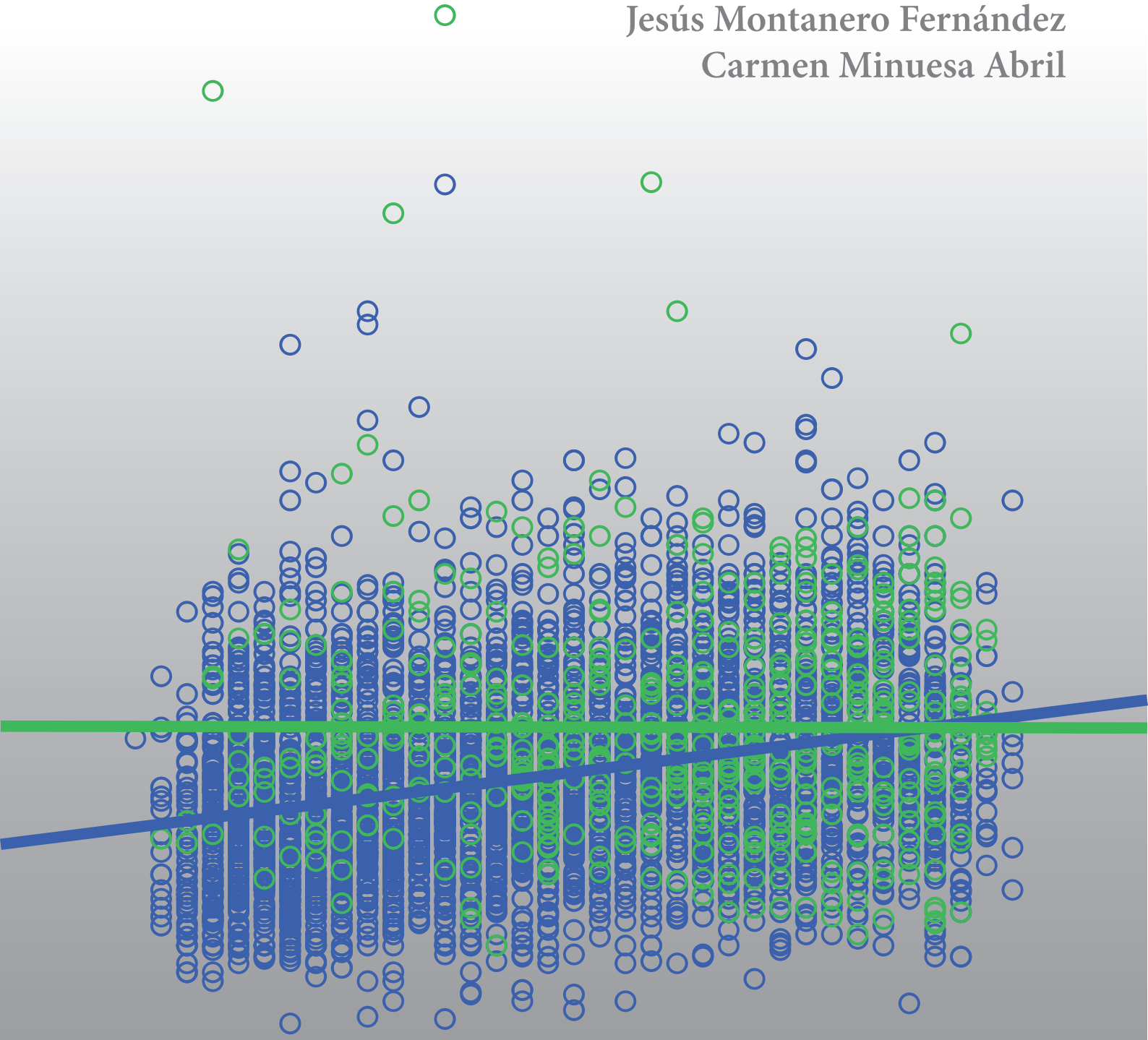
ESTADÍSTICA BÁSICA PARA CIENCIAS DE LA SALUD

Jesús Montanero Fernández
Carmen Minuesa Abril



UNIVERSIDAD DE EXTREMADURA

ÍNDICE

Estadística básica para Ciencias de la Salud

Jesús Montanero Fernández
Carmen Minuesa Abril

Estadística básica para Ciencias de la Salud

UNIVERSIDAD DE EXTREMADURA
UEX

Cáceres
2018

La presente publicación ha sido realizada en el marco de la ayuda FPU13/03213 concedida por el Ministerio de Educación, Cultura y Deporte, y del proyecto GR15013 de la Consejería de Economía e infraestructuras de la Junta de Extremadura, financiado por los Fondos Europeos de Desarrollo Regional.

Tipografía utilizada: Minion Pro (para cubierta) y CMU (páginas iniciales y texto de la obra)

Imagen de cubierta: Figura 2.24 de la obra

Edita:
Universidad de Extremadura. Servicio de Publicaciones
Plaza de Caldereros, 2. 10071 Cáceres (España)
Tel. 927 257 041; Fax 927 257 046
publicac@unex.es
http://www.unex.es/publicaciones

I.S.B.N.: 978-84-697-8323-8

Impreso en España - *Printed in Spain*

Impresión: Dosgraphic, s. l.

PORTADA ÍNDICE

A mis padres, mi hermana María José y Alfonso

Prólogo

El objetivo inicial de este manual es servir de apoyo en el estudio de la materia de Estadística en el Grado en Enfermería de la Universidad de Extremadura (UEx), aunque pensamos que puede ser de utilidad para cualquier estudiante o profesional de Ciencias de la Salud que desee entender y aplicar la Estadística a un nivel básico. Por tanto, nuestra intención no es profundizar en los aspectos más formales de la materia, ni abarcar métodos avanzados que vayan más allá de los contenidos que se imparten en unas 60 horas lectivas en este tipo de asignaturas. Tampoco pretendemos hacer hincapié en cuestiones relativas al cálculo. En lugar de ello, nos esforzaremos en facilitar la comprensión de los conceptos fundamentales, delegando la ejecución de los diferentes algoritmos en un programa estadístico.

El manual está estructurado en tres partes. La primera de ellas está dedicada al análisis de un conjunto concreto de datos; la segunda, a la posible generalización de dicho estudio y, por último, la tercera parte consiste en un tutorial sobre el funcionamiento del programa estadístico SPSS, por el que nos hemos decantado en nuestro caso para ejecutar los diferentes métodos. Dicha elección se debe simplemente a que la UEx dispone actualmente de licencia de red para el mismo y a que lo consideramos un apropiado para que los profesionales de Ciencias de la Salud apliquen las técnicas estadísticas de manera autónoma.

Badajoz, Junio de 2017

Jesús Montanero Fernández
Carmen Minuesa Abril

Índice general

Introducción

El estudio de la Estadística en Ciencias de la Salud, más conocida como Bioestadística, está motivado por la enorme incertidumbre que presentan los diferentes fenómenos a comprender, de ahí la necesidad de diseñar técnicas de recogida y tratamiento de datos con la idea de extraer la mayor información posible acerca de los mismos. Así, la Bioestadística podría entenderse como la metodología a seguir para aprender de las observaciones con el propósito de explicar los fenómenos biomédicos.

Aunque muchas personas puedan considerar esta definición insatisfactoria o decepcionante, el objetivo marcado peca en realidad de ambicioso y está condenado a la derrota en muchos casos. Efectivamente, y según se explica con detalle en [5], el tratamiento racional y objetivo de la información compite, en una batalla que suele perder, con una serie de automatismos psicológicos arraigados en nuestro cerebro, de intuiciones ventajosas desde un punto de vista evolutivo pero erróneas si se analizan matemáticamente. Por ejemplo, la generalizada ilusión y expectación que pueden llegar a generar los sorteos de lotería puede entenderse como un claro ejemplo de derrota de la Estadística.

Desde nuestro punto de vista entenderemos cada fenómeno observable como la suma de una componente cuyas causas están aparentemente controladas en el experimento (componente determinista) y otra sujeta a incertidumbre o azar[1]. El objetivo de la Estadística es, en general, delimitar esta última componente de la mejor manera posible. En todo caso y para clarificar qué entendemos por Estadística, intentaremos acotar el concepto aclarando qué no debería ser la Estadística:

- La Estadística no debería consistir en una serie de procedimientos numéricos innecesarios que deben aplicarse, por imperativo académico, si se quiere publicar un trabajo científico.
- La Estadística tampoco debería consistir en un conjunto de protocolos y algoritmos sofisticados de los que disponemos para convertir en ciencia trabajos que carecen de rigor y profundidad.

En el siguiente apartado comentaremos cuatro nociones estadísticas elementales de carácter transversal en este manual.

[1] ¿Qué entendemos por azar? ¿Existe realmente? Estas preguntas dan pie a una ya vieja discusión científica.

Conceptos básicos

Población: es el objeto del estudio. Se trata pues de un concepto bastante abstracto, aunque en el caso de Ciencias de la Salud seguiremos normalmente la acepción común del término, es decir, un amplio colectivo de individuos.

Carácter y variable: sobre la población se estudiarán uno o varios caracteres. No daremos una definición de carácter sino que lo entenderemos como una noción común. Son ejemplos de caracteres el sexo, la edad, el peso, la talla, el nivel de colesterol, etc. La expresión de un carácter en cada individuo da lugar a una función o aplicación matemática que, en el contexto estadístico se denomina variable aleatoria. Se nombra así porque en un ambiente de incertidumbre toma distintos valores sin que sepamos bien por qué. Según la forma en que se expresan los respectivos caracteres, las variables se clasifican en dos categorías fundamentales:

- **Cuantitativas o numéricas:** se dice que una variable es cuantitativa cuando mide numéricamente el carácter respecto a una unidad de referencia. Son ejemplos de variables cuantitativas la edad medida en años, la concentración de colesterol medida en mg/mm, o la temperatura medida en grados Celsius, la estatura medida en cm, etc.

- **Cualitativas:** se dice que una variable es cualitativa cuando no expresa un carácter de forma numérica sino que distingue entre varias categorías. Son ejemplos de variables cualitativas el sexo si distinguimos entre varón y hembra, el grupo sanguíneo si distinguimos entre A, B, AB y 0, etc.

No obstante, podemos mencionar una tercera categoría que en rigor pertenece a la segunda pero que en la práctica puede recibir el tratamiento estadístico de la primera. Se trata de las variables ordinales, que expresan un carácter cualitativo mediante categorías que admiten un orden natural. Son ejemplos de variables ordinales el grado de una enfermedad (nulo, leve, moderado, severo) o el nivel de dolor de un paciente (bajo, medio, alto). Con frecuencia, se asigna un valor numérico a dichos niveles empezando por 0 ó 1 y acabando en una puntuación máxima, que puede ser 5, 10, etc. Es muy habitual que la puntuación final en una variable de este tipo se obtenga como suma de pequeñas puntuaciones en diferentes apartados, dando lugar a lo que conocemos por escalas ordinales. Así podemos obtener escalas de dolor (EVA), de movilidad (WOMAC), de autonomía (Barthel), de equilibrio (PBS), de consciencia (Glasgow), de agresividad de un tumor (Gleason), de personalidad tipo A, etc. El programa SPSS denomina nominales a las variables cualitativas puras para distinguirlas de estas últimas y, con el mismo fin, denomima de escala a las cuantitativas puras. Es decir, distingue entre variables nominales, ordinales y de escala. Como hemos indicado antes, las ordinales reciben en ocasiones el mismo tratamiento que las nominales (cualitativas) y en otras el de las de escala (numéricas), de ahí que, en el análisis de los datos, si obviamos ciertos métodos muy específicos, sólo distinguiremos entre numéricas y cualitativas.

Ejercicio 1. *Indica otras tres variables nominales, tres ordinales y tres cuantitativas.*

Muestra: ya hemos dicho que sobre una población se va a estudiar un cierto carácter que dará lugar a una variable, denótese por X, y que la población suele ser demasiado grande. Ello nos obliga a contentarnos con estudiar el carácter sobre un subconjunto de n individuos de la población. Dicho subconjunto se dice que es una muestra de tamaño n. Podemos entender por muestra tanto a los n individuos como a los n datos correspondientes a la medición de la variable. En todo caso, la letra n queda reservada para denotar el tamaño de muestra.

Fases y problemas del proceso estadístico

Teniendo en cuenta estas consideraciones, podemos distinguir tres fases en el proceso estadístico:

1. **Muestreo:** selección de la muestra que se analizará.
2. **Estadística Descriptiva:** análisis particular de los datos de la muestra seleccionada.
3. **Inferencia Estadística:** estudio de la posible generalización de los resultados obtenidos en la muestra al global de la población.

Tanto en la primera como en la tercera fase es necesario el concurso del Cálculo de Probabilidades porque, en rigor, sólo a partir de una muestra seleccionada aleatoriamente es posible obtener una extrapolación al global de la población de la que procede, que en tal caso se efectuará en términos probabilísticos. Eso no ocurre en la segunda fase, la descriptiva, que puede desarrollarse muy ampliamente casi de espaldas al concepto de probabilidad. De hecho, al menos en la primera parte de este manual preferimos hablar de proporción sin más, pues es en realidad a lo que nos estamos refiriendo en la mayoría de los estudios en el contexto de las Ciencias de la Salud.

