen del estudio de los residuos o errores en Y, que son las distancias de los puntos del diagrama de dispersión a la curva de regresión construida, medidas en el eje Y, tal y como se muestra en la figura (4.3). Estas distancias, son en realidad, los errores predictivos del modelo sobre los propios valores de la muestra.

Cuanto más pequeños sean los residuos, mejor se ajustará el modelo a la nube de puntos, y por tanto, mejor explicará la relación entre X e Y. Cuando todos los residuos son nulos, la curva de regresión pasa por todos los puntos de la nube, y entonces se dice que la relación es perfecta, o bien que existe una dependencia funcional entre X e Y (figura 4.5(a)). Por contra, cuando los residuos sean grandes, el modelo no explicará bien la relación entre X e Y, y por tanto, sus predicciones no serán fiables (figura 4.5(b)).



(a) Dependencia funcional lineal.

(b) Independencia lineal.

Figura 4.5: Distintos grados de dependencia. En el primer caso, la relación es perfecta y los residuos son nulos. En el segundo caso no existe relación lineal y la pendiente de la recta es nula.

Varianza residual

Una primera medida de correlación, construida a partir de los residuos es la *varianza residual*, que se define como el promedio de los residuos al cuadrado:

$$s_{ry}^2 = \frac{\sum_{i,j} e_{ij}^2 n_{ij}}{n} = \frac{\sum_{i,j} (y_j - f(x_i))^2 n_{ij}}{n}.$$

Cuando los residuos son nulos, entonces $s_{ry}^2 = 0$ y eso indica que hay dependencia funcional. Por otro lado, cuando las variables son independientes, con respecto al modelo de regresión ajustado, entonces los residuos se convierten en las desviaciones de los valores de Y con respecto a su media, y se cumple que $s_{ry}^2 = s_y^2$. Así pues, se cumple que

$$0 \leq s_{ry}^2 \leq s_y^2.$$

Según esto, cuanto menor sea la varianza residual, mayor será la dependencia entre X e Y, de acuerdo al modelo ajustado. No obstante, la varianza tiene como unidades las unidades de Y al cuadrado, y eso dificulta su interpretación.

Coeficiente de determinación

Puesto que el valor máximo que puede tomar la varianza residual es la varianza de Y, se puede definir fácilmente un coeficiente a partir de la comparación de ambas medidas. Surge así el *coeficiente de determinación* que se define como

$$R^2 = 1 - \frac{s_{ry}^2}{s_y^2}.$$

Se cumple que

$$0 \leq R^2 \leq 1,$$

y además no tiene unidades, por lo que es más fácil de interpretar que la varianza residual:

- $R^2 = 0$ indica que existe independencia según el tipo de relación planteada por el modelo de regresión.
- $R^2 = 1$ indica dependencia funcional.

Por tanto, cuanto mayor sea R^2, mejor será el modelo de regresión.

Si multiplicamos el coeficiente de determinación por 100, se obtiene el porcentaje de variabilidad de Y que explica el modelo de regresión. El porcentaje restante corresponde a la variabilidad que queda por explicar y se corresponde con el error predictivo del modelo. Así, por ejemplo, si tenemos un coeficiente de determinación $R^2 = 0.5$, el modelo de regresión explicaría la mitad de la variabilidad de Y, y en consecuencia, si se utiliza dicho modelo para hacer predicciones, estas tendrían la mitad de error que si no se utilizase, y se tomase como valor de la predicción el valor de la media de Y.

Coeficiente de determinación lineal

En el caso de que el modelo de regresión sea lineal, la fórmula del coeficiente de determinación se simplifica y se convierte en

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2},$$

que se conoce como *coeficiente de determinación lineal*.

Coeficiente de correlación

Otra medida de dependencia bastante habitual es el *coeficiente de correlación*, que se define como la raíz cuadrada del coeficiente de determinación:

$$R = \pm\sqrt{1 - \frac{s_{ry}^2}{s_y^2}},$$

tomando la raíz del mismo signo que la covarianza.

La única ventaja del coeficiente de correlación con respecto al coeficiente de determinación, es que tiene signo, y por tanto, además del grado de dependencia entre X e Y, también nos habla de si la relación es directa (signo +) o inversa (signo -). Su interpretación es:

- $R = 0$ indica independencia con respecto al tipo de relación planteada por el modelo de regresión.
- $R = -1$ indica dependencia funcional inversa.
- $R = 1$ indica dependencia funcional directa.

Por consiguiente, cuanto más próximo esté a -1 o a 1, mejor será el modelo de regresión.

Coeficiente de correlación lineal Al igual que ocurría con el coeficiente de determinación, cuando el modelo de regresión es lineal, la fórmula del coeficiente de correlación se convierte en

$$r = \frac{s_{xy}}{s_x s_y},$$

y se llama *coeficiente de correlación lineal*.

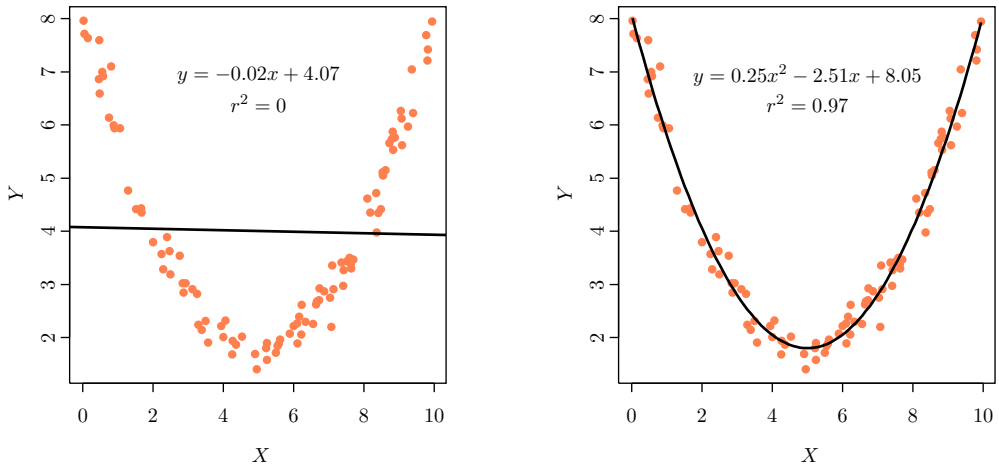
Por último, conviene remarcar que un coeficiente de determinación o de correlación nulo, indica que hay independencia según el modelo de regresión construido, pero puede haber dependencia de otro tipo. Esto se ve claramente en el ejemplo de la figura 4.6.

Fiabilidad de las predicciones

Aunque el coeficiente de determinación o de correlación nos hablan de la bondad de un modelo de regresión, no es el único dato que hay que tener en cuenta a la hora de hacer predicciones.

La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuando mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variablidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones del modelo.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.



(a) Dependencia lineal débil.

(b) Dependencia parabólica fuerte.

Figura 4.6: En la figura de la izquierda se ha ajustado un modelo lineal y se ha obtenido un $R^2 = 0$, lo que indica que el modelo no explica nada de la relación entre X e Y, pero no podemos afirmar que X e Y son independientes. De hecho, en la figura de la derecha se observa que al ajustar un modelo parabólico, $R^2 = 0.97$, lo que indica que casi hay una dependencia funcional parabólica entre X e Y.

Además, hay que tener en cuenta que un modelo de regresión es válido para el rango de valores observados en la muestra, pero fuera de ese rango no tenemos información del tipo de relación entre las variables, por lo que no deberíamos hacer predicciones para valores que estén lejos de los observados en la muestra.

4.2 Ejercicios resueltos

1. Se han medido dos variables A y B en 10 individuos obteniendo los siguientes resultados:

A	0	1	2	3	4	5	6	7	8	9
B	2	5	8	11	14	17	20	23	26	29

Se pide:

(a) Crear las variables A y B e introducir estos datos.

(b) Dibujar el diagrama de dispersión correspondiente.

i

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos...*, elegir la opción *Dispersión simple* y hacer click sobre el botón *Definir*.
ii. Seleccionar la variable B en el campo *Eje Y* del cuadro de diálogo.
iii. Seleccionar la variable A en el campo *Eje X* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.

En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre B y A?

(c) Calcular la recta de regresión de B sobre A.

i

i. Seleccionar el menú *Analizar→Regresión→Lineales...*.
ii. Seleccionar la variable B en el campo *Dependientes* del cuadro de diálogo.
iii. Seleccionar la variable A en el campo *Independientes* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
iv. Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada Coeficientes, y en la columna B de los Coeficientes no estandarizados, encontramos en la primera fila la constante de la recta y en la segunda la pendiente.

(d) Dibujar dicha recta sobre el diagrama de dispersión.

i

i. Editar el gráfico realizado anteriormente haciendo un doble click sobre él.
ii. Seleccionar los puntos haciendo click sobre alguno de ellos.
iii. Seleccionar el menú *Elementos→Linea de ajuste total* (También se podría usar en lugar del menu, la barra de herramientas)
iv. Cerrar la ventana Propiedades.
v. Cerrar el editor de gráficos, cerrando la ventana.

(e) Calcular la recta de regresión de A sobre B y dibujarla sobre el correspondiente diagrama de dispersión.

i

Repetir los pasos de los apartados anteriores pero escogiendo como variable *Dependiente* la variable A, y como variable *Independiente* la variable B.

(f) ¿Son grandes los residuos? Comentar los resultados.

2. En una licenciatura se quiere estudiar la relación entre el número medio de horas de estudio

diarias y el número de asignaturas suspensas. Para ello se obtuvo la siguiente muestra:

Horas	Suspensos	Horas	Suspensos	Horas	Suspensos
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

Se pide:

(a) Crear las variables horas y suspensos e introducir estos datos.

(b) Calcular la recta de regresión de suspensos sobre horas y dibujarla.

i

i. Seleccionar el menú *Analizar→Regresion→Lineales...*.
ii. Seleccionar la variable suspensos en el campo *Dependientes* del cuadro de diálogo.
iii. Seleccionar la variable *horas* en el campo *Independientes* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
iv. Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada Coeficientes, y en la columna B de los Coeficientes no estandarizados, encontramos en la primera fila la constante de la recta y en la segunda la pendiente.
v. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos...*, elegir la opción *Dispersión simple* y hacer click sobre el botón *Definir*.
vi. Seleccionar la variable suspensos en el campo *Eje Y* del cuadro de diálogo.
vii. Seleccionar la variable horas en el campo *Eje X* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
viii. Editar el gráfico realizado haciendo un doble click sobre él.
ix. Seleccionar los puntos haciendo click sobre alguno de ellos.
x. Seleccionar el menú *Elementos→Linea de ajuste total* (También se podría usar en lugar del menu, la barra de herramientas)
xi. Cerrar la ventana Propiedades.
xii. Cerrar el editor de gráficos, cerrando la ventana.

(c) Indicar el coeficiente de regresión de suspensos sobre horas. ¿Cómo lo interpretarías?

i

El coeficiente de regresión es la pendiente de la recta de regresión, que este caso vale -1.23 e indica que por cada hora de estudio adicional se obtienen 1.23 suspensos menos.

(d) La relación lineal entre estas dos variables, ¿es mejor o peor que la del ejercicio anterior? Comentar los resultados a partir las gráficas de las rectas de regresión y sus residuos.

i

La relación lineal entre estas dos variables es peor que la del ejercicio anterior, pues en este caso hay residuos.

(e) Calcular los coeficientes de correlación y de determinación lineal. ¿Es un buen modelo

la recta de regresión? ¿Qué porcentaje de la variabilidad del número de suspensos está explicada por el modelo?

i Observaremos en la ventana de resultados obtenida la tabla denominada Resumen del modelo, y en ella encontramos los valores del coeficiente de correlación lineal R y del coeficiente de determinación lineal R cuadrado.

(f) Utilizar la recta de regresión para predecir el número de suspensos correspondiente a 3 horas de estudio diarias. ¿Es fiable esta predicción?

i
i. Crear una nueva variable valores e introducir los valores de las horas de estudio para los que queremos predecir.
ii. Seleccionar el menú *Transformar→Calcular variable...*
iii. Introducir el nombre de la nueva variable prediccion en el campo *Variable de destino* del cuadro de diálogo.
iv. Introducir la ecuación de la recta en el campo *Expresión numérica*, utilizando los coeficientes calculados anteriormente y la variable valores y hacer click sobre el botón *Aceptar*.

(g) Según el modelo lineal, ¿cuántas horas diarias tendrá que estudiar como mínimo un alumno si quiere aprobarlo todo?.

i Seguir los mismos pasos de los apartados anteriores, pero escogiendo como variable dependiente horas, y como independiente suspensos.

3. Después de tomar un litro de vino se ha medido la concentración de alcohol en la sangre en distintos instantes, obteniendo:

Tiempo después (minutos)	30	60	90	120	150	180	210
Concentración (gramos/litro)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

Se pide:

(a) Crear las variables tiempo y alcohol e introducir estos datos.

(b) Calcular el coeficiente de correlación lineal e interpretarlo.

i
i. Seleccionar el menú *Analizar→Correlaciones→Bivariadas....*
ii. Seleccionar ambas variables en el campo *Variables* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.

(c) Dibujar el diagrama de dispersión junto con la recta ajustada correspondiente a alcohol sobre tiempo. ¿Existe algún individuo con un residuo demasiado grande? Si es así, eliminar dicho individuo de la muestra y volver a calcular el coeficiente de correlación. ¿Ha mejorado el modelo?

i
i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos...*, elegir la opción *Dispersión simple* y hacer click sobre el botón *Definir*.
ii. Seleccionar la variable alcohol en el campo *Eje Y* del cuadro de diálogo.
iii. Seleccionar la variable tiempo en el campo *Eje X* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
iv. Editar el gráfico realizado anteriormente haciendo un doble click sobre él.
v. Seleccionar los puntos haciendo click sobre alguno de ellos.
vi. Seleccionar el menú *Elementos→Linea de ajuste total* (También se podría usar en lugar del menu, la barra de herramientas)

vii. Cerrar la ventana Propiedades.
viii. Cerrar el editor de gráficos, cerrando la ventana.
ix. Si existe algún individuo con un residuo demasiado grande, ir a la ventana del *Editor de datos*, y eliminarlo.
x. Repetir los pasos del apartado anterior.

(d) Si la concentración máxima de alcohol en la sangre que permite la ley para poder conducir es 0.5 g/l, ¿cuánto tiempo habrá que esperar después de tomarse un litro de vino para poder conducir sin infringir la ley? ¿Es fiable esta predicción?

i

i. Seleccionar el menú *Analizar→Regresión→Lineales...*.
ii. Seleccionar la variable tiempo en el campo *Dependientes* del cuadro de diálogo.
iii. Seleccionar la variable alcohol en el campo *Independientes* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
iv. Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada Coeficientes, y en la columna B de los Coeficientes no estandarizados, encontramos en la primera fila la constante de la recta y en la segunda la pendiente.
v. Crear una nueva variable valores e introducir los valores que queremos estudiar.
vi. Seleccionar el menú *Transformar→Calcular variable...*.
vii. Introducir el nombre de la nueva variable prediccion en el campo *Variable de destino* del cuadro de diálogo.
viii. Introducir la ecuación de la recta en el campo *Expresión numérica*, utilizando los coeficientes citados anteriormente y la variable valores y hacer click sobre el botón *Aceptar*.

4.3 Ejercicios propuestos

1. Se determina la pérdida de actividad que experimenta un medicamento desde el momento de su fabricación a lo largo del tiempo, obteniéndose el siguiente resultado:

Tiempo (en años)	1	2	3	4	5
Actividad restante (%)	96	84	70	58	52

Se desea calcular:

(a) La relación fundamental (recta de regresión) entre actividad restante y tiempo transcurrido.

(b) ¿En qué porcentaje disminuye la actividad cada año que pasa?

(c) ¿Cuándo tiempo debe pasar para que el fármaco tenga una actividad del 80%? ¿Cuándo será nula la actividad? ¿Son igualmente fiables estas predicciones?

2. Al realizar un estudio sobre la dosificación de un cierto medicamento, se trataron 6 pacientes con dosis diarias de 2 mg, 7 pacientes con 3 mg y otros 7 pacientes con 4 mg. De los pacientes tratados con 2 mg, 2 curaron al cabo de 5 días, y 4 al cabo de 6 días. De los pacientes tratados con 3 mg diarios, 2 curaron al cabo de 3 días, 4 al cabo de 5 días y 1 al cabo de 6 días. Y de los pacientes tratados con 4 mg diarios, 5 curaron al cabo de 3 días y 2 al cabo de 5 días. Se pide:

(a) Calcular la recta de regresión del tiempo de curación con respecto a la dosis suministrada.

(b) Calcular los coeficientes de regresión. Interpretar los resultados.

(c) Determinar el tiempo esperado de curación para una dosis de 5 mg diarios. ¿Es fiable esta predicción?

(d) ¿Qué dosis debe aplicarse si queremos que el paciente tarde 4 días en curarse? ¿Es fiable la predicción?

5 — Regresión No Lineal

5.1 Fundamentos teóricos

La regresión simple tiene por objeto la construcción de un modelo funcional $y = f(x)$ que explique lo mejor posible la relación entre dos variables Y (variable dependiente) y X (variable independiente) medidas en una misma muestra.

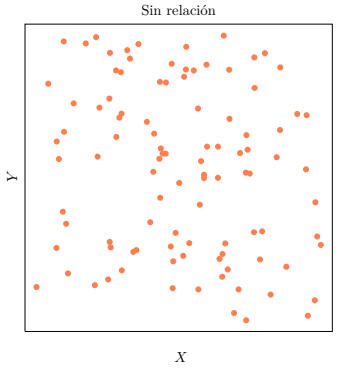
Ya vimos que, dependiendo de la forma de esta función, existen muchos tipos de regresión simple. Entre los más habituales están:

Familia de curvas	Ecuación genérica
Lineal	$y = b_0 + b_1 x$
Cuadrática	$y = b_0 + b_1 x + b_2 x^2$
Cúbica	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Potencia	$y = b_0 \cdot x^{b_1}$
Exponencial	$y = b_0 \cdot e^{b_1 x}$
Logarítmica	$y = b_0 + b_1 \ln x$
Inversa	$y = b_0 + \frac{b_1}{x}$
Compuesto	$y = b_0 b_1^x$
Crecimiento	$y = e^{b_0 + b_1 x}$
G (Curva-S)	$y = e^{b_0 + \frac{b_1}{x}}$

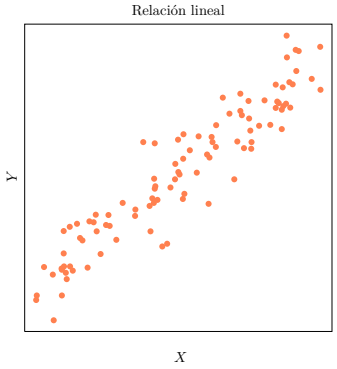
La elección de un tipo de modelo u otro suele hacerse según la forma de la nube de puntos del diagrama de dispersión. A veces estará claro qué tipo de modelo se debe construir, tal y como ocurre en los diagramas de dispersión de la figura 5.1. Pero otras veces no estará tan claro, y en estas ocasiones, lo normal es ajustar los dos o tres modelos que nos parezcan más convincentes, para luego quedarnos con el que mejor explique la relación entre Y y X, mirando el coeficiente de determinación* de cada modelo.