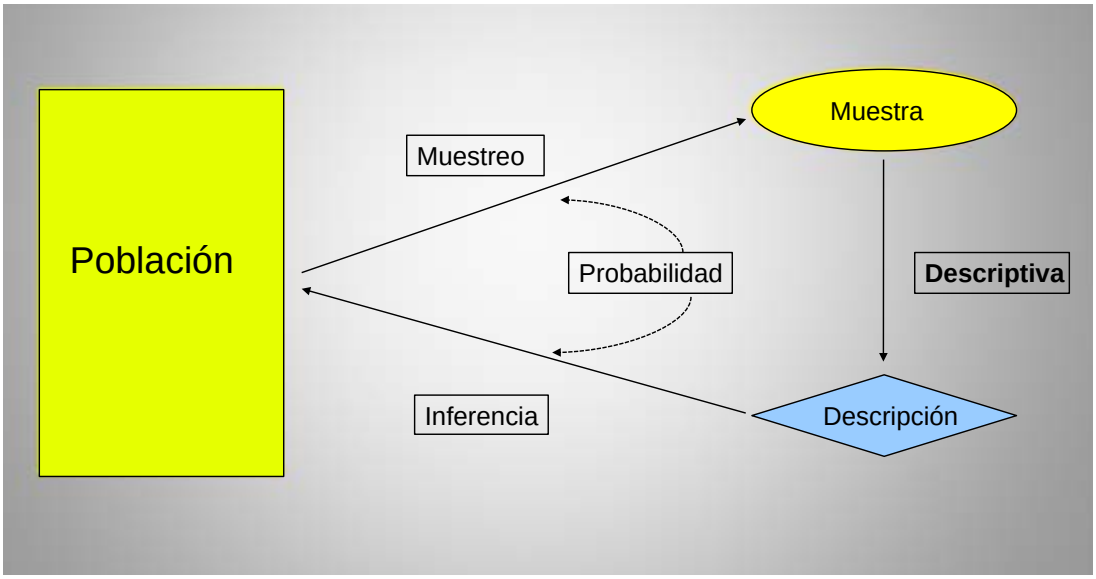


Figura 1: Esquema del proceso estadístico.

Un lector intuitivo o con cierta experiencia en la investigación experimental puede deducir de todo lo anterior dos problemas fundamentales en la aplicación de la Estadística. Tanto es así que entre ambos eclipsan o deberían eclipsar al resto de problemas técnicos que irán surgiendo en el proceso estadístico:

- En primer lugar, en la gran mayoría de los estudios la población a analizar es inabarcable o ni siquiera está bien definida. En todo caso, ¿cómo debería extraerse una muestra de una población para que estuviéramos en condiciones de extrapolar los resultados obtenidos en la misma a la población de la que procede? La respuesta desde el punto de vista técnico ya la conocemos: aleatoriamente. Es decir, deberíamos extraer la muestra de manera similar a un sorteo de lotería en la población a estudiar, lo cual es casi siempre utópico.

- En segundo lugar, hemos de ser muy críticos con las mediciones de las variables, especialmente con aquellas que son de carácter cualitativo u ordinal, más aún con las que están basadas en encuestas, y sin perder de vista a las que consideramos numéricas, porque el hecho de que contemos con un número no nos garantiza que estemos midiendo el parámetro adecuado o que lo midamos correctamente.

Dado que nuestra intención es aplicar la Estadística debemos afrontar con modestia los problemas anteriores teniendo en cuenta, primeramente, que aunque nuestras muestras no se ajusten perfectamente al supuesto teórico de aleatoriedad, la investigación biomédica cuenta con excelentes bases de datos que contienen una información bastante representativa de grandes sectores de la población. Además, gran parte de esa información se obtiene en los grandes centros hospitalarios y de investigación mediante un instrumental capaz de medir con gran precisión y objetividad multitud de variables de posible interés. En ese sentido nos atreveríamos a afirmar que las Ciencias de la Salud es el ámbito ideal de aplicación de la Estadística.

En el extremo opuesto situaríamos los estudios basados en encuestas que se cumplimentan voluntariamente, pues con frecuencia implican un fuerte sesgo en la configuración de la muestra y una medición subjetiva y deficiente de las características a estudiar. Los autores de los estudios suelen ser conscientes de estas limitaciones pero, en ocasiones, confían en que un protocolo estadístico sofisticado obre a modo de piedra filosofal y solucione el problema. Por desgracia, nosotros no nos sentimos capacitados para orientar correctamente a investigadores que diseñen estudios basados en encuestas voluntarias.

Tipos de estudios

Como ya hemos comentado, nuestro objetivo final es explicar un determinado fenómeno biomédico, lo cual nos conduce a relacionar las variables que intervienen en el mismo. En la primera parte del manual nos limitaremos fundamentalmente a un estudio de la relación entre variables desde un punto de vista meramente descriptivo, es decir, sin ánimo de extrapolar los resultados al global de la población. Se trata pues de una Estadística Descriptiva para varias variables; no obstante, y con un carácter meramente preliminar, aprenderemos a describir una única variable de manera aislada en el Capítulo 1. El estudio descriptivo de la relación entre variables puede dar lugar a una amplia casuística

según la naturaleza y cantidad de las variables. Dado que en este manual nos centramos mayormente en el estudio de dos variables y que, a su vez, solo distinguiremos entre dos tipos diferentes, podemos contemplar, a un nivel básico, tres posibilidades (aunque en algunas secciones ampliaremos este esquema):

Variable 1	Variable 2	Problema estadístico
Numérica	Numérica	Correlación numérica: diagrama de dispersión y coeficiente r
Cualitativa	Numérica	Comparación de medias: diagramas de caja comparadas y diferencia de medias
Cualitativa	Cualitativa	Tabla de contingencia: diagrama de barras agrupadas y coeficiente C

Tabla 1: Descriptiva simplificada.

Ejercicio 2. *Se pretende estudiar si existe relación entre el sexo y la estatura. ¿A cuál de los tres tipos de estudio nos estamos refiriendo? ¿Puedes indicar al menos dos ejemplos de cada tipo?*

La extrapolación de estos resultados al global de la población, es decir, la Inferencia Estadística, así como unas nociones mínimas de probabilidad y muestreo, se abordan en la segunda parte del manual, aunque se empiezan a manejar de forma intuitiva en la primera.

PARTE

I

Estadística Descriptiva

1. Estudio de una variable

En un sentido muy amplio, la Estadística Descriptiva es la parte o fase de la Estadística dedicada a la descripción de un conjunto de n datos, entendiendo por descripción la clasificación, representación gráfica y resumen de los mismos. En un contexto más general esos n datos constituirán una muestra de tamaño n extraída de una población, y la descripción de dicha muestra habrá de completarse posteriormente con una inferencia o generalización al total de la población.

El presente capítulo se dedica en su mayoría a la descripción de una variable mientras que los dos siguientes abordan el estudio de la correlación entre dos variables. En todo caso distinguiremos entre la clasificación de los datos en tablas, la representación gráfica y el cálculo de parámetros que resuman la información. A su vez, distinguiremos entre variables cualitativas y cuantitativas. La ejecución de este tipo de análisis mediante el programa estadístico SPSS se ilustra en el Capítulo 6.

1.1. Tablas de frecuencias

La construcción de tablas de frecuencias ha sido hasta hace bien poco la fase preliminar de cualquier estudio descriptivo, utilizándose como medio para la elaboración de gráficos y el cálculo de valores típicos. Hoy en día no se entiende el proceso estadístico sin la utilización de un programa informático que facilite automáticamente los gráficos y cálculos deseados, de ahí que las tablas de frecuencia hayan perdido cierto protagonismo.

Construir una tabla de frecuencias básica equivale a determinar qué valores concretos se dan en la muestra y con qué frecuencia. Se denomina también distribución de frecuencias. Veamos una serie de ejemplos sencillos para distintos tipos de variables. Empezaremos ilustrando una variable cualitativa.

Ejemplo 1. En estudio sobre el grupo sanguíneo realizado con $n = 6313$ individuos se obtuvo la siguiente tabla de frecuencias:

Grupo sanguíneo i	f_i	$\hat{p}_i$
O	2892	0.4580
A	2625	0.416
B	570	0.090
AB	226	0.036
Total	6313	1

Tabla 1.1: Tabla de frecuencias para el grupo sanguíneo.

Nótese que, a la derecha de las frecuencias absolutas, que se denotan por f_i, aparece otra columna donde quedan reflejadas las correspondientes proporciones o frecuencias relativas, que se denotan a su vez por $\hat{p}_i$. En ese caso, el símbolo $\hat{}$ que encontramos encima de p_i hace referencia al hecho de que la proporción es relativa a la muestra, en contraposición con el estudio poblacional que abordaremos en capítulos posteriores. La suma de sus respectivas frecuencias absolutas debe ser igual al número total de datos. Análogamente, la suma de sus frecuencias relativas ha de ser igual a 1, es decir, para una variable cualitativa con k categorías se tiene

$$\sum_{i=1}^{k} f_i = n, \qquad \sum_{i=1}^{k} \hat{p}_i = 1.$$

Ejemplo 2. Las edades en años en un grupo de $n = 25$ estudiantes universitarios son las siguientes: 23, 21, 18, 19, 20, 18, 23, 21, 18, 20, 19, 22, 18, 19, 19, 18, 23, 22, 19, 22 , 21, 18, 24, 24, 20. Estos datos componen la siguiente tabla de frecuencias:

x_i	f_i	$\hat{p}_i$	F_i	H_i
18	6	0.24	6	0.24
19	5	0.20	11	0.44
20	3	0.12	14	0.56
21	3	0.12	17	0.68
22	3	0.12	20	0.80
23	3	0.12	23	0.92
24	2	0.08	25	1
Total	25	1	25	1

Tabla 1.2: Tabla de frecuencias para las edades de alumnos.

Al contrario que en el ejemplo anterior, los datos que obtenemos son numéricos. Se denotará por x_1 el primero de ellos según el orden en que nos llegan los datos, es decir, en nuestro caso $x_1 = 23$. Así se denotará $x_2 = 21$ y sucesivamente hasta llegar a $x_{25} = 20$. Para organizar esta información debemos considerar el valor más pequeño que aparece, en nuestro caso 18. Dicho valor se denotará en lo sucesivo por $\mathtt{x}_1$. Se contabilizará el número de ocasiones en las que se presenta, que será su frecuencia absoluta y se denotará por f_1, que en nuestro caso es 6; el segundo valor es $\mathtt{x}_2 = 19$, que aparece $f_2 = 5$ veces, y así sucesivamente hasta llegar a $\mathtt{x}_7 = 24$, que aparece $f_7 = 2$ veces. Así es como obtenemos la columna de frecuencias absolutas a la que añadimos las frecuencias relativas.

En total, tenemos pues $k = 7$ valores distintos. Nótese que, al tratarse de datos numéricos, existe un orden preestablecido en los mismos, cosa que no sucedía en el ejemplo anterior. Eso nos ha permitido construir otra columna, la de frecuencias absolutas acumuladas, donde se anota, para cada valor $\mathtt{x}_j$, el número F_j total de datos menores o iguales al mismo, es decir,

$$F_j = \sum_{i=1}^{j} f_i.$$

A esta columna le puede ser añadida la de frecuencias relativas acumuladas que resulta de dividir las anteriores por el número total de datos.

1.2. Representación gráfica

El segundo paso del proceso consiste en ilustrar mediante un gráfico lo obtenido en la tabla de frecuencias. Existen varios tipos de gráficos.

Diagrama de sectores: uno de los más utilizados. En el caso del Ejemplo 1, la tabla de frecuencias 1.1 se representa mediante sectores según la Figura 1.1.

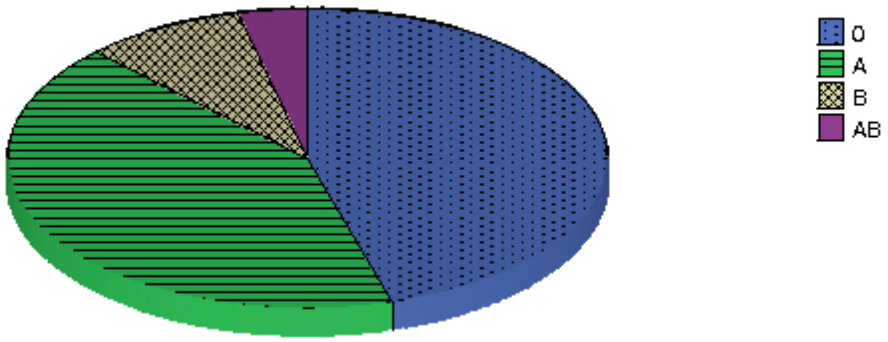


Figura 1.1: Diagrama sectores para el grupo sanguíneo.

Diagrama de barras: para ilustrar la tabla de frecuencias del Ejemplo 2 podríamos escoger también un diagrama de sectores. No obstante, dado el orden natural que existe

en los valores de la variable, se suele optar por otro tipo de gráfico denominado diagrama de barras. En la Figura 1.2 se presenta el diagramas de barras para las frecuencias absolutas.