Ya vimos en la práctica sobre regresión lineal simple, cómo construir rectas de regresión. En el caso de que optemos por ajustar un modelo no lineal, la construcción del mismo puede realizarse siguiendo los mismos pasos que en el caso lineal. Básicamente se trata de determinar los parámetros del modelo que minimizan la suma de los cuadrados de los residuos en Y. En los modelos potencia y exponencial, el sistema aplica transformaciones logarítmicas a las variables y después ajusta un modelo lineal a los datos transformados. En el modelo inverso, el sistema sustituye la variable dependiente por su inverso antes de estimar la ecuación de regresión.

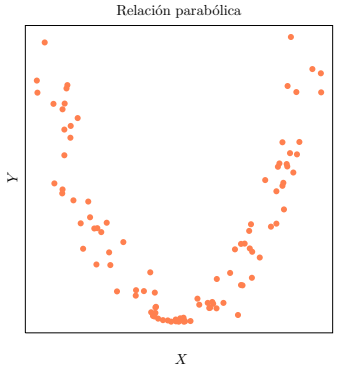
*Ver la práctica de correlación.



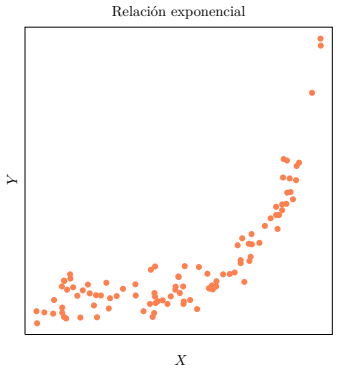
(a) Sin relación.



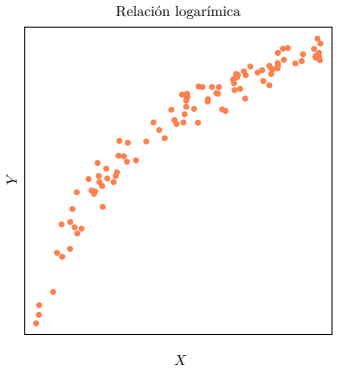
(b) Relación lineal.



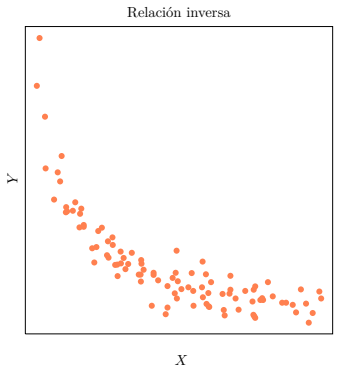
(c) Relación polinómica.



(d) Relación exponencial.



(e) Relación logarítmica.



(f) Relación inversa.

Figura 5.1: Diagramas de dispersión correspondientes a distintos tipos de relaciones entre variables.

5.2 Ejercicios resueltos

1. En un experimento se ha medido el número de bacterias por unidad de volumen en un cultivo, cada hora transcurrida, obteniendo los siguientes resultados:

Horas	0	1	2	3	4	5	6	7	8
Nº Bacterias	25	32	47	65	92	132	190	275	362

Se pide:

(a) Crear las variables horas y bacterias e introducir estos datos.

(b) Dibujar el diagrama de dispersión correspondiente. En vista del diagrama, ¿qué tipo de modelo crees que explicará mejor la relación entre el número de bacterias y el tiempo transcurrido?

i

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos...*, elegir la opción *Dispersión simple* y hacer click sobre el botón *Definir*.
ii. Seleccionar la variable bacterias en el campo *Eje Y* del cuadro de diálogo.
iii. Seleccionar la variable horas en el campo *Eje X* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.

(c) Hacer una comparativa de los distintos modelos de regresión en función del coeficiente de determinación. ¿Qué tipo de modelo es el mejor?

i

i. Seleccionar el menú *Analizar→Regresión→Estimación curvilínea...*.
ii. Seleccionar la variable bacterias en el campo *Dependientes* del cuadro de diálogo.
iii. Seleccionar la variable horas en el campo *Independiente/variable* del cuadro de diálogo.
iv. Desmarcar la opción *Representar los modelos*.
v. Marcar las opciones lineal, cuadrático, cúbico, exponencial y logarítmico, y hacer click sobre el botón *Aceptar*.

(d) En vista de lo anterior, calcular el modelo de regresión que mejor explique la relación entre bacterias y horas.

i

Utilizar los coeficientes que aparecen en el punto anterior y la tabla de la parte de fundamentos teóricos.

(e) Según el modelo anterior, ¿cuántas bacterias habrá al cabo de 3 horas y media del inicio del cultivo? ¿Y al cabo de 10 horas? ¿Son fiables estas predicciones?

i

i. Crear una nueva variable valores e introducir los valores de las horas para los que queremos predecir las bacterias.
ii. Seleccionar el menú *Transformar→Calcular variable...*.
iii. Introducir el nombre de la nueva variable prediccion en el campo *Variable de destino* del cuadro de diálogo.
iv. Introducir la ecuación del mejor modelo en el campo *Expresión numérica*, utilizando los coeficientes obtenidos anteriormente y la variable valores y hacer click sobre el botón *Aceptar*.

(f) Dar una predicción lo más fiable posible del tiempo que tendría que transcurrir para que en el cultivo hubiese 100 bacterias.

i Repetir los pasos del apartado anterior introduciendo la variable horas en el campo *Dependientes* y la variable bacterias en el campo *Independiente/Variables*.

2. Se han medido dos variables S y T en 10 individuos, obteniéndose los siguientes resultados:

$$(-1.5\,,\,2.25)\,,\ (0.8\,,\,0.64)\,,\ (-0.2\,,\,0.04)\,,\ (-0.8\,,\,0.64)\,,\ (0.4\,,\,0.16)\,,$$
$$(0.2\,,\,0.04)\,,\ (-2.1\,,\,4.41)\,,\ (-0.4\,,\,0.16)\,,\ (1.5\,,\,2.25)\,,\ (2.1\,,\,4.41).$$

Se pide:

(a) Crear las variables S y T e introducir estos datos.

(b) Calcular la recta de regresión de T sobre S. Dibujar dicha recta sobre el diagrama de dispersión. ¿Podemos afirmar que S y T son independientes?

i
i. Seleccionar el menú *Analizar→Regresión→Lineales...*.
ii. Seleccionar la variable T en el campo Dependientes del cuadro de diálogo.
iii. Seleccionar la variable S en el campo *Independientes* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
iv. Para escribir la ecuación de la recta, observaremos en la ventana de resultados obtenida, la tabla denominada Coeficientes, y en la columna B de los Coeficientes no estandarizados, encontramos en la primera fila la constante de la recta y en la segunda la pendiente.
v. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos...*, elegir la opción *Dispersión simple* y hacer click sobre el botón *Definir*.
vi. Seleccionar la variable T en el campo *Eje Y* del cuadro de diálogo.
vii. Seleccionar la variable S en el campo *Eje X* del cuadro de diálogo y hacer click sobre el botón *Aceptar*.
viii. Editar el gráfico realizado anteriormente haciendo un doble click sobre él.
ix. Seleccionar los puntos haciendo click sobre alguno de ellos.
x. Seleccionar el menú *Elementos→Linea de ajuste total* (También se podría usar en lugar del menu, la barra de herramientas)
xi. Cerrar la ventana Propiedades.
xii. Cerrar el editor de gráficos, cerrando la ventana.

(c) Hacer una comparativa de los distintos modelos de regresión en función del coeficiente de determinación. ¿Qué tipo de relación existe entre T y S?

i
i. Seleccionar el menú *Analizar→Regresión→Estimación curvilínea...*.
ii. Seleccionar la variable T en el campo *Dependientes* del cuadro de diálogo.
iii. Seleccionar la variable S en el campo Independiente/variable del cuadro de diálogo.
iv. Desmarcar la opción *Representar los modelos*.
v. Marcar las opciones lineal, cuadrático, cúbico y exponencial y hacer click sobre el botón *Aceptar*.

(d) En vista de lo anterior, ajustar el modelo de regresión más apropiado.

i Utilizar los coeficientes que aparecen en el punto anterior y la tabla de la parte de fundamentos teóricos.

5.3 Ejercicios propuestos

1. En un centro dietético se está probando una nueva dieta de adelgazamiento en una muestra de 12 individuos. Para cada uno de ellos se ha medido el número de días que lleva con la

dieta y el número de kilos perdidos desde entonces, obteniéndose los siguientes resultados:

(33 , 3.9), (51 , 5.9), (30 , 3.2), (55 , 6.0), (38 , 4.9), (62 , 6.2),
(35 , 4.5), (60 , 6.1), (44 , 5.6), (69 , 6.2), (47 , 5.8), (40 , 5.3)

Se pide:

(a) Dibujar el diagrama de dispersión. Según la nube de puntos, ¿qué tipo de modelo explicaría mejor la relación entre los kilos perdidos y los días de dieta?

(b) Construir el modelo de regresión que mejor explique la relación entre los kilos perdidos y los días de dieta.

(c) Utilizar el modelo construido para predecir el número de kilos perdidos tras 40 días de dieta y tras 100 días. ¿Son fiables estas predicciones?

2. La concentración de un fármaco en sangre, C en mg/dl, es función del tiempo, t en horas, y viene dada por la siguiente tabla:

t	2	3	4	5	6	7	8
C	25	36	48	64	86	114	168

Se pide:

(a) Según el modelo exponencial, ¿qué concentración de fármaco habría a las 4.8 horas? ¿Es fiable la predicción? Justificar adecuadamente la respuesta.

(b) Según el modelo logarítmico, ¿qué tiempo debe pasar para que la concentración sea de 100 mg/dl?

6 — Intervalos de Confianza para Medias y Proporciones

6.1 Fundamentos teóricos

6.1.1 Inferencia estadística y estimación de parámetros

El objetivo de un estudio estadístico es doble: describir la muestra elegida de una población en la que se quiere estudiar alguna característica, y realizar inferencias, es decir, sacar conclusiones sobre la población de la que se ha extraído dicha muestra.

La metodología que conduce a obtener conclusiones sobre la población, basadas en la información contenida en la muestra, constituye la *Inferencia Estadística*.

Puesto que la muestra contiene menos información que la población, las conclusiones sobre la población serán aproximadas. Por eso, uno de los objetivos de la inferencia estadística es determinar la probabilidad de que una conclusión obtenida a partir del análisis de una muestra sea cierta, y para ello se apoya en la teoría de la probabilidad.

Cuando se desea conocer el valor de alguno de los parámetros de la población, el procedimiento a utilizar es la *Estimación de Parámetros,* que a su vez se divide en *Estimación Puntual*, cuando se da un único valor como estimación del parámetro poblacional considerado, y *Estimación por Intervalos*, cuando interesa conocer no sólo un valor aproximado del parámetro sino también la precisión de la estimación. En este último caso el resultado es un intervalo, dentro del cual estará, con una cierta confianza, el verdadero valor del parámetro poblacional. A este intervalo se le denomina *intervalo de confianza*. A diferencia de la estimación puntual, en la que se utiliza un único estimador, en la estimación por intervalo se emplean dos estimadores, uno para cada extremo del intervalo.

6.1.2 Intervalos de confianza

Dados dos estadísticos muestrales L_1 y L_2, se dice que el intervalo $I = (L_1,\ L_2)$ es un *Intervalo de Confianza* para un parámetro poblacional θ, con *nivel de confianza* $1-\alpha$ (o *nivel de significación* α), si la probabilidad de que los estadísticos que determinan los límites del intervalo tomen valores tales que θ esté comprendido entre ellos, es igual a $1-\alpha$, es decir,

$$P\left(L_1 < \theta < L_2\right) = 1 - \alpha$$

Los extremos del intervalo son variables aleatorias cuyos valores dependen de la muestra considerada. Es decir, los extremos inferior y superior del intervalo serían $L_1\left(X_1, ..., X_n\right)$ y $L_2\left(X_1, ..., X_n\right)$ respectivamente, aunque habitualmente escribiremos L_1 y L_2 para simplificar la notación. Designaremos mediante l_1 y l_2 los valores que toman dichas variables para una muestra determinada $\left(x_1, ..., x_n\right)$.

Cuando en la definición se dice que la probabilidad de que el parámetro θ esté en el intervalo $(L_1,\ L_2)$ es $1-\alpha$, quiere decir que en el $100\,(1-\alpha)$ % de las posibles muestras, el valor de θ estaría en los correspondientes intervalos $(l_1,\ l_2)$.

Una vez que se tiene una muestra, y a partir de ella se determina el intervalo correspondiente $(l_1,\ l_2)$, no tendría sentido hablar de la probabilidad de que el parámetro θ esté en el intervalo $(l_1,\ l_2)$, pues al ser l_1 y l_2 números, el parámetro θ, que también es un número, aunque desconocido, estará o no estará en dicho intervalo, y por ello hablamos de confianza en lugar de probabilidad.

Así, cuando hablemos de un intervalo de confianza para el parámetro θ con nivel de confianza $1-\alpha$, entenderemos que antes de tomar una muestra, hay una probabilidad $1-\alpha$ de que el intervalo que se construya a partir de ella, contenga el valor del parámetro θ. O, dicho de otro modo, si tomásemos todas las posibles muestras del mismo tamaño y calculásemos sus respectivos intervalos, el $100(1-\alpha)\%$ de estos contendrían el verdadero valor del parámetro a estimar (ver figura 6.1).

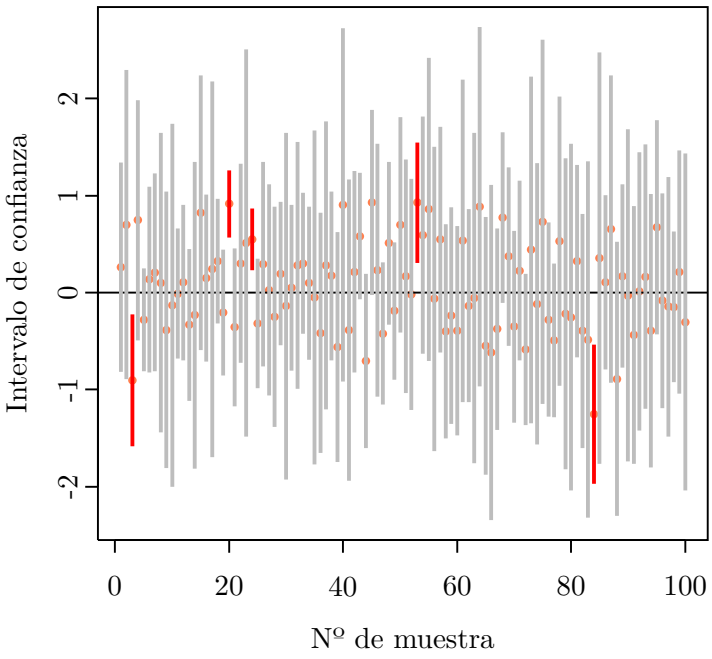


Figura 6.1: Intervalos de confianza del 95% para la media de 100 muestras tomadas de una población normal $N(0,1)$. Como se puede apreciar, de los 100 intervalos, sólo 5 no contienen el valor de la media real $\mu = 0$.

Cuando se realiza la estimación de un parámetro mediante un intervalo de confianza, el nivel de confianza se suele fijar a niveles altos (los más habituales son 0.90, 0.95 ó 0.99), para tener una alta confianza de que el parámetro está dentro del intervalo. Por otro lado, también interesa que la amplitud del intervalo sea pequeña para delimitar con precisión el valor del parámetro poblacional (esta amplitud del intervalo se conoce como *imprecisión* de la estimación). Pero a partir de una muestra, cuanto mayor sea el nivel de confianza deseado, mayor amplitud tendrá el intervalo y mayor imprecisión la estimación, y si se impone que la estimación sea más precisa (menor imprecisión), el nivel de confianza correspondiente será más pequeño. Por consiguiente, hay que llegar a una solución de compromiso entre el nivel de confianza y la precisión de la estimación. No obstante, si con la muestra disponible no es posible obtener un intervalo de amplitud suficientemente pequeña (imprecisión pequeña) con un nivel de confianza aceptable, hay que emplear una muestra de mayor tamaño. Al aumentar el tamaño muestral se consiguen intervalos de menor amplitud sin disminuir el nivel de confianza, o niveles de confianza más altos

manteniendo la amplitud.

Intervalos de confianza para la media

Apoyándose en conclusiones extraídas del Teorema Central del Límite se obtiene que, siempre que las muestras sean grandes (como criterio habitual se considera que son grandes cuando el tamaño muestral, n, sea mayor o igual que 30), e independientemente de la distribución original de la variable de partida X, de media μ y desviación típica σ, la variable

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

sigue una distribución Normal tipificada, $N(0,1)$.

Si la desviación típica σ de la variable de partida es desconocida, se utiliza como estimación la cuasidesviación típica muestral:

$$\hat{S} = \sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}}$$

y con ello, la nueva variable

$$T = \frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}$$

sigue una distribución t de Student con $n-1$ grados de libertad, $T(n-1)$.

Para muestras pequeñas ($n < 30$) también pueden aplicarse los resultados anteriores, siempre y cuando la variable aleatoria de partida X, siga una distribución Normal.

A partir de lo anterior y teniendo en cuenta los tres factores de clasificación expuestos: si la población de partida en la que obtenemos la muestra sigue o no una distribución Normal, si la varianza de dicha población es conocida o desconocida, y si la muestra es grande ($n \geq 30$) o no, pueden deducirse las siguientes expresiones correspondientes a los diferentes intervalos de confianza.

Intervalo de confianza para la media de una población normal con varianza conocida en muestras de cualquier tamaño

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

En la figura 6.2 aparece un esquema explicativo de la construcción de este intervalo.

Intervalo de confianza para la media de una población normal con varianza desconocida en muestras de cualquier tamaño

$$\left(\overline{x} - t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Si las muestras son grandes ($n \geq 30$) el anterior intervalo puede aproximarse mediante:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Intervalo de confianza para la media de una población no normal, varianza conocida y muestras grandes ($n \geq 30$)

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

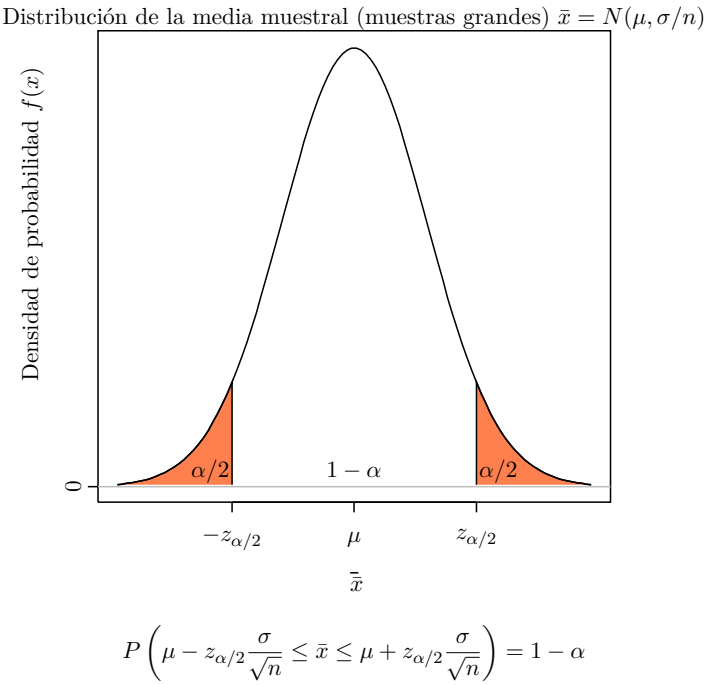


Figura 6.2: Cálculo del intervalo de confianza para la media de una población normal con varianza conocida, a partir de las distribución de la media muestral $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$ para muestras de cualquier tamaño.

Intervalo de confianza para la media de una población no normal, varianza desconocida y muestras grandes ($n \geq 30$)

$$\left(\overline{x} - t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + t_{\alpha/2}^{n-1} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Al tratarse de muestras grandes, el anterior intervalo puede aproximarse por:

$$\left(\overline{x} - z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}},\ \overline{x} + z_{\alpha/2} \cdot \frac{\hat{s}}{\sqrt{n}}\right)$$

Si la población de partida no es normal, y las muestras son pequeñas, no puede aplicarse el Teorema Central del Límite y no se obtienen intervalos de confianza para la media.

Para cualquiera de los anteriores intervalos:

- n es el tamaño de la muestra.
- $\bar{x}$ es la media muestral.
- σ es la desviación típica de la población.
- $\hat{s}$ es la cuasidesviación típica muestral: $\hat{s}^2 = \dfrac{\sum (x_i - \overline{x})^2}{n-1}$.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.
- $t_{\alpha/2}^{n-1}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con $n-1$ grados de libertad.

Intervalos de confianza para la proporción poblacional p

Para muestras grandes ($n \geq 30$) y valores de p (probabilidad de "éxito") cercanos a 0.5, la distribución Binomial puede aproximarse mediante una Normal de media np y desviación típica $\sqrt{np(1-p)}$. En la práctica, para que sea válida dicha aproximación, se toma el criterio de que tanto np como $n(1-p)$ deben ser mayores que 5. Esto hace que también podamos construir intervalos de confianza para proporciones tomando éstas como medias de variables dicotómicas en las que la presencia o ausencia de la característica objeto de estudio ("éxito" ó "fracaso") se expresan mediante un 1 ó un 0 respectivamente.

De este modo, en muestras grandes y con distribuciones binomiales no excesivamente asimétricas (tanto np como $n(1-p)$ deben ser mayores que 5), si denominamos $\widehat{p}$ a la proporción de individuos que presentan el atributo estudiado en la muestra concreta, entonces el intervalo de confianza para la proporción con un nivel de significación α viene dado por:

$$\left(\widehat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}} \ , \ \widehat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p} \cdot (1-\widehat{p})}{n}} \right)$$

donde:

- n es el tamaño muestral.
- $\widehat{p}$ a la proporción de individuos que presentan el atributo estudiado en la muestra concreta.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.

En muestras pequeñas o procedentes de una Binomial fuertemente asimétrica ($np \leq 5$ ó $n(1-p) \leq 5$) no puede aplicarse el Teorema Central del Límite y la construcción de intervalos de confianza debe realizarse a partir de la distribución Binomial.

6.2 Ejercicios resueltos

1. Se analiza la concentración de principio activo en una muestra de 10 envases tomados de un lote de un fármaco, obteniendo los siguientes resultados en mg/mm^3:

$$17.6 - 19.2 - 21.3 - 15.1 - 17.6 - 18.9 - 16.2 - 18.3 - 19.0 - 16.4$$

Se pide:

(a) Crear la variable concentracion, e introducir los datos de la muestra.

(b) Calcular el intervalo de confianza para la media de la concentración del lote con nivel de confianza del 95% (nivel de significación $\alpha = 0.05$).

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Explorar*.
ii. En el cuadro de diálogo que aparece seleccionar la variable concentracion en el campo *Lista de dependientes* y hacer click en el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece habilitar la opción *Descriptivos*, introducir el nivel de confianza deseado en el cuadro *Intervalo de confianza para la media*, y hacer click en el botón *Continuar* y *Aceptar*.

(c) Calcular los intervalos de confianza para la media con niveles del 90% y del 99% (niveles de significación $\alpha = 0.1$ y $\alpha = 0.01$).

i

Repetir los mismos pasos del apartado anterior, cambiando el nivel de confianza para cada intervalo.

(d) Si definimos la precisión del intervalo como la inversa de su amplitud, ¿cómo afecta a la precisión del intervalo de confianza el tomar niveles de significación cada vez más altos? ¿Cuál puede ser la explicación?

(e) Si, para que sea efectivo, el fármaco debe tener una concentración mínima de 16 mg/mm^3 de principio activo, ¿se puede aceptar el lote como bueno? Justificar la respuesta.

2. Una central de productos lácteos recibe diariamente la leche de dos granjas X e Y. Para analizar la calidad de la leche, durante una temporada, se controla el contenido de materia grasa de la leche que proviene de ambas granjas, con los siguientes resultados:

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

(a) Crear las variables grasa y granja, e introducir los datos de la muestra.

(b) Calcular el intervalo de confianza con un 95% de confianza para el contenido medio de materia grasa de la leche sin tener en cuenta si la misma procede de una u otra granja.

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Explorar*.
ii. En el cuadro de diálogo que aparece seleccionar la variable grasa en el campo *Lista de dependientes* y hacer click en el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece habilitar la opción *Descriptivos*, introducir

el nivel de confianza deseado en el cuadro *Intervalo de confianza para la media*, y hacer click en el botón *Continuar* y *Aceptar*.

(c) Calcular los intervalos de confianza con un 95% de confianza para el contenido medio de materia grasa de la leche dividiendo los datos según la granja de procedencia de la leche.

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Explorar*.
ii. En el cuadro de diálogo que aparece seleccionar la variable grasa en el campo *Lista de dependientes*, seleccionar la variable granja en el campo *Lista de factores* y hacer click en el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece habilitar la opción *Descriptivos*, introducir el nivel de confianza deseado en el cuadro Intervalo de confianza para la media, y hacer click en el botón *Continuar* y *Aceptar*.

(d) A la vista de los intervalos obtenidos en el punto anterior, ¿se puede concluir que existen diferencias significativas en el contenido medio de grasa según la procedencia de la leche? Justificar la respuesta.

3. En una encuesta realizada en una facultad, sobre si los alumnos utilizan habitualmente (al menos una vez a la semana) la biblioteca de la misma, se han obtenido los siguientes resultados, en los que se ha anotado 1 si la respuesta ha sido positiva y 0 si ha sido negativa:

Alumno	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Respuesta	0	1	0	0	0	1	0	1	1	1	1	0	1	0	1	0	0	0

Alumno	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Respuesta	1	1	1	0	0	1	0	0	1	1	0	0	1	0	1	0

(a) Crear la variable respuesta e introducir los datos de la muestra.

(b) Calcular el intervalo de confianza con $\alpha = 0.01$ para la media de la variable respuesta.

i

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Explorar*.
ii. En el cuadro de diálogo que aparece seleccionar la variable respuesta en el campo *Lista de dependientes* y hacer click en el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece habilitar la opción *Descriptivos*, introducir el nivel de confianza deseado en el cuadro *Intervalo de confianza para la media*, y hacer click en el botón *Continuar* y *Aceptar*.

(c) ¿Qué interpretación tiene dicho intervalo en términos de proporción de alumnos que habitualmente utilizan la biblioteca?

4. El Ministerio de Sanidad está interesado en la elaboración de un intervalo de confianza para la proporción de personas mayores de 65 años con problemas respiratorios que han sido vacunadas en una determinada ciudad. Para ello, después de preguntar a 200 pacientes mayores de 65 años con problemas respiratorios en los hospitales de dicha ciudad, 154 responden afirmativamente.

(a) Crear la variable respuesta cuyos dos únicos valores serán 0 para las respuestas negativas, y 1 para las positivas, y una segunda variable que podemos denominar frecuencia cuyos valores son las frecuencias absolutas de cada una de las respuestas (46 para la respuesta 0 y 154 para la respuesta 1).

(b) Ponderar los valores de la variable respuesta mediante los pesos introducidos en la variable frecuencia.

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo que aparece activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia y hacer click en el botón *Aceptar*.

(c) Calcular el intervalo de confianza al 95% para la proporción de pacientes vacunados.

i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Explorar*.
ii. En el cuadro de diálogo que aparece seleccionar la variable respuesta en el campo *Lista de dependientes* y hacer click en el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece habilitar la opción *Descriptivos*, introducir el nivel de confianza deseado en el cuadro *Intervalo de confianza para la media*, y hacer click en el botón *Continuar* y *Aceptar*.

(d) Si entre los objetivos del Ministerio se encontraba alcanzar una proporción del al menos un 70% de vacunados en dicho colectivo, ¿se puede concluir que se han cumplido los objetivos? Justificar la respuesta.

6.3 Ejercicios propuestos

1. Para determinar el nivel medio de colesterol en la sangre de una población, se realizaron análisis sobre una muestra de 8 personas, obteniéndose los siguientes resultados:

$$196 - 212 - 188 - 206 - 203 - 210 - 201 - 198$$

Hallar los intervalos de confianza para la media del nivel de colesterol con niveles de significación 0.1, 0.05 y 0.01.

2. Utilizar el fichero Hipertensos Datos Claves para estimar mediante un intervalo de confianza del 95% la presión diastólica inicial media en todos los individuos. Calcular también los intervalos del 95% de confianza correspondientes a los grupos tratados con placebo, IECA y Ca Antagonista + Diurético. ¿Existen diferencias significativas en la presión diastólica incial media de esos grupos?

3. Para tratar un determinado síndrome neurológico se utilizan dos técnicas A y B. En un estudio se tomó una muestra de 60 pacientes con dicho síndrome y se le aplicó la técnica A a 25 de ellos y la técnica B a los 35 restantes. De los pacientes tratados con la técnica A, 18 se curaron, mientras que de los tratados con la técnica B, se curaron 21. Calcular un intervalo de confianza del 95% para la proporción de curaciones con cada técnica.

4. A las siguientes elecciones locales en una ciudad se presentan tres partidos: A, B y C. Con el objetivo de hacer una estimación sobre la proporción de voto que cada uno de ellos obtendrá, se realiza una encuesta a la que responden 300 personas, de las cuales 60 piensan votar a A, 80 a B, 90 a C, 15 en blanco y 55 abstenciones. Calcular un intervalo de confianza para la proporción de votos, sobre el total del censo, de cada uno de los partidos que se presentan.

7 — Intervalos de Confianza para Comparación de Poblaciones

7.1 Fundamentos teóricos

7.1.1 Inferencia estadística y estimación de parámetros

El objetivo de un estudio estadístico es doble: describir la muestra elegida de una población en la que se quiere estudiar alguna característica, y realizar inferencias, es decir, sacar conclusiones sobre la población de la que se ha extraído dicha muestra.

La metodología que conduce a obtener conclusiones sobre la población, basadas en la información contenida en la muestra, constituye la *Inferencia Estadística*.

Puesto que la muestra contiene menos información que la población, las conclusiones sobre la población serán aproximadas. Por eso, uno de los objetivos de la inferencia estadística es determinar la probabilidad de que una conclusión obtenida a partir del análisis de una muestra sea cierta, y para ello se apoya en la teoría de la probabilidad.

Cuando se desea conocer el valor de alguno de los parámetros de la población, el procedimiento a utilizar es la *Estimación de Parámetros,* que a su vez se divide en *Estimación Puntual*, cuando se da un único valor como estimación del parámetro poblacional considerado, y *Estimación por Intervalos*, cuando interesa conocer no sólo un valor aproximado del parámetro sino también la precisión de la estimación. En este último caso el resultado es un intervalo, dentro del cual estará, con una cierta confianza, el verdadero valor del parámetro poblacional. A este intervalo se le denomina *intervalo de confianza*. A diferencia de la estimación puntual, en la que se utiliza un único estimador, en la estimación por intervalo emplearemos dos estimadores, uno para cada extremo del intervalo.

7.1.2 Intervalos de confianza

Dados dos estadísticos muestrales L_1 y L_2, se dice que el intervalo $I = (L_1,\ L_2)$ es un *Intervalo de Confianza* para un parámetro poblacional θ, con *nivel de confianza* $1-\alpha$ (o *nivel de significación* α), si la probabilidad de que los estadísticos que determinan los límites del intervalo tomen valores tales que θ esté comprendido entre ellos, es igual a $1-\alpha$, es decir,

$$P\left(L_1 < \theta < L_2\right) = 1 - \alpha$$

Los extremos del intervalo son variables aleatorias cuyos valores dependen de la muestra considerada. Es decir, los extremos inferior y superior del intervalo serían $L_1\left(X_1, ..., X_n\right)$ y $L_2\left(X_1, ..., X_n\right)$ respectivamente, aunque habitualmente escribiremos L_1 y L_2 para simplificar la notación. Designaremos mediante l_1 y l_2 los valores que toman dichas variables para una muestra determinada $\left(x_1, ..., x_n\right)$.

Cuando en la definición se dice que la probabilidad de que el parámetro θ esté en el intervalo $(L_1,\ L_2)$ es $1-\alpha$, quiere decir que en el $100\,(1-\alpha)$ % de las posibles muestras, el valor de θ estaría en los correspondientes intervalos $(l_1,\ l_2)$.

Una vez que se tiene una muestra, y a partir de ella se determina el intervalo correspondiente $(l_1,\ l_2)$, no tendría sentido hablar de la probabilidad de que el parámetro θ esté en el intervalo $(l_1,\ l_2)$, pues al ser l_1 y l_2 números, el parámetro θ, que también es un número, aunque desconocido, estará o no estará en dicho intervalo, y por ello hablamos de confianza en lugar de probabilidad.

Así, cuando hablemos de un intervalo de confianza para el parámetro θ con nivel de confianza $1-\alpha$, entenderemos que antes de tomar una muestra, hay una probabilidad $1-\alpha$ de que el intervalo que se construya a partir de ella, contenga el valor del parámetro θ.

Cuando se realiza la estimación de un parámetro mediante un intervalo de confianza, el nivel de confianza se suele fijar a niveles altos (los más habituales son 0.90, 0.95 ó 0.99), para tener una alta confianza de que el parámetro está dentro del intervalo. Por otro lado, también interesa que la amplitud del intervalo sea pequeña para delimitar con precisión el valor del parámetro poblacional (esta amplitud del intervalo se conoce como *imprecisión* de la estimación). Pero a partir de una muestra, cuanto mayor sea el nivel de confianza deseado, mayor amplitud tendrá el intervalo y mayor imprecisión la estimación, y si se impone que la estimación sea más precisa (menor imprecisión), el nivel de confianza correspondiente será más pequeño. Por consiguiente, hay que llegar a una solución de compromiso entre el nivel de confianza y la precisión de la estimación. No obstante, si con la muestra disponible no es posible obtener un intervalo de amplitud suficientemente pequeña (imprecisión pequeña) con un nivel de confianza aceptable, hay que emplear una muestra de mayor tamaño. Al aumentar el tamaño muestral se consiguen intervalos de menor amplitud sin disminuir el nivel de confianza, o niveles de confianza más altos manteniendo la amplitud.

Intervalos de confianza para la diferencia de medias

De igual manera a como ocurría con los intervalos de confianza para la media de una variable, apoyándose en conclusiones extraídas del Teorema Central del Límite se puede demostrar que, en muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$), procedentes de poblaciones de dos variables X_1 y X_2, con distribuciones no necesariamente Normales, de medias μ_1 y μ_2 y desviaciones típicas σ_1 y σ_2 respectivamente, la variable

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

sigue una distribución Normal tipificada, $N(0,\ 1)$.

De igual manera, si las varianzas de las variables son desconocidas, utilizando como estimadores muestrales sus correspondientes cuasivarianzas $\hat{S}_1^2$ y $\hat{S}_2^2$, donde

$$\hat{S}_1^2 = \frac{\sum \left(X_{1,i} - \overline{X}_1\right)^2}{n_1 - 1} \quad \text{y} \quad \hat{S}_2^2 = \frac{\sum \left(X_{2,i} - \overline{X}_2\right)^2}{n_2 - 1}$$

entonces la variable

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\hat{S}_1^2}{n_1} + \dfrac{\hat{S}_2^2}{n_2}}}$$

sigue una distribución t de Student, en la que el número de grados de libertad dependerá de si las varianzas, aún siendo desconocidas, pueden considerarse iguales o no.