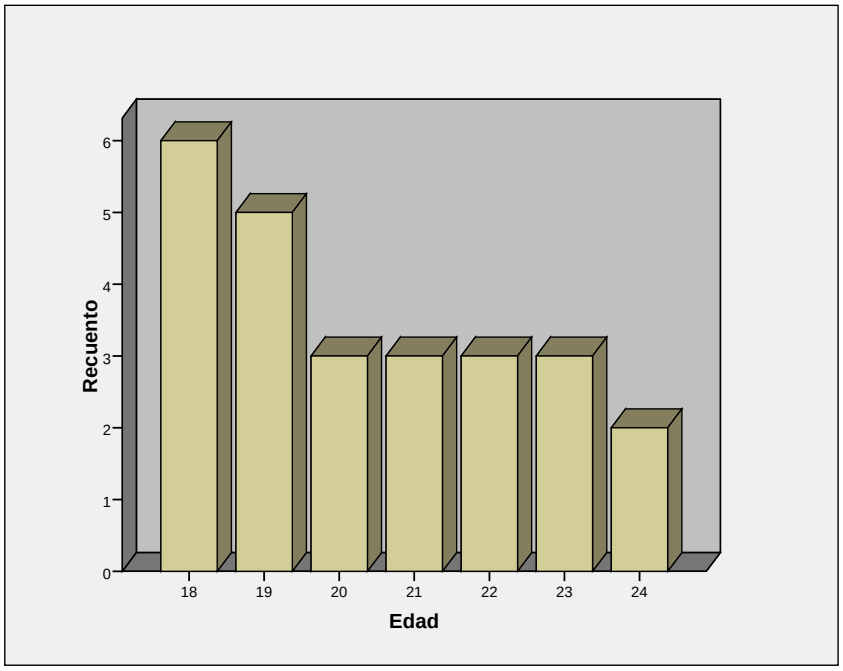


Figura 1.2: Diagrama de barras para edades de alumnos.

Los diagramas de barras para las frecuencias relativas ofrecerían un aspecto idéntico al de los anteriores gráficos pero con diferente escala en el eje OY. Además, se pueden representar líneas que unen las distintas barras y que se denominan polígonos de frecuencia. Los diagramas de barras son también muy recomendables para representar variables cualitativas, especialmente si son de tipo ordinal.

Histograma: dado que la variable estudiada en el Ejemplo 2 admite sólo 7 posibles valores, el diagrama de barras de la Figura 1.2 resulta muy ilustrativo. Imaginemos por un momento qué sucedería si en vez de cuantificar la edad por años cumplidos se midiera por días, o incluso por segundos. En ese caso, lo más probable sería que no hubiera dos estudiantes con la misma edad, con lo que la tabla de frecuencias perdería su sentido último. Consistiría en una larga ordenación vertical de los valores obtenidos donde todos ellos presentarían frecuencia absoluta 1. El diagrama de barras resultante se antojaría claramente mejorable en cuanto a su poder ilustrativo. Esto es lo que entendemos como variable continua, en contraposición con la edad en años, que se consideraría discreta.

Algo parecido ocurriría si, por ejemplo, representamos el diagrama de barras correspondiente a la medición del colesterol sérico (mg/cm^3) en una muestra de $n = 4583$ individuos. Ante tal situación y si nuestra intención es obtener una gráfico que nos ayude a entender fácilmente la distribución de los datos obtenidos, parece razonable empezar por agrupar los datos en clases (intervalos). De esta manera, en la columna de frecuencias absolutas se contabilizará el número de veces que aparece cada clase. Las demás columnas se elaborarán a partir de esta como ya sabemos. Los gráficos resultantes se denominan

histogramas. En el caso del estudio sobre colesterol mencionado anteriormente se obtiene entonces el histograma de frecuencias absolutas que se presenta en la Figura 1.3.

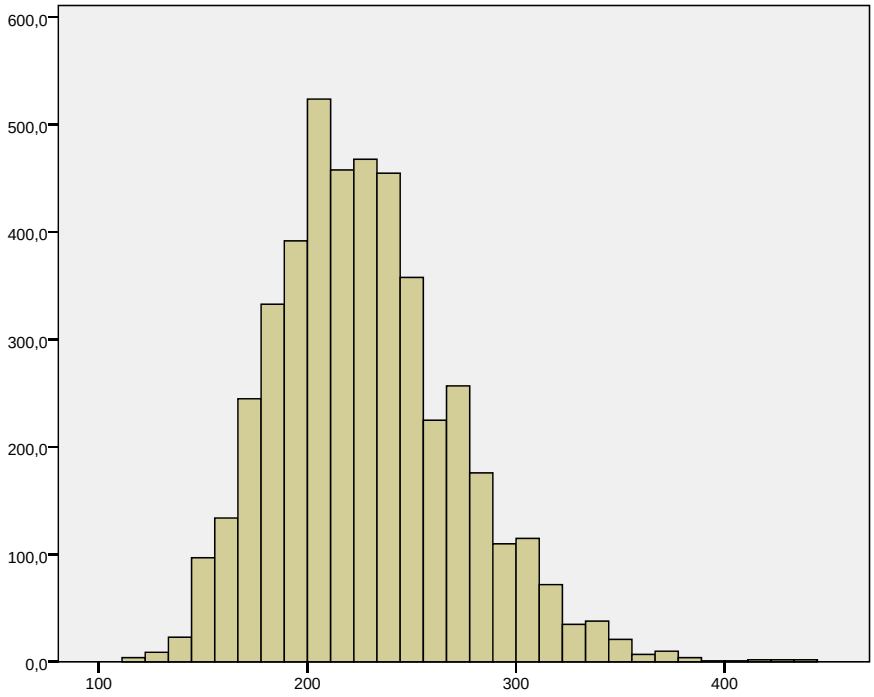


Figura 1.3: Histograma para la colesterolemia.

En definitiva, agrupar en clases significa simplificar, perder una parte de la información en aras de una mejor visión de la misma. Nótese que, en el contexto descriptivo, la distinción que hemos efectuado entre variables continuas y discretas no depende de la naturaleza en sí de la variable sino del tratamiento gráfico que estemos dispuestos a darle. El procedimiento a seguir a la hora de construir las clases y representar los histogramas puede llegar a resultar bastante complejo a la par que puramente convencional. En [10] podemos encontrar un algoritmo perfectamente descrito. En la actualidad, todas las tareas gráficas se realizan mediante programas estadísticos que tienen implementados sus propios algoritmos por lo que no profundizaremos en esta cuestión. Tan sólo destacaremos que el asunto más crucial en lo que respecta al aspecto del gráfico es el número de intervalos que debemos considerar. Parece claro que dicho número debe guardar algún tipo de relación con el número total de datos n. Efectivamente, si el número de intervalos escogido es demasiado pequeño el gráfico resultará excesivamente suave; por contra, si el número de intervalos es demasiado grande el histograma resultará demasiado abrupto. Por eso existen diversos criterios de carácter orientativo para determinar el número de intervalos, como la conocida ley de Sturges, aunque el programa SPSS no la respeta.

Ejercicio 3. *Explica qué te sugiere la Figura 1.3.*

Veamos otro ejemplo de variable que debería tratarse como continua:

Ejemplo 3. La exposición aguda al cadmio produce dolores respiratorios, daños en los riñones y el hígado, y puede ocasionar la muerte. Por esta razón se controla el nivel de polvo de cadmio y de humo de óxido de cadmio en el aire. Este nivel se mide en miligramos de cadmio por metro cúbico de aire. Una muestra de 35 lecturas arroja estos datos:

0.044	0.030	0.052	0.044	0.046
0.020	0.066	0.052	0.049	0.030
0.040	0.045	0.039	0.039	0.039
0.057	0.050	0.056	0.061	0.042
0.055	0.037	0.062	0.062	0.070
0.061	0.061	0.058	0.053	0.060
0.047	0.051	0.054	0.042	0.051

Tabla 1.3: Concentración cadmio.

En este caso sucede también que la variedad de valores posibles es demasiado amplia en relación con el número de datos, es decir, que éstos no se repiten o se repiten demasiado poco como para que merezca la pena construir una tabla de frecuencias con su correspondiente diagrama de barras, de ahí que sea más aconsejable construir un histograma.

Diagrama tallo-hoja: otro tipo de gráfico de gran interés en estas situaciones y que guarda gran similitud con el histograma de frecuencias absolutas es el denominado diagrama tallo-hoja, en el que cada dato se identifica con una cifra de la derecha que indica el valor de las unidades, siendo la correspondiente a su izquierda el valor de las decenas. Así, en la Figura 1.4 podemos encontrar el diagrama de tallo-hoja correspondiente a los datos del Ejemplo 3. También consideraremos más adelante los denominados diagrama de caja o box-plot.

Lectura de cadmio

Frecuencia	Tallo-Hoja
1	2 . 0
6	3 . 007999
9	4 . 022445679
11	5 . 1122345678
7	6 . 0111226
1	7 . 0

Unidad: 0.01

Figura 1.4: Diagrama tallo-hoja para los valores de cadmio.

Ejercicio 4. *Representa el histograma para los datos del Ejemplo 3 haciendo uso de una hoja de cálculo o un programa estadístico. Interpreta el diagrama tallo-hoja de la Figura 1.4.*

Campana de Gauss: para acabar esta sección destacamos que histogramas como el de la Figura 1.3, o incluso diagramas de tallo-hoja como el de la Figura 1.4, sugieren un tipo de curva muy bien caracterizada que denominamos curva normal o campana de Gauss. Concretamente, en casos como éstos solemos afirmar que los datos se distribuyen aproximadamente según un modelo tipo normal. Hablamos de tipo porque no se trata de un modelo único sino de una familia que depende de dos parámetros. Las variables que se ajustan aproximadamente a un modelo normal son relativamente frecuentes en la naturaleza, de ahí que la curva normal desempeñe un papel destacado en la Estadística. Fue estudiada inicialmente por Laplace y Gauss para explicar el comportamiento de los errores en medidas astronómicas. La aplicación de la distribución normal no quedó reducida al campo de la astronomía. Las medidas físicas del cuerpo humano o de un carácter psíquico en una población, las medidas de calidad de productos industriales y de errores en procesos físico-químicos de medición en general, siguen con frecuencia este tipo de distribución. Desde un punto de vista teórico, el denominado Teorema Central del Límite confiere a la distribución normal un papel destacado en la Estadística. Aunque dicho teorema se enunciará más formalmente en la Sección 4.1, en términos intuitivos viene a decirnos lo siguiente:

> Las variables que pueden entenderse como resultado de un fenómeno aditivo tienden a distribuirse según un modelo de distribución tipo normal.

Eso es lo que ocurre precisamente en la denomina máquina de Galton, en la que se deja caer una bola que bajan a izquierda o derecha a través de cuñas colocadas en sucesivos niveles, hasta llegar a un depósito en la parte inferior, tal y como se ilustra (con 12 niveles) en la Figura 1.5. Si repetimos el proceso con una gran cantidad de bolas, ¿cómo se distribuirán las bolas en el depósito inferior?

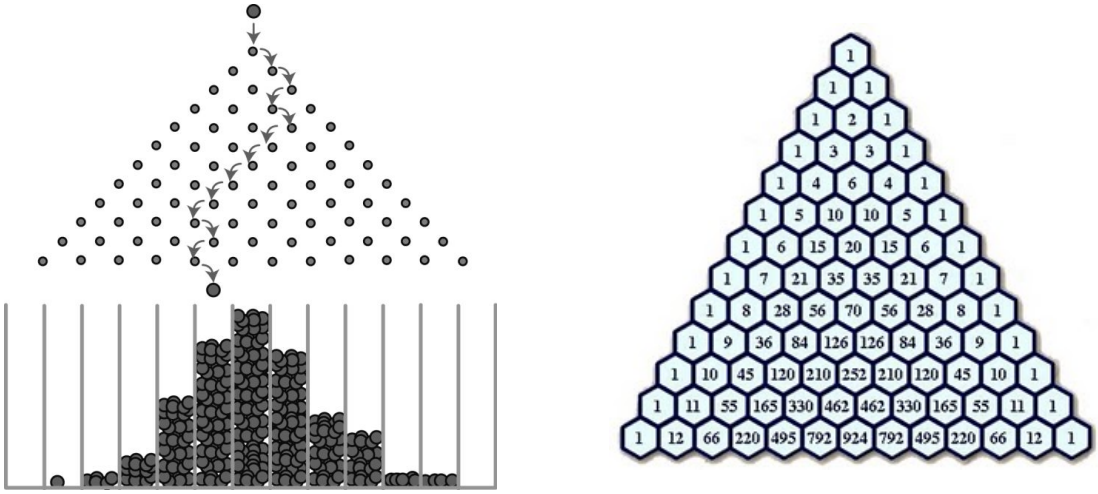


Figura 1.5: Máquina de Galton (izquierda) y triángulo de Pascal (derecha).

La respuesta es que siguen un patrón de distribución de campana de Gauss. Esto se debe a que la posición del depósito donde cae cada bola depende exclusivamente de la suma de veces que la bola cae a la derecha en su recorrido. Es decir, las bolas que quedan en el extremo izquierdo no caen nunca hacia su derecha, todo lo contrario de lo que ocurre con las del extremo opuesto; sin embargo, las bolas que quedan en la posición central suman tantas caídas a su izquierda como a su derecha, y esa circunstancia es mucho más probable que las dos anteriores.

Cuando decimos que es más probable no estamos pensando en un mecanismo inteligente de compensación que funciona a lo largo del recorrido[1]. Efectivamente, partimos del supuesto de que, por simetría[2], todas las trayectorias son equiprobables. No obstante y por pura combinatoria[3], son más numerosas las trayectorias que suman tantas caídas a la izquierda como a la derecha, porque hay muchas formas diferentes de sumar ese resultado. Concretamente, en nuestro caso eso puede ocurrir de $12!/6!6! = 924$ formas diferentes, tal y como se ilustra en el denominado triángulo de Pascal (Figura 1.5, derecha).

Cuando en la naturaleza se observa una variable que se distribuye según un patrón aproximado de campana de Gauss cabe pensar que detrás de lo que se mide exista un fenómeno aditivo en sentido amplio, lo cual no tendría por qué ocurrir necesariamente. Efectivamente, en la Figura 1.6 podemos apreciar un histograma relativo a 97 mediciones de tumores prostáticos, donde se aprecia un modelo de distribución radicalmente diferente al de la campana de Gauss; en este caso, el modelo está caracterizado por un fuerte sesgo o asimetría hacia la derecha (positivo[4]).

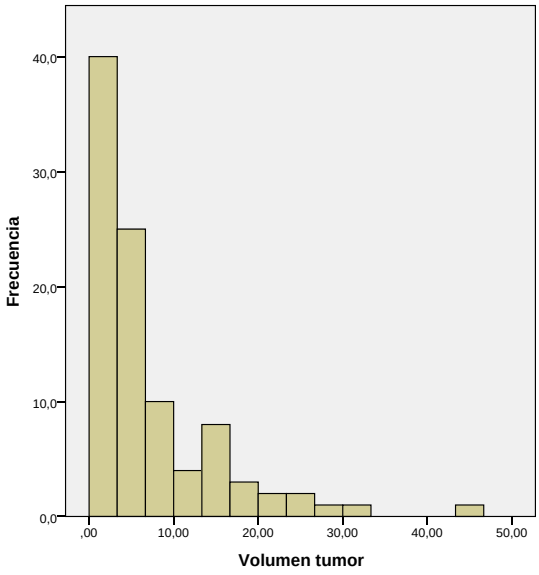


Figura 1.6: Volumen de un tumor de próstata.

[1]Esa es una de las preconcepciones más comunes cuando se trata el concepto de azar.

[2]Este otro es el argumento que se esconde realmente tras la palabra azar.

[3]Advertimos que no es posible profundizar en el concepto de probabilidad sin unas nociones básicas de cálculo combinatorio.

[4]Cuando los valores extremos se encuentran a la izquierda se denomina sesgo negativo.

En algunas ocasiones este modelo de distribución se asocia a fenómenos de tipo multiplicativo. Si ése es el caso, una transformación logarítmica de la variable convertirá el fenómeno en aditivo (ya que el logaritmo del producto es la suma de los logaritmos) y observaremos entonces algo más parecido a una campana de Gauss, como se aprecia en la Figura 1.7.

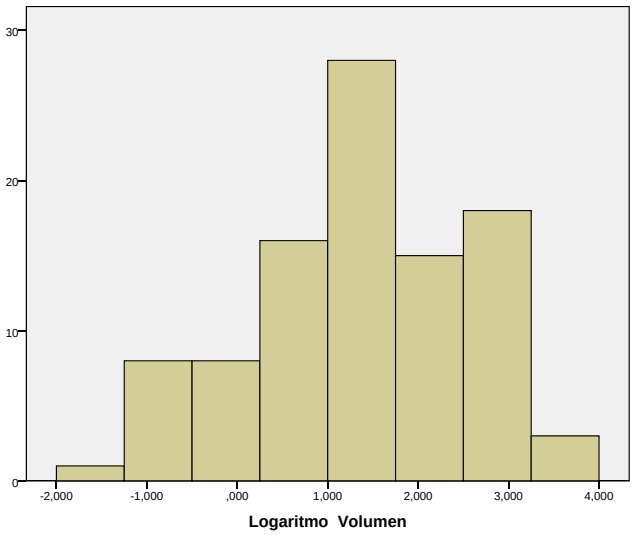


Figura 1.7: Logaritmo del volumen de tumores de próstata.

1.3. Valores típicos

El tercer paso del proceso descriptivo consiste en calcular una serie de números cuyo propósito es sintetizar la información que aportan los n datos de la muestra considerada. Los valores típicos son, precisamente, esos números que pretenden caracterizar la muestra. Esta fase del estudio sólo tiene sentido cuando la variable estudiada es cuantitativa. Distinguiremos entre medidas de centralización, medidas de posición, medidas de dispersión y medidas de forma:

1.3.1. Medidas de centralización

Las medidas de centralización son las más importantes sin duda aunque por sí mismas no suelen bastar para resumir la información. La pregunta puede ser la siguiente: ¿qué número debemos escoger si pretendemos explicar la mayor parte posible de información con un único número? La respuesta es pues un número representativo, un valor central en algún sentido. Los más populares son, sin duda, la media aritmética y la mediana.

Media aritmética: es el valor central en sentido aritmético. Se obtiene sumando los n datos de la muestra y dividiéndolos por el tamaño de esta, es decir,

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

donde cada dato x_i aparece en el sumatorio tantas veces como se repita en la muestra, es decir, si los datos están agrupados en una tabla de frecuencias, se puede calcular también de la forma:

$$\overline{x} = \frac{\sum_{i=1}^{k} \mathtt{x}_i f_i}{n} = \sum_{i=1}^{k} \mathtt{x}_i \hat{p}_i. \tag{1.1}$$

Como podemos apreciar en la expresión anterior, a cada dato $\mathtt{x}_i$ se le asigna un peso $\hat{p}_i$ equivalente a la proporción que representa en la muestra. Podemos establecer una analogía entre la media aritmética y el concepto físico de centro de gravedad, es decir, la media aritmética puede entenderse como el centro de gravedad de los datos de la muestra, y como tal puede verse muy afectada ante la presencia de valores extremos.

En el Ejemplo 2 tenemos una edad media de $\overline{x} = 20.36$ años para los estudiantes de la muestra. La media se expresa, lógicamente, en las mismas unidades que los datos originales. Indicar dicha unidad es aconsejable. El hecho de que los datos aparezcan agrupados en intervalos, como ocurre con los valores de colesterol que se ilustran en la Figura 1.3, no debe afectar al cálculo de la media. Es decir, la media debe calcularse a partir de los datos originales sin agrupar. En ese ejemplo, obtenemos precisamente un colesterol medio de $\overline{x} = 228.18$ mg/ml.

Ejercicio 5. *¿Qué le sucede a la media aritmética si a todos los datos les sumamos una misma cantidad k? ¿Y si los multiplicamos por una misma cantidad k?*

Ejercicio 6. *¿Es cierto que sumar n datos es equivalente a sumar la media de los mismos n veces?*

Ejercicio 7. *Averigua qué entendemos por esperanza de vida.*

Media truncada: es la media aritmética que se obtiene una vez se han excluido el 5 % de datos más extremos.

Media ponderada: se obtiene de manera similar a la media según la expresión (1.1) pero ponderando cada dato $\mathtt{x}_i$ al gusto de quien la calcula. Desde ese punto de vista, la media aritmética puede entenderse como una media ponderada en la que se considera, para cada dato $\mathtt{x}_i$, la ponderación correspondiente a la frecuencia relativa del mismo en la muestra. El Ejercicio 21 puede servirnos como ejemplo para entender cómo se calcula. La media ponderada no destaca especialmente por su interés científico, sino por su uso en ámbitos académicos a la hora de calificar asignaturas, por lo que no profundizaremos más en este parámetro.

Mediana: es el valor central $\tilde{x}$ en el sentido del orden, es decir, aquel que quedaría en el medio una vez ordenados los datos de menor a mayor, repitiéndose si es necesario tantas veces como aparezcan en la muestra. Para calcularla basta pues con ordenar los datos y determinar la posición del medio. Si el número de datos n es impar no cabe duda de que la mediana es el dato que ocupa la posición $\frac{n+1}{2}$. Si n es par tenemos un conflicto que puede resolverse mediante un convenio: definir la mediana como la semisuma de los datos que ocupen las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$. En este proceso puede ser de utilidad la columna

de las frecuencias absolutas acumuladas o un diagrama tallo-hoja. De todas formas, lo ideal es delegar el cálculo de media o mediana en un programa estadístico. Si es así, todos estos detalles resultan irrelevantes. En el Ejemplo 2, el valor mediano es 20, que ocupa la posición 13. Para los datos del colesterol (Figura 1.3) es $\tilde{x} = 225$, muy similar a la media. Sin embargo, para los datos de la Figura 1.6, tenemos $\overline{x} = 7.00$ y $\tilde{x} = 4.25$.

Ejercicio 8. *¿A qué se debe esta última diferencia?*

Al contrario de lo que sucede con la media, la mediana es robusta en el sentido de que no se ve afectada por la presencia de valores extremos. Efectivamente, es obvio que podemos reemplazar el valor mayor de la muestra por otro mucho más grande sin que ello afecte a la mediana. Esta cualidad podría considerarse negativa por denotar un carácter menos informativo que la media pero también puede resultar positiva cuando una clara asimetría con presencia de valores extremos (sesgo) desplaza fuertemente la media restándole representatividad, como sucede precisamente en la Figura 1.6.

Ejercicio 9. *¿Qué relación se da entre la media y la mediana si el sesgo es positivo, es decir, cuál es mayor? ¿Qué relación se dará entre la media y la mediana si la distribución es normal?*

Ejercicio 10. *Calcula la media y la mediana del siguiente conjunto de datos: 8,0,10,9,9.*

1.3.2. Medidas de posición

Las medidas de posición son una serie de números que dividen la muestra ordenada en partes con la misma cantidad de datos. La principal medida de posición ya la hemos estudiado: la mediana, pues divide la muestra en dos mitades. Efectivamente, sabemos que el 50 % de los datos debe ser inferior a la mediana y el resto superior.

Cuartiles: si pretendemos dividir la muestra ordenada en cuatro partes iguales obtenemos los denominados cuartiles, que se denotan por Q_1, Q_2 y Q_3. El primero deja a su izquierda (o debajo, según se prefiera) el 25 % de los datos; el segundo deja a la izquierda el 50 %, por lo que se trata de la propia mediana; el tercero deja a la derecha el 25 %. Respecto al cálculo de Q_1 y Q_3, lo ideal es decantarse por el uso de un programa estadístico. Si no se cuenta con él convenimos, por ejemplo, lo siguiente: para una muestra de tamaño n y ordenada de menor a mayor Q_1 será el dato que tenga por posición la parte entera de $n/4$ y Q_3 será el dato que ocupe esa posición pero contando desde el final.

Percentiles: si dividimos la muestra en 100 partes iguales, obtendremos los percentiles, que van de p_1 a p_{99}. De nuevo, la mediana coincide con el percentil 50 y los cuartiles Q_1 y Q_3 con p_{25} y p_{75}, respectivamente. Los percentiles se utilizan mucho en pediatría para analizar el crecimiento de los recién nacidos. Hemos de tener en cuenta que sólo para una muestra amplia, la cual hace imprescindible el uso de un programa estadístico, tiene sentido considerar divisiones finas de la misma. Por ello, si contamos con pocos datos es absurdo hablar de percentiles.

En general, podemos hablar de los cuantiles. Dado un valor γ en el intervalo $(0, 1)$, el cuantil γ se define como el valor que deja a su izquierda el $\gamma \times 100$ % de los datos. De esta forma, la mediana es el cuantil 0.50 y el percentil p_{95}, el 0.95, por ejemplo.

1.3.3. Medidas de dispersión

Las medidas de dispersión tienen por objeto completar la información que aportan las medidas de centralización pues miden el grado de dispersión de los datos o, lo que es lo mismo, la variabilidad de la muestra. Las fundamentales son la desviación típica y el rango intercuartílico.

Rango: es el más inmediato pues expresa la diferencia entre el valor mayor y el menor. En el Ejemplo 2 es igual a $24 - 18$, es decir, 6 años de diferencia entre el alumno mayor y el más joven.