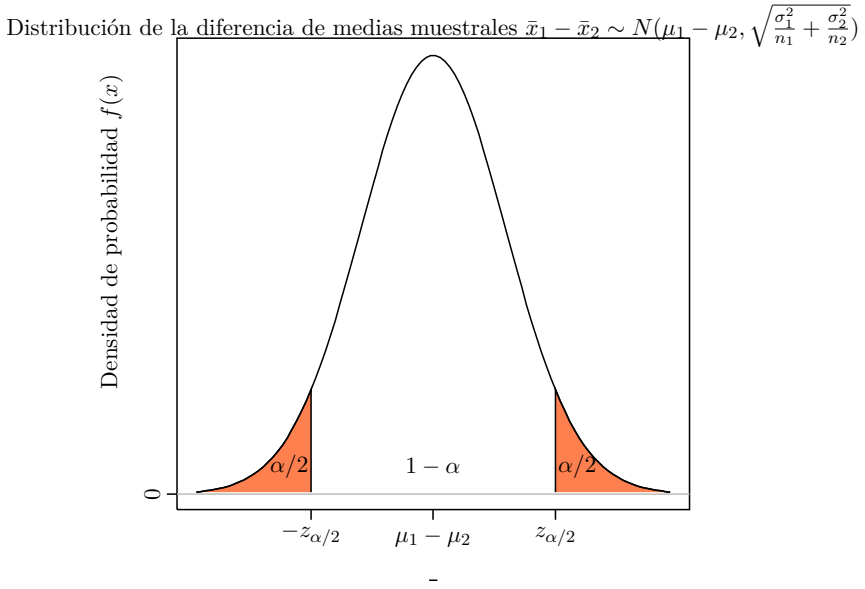
Para muestras pequeñas ($n_1 < 30$ ó $n_2 < 30$), las distribuciones anteriores son también aplicables siempre que las variables de partida sigan distribuciones Normales.

A partir de todo ello y teniendo en cuenta los tres factores de clasificación comentados: si las poblaciones de partida en las que obtenemos las muestras siguen o no distribuciones Normales, si las varianzas de dichas poblaciones son conocidas o desconocidas, y si la muestras son grandes o no, obtenemos las siguientes expresiones correspondientes a los diferentes intervalos de confianza.

Intervalo de confianza para la diferencia de dos medias en poblaciones normales, con varianzas poblacionales conocidas, independientemente del tamaño de la muestra

$$\left(\overline{x}_1 - \overline{x}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

En la figura 7.1 aparece un esquema explicativo de la construcción de este intervalo.



$$P\left(\mu_1 - \mu_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \bar{x}_1 - \bar{x}_2 \leq \mu_1 - \mu_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

Figura 7.1: Cálculo del intervalo de confianza para la diferencia de medias en poblaciones normales con varianzas conocidas a partir de la distribución de la diferencia de medias muestrales $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$.

Intervalo de confianza para la diferencia de dos medias en poblaciones normales, con varianzas poblacionales desconocidas, independientemente del tamaño de la muestra

Si aún siendo desconocidas, las varianzas pueden considerarse iguales, el intervalo es:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{n_1+n_2-2} \cdot \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

donde s_p^2 es una cuasivarianza ponderada:

$$s_p^2 = \frac{(n_1 - 1) \cdot \hat{s}_1^2 + (n_2 - 1) \cdot \hat{s}_2^2}{n_1 + n_2 - 2}$$

Si las varianzas, desconocidas, no pueden considerarse iguales, el intervalo es:

$$\left(\overline{x}_1 - \overline{x}_2 - t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\ ,\ \overline{x}_1 - \overline{x}_2 + t_{\alpha/2}^{\nu} \cdot \sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}\right)$$

donde ν es el número entero que se obtiene redondeando por defecto el valor de la expresión:

$$\frac{\left(\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1+1}+\frac{\left(\frac{\hat{s}_2^2}{n_2}\right)^2}{n_2+1}}-2$$

Si los tamaños muestrales son grandes ($n_1 \geq 30$ y $n_2 \geq 30$) las $t_{\alpha/2}^{\nu}$ y $t_{\alpha/2}^{n_1+n_2-2}$ pueden sustituirse por $z_{\alpha/2}$.

Intervalo de confianza para la diferencia de dos medias en poblaciones no normales, y muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$)

En este caso, como ya sucedía con la media muestral, los intervalos para la diferencia de medias son los mismos que sus correspondientes en poblaciones normales y, de nuevo, habría que distinguir si las varianzas son conocidas o desconocidas (iguales o diferentes), lo cual se traduce en que sus correspondientes fórmulas son las mismas que las dadas en los párrafos anteriores. No obstante, por tratarse de muestras grandes, también es válida la aproximación de $t_{\alpha/2}^{\nu}$ y $t_{\alpha/2}^{n_1+n_2-2}$ por $z_{\alpha/2}$, y habitualmente tan sólo se distingue entre varianzas conocidas y desconocidas.

Para varianzas conocidas:

$$\left(\overline{x}_1-\overline{x}_2-z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\ ,\ \overline{x}_1-\overline{x}_2+z_{\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Y para varianzas desconocidas:

$$\left(\overline{x}_1-\overline{x}_2-z_{\alpha/2}\cdot\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\ ,\ \overline{x}_1-\overline{x}_2+z_{\alpha/2}\cdot\sqrt{\frac{\hat{s}_1^2}{n_1}+\frac{\hat{s}_2^2}{n_2}}\right)$$

Si las poblaciones de partida no son normales y las muestras son pequeñas, no puede aplicarse el Teorema Central de Límite y no se obtienen intervalos de confianza para la diferencia de medias.

Para cualquiera de los anteriores intervalos:

- n_1 y n_2 son los tamaños muestrales.
- $\overline{x}_1$ y $\overline{x}_2$ son las medias muestrales.
- σ_1 y σ_2 son las desviaciones típicas poblacionales.
- $\hat{s}_1$ y $\hat{s}_2$ son las cuasidesviaciones típicas muestrales: $\hat{s}_1^2=\dfrac{\sum\left(x_{1,i}-\overline{x}_1\right)^2}{n_1-1}$, y análogamente $\hat{s}_2^2$.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.
- $t_{\alpha/2}^{n_1+n_2-1}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con n_1+n_2-1 grados de libertad.
- $t_{\alpha/2}^{\nu}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución t de Student con ν grados de libertad.

Intervalos de confianza para la media de la diferencia en datos emparejados

En muchas ocasiones hay que estudiar una característica en una población en dos momentos distintos, para estudiar cómo evoluciona con el tiempo, o para analizar la incidencia de algún hecho ocurrido entre dichos momentos.

En estos casos se toma una muestra aleatoria de la población y en cada individuo de la misma se observa la característica objeto de estudio en los dos momentos citados. Así se tienen

dos conjuntos de datos que no son independientes, pues los datos están emparejados para cada individuo. Por consiguiente, no se pueden aplicar los procedimientos vistos anteriormente, ya que se basan en la independencia de las muestras.

El problema se resuelve tomando para cada individuo la diferencia entre ambas observaciones. Así, la construcción del intervalo de confianza para la diferencia de medias, se reduce a calcular el intervalo de confianza para la media de la variable diferencia. Además, si cada conjunto de observaciones sigue una distribución Normal, su diferencia también seguirá una distribución Normal.

Intervalos de confianza para la diferencia de dos proporciones poblacionales p_1 y p_2

Para muestras grandes ($n_1 \geq 30$ y $n_2 \geq 30$) y valores de p_1 y p_2 (probabilidad de "éxito") cercanos a 0.5, las correspondientes distribuciones Binomiales pueden aproximarse mediante distribuciones Normales de medias respectivas n_1p_1 y n_2p_2, y desviaciones típicas respectivas $\sqrt{n_1p_1(1-p_1)}$ y $\sqrt{n_2p_2(1-p_2)}$. En la práctica, para que sea válida dicha aproximación, se toma el criterio de que tanto n_1p_1 y n_2p_2 como $n_1(1-p_1)$ y $n_2(1-p_2)$ deben ser mayores que 5. Lo anterior hace que también podamos construir intervalos de confianza para la diferencia de proporciones tomando éstas como medias de variables dicotómicas en las que la presencia o ausencia de la característica objeto de estudio ("éxito" ó "fracaso") se expresan mediante un 1 ó un 0 respectivamente.

De este modo, en muestras grandes y con distribuciones Binomiales no excesivamente asimétricas (tanto n_1p_1 y n_2p_2 como $n_1(1-p_1)$ y $n_2(1-p_2)$ deben ser mayores que 5), si denominamos $\widehat{p}_1$ y $\widehat{p}_2$ a la proporción de individuos que presentan el atributo estudiado en la primera y segunda muestras respectivamente, entonces el intervalo de confianza para la diferencia de proporciones con un nivel de significación α viene dado por:

$$\left(\widehat{p}_1 - \widehat{p}_2 - z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}} \ , \ \widehat{p}_1 - \widehat{p}_2 + z_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}_1 \cdot (1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2 \cdot (1-\widehat{p}_2)}{n_2}} \right)$$

donde:

- n_1 y n_2 son los respectivos tamaños muestrales.
- $\widehat{p}_1$ y $\widehat{p}_2$ son las proporciones de individuos que presentan los atributos estudiados en sus respectivas muestras.
- $z_{\alpha/2}$ es el valor que deja a su derecha una probabilidad $\alpha/2$ en una distribución Normal tipificada.

En muestras pequeñas o procedentes de unas distribuciones Binomiales fuertemente asimétricas ($n_1p_1 \leq 5$, $n_2p_2 \leq 5$, $n_1(1-p_1) \leq 5$ ó $n_2(1-p_2) \leq 5$) no puede aplicarse el Teorema Central del Límite y la construcción de intervalos de confianza debe realizarse basándose en la distribución Binomial.

Intervalo de confianza para la razón de dos varianzas σ_1^2 y σ_2^2 de poblaciones normales

Como ya hemos visto en la sección de los intervalos de confianza para la diferencia de dos medias en poblaciones normales con varianzas desconocidas, los mismos dependen de si las varianzas, aún siendo desconocidas, pueden considerarse iguales o no. Para dar respuesta a esta cuestión, previa al cálculo del intervalo para la diferencia de medias, se construye un intervalo para la razón (cociente) de varianzas de ambas poblaciones. Para ello tenemos en cuenta que si partimos de dos variables X_1 y X_2 que siguen distribuciones normales con varianzas σ_1^2 y σ_2^2 respectivamente, y tomamos muestras de tamaños n_1 y n_2 de las respectivas poblaciones se tiene que la variable

$$F = \frac{\frac{\hat{S}_1^2}{\sigma_1^2}}{\frac{\hat{S}_2^2}{\sigma_2^2}}$$

sigue una distribución F de Fisher de $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador.

De lo anterior se deduce que el intervalo de confianza con nivel de significación α para $\frac{\sigma_2^2}{\sigma_1^2}$ es

$$\left(\frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{1-\alpha/2}^{(n_1-1,n_2-1)}, \ \frac{\hat{s}_2^2}{\hat{s}_1^2} \cdot F_{\alpha/2}^{(n_1-1,n_2-1)} \right)$$

Si dentro del intervalo de confianza obtenido está el número 1 (el cociente de varianzas vale la unidad), no habrá, por tanto, evidencia estadística suficiente, con un nivel de significación α, para rechazar que las varianzas sean iguales.

7.2 Ejercicios resueltos

1. Para ver si una campaña de publicidad sobre un fármaco ha influido en sus ventas, se tomó una muestra de 8 farmacias y se midió el número de unidades de dicho fármaco vendidas durante un mes, antes y después de la campaña, obteniéndose los siguientes resultados:

Antes	147	163	121	205	132	190	176	147
Después	150	171	132	208	141	184	182	145

   (a) Crear las variables antes y despues e introducir los datos de la muestra.

   (b) Obtener un resumen estadístico en el que aparezcan la media y la desviación típica de ambas variables. A la vista de los resultados: ¿son las medias diferentes?, ¿ha aumentado la campaña el nivel de ventas?, ¿crees que los resultados son estadísticamente significativos?

   i. Seleccionar el menú *Analizar→Estadísticos descriptivos→Frecuencias*.
   ii. En el cuadro de diálogo que aparece seleccionar las dos variables en el campo *Variables* y hacer click en el botón *Estadísticos*.
   iii. En el cuadro de diálogo que aparece activar los estadísticos deseados y hacer click en el botón *Continuar* y *Aceptar*.

   (c) Obtener los intervalos de confianza para la media de la diferencia entre ambas variables con niveles de significación 0.05 y 0.01.

   i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras relacionadas*.
   ii. En el cuadro de diálogo que aparece seleccionar ambas variables en el campo *Variables emparejadas* y hacer click en el botón *Opciones*.
   iii. En el cuadro de diálogo que aparece introducir el nivel de confianza deseado en el campo *Porcentaje del intervalo de confianza* y hacer click en el botón *Continuar* y *Aceptar*.

   (d) Crear la variable diferencia, que se obtiene como antes-despues.

   i. Seleccionar el menú *Transformar→Calcular variable*.
   ii. En el cuadro de diálogo que aparece introducir el nombre de la nueva variable diferencia en el campo *Variable de destino*, en el campo *Expresión numérica* introducir antes-despues y hacer click en el botón *Aceptar*.

   (e) Calcular el intervalo de confianza para la media de la variable diferencia con un nivel de significación del 0.05 y comparar el intervalo obtenido con el del apartado anterior.

   i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para una muestra*.
   ii. En el cuadro de diálogo que aparece seleccionar la variable diferencia en el campo *Variables para contrastar* y hacer click en el botón *Opciones*.
   iii. En el cuadro de diálogo que aparece introducir el nivel de confianza deseado en el campo *Porcentaje del intervalo de confianza* y hacer click en el botón *Continuar* y *Aceptar*.

   (f) ¿Existen pruebas suficientes para afirmar con un 95% de confianza que la campaña de publicidad ha aumentado las ventas? ¿Y si cambiamos los dos últimos datos de la variable despues y ponemos 190 en lugar de 182 y 165 en lugar de 145? Observar qué le ha sucedido al intervalo para la diferencia de medias y darle una explicación.

2. Una central de productos lácteos recibe diariamente la leche de dos granjas X e Y. Para analizar la calidad de la leche, durante una temporada, se controla el contenido de materia grasa de la leche que proviene de ambas granjas, con los siguientes resultados:

X		Y	
0.34	0.34	0.28	0.29
0.32	0.35	0.30	0.32
0.33	0.33	0.32	0.31
0.32	0.32	0.29	0.29
0.33	0.30	0.31	0.32
0.31	0.32	0.29	0.31
		0.33	0.32
		0.32	0.33

   (a) Crear las variables grasa y granja, e introducir los datos de la muestra.

   (b) Calcular el intervalo de confianza con un 95% de confianza para la diferencia en el contenido medio de materia grasa de la leche procedente de ambas granjas.

i

   i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras independientes*.
   ii. En el cuadro de dialogo que aparece seleccionar la variable grasa en el campo *Variables para contrastar*, seleccionar la variable granja en el campo *Variable de agrupación* y hacer click sobre el botón *Definir grupos*.
   iii. En el cuadro de diálogo que aparece introducir en el campo *Grupo 1* el valor de la variable granja correspondiente a la granja X y en el campo *Grupo 2* el correspondiente a la granja Y, y hacer click sobre el botón *Continuar*.
   iv. En el cuadro de diálogo inicial hacer click sobre el botón *Opciones*.
   v. En el cuadro de diálogo que aparece introducir el nivel de confianza deseado en el campo *Porcentaje del intervalo de confianza* y hacer click en el botón *Continuar* y *Aceptar*.

   (c) A la vista de los intervalos obtenidos en el punto anterior, ¿se puede concluir que existen diferencias significativas en el contenido medio de grasa según la procedencia de la leche?

   (d) De los dos intervalos de confianza que se obtienen, uno da como supuesta la igualdad de varianzas y el otro no. En este ejemplo concreto, ¿sería metodológicamente correcto considerar adecuado el que supone igualdad de varianzas?

i

Entre los resultados aparece la prueba de Levene para contrastar la igualdad de varianzas. Se trata de un contraste de hipótesis cuyo resultado final es un p-valor, denominado Sig. por el programa, que aprenderemos a interpretar en prácticas sucesivas. Por ahora es suficiente tener en cuenta que cuando el p-valor es mayor que el nivel de significación elegido no podemos rechazar la hipótesis de igualdad de varianzas.

3. Un profesor universitario ha tenido dos grupos de clase a lo largo del año: uno con horario de mañana y otro de tarde. En el de mañana, sobre un total de 80 alumnos, han aprobado 55; y en el de tarde, sobre un total de 90 alumnos, han aprobado 32.

   (a) Crear las variables: grupo, cuyos valores serán 0 (mañana) y 1 (tarde); aprobado, cuyos valores serán 1 (aprobado) y 0 (suspenso); y frecuencia, cuyos valores serán el número de aprobados y suspensos en cada grupo.

   (b) Ponderar los valores de la variable aprobado mediante los pesos de la variable frecuencia.

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable de frecuencia* y hacer click en el botón *Aceptar*.

(c) Calcular el intervalo para la diferencia de proporciones de alumnos aprobados en cada grupo.

i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras independientes*.
ii. Seleccionar la variable aprobado en el campo *Varaibles para contrastar*, la variable grupo en el campo *Variable de agrupación* y hacer click en el botón *Definir grupos*.
iii. En el cuadro de diálogo que aparece introducir en el campo *Grupo 1* el valor de la variable grupo correspondiente al grupo de mañana y en el campo *Grupo 2* el correspondiente al grupo de tarde, y hacer click sobre el botón *Continuar*.
iv. En el cuadro de diálogo inicial hacer click sobre el botón *Opciones*.
v. En el cuadro de diálogo que aparece introducir el nivel de confianza deseado en el campo *Porcentaje del intervalo de confianza* y hacer click en el botón *Continuar* y *Aceptar*.

(d) Suponiendo que el resto de los factores (temario, complejidad de examen, nivel previo de conocimientos, expediente académico previo de los alumnos,...) no han influido en el aprobado o suspenso en la asignatura, ¿se puede concluir que el factor horario ha sido determinante en la proporción de suspensos? Justificar la respuesta.