Varianza: nos da una medida de dispersión relativa al tamaño muestral de los distintos datos respecto a la media aritmética $\overline{x}$. Una primera definición es la siguiente:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}.$$

El hecho de elevar las diferencias respecto a $\overline{x}$ al cuadrado se debe a que, como es fácil de comprobar, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, pues al sumarse los datos superiores a la derecha de la media se anulan con los inferiores. Se podría haber optado por considerar el valor absoluto de las diferencias, lo cual daría lugar a lo que se conoce como desviación media, pero eso conllevaría numerosas inconvenientes técnicos. Si los datos están tabulados, la expresión anterior equivale a la siguiente:

$$s^2 = \sum_{i=1}^{k}(\mathtt{x}_i - \overline{x})^2 \hat{p}_i. \tag{1.2}$$

No obstante, con vista a una posterior Inferencia Estadística aparecerá dividida por $n-1$ en lugar de n. Suele denominarse en tal caso varianza insesgada o cuasi-varianza. En la segunda parte del manual y si no se especifica lo contrario, cada vez que hablemos de varianza nos estaremos refiriendo a la insesgada. El hecho de dividir por $n-1$ en lugar de n el contexto de la Inferencia Estadística es apenas apreciable cuando n es grande, por lo que no debe desviar nuestra atención de la esencia del parámetro. El cálculo de la varianza lo realizaremos mediante un programa estadístico o en su defecto, con una calculadora. En el Ejemplo 2, de las edades en años de 25 alumnos, se obtiene una varianza $s^2 = 4.157$ años^2.

Desviación típica: podemos observar que en la varianza anterior las unidades originales se perdieron por la necesidad de elevar al cuadrado las diferencias. Para recuperarlas basta con efectuar la raíz cuadrada de la varianza obteniendo lo que denominamos desviación típica, que se denotará por s. Así pues,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

Igualmente, en la Inferencia Estadística, se utilizará la cuasi-desviación típica que se obtiene a partir de la cuasi-varianza. En el ejemplo de las edades tenemos $s = 2.039$ años.

En el caso del colesterol (Figura 1.3) la desviación típica es $s = 44.82$ mg/ml y, en el del volumen del tumor de próstata (Figura 1.6), $s = 7.89$ en la unidades correspondientes.

Ejercicio 11. *¿Puede ser negativa la desviación típica? ¿Cómo se interpreta una desviación típica nula?*

Ejercicio 12. *¿Qué le sucede a la desviación típica si a todos los datos les sumamos una misma cantidad k? ¿Y si los multiplicamos por una misma cantidad k?*

Ejercicio 13. *Se denomina tipificación o estandarización a la acción de restar a cada dato x_i de la muestra la media aritmética y, posteriormente, dividir el resultado entre la desviación típica, es decir, calcular*

$$z_i = \frac{x_i - \overline{x}}{s}. \tag{1.3}$$

¿Cuáles serán entonces la media y la desviación típica de los datos tipificados? ¿En qué dimensiones se expresarán?

La desviación típica funciona como complemento de la media dado que, mientras la última indica el centro aritmético de los datos, la primera expresa el grado de dispersión respecto a dicho centro. De esta forma, el par de números $(\overline{x}, s)$ pretende resumir la información contenida en los n datos de la muestra. En concreto, si nuestros datos se distribuyeran según una distribución normal, el mero conocimiento de $\overline{x}$ y s permitiría reproducir con exactitud el histograma. Así, ocurre, por ejemplo, que entre los valores $\overline{x} - s$ y $\overline{x} + s$ se encuentra ua proporción muy cercana al 68 % de los datos, o que entre $\overline{x} - 2 \cdot s$ y $\overline{x} + 2 \cdot s$ se encuentra una proporción muy cercana al 95 %. Efectivamente, dado que el histograma de la Figura 1.3 se asemeja bastante a una campana de Gauss, la mayor parte de los datos (95 %) debe estar comprendida aproximadamente, según lo que ya sabemos, en el intervalo $228 \pm 2 \cdot 45$, es decir, entre 138 y 318, cosa que podemos verificar gráficamente. No ocurre lo mismo con los datos del gráfico de la Figura 1.6. En ese sentido afirmamos que el par $(\overline{x}, s)$ resume perfectamente la información contenida en una muestra cuando los datos de la misma se distribuyen según una curva normal. Entendemos también que, a medida que nos alejamos de dicho modelo, el par anterior pierde su capacidad de síntesis. De hecho, sabemos que en determinadas situaciones la media aritmética puede considerarse menos representativa que la mediana. En tal caso necesitamos una medida de dispersión que complemente dicho valor central.

Rango intercuartílico: pretende ser un complemento adecuado a la mediana. Está basado, al igual que esta, en el orden de los datos y se define mediante $R_I = Q_3 - Q_1$. En el caso de los datos del ejemplo de las edades, obtenemos $R_I = 2$. Para los datos de la Figura 1.6 obtenemos $R_I = 7.03$.

Coeficiente de variación: se trata de un coeficiente adimensional relacionado con la media y la desviación típica que es de gran utilidad para comparar la dispersión de distintos grupos de datos, dado que nos da una medida de la dispersión de los datos relativa al orden de magnitudes que estos presentan. Concretamente, se define mediante

$$C.V. = \frac{s}{\overline{x}} \times 100.$$

Ejercicio 14. *Se tienen 30 datos numéricos correspondientes a la medición del peso en kg de 30 individuos. ¿En qué dimensiones se expresarán la media aritmética, varianza, desviación típica y coeficiente de variación?*

Ejercicio 15. *Considera los dos grupos de datos (a) y (b) siguientes: (a) 1.80, 1.79, 1.77, 1.83, 1.52. (b) 180, 179, 177, 183, 152. ¿Tienen la misma media? ¿Tienen la misma desviación típica? ¿Tienen en común algún parámetro descriptivo de los considerados anteriormente?*

1.3.4. Medidas de forma

Coeficiente de asimetría: indica el grado de asimetría o sesgo que se da en la distribución de los datos. Se define mediante

$$g_1 = \frac{m_3}{s^3}, \quad \text{donde} \quad m_k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^k}{n}, \quad k = 1, 2, 3 \dots$$

Distinguimos a grandes rasgos tres situaciones:

1. $g_1 > 0$: distribución asimétrica de los datos con sesgo positivo (Figura 1.6).
2. $g_1 < 0$: distribución asimétrica con sesgo negativo.
3. $g_1 = 0$: distribución simétrica.

Coeficiente de aplastamiento o de Curtosis: expresa el grado de aplastamiento de una distribución simétrica respecto al que correspondería a una distribución normal con su media y desviación típica, de manera que un valor 0 equivale a una campana de Gauss, mientras que un valor negativo indica un aplastamiento excesivo y un valor positivo indica un apuntamiento.

1.4. Otros gráficos y tablas

Diagrama de caja: a partir de los cuartiles y el rango intercuartílico podemos construir un gráfico denominado de caja o box-plot. Se trata de una caja cuyos bordes son los cuartiles primero y tercero, por lo que su longitud coincide con el rango intercuartílico. En su interior se marca la mediana con una línea gruesa. A partir del rango intercuartílico se determina qué valores se considerarán extremos: concretamente aquellos que disten de los cuartiles Q_1 o Q_3, según corresponda, más de 1.5 veces el rango intercuartílico. Se marcan con unas vallas los valores no extremos más próximos a dichos límites de manera que los que queden fuera de las mismas serán los datos extremos, que representarán mediante círculos o asteriscos según el grado de extremismo que alcancen. En la Figura 1.8 se representa el diagrama de caja correspondiente al histograma de la Figura 1.6.

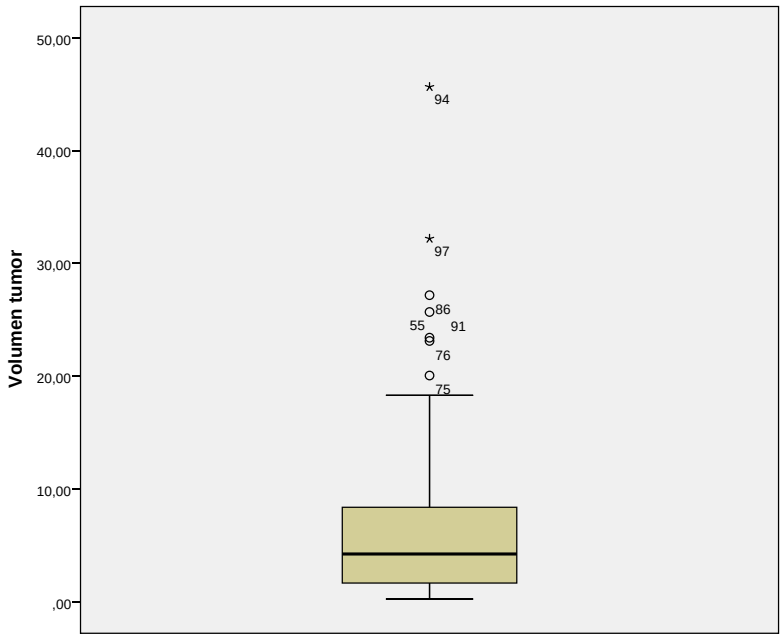


Figura 1.8: Box plot para el volumen de tumores de próstata.

Llegados a este punto hemos estudiado ya una amplia variedad de valores típicos. Recordemos que éstos tienen como función resumir la información que aporta la muestra. Ahora bien, un resumen ideal debería verificar simultáneamente dos condiciones en ocasiones incompatibles: exhaustividad y brevedad. En ese sentido, aconsejamos lo siguiente:

> Si pretendemos resumir lo mejor posible la información contenida en la muestra debemos escoger al menos una medida de centralización junto con otra de dispersión. Lo más frecuente es considerar el par $(\overline{x}, s)$. Esta opción es la ideal en el caso de que los datos se distribuyan según una curva normal.

A medida que nos diferenciamos de ese modelo de distribución, el par anterior pierde su capacidad de síntesis, por lo que debe añadirse al resumen de los datos la mediana como medida de centralización y, si es posible, el rango intercuartílico como medida de dispersión. Nos decantaremos por esta opción preferiblemente cuando observemos una fuerte asimetría con presencia de valores extremos. Esta elección debería ir acompañada del uso de técnicas no paramétricas en la posterior inferencia (Capítulo 5). Por último, el tamaño de muestra nunca debe faltar en un resumen adecuado.

Estas normas no dejan de ser orientativas, porque en la redacción de trabajos científicos prima la capacidad de síntesis, de manera que debemos intentar elaborar tablas que recojan la máxima información en el mínimo espacio y escoger sólo los gráficos que resulten más esclarecedores. La Tabla 1.4 es un ejemplo extraído de una publicación sobre duración de las bajas laborales en España (véase [3]).

Diagnosis	*N*	Mean	*SD*	Median	Over 15 days (%)
Gastroenteritis	40,780	3.8	8.0	3	1.9
Noninfective gastroenteritis	16,342	4.5	11.5	3	2.4
Tonsillitis strep throat	10,374	4.8	5.7	4	2.2
Acute pharyngitis	29,449	4.8	8.3	3	2.5
Tonsillitis not strep throat	16,373	5.4	6.5	4	2.7
Cold	26,171	5.4	9.0	4	3.5
Flu	24,169	6.3	7.5	5	3.9
Diarrhea	3,909	6.4	19.4	3	4.4
Others flu types	7,137	6.6	7.7	5	4.2
Bronchitis not acute or chronic	3,461	10.8	20.6	7	13.4
Migraine	3,202	10.8	30.1	2	11.7
Acute or chronic bronchitis	11,482	11.7	20.6	8	15.9
Headache	3,622	17.8	40.2	4	20.6
Renal colic	7,850	20.0	36.0	8	29.0
Ankle strain	6,571	24.1	30.7	15	48.7
Giddiness	9,379	24.7	46.8	8	30.2
Dorsalgia	4,144	32.1	49.8	11	42.5
Low back pain	48,933	35.2	49.9	15	49.7
Cervical pain	17,886	50.5	58.0	31	64.5
Inguinal hernia	6,044	52.8	41.5	43	95.0
Other maternal disorder related to pregnancy	3,822	58.2	46.6	46	88.3
Sciatica	21,801	59.9	71.9	32	65.8
Cervical strain	4,156	61.1	50.2	50	84.1
Anxiety	19,857	61.8	73.9	32	66.7
Threatened abortion	6,986	69.0	72.3	38	74.7
Carpal tunnel syndrome	3,859	78.0	67.2	57	93.4
Depression	6,437	82.4	87.1	49	74.3
Medical meniscus injury	2,801	90.2	75.6	64	94.7
Adjustment disorders	3,079	107.0	86.6	82	92.3
Total	370,076	7.0	48.9	7	30.7

Tabla 1.4: Ejemplo de tabla descriptiva.

En el Capítulo 6 se proporcionan algunas indicaciones para realizar tablas de este tipo mediante SPSS.

Otras cuestiones propuestas

Ejercicio 16. *Se midió, a través de cierto aparato, una determinada variable bioquímica, obteniendo un total de 146 datos numéricos, que presentaron una media aritmética de 4.2 y una desviación típica de 1.1, en las unidades de medida correspondientes. Tras representar el histograma de frecuencias absolutas, se comprobó que los datos configuraban aproximadamente una campana de Gauss.*

(a) Indica un intervalo que contenga aproximadamente al 95 % de los datos.

(b) *Se averigua posteriormente que el aparato de medida comete un error sistemático consistente en indicar, en todo caso, media unidad menos que el verdadero valor de la variable. ¿Cuáles serán entonces la media aritmética y desviación típica de los 146 verdaderos valores?*

Ejercicio 17. *Se mide cierta variable sobre una muestra de 10 individuos, obteniéndose los siguientes datos.*

$$4 \quad 5 \quad 4.5 \quad 3.9 \quad 5.2 \quad 4 \quad 5.2 \quad 5.3 \quad 23 \quad 4.1$$

Indica una medida de centralización y otra de dispersión adecuadas.

Ejercicio 18. *Indica dos grupos, de 5 datos cada uno, que presenten...*

(a) *La misma media pero distinta desviación típica.*

(b) *La misma desviación típica pero distinta media.*

(c) *La misma mediana y distinta media.*

(d) *La misma media y distinta mediana.*

Ejercicio 19. *Los individuos A y B manejan un ecógrafo. Se pretende dilucidar cuál de los dos tiene mayor precisión a la hora de efectuar mediciones. Para ello se asignó al individuo A la medición de un mismo objeto en 10 ocasiones diferentes, anotándose los resultados. Al individuo B se le asigna un objeto diferente que mide en otras 10 ocasiones. Razona qué parámetro (o parámetros) estadístico consideras más apropiado para efectuar la comparación.*

Ejercicio 20. *Razona si son verdaderas o falsas cada una de las siguientes afirmaciones:*

(a) *Si una muestra de datos presenta media 0, su desviación típica será pequeña.*

(b) *Cuanto mayor es el tamaño de la muestra, mayor es su varianza.*

(c) *Cuanto mayor es el tamaño de la muestra, mayor es su media.*

(d) *Si $g_1 \simeq 0$ la media y la mediana deben ser parecidas.*

Ejercicio 21. *La calificación final de cierta asignatura consiste en la media ponderada entre los resultados de tres exámenes, A, B y C, a los que se les asigna unos pesos del 50 %, 30 % y 20 %, respectivamente. Indica la calificación final que corresponde a cada uno de los tres alumnos de la Tabla 1.5.*

Alumno	Examen A	Examen B	Examen C	Calificación final
Alumno 1	7	3	10	
Alumno 2	2	8	5	
Alumno 3	5.1	5.1	5.1	

Tabla 1.5: Calificaciones.

Ejercicio 22. *Se ha desarrollado una nueva vacuna contra la difteria para aplicarla a niños. El nivel de protección estándar obtenido por antiguas vacunas es de 10 $\mu g/ml$ un mes después de la inmunización. Se han obtenido estos datos*[5] *del nivel de protección de la nueva vacuna al transcurrir un mes:*

12.5	13.5	13	13.5	13
12.5	13.5	14	13.5	13
13	14	14.5	13	12
13.5	13.5	12.5	12.5	12.5

(a) *Representa el diagrama de barras para las frecuencias relativas acumuladas.*

(b) *Calcula la media, mediana, desviación típica y rango intercuartílico.*

(c) *¿Qué proporción de datos son inferiores o iguales a 13?*

Ejercicio 23. *Considera los datos del Ejemplo 3.*

(a) *Obtén mediante una calculadora o un programa estadístico los valores de la media aritmética, la desviación típica y el coeficiente de variación.*

(b) *Obtén, a partir del diagrama tallo-hoja, la mediana y el rango intercuartílico.*

(c) *Indica un par de números que resuman lo mejor posible esos 35 datos.*

(d) *Razona cuál debe ser el signo del coeficiente de simetría.*

Ejercicio 24. *Indica qué tiene que ocurrir exactamente para que, en una muestra de 40 datos de cierta variable numérica, obtengamos como media aritmética y desviación típica los siguientes valores: $\overline{x} = 23.1$, $s = 0$.*

Ejercicio 25. *Describe de manera concisa qué podemos decir de un individuo varón cuya estatura en metros tipificada (respecto a la media y desviación típica de los varones de su franja de edad y su comunidad autónoma) sea igual a -0,02. ¿Cuál será el valor tipificado de su estatura si la medimos en centímetros?*

Ejercicio 26. *En la Figura 1.9 y en la Tabla 1.6 se describe el consumo acumulado de tabaco medido en 452 sudafricanos. Comenta los aspectos más destacados de la distribución de los datos y selecciona un par de parámetros que resuman lo mejor posible la información que contiene la muestra.*

[5] Basado en un informe del *Journal of Family Practice*, enero 1990.

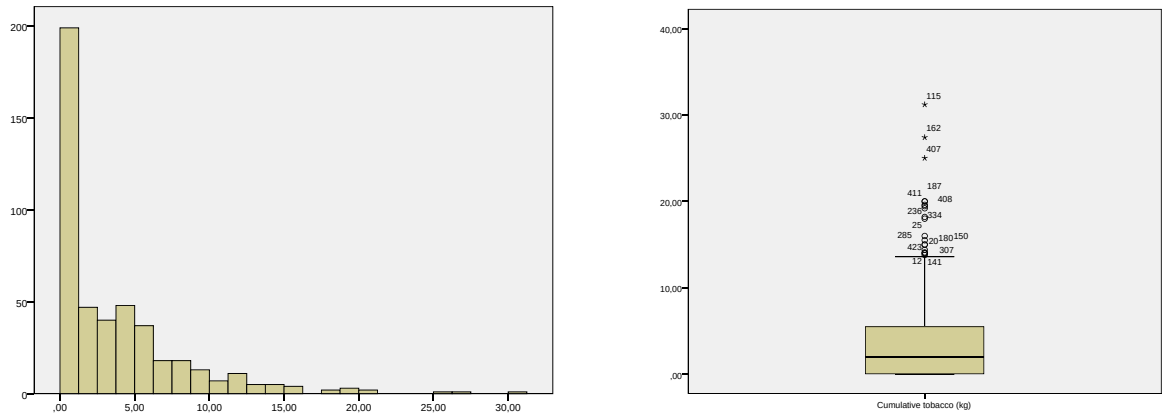


Figura 1.9: Consumo de tabaco en Sudáfrica.

Descriptivos

		Estadístico
Media		3,6356
95% de intervalo de confianza para la media	Límite inferior	3,2157
	Límite superior	4,0556
Media recortada al 5%		3,0670
Mediana		2,0000
Varianza		21,096
Desviación estándar		4,59302
Mínimo		,00
Máximo		31,20
Rango		31,20
Rango intercuartil		5,45
Asimetría		2,079
Curtosis		5,968

Tabla 1.6: Consumo de tabaco en Sudáfrica.

Ejercicio 27. *Tipifica los valores correspondientes al peso en* kg *de 10 personas: 35, 92, 71, 64, 72, 101, 45, 83, 60, 72. ¿Cómo se interpreta una puntuación tipificada positiva? ¿Y negativa? ¿Cuáles serán las puntuaciones tipificadas de los mismos datos expresados en gramos?*

Ejercicio 28. *Cuando los datos de una variable se ajustan aproximadamente a un modelo de distribución normal, la distribución de las puntuaciones tipificadas sigue a su vez un modelo de distribución que se denomina normal estándar, cuya media es 0 y cuya desviación típica es 1. El modelo se denota por* $N(0,1)$*. Es frecuente, en general, calificar como extremos a los datos más alejados del centro de la distribución hasta completar un 5 %. Si la distribución es del tipo campana de Gauss, serán entonces calificados como extremos los datos cuya distancia a la media sea superior al doble de la desviación típica. ¿Por qué? ¿Cómo debe ser la puntuación tipificada de un dato extremo en una campana de Gauss, es decir, qué caracteriza a los valores extremos en una distribución normal estándar?*

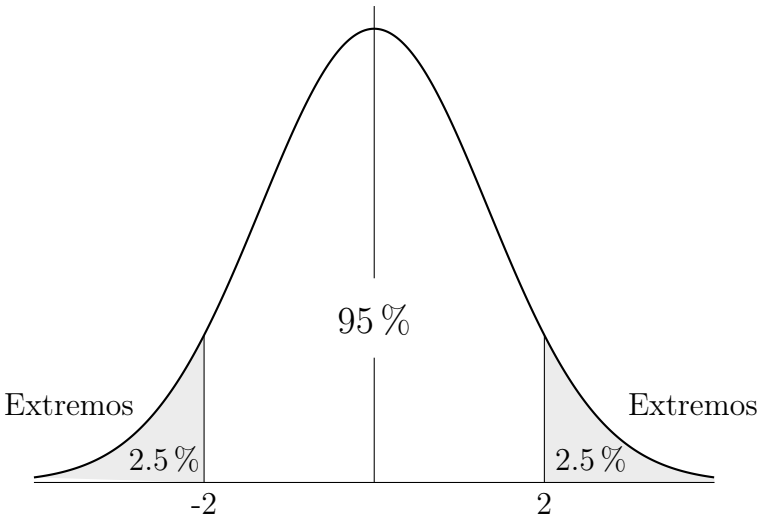


Figura 1.10: Distribución normal $N(0, 1)$.

Ejercicio 29. *Un total de 100 jugadores lanza tres dados cada uno y suman sus puntuaciones, obteniéndose 100 números entre el 3 y el 18 cuyo histograma se representa en la Figura 1.11. ¿Cómo se explica a nivel intuitivo que los datos se ajusten aproximadamente a una curva normal? Según el gráfico, ¿cuál es aproximadamente el valor de la media? ¿Y el de la mediana? ¿Y el de la desviación típica?*

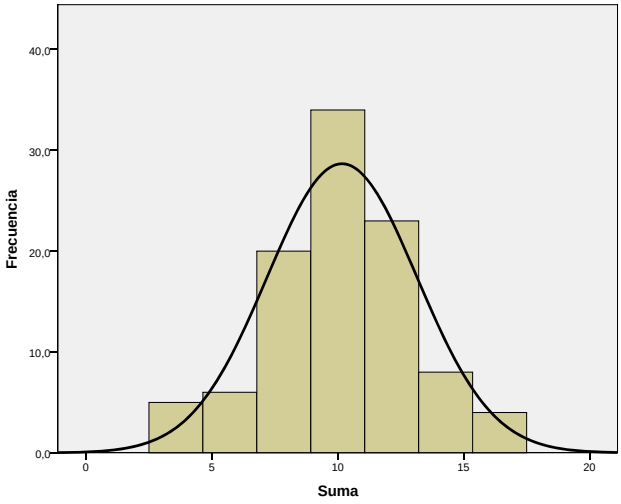


Figura 1.11: Suma de tres dados tras los lanzamientos de $n = 100$ jugadores.

Ejercicio 30. *En la Figura 1.12 se muestra el histograma correspondiente a la edad de 160 enfermos coronarios. Razona brevemente si la media aritmética será menor, mayor o aproximadamente igual que la mediana. Representa esquemáticamente un diagrama de caja posible para estos datos.*

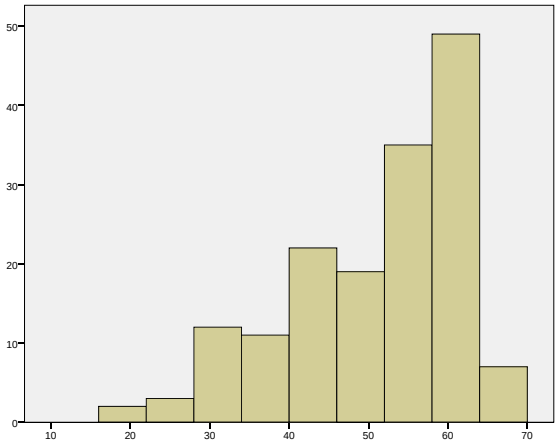


Figura 1.12: Edad de enfermos coronarios.

2. Relación entre variables numéricas

En este capítulo iniciamos la parte realmente interesante del estudio estadístico. Entendemos que existe relación o dependencia entre dos variables cuando un cambio en el valor de una de ellas se asocia a un cambio en el de la otra. La situación contraria, es decir, la ausencia de relación, se denomina independencia. Por ejemplo, nada nos hace pensar que un valor mayor o menor en la última cifra del DNI se asocie a un valor mayor o menor en la concentración de colesterol en sangre, por lo que, en principio, podemos pensar que ambas variables son independientes. Por contra, si observamos la Tabla 1.4 podemos comprobar cómo los cambios en el diagnóstico médico se asocian a cambios en los tiempos medios (y medianos) de baja de los trabajadores, por lo que podemos pensar que ambas variables, diagnóstico y duración de la baja, están relacionadas. Recordemos que, tal y como indicamos en la Tabla 1, a nivel muy básico podemos distinguir tres tipos distintos de relaciones. En este capítulo nos centraremos principalmente en la relación entre dos variables numéricas y trataremos muy brevemente el estudio de la relación entre una variable cualitativa y otra numérica, que se abordará de manera más exhaustiva en la segunda parte del manual. El estudio de la relación entre variables cualitativas lo abordaremos en el siguiente capítulo.

Hemos de precisar que la evidencia de una dependencia o asociación estadística no equivale a la existencia de una relación causa-efecto. Esta última vinculación tiene implicaciones más profundas que, desde una perspectiva estadística, sólo pueden ser analizadas, si acaso, en estudios multifactoriales que apenas estudiaremos aquí (véase Sección 5.5.2).