7.3 Ejercicios propuestos

1. Se ha realizado un estudio para investigar el efecto del ejercicio físico en el nivel de colesterol en la sangre. En el estudio participaron once personas, a las que se les midió el nivel de colesterol antes y después de desarrollar un programa de ejercicios. Los resultados obtenidos fueron los siguientes:

Nivel Previo	182	232	191	200	148	249	276	213	241	280	262
Nivel Posterior	198	210	194	220	138	220	219	161	210	213	226

(a) Hallar el intervalo de confianza del 90% para la diferencia del nivel medio de colesterol antes y después del programa de ejercicios.

(b) A la vista de dicho intervalo, ¿se concluye que el ejercicio físico disminuye el nivel de colesterol con una confianza del 90%?

2. Dos químicos A y B realizan respectivamente 14 y 16 determinaciones de la actividad radiactiva de una muestra de material. Sus resultados en Curios fueron:

A		B	
263.36	254.68	286.53	254.54
248.64	276.32	284.55	286.30
243.64	256.42	272.52	282.90
272.68	261.10	283.85	253.75
287.33	268.41	252.01	245.26
287.26	282.65	275.08	266.08
250.97	284.27	267.53	252.05
		253.82	269.81

(a) Calcular el intervalo de confianza para la diferencia de las medias de actividad detectada por cada uno de los químicos con un 95% de confianza.

(b) ¿Se puede decir que existen diferencias significativas en la media de actividad detectada por cada químico?

3. Utilizar el fichero Hipertensos Datos Claves para calcular el intervalo de confianza del 95% para la comparación de la presión sistólica media inicial y final, en cada uno de los tratamientos. ¿Qué tratamiento ha sido más efectivo para reducir la presión?

4. En una encuesta realizada por la Junta de Comunidades de Castilla la Mancha, en los dos hospitales de una ciudad, se pregunta a los pacientes hospitalizados cuando salen del hospital por si consideran que el trato recibido ha sido correcto. En el primero de ellos se pregunta a 200 pacientes y 140 responden que sí, mientras que en el segundo, se pregunta a 300 pacientes y 160 responden que sí.

   (a) Calcular el intervalo de confianza para la diferencia de proporciones de pacientes satisfechos con el trato recibido.

   (b) ¿Hay pruebas significativas de que el trato recibido en un hospital es mejor que en el otro?

8 — Contraste de Hipótesis

8.1 Fundamentos teóricos

8.1.1 Inferencia estadística y contrastes de hipótesis

Cualquier afirmación o conjetura que determina, parcial o totalmente, la distribución de una población se realiza mediante una *Hipótesis Estadística*.

En general, nunca se sabe con absoluta certeza si una hipótesis es cierta o falsa, ya que, para ello tendríamos que medir a todos los individuos de la población. Las decisiones se toman sobre una base de probabilidad y los procedimientos que conducen a la aceptación o rechazo de la hipótesis forman la parte de la Inferencia Estadística que se denomina *Contraste de Hipótesis*.

Una hipótesis se contrasta comparando sus predicciones con la realidad que se obtiene de las muestras: si coinciden, dentro del margen de error probabilísticamente admisible, mantendremos la hipótesis; en caso contrario, la rechazamos y buscaremos nuevas hipótesis capaces de explicar los datos observados.

8.1.2 Tipos de contrastes de hipótesis

Los contrastes de hipótesis se clasifican como:

- *Contrastes Paramétricos*. Que a su vez son de dos tipos según que:
  - Se contraste un valor concreto o intervalo para los parámetros de la distribución de una variable aleatoria. Por ejemplo: podemos contrastar la hipótesis de que la media del nivel de colesterol en sangre en una población es de 180 mg/dl.
  - Se comparen los parámetros de las distribuciones de dos o más variables. Por ejemplo: podemos contrastar la hipótesis de que la media del nivel de colesterol en sangre es más baja en las personas que ingieren por debajo de una cierta cantidad de grasas en su dieta.
- *Contrastes No Paramétricos*. En los que se contrastan las hipótesis que se imponen como punto de partida en los contrastes paramétricos, y que reciben el nombre de *Hipótesis Estructurales*. Entre ellas, el modelo de distribución de los datos y la independencia de los mismos. Por ejemplo: en muchos de los contrastes paramétricos se exige como hipótesis de partida que los datos muestrales provengan de una población normal, pero precisamente éste sería el primer contraste al que habría que dar respuesta, puesto que si los datos no provienen de una población normal, las conclusiones obtenidas gracias a los contrastes paramétricos derivados pueden ser completamente erróneas.

8.1.3 Elementos de un contraste

Hipótesis nula e hipótesis alternativa

El primer punto en la realización de un contraste de hipótesis es la formulación de la Hipótesis Nula y su correspondiente Hipótesis Alternativa.

Llamaremos *Hipótesis Nula* a la hipótesis que se contrasta. Se suele representar como H_0 y representa la hipótesis que mantendremos a no ser que los datos observados en la muestra indiquen su falsedad, en términos probabilísticos.

El rechazo de la hipótesis nula lleva consigo la aceptación implícita de la *Hipótesis Alternativa*, que se suele representar como H_1. Para cada H_0 tenemos dos H_1 diferentes según que el contraste sea de tipo *Bilateral*, si desconocemos la dirección en que H_0 puede ser falsa, o *Unilateral*, si sabemos en qué dirección H_0 puede ser falsa. Y el Unilateral se clasifica como *Con Cola a la Derecha*, si en H_1 sólo englobamos valores mayores del parámetro para el que planteamos el contraste que el que aparecen en H_0, y *Con Cola a la Izquierda*, si en H_1 sólo englobamos valores menores.

En la siguiente tabla se formulan tanto H_0 como H_1 en contrastes paramétricos, para un parámetro cualquiera, que denominaremos θ, de una población, y para la comparación de dos parámetros, θ_1 y θ_2, de dos poblaciones.

	H_0	H_1
Bilateral en una población	$\theta = \theta_0$	$\theta \neq \theta_0$
Unilateral en una población	$\theta = \theta_0$	$\theta > \theta_0$ (Cola a la dcha.) ó $\theta < \theta_0$ (Cola a la Izda.)
Bilateral en dos poblaciones	$\theta_1 = \theta_2$	$\theta_1 \neq \theta_2$
Unilateral en dos poblaciones	$\theta_1 = \theta_2$	$\theta_1 > \theta_2$ ó $\theta_1 < \theta_2$

Ejemplo 5. *Supongamos que, gracias a datos previos, conocemos que la media del nivel de colesterol en sangre en una determinada población es 180 mg/dl, y suponemos que la aplicación de una cierta terapia ha podido influir (ya sea para aumentar o para disminuir) en dicha media. Para formular H_0 debemos tener en cuenta que la hipótesis nula siempre es conservadora, es decir, no cambiaremos nuestro modelo si no hay evidencias probabilísticamente fuertes de que ha dejado de ser válido. Según esto, la hipótesis nula será que la media no ha cambiado:*

$$H_0 : \mu = 180.$$

Una vez fijada la hipótesis nula, para formular la hipótesis alternativa debemos tener en cuenta que se trata de un contraste bilateral, ya que no conocemos, a priori, el sentido de la variación de la media (si será mayor o menor de 180 mg/dl). Por tanto, la hipótesis alternativa es que la media es distinta de 180 mg/dl:

$$H_1 : \mu \neq 180.$$

Por otro lado, si presumimos que la aplicación de la terapia ha servido para disminuir el nivel de colesterol, estamos ante un contraste unilateral en el que la hipótesis nula sigue siendo que la media no ha cambiado, y la alternativa es que ha disminuido:

$$H_0 : \mu = 180,$$
$$H_1 : \mu < 180.$$

Normalmente, el objetivo del investigador es rechazar la hipótesis nula para probar la certeza de la hipótesis alternativa, y esto sólo lo hará cuando haya pruebas suficientemente significativas de la falsedad de H_0. Si los datos observados en la muestra no aportan estas pruebas, entonces se mantiene la hipótesis nula, y en este sentido se dice que es la hipótesis conservadora. Pero conviene aclarar que aceptar la hipótesis nula no significa que sea cierta, sino que no tenemos información suficiente o evidencia estadística para rechazarla.

Errores en un contraste. Nivel de significación y potencia

Como ya hemos comentado, la aceptación o rechazo de H_0 siempre se realiza en términos probabilísticos, a partir de la información obtenida en la muestra. Esto supone que nunca tendremos absoluta seguridad de conocer la certeza o falsedad de una hipótesis, de modo que al aceptarla o rechazarla es posible que nos equivoquemos.

Los errores que se pueden cometer en un contraste de hipótesis son de dos tipos:

- **Error de tipo I**: se produce cuando rechazamos H_0 siendo correcta.
- **Error de tipo II**: se produce cuando aceptamos H_0 siendo falsa.

La probabilidad de cometer un error de tipo I se conoce como *Nivel de Significación* del contraste y se designa por

$$\alpha = P(\text{Rechazar } H_0 | H_0 \text{ es cierta}).$$

Y la probabilidad de cometer un error de tipo II se designa por

$$\beta = P(\text{Aceptar } H_0 | H_0 \text{ es falsa}).$$

Así pues, al realizar un contraste de hipótesis, pueden darse las cuatro situaciones que aparecen esquematizadas en el cuadro 8.1.

	Realidad	
Decisión	H_0 cierta	H_0 falsa
Aceptar H_0	Decisión correcta (Probabilidad $1-\alpha$)	Error de Tipo II (Probabilidad β)
Rechazar H_0	Error de Tipo I (Probabilidad α)	Decisión correcta (Probabilidad $1-\beta$)

Cuadro 8.1: Tipos de errores en un contraste de hipótesis.

Puesto que lo interesante en un contraste es rechazar la hipótesis nula, lo que más interesa controlar es el riesgo de equivocación si se rechaza, es decir el error del tipo I. Por tanto, α se suele fijar a niveles bajos, ya que cuanto más pequeño sea, mayor seguridad tendremos al rechazar la hipótesis nula. Los niveles más habituales a los que se fija α son 0.1, 0.05 y 0.01.

Una vez controlado el error de tipo I, también es interesante controlar el error del tipo II. Ahora bien, el valor de β se calcula partiendo de que la hipótesis nula es falsa, es decir $\theta \neq \theta_0$ (o $\theta_1 \neq \theta_2$ en el caso de dos poblaciones), pero esto engloba infinitas posibilidades, de manera que para poder calcularlo no queda más remedio que fijar H_1 dando un único valor al parámetro. En este caso, se define la *Potencia del Contraste* como la probabilidad de rechazar H_0 cuando la hipótesis alternativa fijada es verdadera, y vale $1-\beta$. Resulta evidente que para un nivel de significación dado, un contraste será mejor cuanta más potencia tenga.

Como la potencia depende del valor del parámetro fijado en la hipótesis alternativa, se puede definir una función para la potencia como

$$\text{Potencia}(x) = P(\text{Rechazar } H_0 | \theta = x),$$

que indica la probabilidad de rechazar H_0 para cada valor del parámetro θ. Esta función se conoce como *curva de potencia*.

Por otro lado, β también depende de α ya que al disminuir α, cada vez es más difícil rechazar H_0, y por tanto, la probabilidad de aceptar la hipótesis nula siendo falsa aumenta. En consecuencia, y como veremos más adelante, la única forma de disminuir β y ganar potencia, una vez fijado α, es aumentando el tamaño de la muestra.

Estadístico del contraste y regiones de aceptación y rechazo

La decisión entre la aceptación o el rechazo de H_0 que se plantea en el contraste, se realiza a partir de un estadístico en el muestreo, relacionado con el parámetro o característica que queremos contrastar, y cuya distribución debe ser conocida suponiendo cierta H_0 y una vez fijado el tamaño de la muestra. Este estadístico recibe el nombre de *Estadístico del Contraste*.

Para cada muestra el estadístico del contraste toma un valor concreto que recibe el nombre de *estimación del estadístico*. A partir de esta estimación tomaremos la decisión de aceptar o rechazar la hipótesis nula. Si la estimación difiere demasiado del valor que propone H_0 para el parámetro, entonces rechazaremos H_0, mientras que si no es demasiado diferente la aceptaremos.

La magnitud de la diferencia que estamos dispuestos a tolerar entre la estimación y el valor del parámetro para mantener la hipótesis nula, depende de la probabilidad de error de tipo I que estemos dispuestos a asumir. Si α es grande, pequeñas diferencias pueden ser suficientes para rechazar H_0, mientras que si α es muy pequeño, sólo rechazaremos H_0 cuando la diferencia entre el estimador y el valor del parámetro según H_0, sea muy grande. De esta manera, al fijar el nivel de significación α, el conjunto de valores que puede tomar el estadístico del contraste queda divido en dos partes: la de las estimaciones que conducirían a la aceptación de H_0, que se denomina *Región de Aceptación*, y la de las estimaciones que conducirían al rechazo de H_0, que se denomina *Región de Rechazo*.

Si llamamos al estadístico del contraste $\hat{\theta}$, entonces, dependiendo de si el contraste es unilateral o bilateral, tendremos las siguientes regiones de aceptación y rechazo:

Contraste	Región de Aceptación	Región de Rechazo
$H_0: \ \theta = \theta_0$ $H_1: \ \theta \neq \theta_0$	$\{\hat{\theta}_{1-\alpha/2} \leq \hat{\theta} \leq \hat{\theta}_{\alpha/2}\}$	$\{\hat{\theta} < \hat{\theta}_{1-\alpha/2}\} \cup \{\hat{\theta} > \hat{\theta}_{\alpha/2}\}$
$H_0: \ \theta = \theta_0$ $H_1: \ \theta < \theta_0$	$\{\hat{\theta} \geq \hat{\theta}_{1-\alpha}\}$	$\{\hat{\theta} < \hat{\theta}_{1-\alpha}\}$
$H_0: \ \theta = \theta_0$ $H_1: \ \theta > \theta_0$	$\{\hat{\theta} \leq \hat{\theta}_{\alpha}\}$	$\{\hat{\theta} > \hat{\theta}_{\alpha}\}$

donde $\hat{\theta}_{1-\alpha}$ y $\hat{\theta}_{\alpha}$ son valores tales que $P(\hat{\theta} < \hat{\theta}_{1-\alpha} | \theta = \theta_0) = \alpha$ y $P(\hat{\theta} > \hat{\theta}_{\alpha} | \theta = \theta_0) = \alpha$, tal y como se muestra en las figuras 8.1 y 8.2.

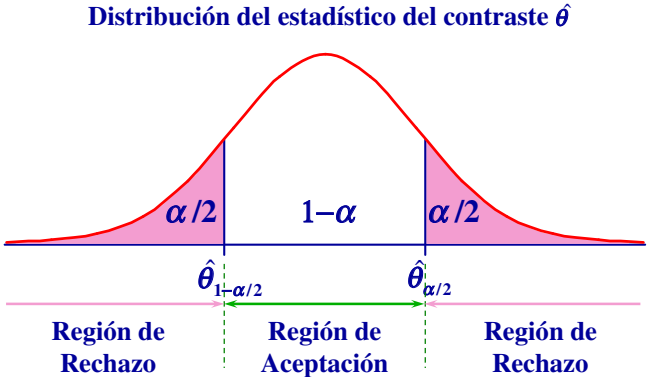


Figura 8.1: Regiones de aceptación y rechazo en un contraste bilateral.

En resumen, una vez que tenemos el estadístico del contraste y hemos fijado el nivel de significación α, las regiones de aceptación y rechazo quedan delimitadas, y ya sólo queda tomar una muestra, calcular a partir de ella la estimación del estadístico delcontraste, y aceptar o rechazar la hipótesis nula, dependiendo de si la estimación cae en la región de aceptación o en la de rechazo respectivamente.

El p-valor de un contraste

Aunque ya disponemos de los elementos necesarios para realizar un contraste de hipótesis que nos permita tomar una decisión respecto a aceptar o rechazar la hipótesis nula, en la práctica,

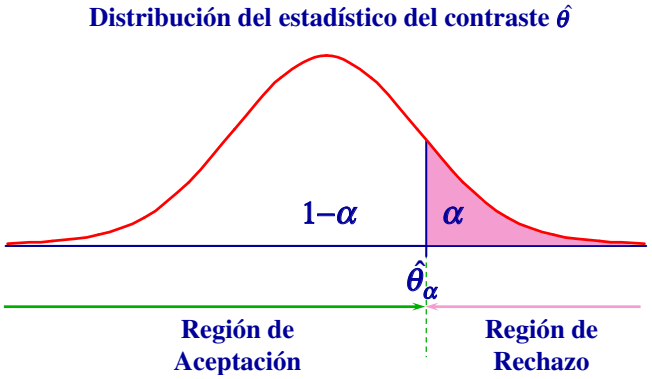


Figura 8.2: Regiones de aceptación y rechazo en un contraste unilateral.

la decisión que se toma suele acompañarse del grado de confianza que tenemos en la misma. Si por ejemplo, tenemos una región de rechazo $\{\hat{\theta} > \hat{\theta}_\alpha\}$, siempre que la estimación del estadístico del contraste caiga dentro de esta región rechazaremos H_0, pero obviamente, si dicha estimación es mucho mayor que $\hat{\theta}_\alpha$ tendremos más confianza en el rechazo que si la estimación está cerca del límite entre las regiones de aceptación y rechazo $\hat{\theta}_\alpha$. Por este motivo, al realizar un contraste, también se calcula la probabilidad de obtener una discrepancia mayor o igual que la observada entre el valor del parámetro, suponiendo cierta H_0, y la estimación que se obtiene de los datos muestrales. Esta probabilidad se conoce como *p-valor del contraste*, y en cierto modo, expresa la confianza que se tiene al tomar la decisión en el contraste, ya que si H_0 es cierta y el p-valor es pequeño, es porque bajo la hipótesis nula resulta poco probable encontrar una discrepancia como la observada, y por tanto, tendremos bastante seguridad a la hora de rechazar H_0. En general, cuanto más próximo esté p a 1, mayor seguridad existe al aceptar H_0, y cuanto más próximo esté a 0, mayor seguridad hay al rechazarla.