Para llevar a cabo el estudio de relación entre dos variables numéricas es preciso efectuar un análisis previo de las mismas por separado, según vimos en el capítulo anterior. Dado que nos encontramos en un contexto descriptivo, el análisis de las relaciones consiste fundamentalmente en representarlas gráficamente y calcular los respectivos valores típicos. Así pues, supongamos que contamos con n individuos o unidades experimentales sobre los que se miden numéricamente dos caracteres, dando lugar a sendas variables cuantitativas X e Y. De la medición de dichos caracteres sobre las unidades experimentales resultarán n pares de datos numéricos, que se denotarán así: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. La primera componente del par (x_i, y_i), es decir, el valor x_i, corresponde a la medición de X en la i-ésima unidad experimental y la segunda corresponde a la variable Y. Veamos un ejemplo

de carácter didáctico con una pequeña muestra de tamaño $n = 12$.

Ejemplo 4. Se indica a continuación el peso, X, (kg) y la estatura, Y, (cm) de 12 personas:

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
X	80	45	63	94	24	75	56	52	61	34	21	78
Y	174	152	160	183	102	183	148	152	166	140	98	160

Tabla 2.1: Peso y altura de 12 personas.

El estudio debe empezar con una estadística descriptiva de cada variable por separado, que podría incluir sendos histogramas, así como al menos una medida de centralización y otra de dispersión (en principio estamos considerando la media y la desviación típica). A continuación, nos dedicaremos al estudio descriptivo de la relación entre ambas variables. En el caso numérico continuo las tablas de frecuencia no tienen interés ya que las parejas de datos no suelen repetirse. No ocurrirá lo mismo en el estudio de dos variables cualitativas.

2.1. Diagrama de dispersión

Así pues, lo primero que nos interesa realmente es la representación gráfica de la muestra. Esta tarea debe realizarse con un programa estadístico aunque, en este caso y dado el escaso tamaño de la misma, podríamos hacerlo nosotros mismos. El gráfico más adecuado para apreciar la relación entre dos variables numéricas es el denominado diagrama de dispersión o nube de puntos, que consiste en identificar cada unidad experimental (x_i, y_i) con el punto del plano que tenga por coordenadas x_i para el eje OX e y_i para OY. De esta forma, los datos anteriores se ilustran en la Figura 2.1.

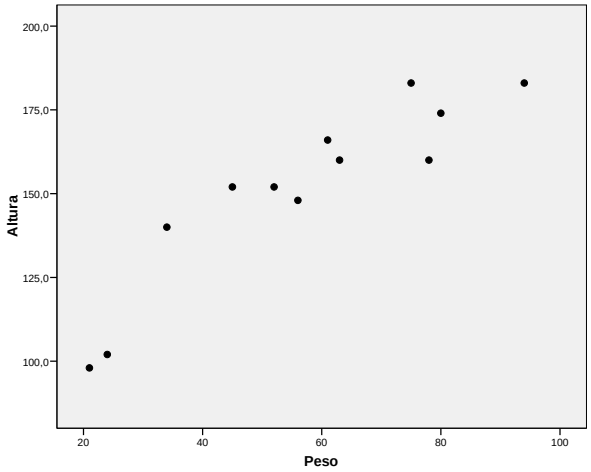


Figura 2.1: Diagrama de dispersión para las variables altura y peso.

En el diagrama de la Figura 2.2 se aprecia la relación entre la presión diastólica y la sistólica medidas en $n = 403$ adultos afroamericanos.

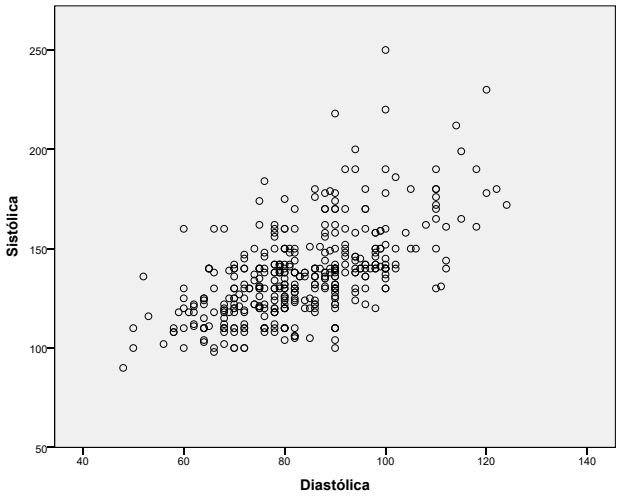


Figura 2.2: Diagrama de dispersión para las variables presión diastólica y presión sistólica.

En ambos casos se observa en la muestra una relación directa o positiva, es decir, que un incremento en los valores de una variable se asocia al incremento de la otra. Para llegar a una conclusión de este tipo es indiferente cuál de las dos variables se identifique con el eje OX. En general, podemos afirmar que tal decisión es intrascendente cuando se trata de un problema de correlación, es decir, cuando estamos interesados simplemente en medir el sentido y la intensidad de una posible relación. No ocurrirá lo mismo cuando estemos ante un problema de regresión, como veremos más adelante.

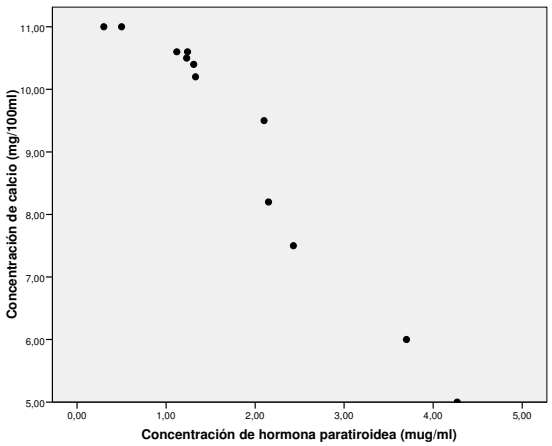


Figura 2.3: Diagrama de dispersión para las variables concentración de hormona paratiroidea, [Pth], y concentración de calcio, [Ca].

El diagrama de dispersión de la Figura 2.3 corresponde a $n = 12$ mediciones de las

concentraciones de hormona paratiroidea ($\mu g/ml$) y de calcio en sangre ($mg/100ml$). En este caso se observa una relación inversa o negativa, pues el aumento en la concentración de la hormona se asocia a una disminución del calcio en sangre. Podemos resaltar que en los tres ejemplos considerados la relación entre el incremento de la variable X y el correspondiente incremento (posiblemente negativo) de Y es constante. Dicho de una manera más gráfica, las nubes que observamos se agrupan en torno a una línea recta, que puede ser creciente o decreciente, según el signo de la relación, y que será plana cuando la relación sea nula. Este tipo de relación se denomina lineal y es el objeto principal de estudio en este capítulo. Con ello no queremos decir que sea la única relación posible, aunque sí es la más sencilla. Además, más adelante veremos que, en la práctica, puede servirnos como referencia para abordar problemas en los que las relaciones que se observan no son lineales.

2.2. Coeficientes de correlación y determinación

Abordamos a continuación el cálculo de valores típicos. En primer lugar, necesitamos conocer la media y desviación típica de cada una de las variables por separado, es decir,

$$\overline{x} = \frac{\sum_i x_i}{n}, \qquad s_x = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n}},$$

$$\overline{y} = \frac{\sum_i y_i}{n}, \qquad s_y = \sqrt{\frac{\sum_i (y_i - \overline{y})^2}{n}}.$$

En el Ejemplo 4 correspondiente a los datos de peso (X) y altura (Y) de 12 individuos se tiene:

$$\overline{x} = 56.92kg, \quad s_x = 22.96kg, \quad \overline{y} = 151.5cm, \quad s_y = 27.47cm.$$

En segundo lugar, nos interesa calcular un valor típico que exprese el grado de relación (o correlación) lineal entre ambas variables observado en la muestra. Al contrario que los parámetros anteriores, dicho valor debe conjugar las informaciones que aportan ambas variables.

Covarianza: la covarianza muestral es una primera medida del grado de correlación y se define mediante

$$s_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n}.$$

La covarianza, que en el caso del Ejemplo 4 se expresará en $kg{\cdot}cm$, puede ser tanto positiva como negativa, pero puede probarse que debe estar comprendida entre los siguientes valores:

$$-\, s_x \cdot s_y \;\leq\; s_{xy} \leq \; +\, s_x \cdot s_y.$$

En ese caso y teniendo en cuenta las desviaciones típicas calculadas antes para el Ejemplo 4, s_{xy} debe estar comprendida entre -630.71 y 630.71. A través del programa estadístico obtenemos su valor concreto en este caso, que es $s_{xy} = 577.86\ kg \cdot cm$. Según eso, en la Figura 2.1 se observa una alto grado de correlación lineal positiva. Observando bien la Figura 2.4 podremos entender el porqué.

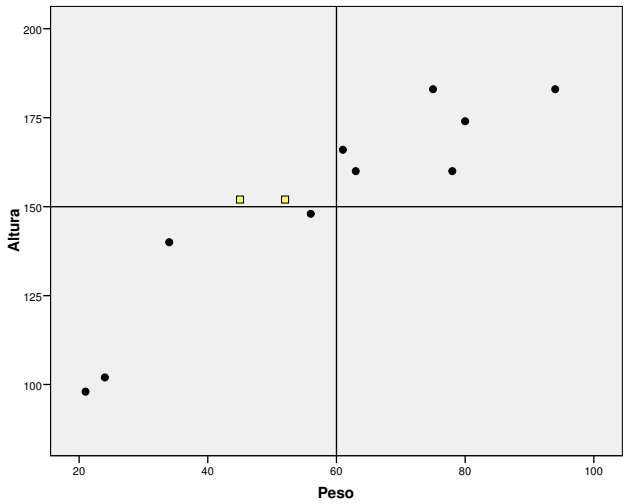


Figura 2.4: Covarianza.

Efectivamente, las líneas de referencia se corresponden con las medias $\overline{x}$ y $\overline{y}$ y determinan cuatro cuadrantes. Los puntos que se encuentran en los cuadrantes superior derecho e inferior izquierdo aportan sumandos positivos a la expresión $\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$ mientras los que se encuentran en los restantes aportan sumandos negativos. En este caso abunda claramente lo primero, razón por la cual la suma resultante será un número positivo y bastante grande. En general, podríamos decir:

- Un valor positivo de s_{xy} significa una tendencia creciente en la nube de puntos, es decir, si los valores de X crecen, los de Y también. Existirá por tanto correlación directa entre ambas variables, según la muestra. El caso extremo $s_{xy} = +s_x \cdot s_y$ representa una correlación lineal perfecta, es decir, que la nube de puntos esté incluida en una única recta, que será además creciente (véase Figura 2.5, izquierda).

- Un valor negativo de s_{xy} significa una tendencia decreciente en la nube de puntos, es decir, si los valores de X crecen, los de Y decrecen. Existirá por tanto correlación inversa entre ambas variables, según la muestra. El caso extremo $s_{xy} = -s_x \cdot s_y$ representa una correlación lineal perfecta, es decir, que la nube de puntos esté incluida en una única recta, que será además decreciente (véase Figura 2.5, derecha).

- El caso $s_{xy} \simeq 0$ se traduce, por contra, en la ausencia de relación lineal en los datos de la muestra (véase Figura 2.5, centro).

Para evaluar qué entendemos por grande o pequeño cuando hablamos de la covarianza hemos de tener en cuenta la cota máxima que se puede alcanzar, es decir, $s_x \cdot s_y$. Dicha cota no es universal, de hecho, un cambio de unidades (pasar de centímetros a metros, por ejemplo), hace variar tanto las desviaciones típicas como la covarianza. Este hecho complica la interpretación del parámetro s_{xy}. Nos interesa pues otro parámetro que se interprete de forma análoga pero que sea adimensional.

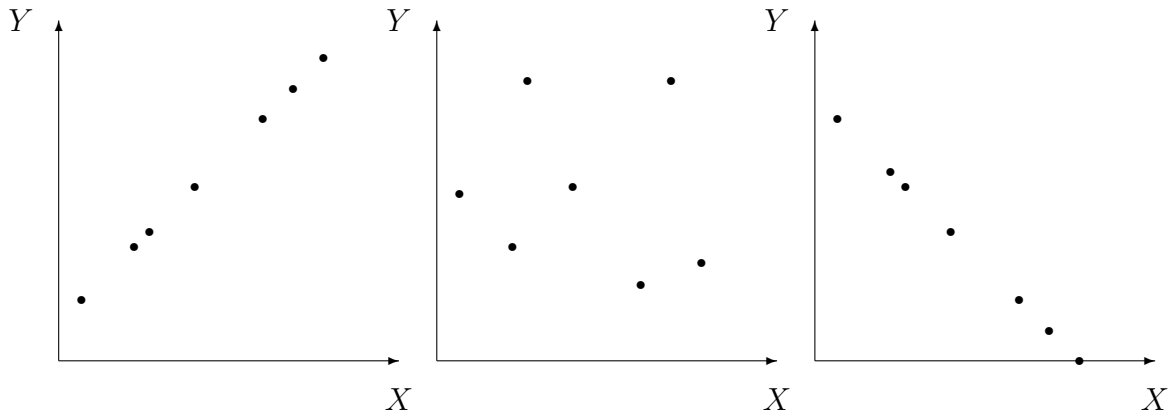


Figura 2.5: Caso $s_{xy} = s_x s_y$ (izquierda); caso $s_{xy} \simeq 0$ (centro); caso $s_{xy} = -s_x s_y$ (derecha).