El cálculo del p-valor dependerá de si el contraste es bilateral o unilateral, y en este último caso de si es unilateral con cola a la derecha o con cola a la izquierda. El p-valor que se obtiene para los diferentes tipos de contrastes es el que aparece en la tabla siguiente:

Contraste	p-valor
Bilateral	$2P(\hat{\theta} > \hat{\theta}_0 \mid H_0 \text{ es cierta})$ si $\hat{\theta} > \hat{\theta}_0$ o $2P(\hat{\theta} < \hat{\theta}_0 \mid H_0 \text{ es cierta})$ si $\hat{\theta} < \hat{\theta}_0$
Unilateral con cola a la derecha	$P(\hat{\theta} > \hat{\theta}_0 \mid H_0 \text{ es cierta})$
Unilateral con cola a la izquierda	$P(\hat{\theta} < \hat{\theta}_0 \mid H_0 \text{ es cierta})$

En la figura 8.3 se observa que el p-valor es el área de la cola de la distribución (o colas si se trata de un contraste bilateral) definida a partir del estadístico del contraste.

Una vez calculado el p-valor, si hemos fijado el nivel de significación α y han quedado delimitadas las regiones de aceptación y rechazo, el que la estimación caiga dentro de la región de rechazo es equivalente a que $p < \alpha$, mientras que si cae dentro de la región de aceptación, entonces $p \geq \alpha$. Esta forma de abordar los contrastes, nos da una visión más amplia, ya que nos da información de para qué niveles de significación puede rechazarse la hipótesis nula, y para cuales no se puede.

Contrastes y estadísticos de contraste

Apoyándose en las distintas distribuciones en el muestreo comentadas en las prácticas sobre intervalos de confianza, a continuación se presentan las fórmulas para los principales estadísticos de contraste.

Contraste para la media de una población normal con varianza conocida

- Hipótesis Nula: $H_0 : \mu = \mu_0$

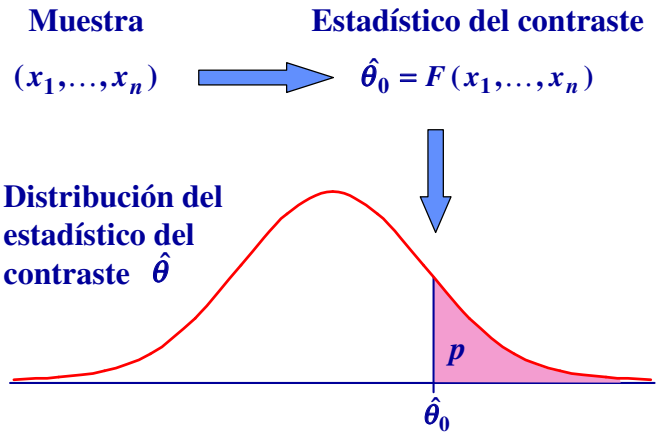


Figura 8.3: El p-valor de un contraste unilateral con cola a la derecha.

- Estadístico del contraste:
$$\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$$
que sigue una distribución normal tipificada, $N(0,1)$.

Este contraste también es válido para la media de una población no normal, siempre y cuando las muestras sean grandes ($n \geq 30$), con varianza conocida; y para la media de la diferencia de datos emparejados, siempre y cuando la variable diferencia siga una distribución normal con varianza conocida, o una distribución cualquiera si la muestra es grande.

Contraste para la media de una población normal con varianza desconocida

- Hipótesis Nula: $H_0 : \mu = \mu_0$
- Estadístico del contraste:
$$\frac{\bar{X}-\mu_0}{\hat{s}/\sqrt{n}}$$
que sigue una distribución t de Student con $n-1$ grados de libertad, $T(n-1)$.

Este contraste también es válido para la media de una población no normal en muestras grandes ($n \geq 30$), con varianza desconocida; y para la media de la diferencia de datos emparejados, siempre y cuando la variable diferencia siga una distribución normal con varianza desconocida, o una distribución cualquiera si la muestra es grande.

Contraste para la proporción en muestras grandes y distribuciones simétricas (tanto np como $n(1-p)$ deben ser mayores que 5)

- Hipótesis Nula: $H_0 : p = p_0$
- Estadístico del contraste:
$$\frac{\hat{p}-p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$
que sigue una distribución normal tipificada, $N(0,1)$.

Contraste para la varianza de una población normal

- Hipótesis Nula: $H_0 : \sigma^2 = \sigma_0^2$
- Estadístico del contraste:
$$\frac{(n-1)\hat{s}^2}{\sigma_0^2}$$
que sigue una distribución Chi-cuadrado con $n-1$ grados de libertad.

Contraste para la diferencia de medias de poblaciones normales con varianzas conocidas

- Hipótesis Nula: $H_0 : \mu_1 = \mu_2$
- Estadístico del contraste:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

que sigue una distribución normal tipificada, $N(0, 1)$.

Este contraste también es válido para la diferencia de medias de dos poblaciones no normales, siempre y cuando las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), con varianzas conocidas.

Contraste para la diferencia de medias de poblaciones normales con varianzas desconocidas

- Hipótesis Nula: $H_0 : \mu_1 = \mu_2$
- Estadístico del contraste:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}}}$$

que sigue una distribución t de Student con ν grados de libertad, donde ν es el número entero más próximo al valor de la expresión:

$$\frac{\left(\frac{\hat{s}_1^2}{n_1} + \frac{\hat{s}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{s}_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{\hat{s}_2^2}{n_2}\right)^2}{n_2 + 1}} - 2.$$

Este contraste también es válido para la diferencia de medias de dos poblaciones no normales, siempre y cuando las muestras sean grandes ($n_1 \geq 30$ y $n_2 \geq 30$), con varianzas desconocidas.

Contraste para la diferencia de proporciones en muestras grandes y distribuciones simétricas (n_1p_1, $n_1(1-p_1)$, n_2p_2, $n_2(1-p_2)$ deben ser mayores que 5)

- Hipótesis Nula: $H_0 : p_1 = p_2$
- Estadístico del contraste:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

que sigue una distribución normal tipificada, $N(0, 1)$.

Contraste para la igualdad de varianzas de poblaciones normales

- Hipótesis Nula: $H_0 : \sigma_1^2 = \sigma_2^2$
- Estadístico del contraste:

$$\frac{\hat{S}^2_{1,n_1-1}}{\hat{S}^2_{2,n_2-1}}$$

que sigue una distribución F de Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad.

8.2 Ejercicios resueltos

1. Se analiza la concentración de principio activo en una muestra de 10 envases tomados de un lote de un fármaco, obteniendo los siguientes resultados en mg/mm^3:

$$17.6 - 19.2 - 21.3 - 15.1 - 17.6 - 18.9 - 16.2 - 18.3 - 19.0 - 16.4$$

Se pide:

(a) Crear la variable concentración, e introducir los datos de la muestra.

(b) Realizar el contraste de hipótesis bilateral: $H_0 : \mu = 18$ y $H_1 : \mu \neq 18$ con un nivel de significación 0.05.

i

i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para una muestra*.

ii. En el cuadro de diálogo que aparece, seleccionar la variable concentración en el campo *Variables para contrastar*, introducir el valor de la media en la hipótesis nula (en este caso 18) en el campo *Valor de prueba* y hacer click en el botón *Opciones*.

iii. En el cuadro de diálogo que aparece, introducir el nivel de confianza del contraste en el campo *Porcentaje del Intervalo de conficanza* y hacer click sobre el botón *Continuar* y *Aceptar*.
Aunque el nivel de confianza no afecta al p-valor del contraste, que el programa denomina sig. (bilateral), si afecta al intervalo de confianza que se muestra junto con el p-valor.

(c) De igual manera realizar los contrastes bilaterales: $H_0 : \mu = 19.5$ y $H_1 : \mu \neq 19.5$ con niveles de significación 0.05 y 0.01. ¿Cómo afecta la disminución en el nivel de significación en la facilidad para rechazar H_0?

i Seguir los mismos pasos del apartado anterior.

(d) Realizar los contrastes de hipótesis unilaterales: $H_0 : \mu = 17$ y $H_1 : \mu > 17$, y $H_0 : \mu = 17$ y $H_1 : \mu < 17$ con un nivel de significación de 0.1.

i Repetir los mismos pasos de los apartados anteriores pero teniendo en cuenta que lo que el programa denomina Sig.(bilateral) es el p-valor del contraste bilateral; por lo tanto, para el contraste unilateral de mayor el p-valor será Sig.(bilateral)/2, y para el contraste unilateral de menor el p-valor será 1-(Sig.(bilateral)/2).

(e) Si el fabricante del lote asegura haber aumentado la concentración de principio activo con respecto a anteriores lotes, en los que la media era de 17.5 mg/mm^3, con un nivel de confianza del 95%; ¿aceptamos o rechazamos lo dicho por el fabricante?

2. Varios investigadores desean saber si es posible concluir que dos poblaciones de niños difieren respecto a la edad promedio en la cual pueden caminar por sí solos. Los investigadores obtuvieron los siguientes datos para la edad al comenzar a andar (expresada en meses):

Muestra en la población A: $9.5 - 10.5 - 9.0 - 9.8 - 10.0 - 13.0 - 10.0 - 13.5 - 10.0 - 9.8$

Muestra en la población B: $12.5 - 9.5 - 13.5 - 13.8 - 12.0 - 13.8 - 12.5 - 9.5 - 12.0 - 13.5 - 12.0 - 12.0$

(a) Crear las variables población y edad.

(b) Realizar un contraste de hipótesis, con un nivel de significación de 0.05, para dar respuesta a la conclusión que buscan los investigadores.

i Se trata de un contraste bilateral de comparación de medias en poblaciones independientes.

i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras independientes*.
ii. En el cuadro de diálogo que aparece, seleccionar la variable edad en el campo *Variables para contrastar*, la variable población en el campo *Variable de agrupación* y hacer click sobre el botón *Definir grupos*.
iii. En el cuadro de diálogo que aparece, escribir el valor de la variable población correspondiente a la población A en el campo *Grupo 1* y el correspondiente a la población B en el campo *Grupo 2*, y hacer click sobre el botón *Continuar* y *Aceptar*.

3. Algunos investigadores han observado una mayor resistencia de las vías respiratorias en fumadores que en no fumadores. Para confirmar dicha hipótesis, se realizó un estudio para comparar el porcentaje de retención traqueobronquial en las mismas personas cuando aún eran fumadoras y transcurrido un año después de dejarlo. Los resultados se indican en la tabla siguiente:

Porcentaje de retención	
Cuando fumaba	Transcurrido un año sin fumar
60,6	47,5
12,0	13,3
56,0	33,0
75,2	55,2
12,5	21,9
29,7	27,9
57,2	54,3
62,7	13,9
28,7	8,9
66,0	46,1
25,2	29,8
40,1	36,2

(a) Crear las variables antes y después e introducir los datos.
(b) Plantear el contraste de hipótesis adecuado para confirmar o rechazar la hipótesis de los investigadores.

i Se trata de un contraste unilateral ($H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 > \mu_2$) de igualdad de medias en datos pareados.

i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras relacionadas*.
ii. En el cuadro de diálogo que aparece, seleccionar ambas variables y pasarlas al campo *Variables emparejadas*.

Al tratarse de un contraste unilateral de mayor el p-valor será Sig.(bilateral)/2.

4. Un profesor universitario ha tenido dos grupos de clase a lo largo del año: uno con horario de mañana y otro de tarde. En el de mañana, sobre un total de 80 alumnos, han aprobado 55; y en el de tarde, sobre un total de 90 alumnos, han aprobado 32.

(a) Crear las variables: grupo, cuyos valores serán mañana y tarde; calificación, cuyos valores serán 1 (aprobado) y 0 (suspenso); y frecuencia, cuyos diferentes valores son el número de aprobados y suspensos en cada grupo.
(b) Ponderar los datos mediante la variable frecuencia.

i

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable de frecuencia* y hacer click en el botón *Aceptar*.

(c) Realizar un contraste de hipótesis para determinar si el factor horario ha sido o no determinante en la proporción de suspensos con un nivel de significación 0.05

i

Se trata de un contraste bilateral de comparación de medias en poblaciones independientes.

i. Seleccionar el menú *Analizar→Comparar medias→Prueba T para muestras independientes*.
ii. Seleccionar la variable calificación en el campo *Variables para contrastar*, la variable grupo en el campo *Variable de agrupación* y hacer click en el botón *Definir grupos*.
iii. En el cuadro de diálogo que aparece introducir en el campo *Grupo 1* el valor de la variable grupo correspondiente al grupo de mañana y en el campo *Grupo 2* el correspondiente al grupo de tarde, y hacer click sobre el botón *Continuar*.

8.3 Ejercicios propuestos

1. Un grupo de médicos intenta probar que un programa de ejercicios moderadamente activos puede beneficiar a los pacientes que han sufrido previamente un infarto de miocardio. Para ello escogieron once individuos para participar en el estudio y midieron su capacidad de trabajo, entendida como el tiempo que se tarda en alcanzar una tasa de 160 latidos por minuto mientras se camina con una cierta velocidad sobre una cinta de andar, al comienzo del estudio y después de 25 semanas de ejercicio controlado. Los resultados, expresados en minutos, fueron los siguientes:

Sujeto	Antes	Después
1	7,6	14,7
2	9,9	14,1
3	8,6	11,8
4	9,5	16,1
5	8,4	14,7
6	9,2	14,1
7	6,4	13,2
8	9,9	14,9
9	8,7	12,2
10	10,3	13,4
11	8,3	14,0

¿Sostienen estos datos el argumento de los investigadores?

2. Se acepta generalmente que existen diferencias ligadas al sexo relacionadas con la respuesta a la tensión producida por el calor. Para comprobarlo, se sometió a un grupo de 10 hombres y 8 mujeres a un programa de ejercicios duros, en un medio con temperatura alta (40 ºC) y sin posibilidad de beber. La variable de interés medida fue el porcentaje de peso corporal perdido. Se obtuvieron los datos siguientes:

Varones		Mujeres	
2,9	3,7	3,0	3,8
3,5	3,8	2,5	4,1
3,9	4,0	3,7	3,6
3,8	3,6	3,3	4,0
3,6	3,7		

Según los datos recogidos, ¿hay diferencias en el porcentaje medio de peso corporal perdido como respuesta al ejercicio físico desarrollado en alta temperatura entre hombres y mujeres?

9 — Análisis de la Varianza de 1 Factor

9.1 Fundamentos teóricos

El *Análisis de la Varianza con un Factor* es una técnica estadística de contraste de hipótesis, que sirve para comparar las medias una variable cuantitativa, que suele llamarse *variable dependiente* o *respuesta*, en distintos grupos o muestras definidas por una variable cualitativa, llamada *variable independiente* o *factor*. Las distintas categorías del factor que definen los grupos a comparar se conocen como *niveles* o *tratamientos* del factor.

Se trata, por tanto, de una generalización de la *prueba T para la comparación de medias de dos muestras independientes*, para diseños experimentales con más de dos muestras. Y se diferencia de un análisis de regresión simple, donde tanto la variable dependiente como la independiente eran cuantitativas, en que en el análisis de la varianza de un factor, la variable independiente o factor es una variable cualitativa.

Un ejemplo de aplicación de esta técnica podría ser la comparación del nivel de colesterol medio según el grupo sanguíneo. En este caso, la dependiente o factor es el grupo sanguíneo, con cuatro niveles (A, B, O, AB), mientras que la variable respuesta es el nivel de colesterol.

Para comparar las medias de la variable respuesta según los diferentes niveles del factor, se plantea un contraste de hipótesis en el que la hipótesis nula, H_0, es que la variable respuesta tiene igual media en todos los niveles, mientras que la hipótesis alternativa, H_1, es que hay diferencias estadísticamente significativas entre al menos dos de las medias. Dicho contraste se realiza mediante la descomposición de la varianza total de la variable respuesta; de ahí procede el nombre de esta técnica: *ANOVA* (Analysis of Variance en inglés).

9.1.1 El contraste de ANOVA

La notación habitual en ANOVA es la siguiente:

k es el número de niveles del factor.

n_i es el tamaño de la muestra aleatoria correspondiente al nivel i-ésimo del factor.

$n = \sum_{i=1}^{k} n_i$ es el número total de observaciones.

X_{ij} $(i = 1, ..., k;\ j = 1, ..., n_i)$ es una variable aleatoria que indica la respuesta de la j-ésimo individuo al i-ésimo nivel del factor.

x_{ij} es el valor concreto, en una muestra dada, de la variable X_{ij}.

Niveles del Factor			
1	2	$\cdots$	k
X_{11}	X_{21}	$\cdots$	X_{k1}
X_{12}	X_{22}	$\cdots$	X_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1n_1}	X_{2n_2}	$\cdots$	X_{kn_k}

μ_i es la media de la población del nivel i.

$\bar{X}_i = \sum_{j=1}^{n_i} X_{ij}/n_i$ es la variable media muestral del nivel i, y estimador de μ_i.

$\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$ es la estimación concreta para una muestra dada de la variable media muestral del nivel i.

μ es la media global de la población (incluidos todos los niveles).

$\bar{X} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}/n$ es la variable media muestral de todas las respuestas, y estimador de μ.

$\bar{x} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}/n$ es la estimación concreta para una muestra dada de la variable media muestral.

Con esta notación podemos expresar la variable respuesta mediante un modelo matemático que la descompone en componentes atribuibles a distintas causas:

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i),$$

es decir, la respuesta j-ésima en el nivel i-ésimo puede descomponerse como resultado de una media global, más la desviación con respecto a la media global debida al hecho de que recibe el tratamiento i-ésimo, más una nueva desviación con respecto a la media del nivel debida a influencias aleatorias.

Sobre este modelo se plantea la hipótesis nula: las medias correspondientes a todos los niveles son iguales; y su correspondiente alternativa: al menos hay dos medias de nivel que son diferentes.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_1 : \mu_i \neq \mu_j \text{ para algún } i \neq j.$$

Para poder realizar el contraste con este modelo es necesario plantear ciertas hipótesis estructurales (supuestos del modelo):

Independencia Las k muestras, correspondientes a los k niveles del factor, representan muestras aleatorias independientes de k poblaciones con medias $\mu_1 = \mu_2 = \cdots = \mu_k$ desconocidas.

Normalidad Cada una de las k poblaciones es normal.

Homocedasticidad Cada una de las k poblaciones tiene la misma varianza σ^2.

Teniendo en cuenta la hipótesis nula y los supuestos del modelo, si se sustituye en el modelo las medias poblacionales por sus correspondientes estimadores muestrales, se tiene

$$X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i),$$

o lo que es lo mismo,

$$X_{ij} - \bar{X} = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i).$$

Elevando al cuadrado y teniendo en cuenta las propiedades de los sumatorios, se llega a la ecuación que recibe el nombre de *identidad de la suma de cuadrados*:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2,$$

donde:

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2$ recibe el nombre de *suma total de cuadrados*, (STC), y es la suma de cuadrados de las desviaciones con respecto a la media global; por lo tanto, una medida de la variabilidad total de los datos.

$\sum_{j=1}^{k} n_i(\bar{X}_i-\bar{X})^2$ recibe el nombre de *suma de cuadrados de los tratamientos o suma de cuadrados intergrupos*, $(SCInter)$, y es la suma ponderada de cuadrados de las desviaciones de la media de cada nivel con respecto a la media global; por lo tanto, una medida de la variabilidad atribuida al hecho de que se utilizan diferentes niveles o tratamientos.

$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2$ recibe el nombre de *suma de cuadrados residual o suma de cuadrados intragrupos*, $(SCIntra)$, y es la suma de cuadrados de las desviaciones de las observaciones con respecto a las medias de los sus respectivos niveles o tratamientos; por lo tanto, una medida de la variabilidad en los datos atribuida a las fluctuaciones aleatorias dentro del mismo nivel.

Con esta notación la identidad de suma de cuadrados se expresa:

$$SCT = SCInter + SCIntra$$

Y un último paso para llegar al estadístico que permitirá contrastar H_0, es la definición de los *Cuadrados Medios*, que se obtienen al dividir cada una de las sumas de cuadrados por sus correspondientes grados de libertad. Para SCT el número de grados de libertad es $n-1$; para $SCInter$ es $k-1$; y para $SCIntra$ es $n-k$. Por lo tanto,

$$CMT = \frac{SCT}{n-1}$$
$$CMInter = \frac{SCInter}{k-1}$$
$$CMIntra = \frac{SCIntra}{n-k}$$

Y se podría demostrar que, en el supuesto de ser cierta la hipótesis nula y los supuestos del modelo, el cociente:

$$\frac{CMInter}{CMIntra}$$

sigue una distribución F de Fisher con $k-1$ y $n-k$ grados de libertad.

De esta forma, si H_0 es cierta, el valor del cociente para un conjunto de muestras dado, estará próximo a 0 (aún siendo siempre mayor que 0); pero si no se cumple H_0 crece la variabilidad intergrupos y la estimación del estadístico crece. En definitiva, realizaremos un contraste de hipótesis unilateral con cola a la derecha de igualdad de varianzas, y para ello calcularemos el p-valor de la estimación de F obtenida y aceptaremos o rechazaremos en función del nivel de significación fijado.

Tabla de ANOVA

Todos los estadísticos planteados en el apartado anterior se recogen en una tabla denominada Tabla de ANOVA, en la que se ponen los resultados de las estimaciones de dichos estadísticos en las muestras concretas objeto de estudio. Esas tablas también son las que aportan como resultado de cualquier ANOVA los programas estadísticos, que suelen añadir al final de la tabla el p-valor del estadístico F calculado, y que permite aceptar o rechazar la hipótesis nula de que las medias correspondientes a todos los niveles del factor son iguales.

	Suma de cuadrados	Grados de libertad	Cuadrados medios	Estadístico F	p-valor
Intergrupos	$SCInter$	$k-1$	$CMInter = \frac{SCInter}{k-1}$	$f = \frac{CMInter}{CMIntra}$	$P(F > f)$
Intragrupos	$SCIntra$	$n-k$	$CMIntra = \frac{SCIntra}{n-k}$		
Total	SCT	$n-1$			

9.1.2 Test de comparaciones múltiples y por parejas

Una vez realizado el ANOVA de un factor para comparar las k medias correspondientes a los k niveles o tratamientos del factor, se puede concluir aceptando la hipótesis nula, en cuyo caso se da por concluido el análisis de los datos en cuanto a detección de diferencias entre los niveles, o rechazándola, en cuyo caso es natural continuar con el análisis para tratar de localizar con precisión dónde está la diferencia, cuáles son los niveles cuyas respuestas son estadísticamente diferentes.

En el segundo caso, hay varios métodos que permiten detectar las diferencias entre las medias de los diferentes niveles, y que reciben el nombre de *test de comparaciones múltiples*. A su vez este tipo de test se suelen clasificar en:

Test de comparaciones por parejas Su objetivo es la comparación una a una de todas las posibles parejas de medias que se pueden tomar al considerar los diferentes niveles. Su resultado es una tabla en la que se reflejan las diferencias entre todas las posibles parejas y los intervalos de confianza para dichas diferencias, con la indicación de si hay o no diferencias significativas entre las mismas. Hay que aclarar que los intervalos obtenidos no son los mismos que resultarían si se considera cada pareja de medias por separado, ya que el rechazo de H_0 en el contraste general de ANOVA implica la aceptación de una hipótesis alternativa en la que están involucrados varios contrastes individuales a su vez; y si queremos mantener un nivel de significación α en el general, en los individuales debemos utilizar un α' considerablemente más pequeño.

Test de rango múltiple Su objetivo es la identificación de subconjuntos homogéneos de medias que no se diferencian entre sí.

Para los primeros se puede utilizar el test de Bonferroni; para los segundos, el test de Duncan; y para ambas categorías a la vez los test HSD de Tukey y Scheffé.

9.2 Ejercicios resueltos

1. Se realiza un estudio para comparar la eficacia de tres programas terapéuticos para el tratamiento del acné. Se emplean tres métodos:

   (a) Lavado, dos veces al día, con cepillo de polietileno y un jabón abrasivo, junto con el uso diario de 250 mg de tetraciclina.

   (b) Aplicación de crema de tretinoína, evitar el sol, lavado dos veces al día con un jabón emulsionante y agua, y utilización dos veces al día de 250 mg de tetraciclina.

   (c) Evitar el agua, lavado dos veces al día con un limpiador sin lípidos y uso de crema de tretinoína y de peróxido benzoílico.

   En el estudio participan 35 pacientes. Se separó aleatoriamente a estos pacientes en tres subgrupos de tamaños 10, 12 y 13, a los que se asignó respectivamente los tratamientos I, II, y III. Después de 16 semanas se anotó para cada paciente el porcentaje de mejoría en el número de lesiones.

Tratamiento					
I		II		III	
48.6	50.8	68.0	71.9	67.5	61.4
49.4	47.1	67.0	71.5	62.5	67.4
50.1	52.5	70.1	69.9	64.2	65.4
49.8	49.0	64.5	68.9	62.5	63.2
50.6	46.7	68.0	67.8	63.9	61.2
		68.3	68.9	64.8	60.5
				62, 3	

   (a) Crear las variables tratamiento y mejora e introducir los datos de la muestra.

   i Para cualquier ANOVA, a pesar de que la variable tratamiento es cualitativa, conviene que se introduzca como cuantitativa, ya que SPSS sólo admite como factor de clasificación variables cuantitativas. Si hubiese que mostrar los diferentes niveles de la variable factor como cualidades, bastará con asignarles etiquetas).

   (b) Dibujar el diagrama de dispersión. ¿Qué conclusiones sacas de la nube de puntos?

   i
   i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos*.
   ii. En el cuadro de diálogo que aparece, seleccionar la opción *Dispersión simple* y hacer click en el botón *Definir*.
   iii. En el siguiente cuadro de diálogo que aparece, seleccionar la variable mejora al campo *Eje Y* y la variable tratamiento al campo *Eje X*, y hacer click sobre el botón *Aceptar*.

   (c) Obtener la tabla de ANOVA correspondiente al problema. ¿Se puede concluir que los tres tratamientos tienen el mismo efecto medio con un nivel de significación de 0.05?

   i
   i. Seleccionar el menú *Analizar→Comparar medias→ANOVA de un factor*.
   ii. En el cuadro de diálogo que aparece, seleccionar la variable mejora al campo *Lista de dependientes* y la variable tratamiento al campo *factor*, y hacer click sobre el botón *Aceptar*.

   (d) Obtener la tabla de ANOVA correspondiente al problema pero que además muestre los intervalos de confianza de los 3 diferentes tratamientos, con una significación de 0.05, y diversos estadísticos de cada unos de ellos.

Repetir los mismos pasos del apartado anterior, haciendo click en el botón *Opciones* del último cuadro de diálogo y activar la opción *Descriptivos*.

(e) Si se puede concluir que los tratamientos no han tenido el mismo efecto medio, ¿entre qué parejas de tratamientos hay diferencias estadísticamente significativas?

Repetir los mismos pasos del apartado anterior, haciendo click en el botón *Post hoc* del último cuadro de diálogo y activar la opción de *Bonferroni* con un nivel de significación de 0.05.

(f) Si se puede concluir que los tratamientos no han tenido el mismo efecto medio, ¿cuáles son los grupos homogéneos (grupos con un comportamiento similar en cuanto a mejora se refiere) de tratamientos que se pueden establecer?

Repetir los mismos pasos del apartado anterior, activando la opción de *Duncan*.

2. Se sospecha que hay diferencias en la preparación del examen de selectividad entre los diferentes centros de bachillerato de una ciudad. Con el fin de comprobarlo, de cada uno de los 5 centros, se eligieron 8 alumnos al azar, con la condición de que hubieran cursado las mismas asignaturas, y se anotaron las notas que obtuvieron en el examen de selectividad. Los resultados fueron:

Centros				
1	2	3	4	5
5.5	6.1	4.9	3.2	6.7
5.2	7.2	5.5	3.3	5.8
5.9	5.5	6.1	5.5	5.4
7.1	6.7	6.1	5.7	5.5
6.2	7.6	6.2	6.0	4.9
5.9	5.9	6.4	6.1	6.2
5.3	8.1	6.9	4.7	6.1
6.2	8.3	4.5	5.1	7.0

(a) Crear las variables nota y centro e introducir los datos de la muestra.

(b) Dibujar el diagrama de dispersión. ¿Qué conclusiones sacas sobre la nota media de selectividad en los distintos centros?

i. Seleccionar el menú *Gráficos→Cuadros de diálogo antiguos→Dispersión/Puntos*.
ii. En el cuadro de diálogo que aparece, seleccionar la opción *Dispersión simple* y hacer click en el botón *Definir*.
iii. En el siguiente cuadro de diálogo que aparece, seleccionar la variable nota al campo *Eje Y* y la variable centro al campo *Eje X*, y hacer click sobre el botón *Aceptar*.

(c) Realizar el contraste de ANOVA. ¿Se puede confimar la sospecha de que hay diferencias entre las notas medias de los centros?

i. Seleccionar el menú *Analizar→Comparar medias→ANOVA de un factor*.
ii. En el cuadro de diálogo que aparece, seleccionar la variable nota al campo *Lista de dependientes* y la variable centro al campo *factor*, y hacer click sobre el botón *Aceptar*.

(d) ¿Qué centros son los mejores en la preparación de la selectividad?

i Repetir los mismos pasos del apartado anterior, haciendo click en el botón *Post hoc* del último cuadro de diálogo y activar las opciones de *Bonferroni*, para ver los intervalos de diferencias entre centros, y de *Duncan* para establecer grupos de comportamiento homogéneo.

9.3 Ejercicios propuestos

1. Se midió la frecuencia cardíaca (latidos por minuto) en cuatro grupos de adultos; controles normales (A), pacientes con angina (B), individuos con arritmias cardíacas (C) y pacientes recuperados del infarto de miocardio (D). Los resultados son los siguientes:

A	B	C	D
83	81	75	61
61	65	68	75
80	77	80	78
63	87	80	80
67	95	74	68
89	89	78	65
71	103	69	68
73	89	72	69
70	78	76	70
66	83	75	79
57	91	69	61

   ¿Proporcionan estos datos la suficiente evidencia para indicar una diferencia en la frecuencia cardiaca media entre esos cuatro tipos de pacientes?. Considerar $\alpha = 0.05$.

2. Se midió la frecuencia respiratoria (inspiraciones por minuto) en ocho animales de laboratorio y con tres niveles diferentes de exposición al monóxido de carbono. Los resultados son los siguientes:

Nivel de exposición		
Bajo	Moderado	Alto
36	43	45
33	38	39
35	41	33
39	34	39
41	28	33
41	44	26
44	30	39
45	31	29

   Con base en estos datos, ¿es posible concluir que los tres niveles de exposición, en promedio, tienen un efecto diferente sobre la frecuencia respiratoria? Tomar $\alpha = 0,05$.

10 — Contrastes Basados en el Estadístico χ^2

10.1 Fundamentos teóricos

Existen multitud de situaciones en el ámbito de la salud, o en cualquier otro ámbito, en las que el investigador está interesado en determinar posibles relaciones entre variables cualitativas. Un ejemplo podría ser el estudio de si existe relación entre las complicaciones tras una intervención quirúrgica y el sexo del paciente, o bien el hospital en el que se lleva a cabo la intervención. En este caso, todas las técnicas de inferencia vistas hasta ahora para variables cuantitativas no son aplicables, y para ello utilizaremos un contraste de hipótesis basado en el estadístico χ^2 (Chi-cuadrado).

Sin embargo, aunque éste sea su aspecto más conocido, el uso del test no se limita al estudio de la posible relación entre variables cualitativas, y también se aplica para comprobar el ajuste de la distribución muestral de una variable, ya sea cualitativa o cuantitativa, a su hipotético modelo teórico de distribución.

En general, este tipo de tests consiste en tomar una muestra y observar si hay diferencia significativa entre las *frecuencias observadas* y las especificadas por la ley teórica del modelo que se contrasta, también denominadas *frecuencias esperadas*.

Podríamos decir que existen dos grandes bloques de aplicaciones básicas en el uso del test de la χ^2:

1. **Test de ajuste de distribuciones**. Es un contraste de significación para saber si los datos de la población, de la cual hemos extraído una muestra, son conforme a una ley de distribución teórica que sospechamos que es la correcta.

   Por ejemplo: disponemos de 400 datos que, a priori, siguen una distribución de probabilidad uniforme, pero ¿es estadísticamente cierto que se ajusten a dicho tipo de distribución?

2. **Test para tablas de contingencia.** En las que se parte de la tabla de frecuencias bidimensional para las distintas modalidades de las variables cualitativas. Aunque muy a menudo el test de la χ^2 aplicado en tablas de contingencia se denomina prueba de independencia, en realidad se aplica en dos diseños experimentales diferentes, que hacen que se clasifique en dos bloques diferentes:

   (a) **Prueba de independencia**. Mediante la que el investigador pretende estudiar la relación entre dos variables cualitativas en una población.

   Por ejemplo: tenemos una muestra de 200 enfermos (el investigador tan sólo controla el total en una muestra) operados de apendicitis en 4 hospitales diferentes y queremos ver si hay relación entre la posible infección postoperatoria y el hospital en el que el paciente ha sido operado.

(b) **Prueba de homogeneidad**. Mediante la que el investigador pretende ver si la proporción de una determinada característica es la misma en poblaciones, tal vez, diferentes.

Por ejemplo: tenemos dos muestras diferentes, una de ellas de 100 individuos VIH positivos, y otra de 600 VIH negativos (el investigador controla el total en ambas muestras), y queremos analizar si la proporción de individuos con problemas gastrointestinales es la misma en ambas.

Por último, aunque el test de la Chi-cuadrado es muy importante en el análisis de las relaciones entre variables cualitativas, su aplicación puede conducir a errores en determinadas situaciones; sobre todo cuando los tamaños muestrales son pequeños, lo cual conduce a que en algunas categorías apenas tengamos individuos y ello invalida los supuestos de aplicación del test; y también cuando tenemos variables cualitativas analizadas en los mismos individuos pero en diferentes tiempos, es decir, mediante datos pareados. Para el primer caso, cuando el número de individuos en alguna categoría es muy pequeño, se utiliza el test Exacto de Fisher, mientras que en el segundo, con datos pareados, se utiliza el test de McNemar.

10.1.1 Contraste χ^2 de Pearson para ajuste de distribuciones

Es el contrate de ajuste más antiguo y es válido para todo tipo de distribuciones. Para analizar una muestra de una variable agrupada en categorías (aunque sea cuantitativa), evaluando una hipótesis previa sobre probabilidad de cada modalidad o categoría, se realiza un contraste de hipótesis Chi-cuadrado de bondad de ajuste.

El contraste se basa en hacer un recuento de los datos y comparar las frecuencias observadas de cada una de las modalidades con las frecuencias esperadas por el modelo teórico que se contrasta. De este modo, se calcula el estadístico:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

donde O_i son las frecuencias observadas en la muestra en la modalidad i, y E_i son las frecuencias esperadas para la misma modalidad según el modelo teórico. Las frecuencias esperadas se calculan multiplicando el tamaño de la muestra por la probabilidad de la correspondiente modalidad según el modelo teórico, es decir $E_i = np_i$, siendo p_i la probabilidad de la modalidad i.

Si la población de la que se ha obtenido la muestra sigue el modelo de distribución teórica, el estadístico anterior se distribuye como χ^2 con $k-1$ grados de libertad, donde k es el número de modalidades de la variable. Un valor del estadístico χ^2 grande indica que las distribuciones de las frecuencias observadas y esperadas son bastantes diferentes, mientras que un valor pequeño del estadístico indica que hay poca diferencia entre ellas.

La prueba χ^2 de bondad del ajuste es válida si todas las frecuencias esperadas son mayores o iguales que 1 y no más de un 20% de ellas tienen frecuencias esperadas menores que 5. Si no se cumple lo anterior, entonces las categorías implicadas deben combinarse con categorías adyacentes para garantizar que todas cumplen la condición. Si las categorías corresponden a variables cuantitativas categorizadas, no tienen necesariamente que corresponder a la misma amplitud de variable.

10.1.2 Contraste χ^2 en tablas de contingencia

Como ya hemos visto, el contraste de la χ^2 en tablas de contingencia sirve para establecer relaciones entre variables cualitativas (o cuantitativas categorizadas), entre las que no puede realizarse un análisis de regresión y correlación, y tanto para determinar independencia entre variables, como homogeneidad entre poblaciones (igual proporción de una determinada característica). Para ello, describimos el proceso metodológico en el caso de independencia entre variables, que en la práctica, y aunque conceptualmente son casos diferentes, es el mismo también para la homogeneidad entre poblaciones.

Por tablas de contingencia se entiende aquellas tablas de doble entrada donde se realiza una clasificación de la muestra de acuerdo a un doble criterio de clasificación. Por ejemplo, la clasificación de unos individuos de acuerdo a su sexo y su grupo sanguíneo crearía una tabla donde

cada celda de la tabla representaría la frecuencia bivariante de las características correspondientes a su fila y columna (por ejemplo mujeres de grupo sanguíneo A). Si se toma una muestra aleatoria de tamaño n en la que se miden ambas variables y se representan las frecuencias de los pares observados en una tabla bidimensional, tenemos:

X/Y	y_j	
x_i	n_{ij}	n_i
	n_j	n

Donde n_{ij} es la frecuencia absoluta del par (x_i, y_j), n_i es la frecuencia marginal de la modalidad x_i y n_j es la frecuencia marginal de la modalidad y_j. Dichas frecuencias aparecen en los márgenes de la tabla de contingencia sumando las frecuencias por filas y columnas, y por ello se conocen como frecuencias marginales.

Siguiendo un procedimiento parecido al del apartado anterior, se comparan las frecuencias observadas en la muestra (frecuencias reales) con las frecuencias esperadas (frecuencias teóricas). Para ello, calculamos la probabilidad de cada casilla de la tabla teniendo en cuenta que si ambas variables son independientes la probabilidad de cada celda surge como un producto de probabilidades (probabilidad de la intersección de dos sucesos independientes) $p_{ij} = p_i p_j = \frac{n_i}{n}\frac{n_j}{n}$. De este modo, obtenemos la frecuencia esperada como:

$$E_i = np_{ij} = n\frac{n_i}{n}\frac{n_j}{n} = \frac{n_i n_j}{n},$$

Y con ello se calcula el estadístico de la Chi-cuadrado de Pearson:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

En el caso de que X e Y fuesen independientes, este estadístico presenta una distribución Chi-cuadrado con $(f-1)(c-1)$ grados de libertad, donde f es el número de filas de la tabla de contingencia y c el número de columnas. Un valor del estadístico Chi-cuadrado grande indica que las distribuciones de las frecuencias observadas y esperadas son bastantes diferentes, y por lo tanto falta de independencia; mientras que un valor pequeño del estadístico indica que hay poca diferencia entre ellas, lo cual nos indica que son independientes.

Este test es adecuado si las frecuencias esperadas para cada celda valen como mínimo 1 y no más de un 20% de ellas tienen frecuencias esperadas menores que 5. En el caso de una tabla 2x2, estas cifras se alcanzan sólo cuando ninguna frecuencia esperada es menor que 5. Si esto no se cumple, puede, entre otras, utilizarse una prueba para pequeñas muestras llamada prueba exacta de Fisher.

10.1.3 Test exacto de Fisher

Este test se puede utilizar cuando no se cumplan las condiciones necesarias para aplicar el test de la Chi cuadrado (más de un 20% de las frecuencias esperadas para cada celda son menores que 5). Aunque, dada la gran cantidad de cálculos necesarios para llegar al resultado final del test, los programas de Estadística sólo lo calculaban para tablas de contingencia 2x2, en versiones actuales pueden incorporar módulos específicos que amplían la capacidad del núcleo del programa base, y que sí que permiten evaluar el test de Fisher en tablas con más categorías. Por ejemplo, en PASW, si se ha instalado el módulo de Pruebas Exactas, se puede calcular el test Exacto de Fisher en tablas generales con F categorías en las filas y C categorías en las columnas.

El test Exacto de Fisher está basado en la distribución exacta de los datos y no en aproximaciones asintóticas, y presupone que los marginales de la tabla de contingencia están fijos. El procedimiento para su cálculo consiste en evaluar la probabilidad asociada, bajo el supuesto de independencia, a todas las tablas que se pueden formar con los mismos totales marginales que los

datos observados y variando las frecuencias de cada casilla para contemplar todas las situaciones en las que hay un desequilibrio de proporciones tan grande o más que en la tabla analizada. Para el cálculo de la probabilidad asociada a cada tabla se utiliza la función de probabilidad de una variable discreta hipergeométrica.

Aunque generalmente el test Exacto de Fisher es más conservador que la Chi cuadrado (resulta más complicado que detecte diferencias estadísticamente significativas entre las proporciones), no obstante tiene la ventaja de que se puede aplicar sin ninguna restricción en las frecuencias de las casillas de la tabla de contingencia.

10.1.4 Test de McNemar para datos emparejados

Hasta ahora hemos supuesto que las muestras a comparar eran independientes, es decir dos grupos diferentes en los que se había mirado una determinada característica. Por lo tanto, hemos realizado comparaciones de proporciones de individuos que presentan una determinada característica en dos grupos distintos, pero también nos podemos plantear comparar la proporción de individuos que presentan esa característica en un mismo grupo de individuos pero analizados en dos momentos diferentes. En este último caso se habla comparación de proporciones en datos emparejados, pareados o apareados.

Por ejemplo, si queremos ver si existen o no diferencias en la mejora de los síntomas de una determinada enfermedad, y para ello aplicamos dos fármacos distintos a un grupo de individuos en dos momentos diferentes en los que hayan contraído la misma enfermedad. En este caso, podría pensarse que resultaría adecuado aplicar tanto la chi cuadrado como el test exacto de Fisher para determinar si existe diferencias entre ambos fármacos en la proporción de pacientes curados, pero aquí hay una diferencia fundamental con los casos anteriores y es que sólo tenemos un grupo de pacientes y no dos. En este tipo de estudios se reduce considerablemente la variabilidad aleatoria, ya que es un mismo individuo el que se somete a los dos tratamientos, y el que manifieste mejoría en los síntomas no dependerá de otros factores tan importantes como, por ejemplo, la edad, el sexo o el tipo de alimentación, que pueden influir pero que tal vez no se controlen adecuadamente en un diseño de grupos independientes. Al reducir la variabilidad aleatoria mediante datos emparejados, pequeñas diferencias entre las proporciones pueden llegar a ser significativas, incluso con tamaños muestrales pequeños, lo cual se traduce en que este tipo de diseños del experimento resultan más eficientes a la hora de obtener resultados estadísticamente significativos.

No obstante, nuevos diseños implican nuevas formas de tratar los datos, y el procedimiento más adecuado es el que se utiliza en el test de McNemar para datos emparejados. Para su aplicación en nuestro ejemplo, se debería construir una tabla con 4 casillas en las que se contabilicen: las personas que han obtenido una mejoría de los síntomas con los dos fármacos, los que han obtenido con el primero y no con el segundo, los que han obtenido con el segundo y no con el primero y los que no han obtenido mejoría con ninguno.

Mejoría 1º \ Mejoría 2º	Sí	No	Totales
Sí	a	b	$a+b$
No	c	d	$c+d$
Totales	$a+c$	$b+d$	$n=a+b+c+d$

Con ello, la proporción muestral de pacientes que han experimentado mejoría con el medicamento 1 vale: $\widehat{p}_1 = (a+b)/n$, e igualmente con el 2: $\widehat{p}_2 = (a+c)/n$, y podemos plantear el contraste cuya hipótesis nula es que no hay diferencia de proporciones poblacionales entre ambos medicamentos: $H_0 : p_1 = p_2$, que puede realizarse sin más que tener en cuenta el oportuno intervalo de confianza para la diferencia de proporciones, o también que, en el supuesto de igualdad de proporciones

$$z = \frac{b-c}{\sqrt{b+c}},$$

es un estadístico que sigue una distribución normal tipificada, y

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

,

es un estadístico que sigue una distribución Chi-cuadrado con un grado de libertad. Con cualquiera de ellos, se podría calcular el p-valor del contraste.

10.2 Ejercicios Resueltos

1. Dadas dos parejas de genes Aa y Bb, la descendencia del cruce efectuado según las leyes de Mendel, debe estar compuesto del siguiente modo:

Fenotipo	Frecuencias Relativas
AB	$9/16 = 0{,}5625$
Ab	$3/16 = 0{,}1875$
aB	$3/16 = 0{,}1875$
ab	$1/16 = 0{,}0625$

Elegidos 300 individuos al azar de cierta población, se observa la siguiente distribución de frecuencias:

Fenotipo	Frecuencias Observadas
AB	165
Ab	47
aB	67
ab	21

Se pide

(a) Crear las variables fenotipo y frecuencia e introducir los datos de la muestra.

(b) Ponderar los datos mediante la variable frecuencia.

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable frecuencia* y hacer click en el botón *Aceptar*.

(c) Comprobar si esta muestra cumple las leyes de Mendel.

i. Seleccionar el menú Analizar→Pruebas no paramétricas→Cuadro de diálogo antiguos→Chi-cuadrado.
ii. En el cuadro de diálogo que aparece seleccionar la variable fenotipo al campo *Lista Contrastar Variables*, y en Valores esperados marcar la opción *valores* e introducir las proporciones según las leyes de Mendel y siguiendo el orden en el que aparecen los fenotipos, y hacer click sobre el botón *Aceptar*.

(d) A la vista de los resultados del contraste, ¿se puede aceptar que se cumplen las leyes de Mendel en los individuos de dicha población?

2. En un estudio sobre úlceras pépticas se determinó el grupo sanguíneo de 1655 pacientes ulcerosos y 10000 controles, los datos fueron:

	O	A	B	AB
Paciente	911	579	124	41
Controles	4578	4219	890	313

(a) Crear las variables participantes, grupo_sanguíneo y frecuencia e introducir los datos.

(b) Ponderar los datos mediante la variable frecuencia.

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable frecuencia* y hacer click

en el botón *Aceptar*.

(c) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

i

i. Seleccionar el menú *Analizar→Estadísticos Descriptivos→Tablas de contingencia*.
ii. En el cuadro de diálogo que aparece, seleccionar la variable participantes al campo *Filas* y la variable grupo_sanguíneo al campo *Columnas*, y hacer click sobre el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece, marcar la casilla Chi-cuadrado y hacer click en el botón *Continuar*.
iv. En el cuadro de diálogo inicial, hacer click sobre el botón *Casillas*.
v. En el cuadro de diálogo que aparece, marcar las casillas Frecuencias observadas y Esperadas, y hacer click sobre el botón *Continuar* y Aceptar.

(d) A la vista de los resultados del contraste, ¿existe alguna relación entre el grupo sanguíneo y la úlcera péptica?, es decir, ¿se puede concluir que la proporción de pacientes y de controles es diferente dependiendo del grupo sanguíneo?

3. Mitchell et al. (1976, Annals of Human Biology), partiendo de una muestra de 478 individuos, estudiaron la distribución de los grupos sanguíneos en varias regiones del sur-oeste de Escocia, obteniendo los resultados que se muestran:

	Eskdale	Annandale	Nithsdale	
A	33	54	98	185
B	6	14	35	55
O	56	52	115	223
AB	5	5	5	15
	100	125	253	478

(a) Crear las variables grupo_sanguíneo, región y frecuencia e introducir los datos.

(b) Ponderar el estudio, por la variable frecuencia

i

i. Seleccionar el menú *Datos→Ponderar casos*.
ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable frecuencia* y hacer click en el botón *Aceptar*.

(c) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

i

i. Seleccionar el menú *Analizar→Estadísticos Descriptivos→Tablas de contingencia*.
ii. En el cuadro de diálogo que aparece, seleccionar la variable grupo_sanguíneo al campo *Filas* y la variable región al campo *Columnas*, y hacer click sobre el botón *Estadísticos*.
iii. En el cuadro de diálogo que aparece, marcar la casilla Chi-cuadrado y hacer click en el botón *Continuar*.
iv. En el cuadro de diálogo inicial, hacer click sobre el botón *Casillas*.
v. En el cuadro de diálogo que aparece, marcar las casillas Frecuencias observadas y Esperadas, y hacer click sobre el botón *Continuar* y Aceptar.

(d) En vista de los resultados del contraste, ¿se distribuyen los grupos sanguíneos de igual manera en las diferentes regiones?

4. En un estudio para saber si el habito de fumar está relacionado con el sexo, se ha preguntado a 26 personas. De los 9 hombres consultados 2 respondieron que fumaban, mientras que de las 17 mujeres consultadas, 6 fumaban. ¿Podemos afirmar que existe relación entre ambas variables?

   (a) Crear las variables sexo, fuma y frecuencia e introducir los datos.

   (b) Ponderar el estudio, por la variable frecuencia

   i

   i. Seleccionar el menú *Datos→Ponderar casos*.
   ii. En el cuadro de diálogo resultante activar la opción *Ponderar casos mediante*, seleccionar la variable frecuencia en el campo *Variable frecuencia* y hacer click en el botón *Aceptar*.

   (c) Construir la tabla de contingencia y realizar el contraste Chi-cuadrado.

   i

   i. Seleccionar el menú *Analizar→Estadísticos Descriptivos→Tablas de contingencia*.
   ii. En el cuadro de diálogo que aparece, seleccionar la variable sexo al campo *Filas* y la variable fuma al campo *Columnas*, y hacer click sobre el botón *Estadísticos*.
   iii. En el cuadro de diálogo que aparece, marcar la casilla Chi-cuadrado y hacer click en el botón *Continuar*.
   iv. En el cuadro de diálogo inicial, hacer click sobre el botón *Casillas*.
   v. En el cuadro de diálogo que aparece, marcar las casillas Frecuencias observadas y Esperadas, y hacer click sobre el botón *Continuar* y Aceptar.

   (d) En vista de los resultados del contraste, ¿se distribuyen los fumadores de igual manera en ambos sexos?

   i

   En este caso el procedimiento a seguir es igual que para la Chi cuadrado, pero vemos que ahora no se cumplen las condiciones para poder aplicar esta prueba, por eso nos tendremos que fijar en el Estadístico exacto de Fisher, que si podemos aplicar, teniendo en cuenta si estamos realizando un contraste bilateral o unilateral.

5. Para probar la eficacia de dos fármacos diferentes contra las migrañas, se seleccionaron a 20 personas que padecían migrañas habitualmente, y se les dió a tomar a cada uno los fármacos en momentos diferentes. Luego se les preguntó si habían obtenido mejoría o no con el fármaco tomado. Los resultados fueron los siguientes:

	1	2	3	4	5	6	7	8	9	10
Fármaco 1	Sí	Sí	Sí	Sí	Sí	No	Sí	No	Sí	Sí
Fármaco 2	No	No	Sí	No	Sí	Sí	No	No	No	No

	11	12	13	14	15	16	17	18	19	20
Fármaco 1	Sí	No	Sí	No	Sí	Sí	Sí	No	Sí	Sí
Fármaco 2	Sí	No	Sí	No	No	Sí	No	Sí	No	No

   (a) Crear las variables Mejora_Farmaco1, y Mejora_Farmaco2 e introducir los datos.

   (b) Construir la tabla de contingencia y realizar el contraste de McNemar.

   i

   i. Seleccionar el menú *Analizar→Estadísticos Descriptivos→Tablas de contingencia*.
   ii. En el cuadro de diálogo que aparece, seleccionar la variable Mejora_Farmaco1 al campo *Filas* y la variable Mejora_Farmaco2 al campo *Columnas*, y hacer click sobre el botón *Estadísticos*.

iii. En el cuadro de diálogo que aparece, marcar la casilla McNemar y hacer click en el botón *Continuar*.

iv. En el cuadro de diálogo inicial, hacer click sobre el botón *Casillas*.

v. En el cuadro de diálogo que aparece, marcar las casillas Frecuencias observadas y Esperadas, y hacer click sobre el botón *Continuar* y Aceptar.

(c) En vista de los resultados del contraste, ¿podemos afirmar que existen diferencias significativas entre los dos fármacos?

i Otra forma de realizar este mismo contraste, sería seleccionado el menú *Analizar→Pruebas no paramétricas→Cuadros de diálogo antiguos→2 muestras relacionadas*. Luego pasar las dos variables a contrastar al cuadro Contrastar pares, marcar la casilla McNemary hacer click sobre el botón Aceptar.

10.3 Ejercicios propuestos

1. Supongamos que queremos comprobar si un dado está bien equilibrado o no. Lo lanzamos 1200 veces, y obtenemos los siguientes resultados:

Número	Frecuencias de aparición
1	120
2	275
3	95
4	310
5	85
6	315

   (a) A la vista de los resultados, ¿se puede aceptar que el dado está bien equilibrado?

   (b) Nos dicen que, en este dado, los números pares aparecen con una frecuencias 3 veces superior a la de los impares. Contrastar dicha hipótesis.

2. Se realiza un estudio en una población de pacientes críticos hipotéticos y se observan, entre otras, dos variables, la evolución (si sobreviven SV o no NV) y la presencia o ausencia de coma, al ingreso. Se obtienen los siguientes resultados:

	No coma	Coma	
SV	484	37	521
NV	118	89	207
	602	126	728

   Nos preguntamos: ¿es el coma al ingreso un factor de riesgo para la mortalidad?

3. La recuperación producida por dos tratamientos distintos A y B, se clasifican en tres categorías: muy buena, buena y mala. Se administra el tratamiento A a 32 pacientes y el B a otros 28. De las 22 recuperaciones muy buenas, 10 corresponden al tratamiento A; de las 24 recuperaciones buenas, 14 corresponden al tratamiento A y de las 14 que tienen una mala recuperación, 8 corresponden al tratamiento A. ¿Son igualmente efectivos ambos tratamientos para la recuperación de los pacientes?

4. Para contrastar la hipótesis de que las mujeres tienen más éxito en sus estudios que los hombres, se ha tomado una muestra de 10 chicos y otra de 10 chicas que han sido examinados por un profesor que aprueba siempre al 40% de los alumnos presentados a examen. Teniendo en cuenta que sólo aprobaron 2 chicos, utiliza el test de hipótesis más adecuado para decidir si la citada hipótesis es cierta.

5. Se ha preguntado a los 150 alumnos de un curso, si estaban de acuerdo o no, con la metodología de enseñanza de dos profesores, distintos que les han dado clase en la asignatura de bioestadística. Los resultados se recogen en la siguiente tabla:

Profesor 1 \ Profesor 2	Opinión favorable	Opinión desfavorable
Opinión favorable	37	48
Opinión desfavorable	44	21

¿Podemos afirmar que existe diferente opinión por parte de los alumnos, sobre los dos profesores?