Coeficiente de correlación lineal de Pearson: supone una medida adimensional de grado de correlación lineal observado en la muestra y se define como sigue:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}.$$

Este parámetro, que se denota normalmente de forma simplificada por r, se interpreta en los mismos términos que la covarianza con la salvedad de que se encuentra en todo caso entre -1 y 1 y alcanza esos valores cuando se da en la muestra una correlación lineal perfecta, bien sea inversa o directa, respectivamente. La proximidad a 0 indica que en la muestra se observa escasa correlación lineal. Así, a los datos del Ejemplo 4 le corresponde $r = 0.916$.

Ejercicio 31. *¿En qué dimensiones se expresará el coeficiente r en el Ejemplo 4?*

Ejercicio 32. *¿Cómo se interpetaría un valor $r = -1.2$?*

Ejercicio 33. *¿Qué le sucede a r si permutamos las variables en el Ejemplo 4, es decir, si identificamos el peso con el eje OY y la altura con el eje OX?*

Coeficiente de determinación: no es más que el cuadrado del anterior, es decir, r^2_{xy}. Como veremos más adelante, goza de una interpretación aún más clara que r. En el caso del Ejemplo 4 tenemos $r^2 = 0.839$.

A la Figura 2.2 le corresponde un coeficiente de correlación r =0.597, lo cual expresa una correlación positiva pero no tan fuerte como la observada en el Ejemplo 4, cosa que debe quedar clara si en el diagrama de dispersión trazamos las líneas de referencia que pasan por las medias, como vemos en la Figura 2.6.

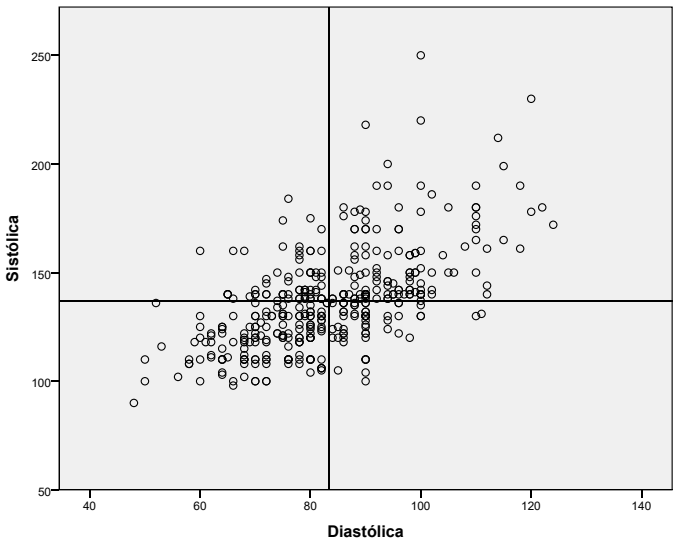


Figura 2.6: Diagrama de dispersión de las variables presión sistólica y presión diastólica.

2.3. Regresión lineal

En el caso de que se observe una fuerte correlación lineal entre los datos de X y los de Y puede ser interesante obtener una ecuación que permita relacionar de manera aproximada ambas variables. Esto es de especial interés cuando una de las variables puede medirse de manera sencilla pero otra no. Efectivamente, si entre ambas existe un alto grado de correlación, el valor de la primera puede utilizarse para pronosticar con mayor o menor fiabilidad el de la segunda. Por ejemplo, la longitud del fémur (mm) en un feto de 26 semanas puede medirse de forma sencilla mediante un ecógrafo. Si dicha longitud correlaciona con el peso (gr), podemos hacer uso de la misma para predecirlo. En nuestro caso, dado que estamos considerando por el momento relaciones exclusivamente lineales, la ecuación que buscamos será del tipo:

$$Y = B_0 + B_1 X,$$

y se denomina ecuación de regresión lineal simple (muestral). Se corresponde obviamente con un recta de pendiente B_1 y término independiente B_0. Parece lógico pensar que la recta idónea será la que mejor se ajuste a nuestra nube de puntos, aunque habrá que especificar primeramente que entendemos por "ajuste". En nuestro caso utilizaremos un criterio muy utilizado en Matemáticas conocido como el de criterio de Mínimos Cuadrados, cuya conveniencia fue argumentada hace casi dos siglos por el propio Gauss. A continuación explicamos en qué consiste dicho criterio.

Como hemos dicho, una recta en el plano puede expresarse de la forma $Y = B_0 + B_1 X$. Dada una unidad experimental de la muestra (x_i, y_i), al valor x_i correspondiente a la variable X (abcisas) le corresponde, según la recta anterior, el valor $B_0 + B_1 x_i$ para la variable Y (ordenadas). La diferencia entre dicho valor y el que realmente corresponde

a la variable Y, es decir, y_i, se entiende como el error cometido al intentar explicar y_i mediante la ecuación anterior. El método de Mínimos Cuadrados propone cuantificar el error total mediante la suma de los cuadrados de los errores particulares, como ocurre en el cálculo de la varianza, es decir,

$$\sum_{i=1}^{n}[y_i - (B_0 + B_1 x_i)]^2.$$

La recta que minimice dicho error será la solución deseada. Puede probarse que, en general, adopta los siguientes parámetros:

$$\begin{aligned} B_1 &= s_{xy}/s_x^2, \\ B_0 &= \overline{y} - B_1\overline{x}. \end{aligned}$$

En la Figura 2.7 se muestra el diagrama de dispersión simple para el peso y la longitud de fémur de 40 fetos de 26 semanas, así como la recta de regresión lineal correspondiente a esta muestra concreta de datos, cuya ecuación resulta ser

$$\text{Peso} = -29.1 + 13.1\text{Fémur}.$$

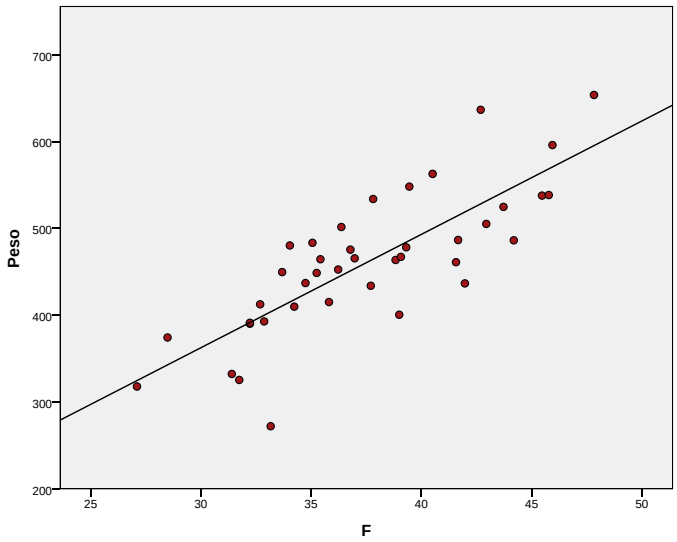


Figura 2.7: Diagrama de dispersión de las variables longitud de fémur y peso y recta de regresión.

A la vista del gráfico anterior cabe realizar tres observaciones:

- El signo de B_1 es el que le otorga la covarianza s_{xy}, que a su vez coincide con el de r. Es decir, que si la correlación es directa, la recta de regresión tiene pendiente positiva; si es inversa, negativa, y si es nula, la pendiente de la recta también lo será.
- En todo caso, la recta pasará por el punto $(\overline{x}, \overline{y})$. Por decirlo de alguna forma, pasa por el centro de la nube de puntos.

- La recta de regresión puede calcularse siempre, independientemente del grado de correlación existente entre las variables.

Ejercicio 34. *¿Es importante determinar qué variable identificamos con el eje OX antes de calcular la ecuación de la recta de regresión o, por el contrario, resulta indiferente cuál de las dos desempeña ese papel?*

Ejercicio 35. *¿Qué peso predecirías a un feto cuyo fémur mide 35mm?*

Ejercicio 36. *Según la ecuación de regresión, ¿cuántos gramos aumenta o disminuye por término medio el peso del feto por cada milímetro más de fémur?*

En la Figura 2.8 se representa la recta de regresión lineal correspondiente a la muestra del Ejemplo 4, cuya ecuación resulta ser Altura=89.11+1.10Peso. En este caso, el interés práctico de la ecuación es discutible pues ambas variables pueden medirse trivialmente.

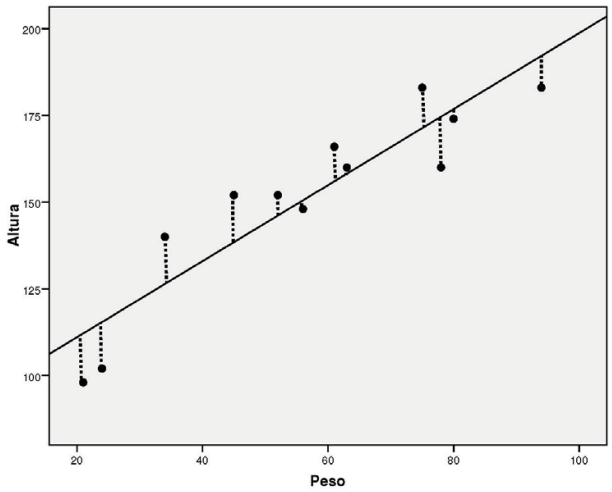


Figura 2.8: Diagrama de dispersión de las variables peso y altura y recta de regresión.

Varianza residual: en la Figura 2.8 hemos marcado para cada punto una linea discontinua que expresa el error cometido por la recta en su predicción. Desde un punto de vista numérico, en la primera columna de la Tabla 2.2 se muestran los valores de X para los 12 datos de la figura; en la segunda, los correspondientes valores de Y; en la tercera, los valores de las ordenadas que se obtienen según la recta de regresión $y = 89.11 + 1.10x$; en la cuarta columna tenemos precisamente las diferencias al cuadrado entre los valores reales de Y y sus predicciones, de manera que su suma cuantifica el error cometido por la recta de regresión. La suma de esos errores divida entre n se denomina varianza residual. La varianza residual viene a expresar pues la parte de la variabilidad de los datos de Y no explicada por la variabilidad de los datos de X mediante la recta de regresión lineal. Por último, en la quinta columna de la tabla aparecen los cuadrados de las diferencias entre los valores reales de Y y su media. La suma dividida entre n es la varianza (total) s_y^2.

x_i	y_i	$(B_0 + B_1 x_i)$	$[y_i - (B_0 + B_1 x_i)]^2$	$(y_i - \overline{y})^2$
80	174	176.80	7.86	506.25
45	152	138.44	183.94	0.25
63	160	158.17	3.36	72.25
94	183	192.15	83.70	992.25
24	102	115.42	180.05	2450.25
75	183	171.32	136.37	992.25
56	148	150.50	6.23	12.25
52	152	146.11	34.69	0.25
61	166	155.98	100.48	210.25
34	140	126.38	185.51	132.25
21	98	112.12	199.66	2862.25
78	160	174.61	213.47	72.25
	$\overline{y} = 151.5$		**1335.32**	**8303.00**

Tabla 2.2: Tabla para el cálculo de la varianza residual.

El cociente entre la varianza residual y la total se entiende pues como la proporción de variabilidad total de Y que no es explicada la regresión, en nuestro caso $1335/8303 = 0.161$. Parece lógico que este valor guarde alguna relación con el coeficiente de correlación $r = 0.91$ y, efectivamente, ocurre en este caso que $0.161 = 1 - r^2$. Puede probarse sin mucha dificultad que esa igualdad se verifica en general y que, por lo tanto, $1 - r^2$ es la proporción de la variabilidad de Y no explicada linealmente por X, es decir:

> Podemos interpretar el coeficiente de determinación r^2 como la proporción de variabilidad de Y que sí es explicada linealmente por X.

Así pues, en este caso, el peso explica un 83.9 % (valor de r^2) de la variabilidad de la altura (mediante la recta de regresión). Recíprocamente, la altura explica un 83.9 % de la variabilidad del peso o, mejor, la altura y el peso comparten un 83.9 % de su variabilidad, hecho que se pretende ilustrar esquemáticamente en la Figura 2.9.

En el caso de la predicción del peso de fetos mediante la longitud de su fémur, la muestra aporta un valor de $r^2 = 0.643$ ($r = 0.802$), lo cual se traduce en que, en esta muestra concreta, la recta de regresión permite explicar a partir de la longitud del fémur un 64.3 % de la variabilidad del peso o, lo que es lo mismo, que conlleva un 35.7 % de error global. Obviamente, r^2 mide globalmente la fiabilidad de las predicciones. En la segunda parte ampliaremos este estudio valorando dicha fiabilidad de manera más precisa (véase ecuación (5.4)), aunque podemos adelantar que el margen de error atribuible a una

predicción concreta efectuada a partir de la recta de regresión depende principalmente de los valores de r^2 y n.

Ejercicio 37. *¿En qué sentido crees que influirán los valores de r^2 y n en el error cometido por la recta de regresión?*

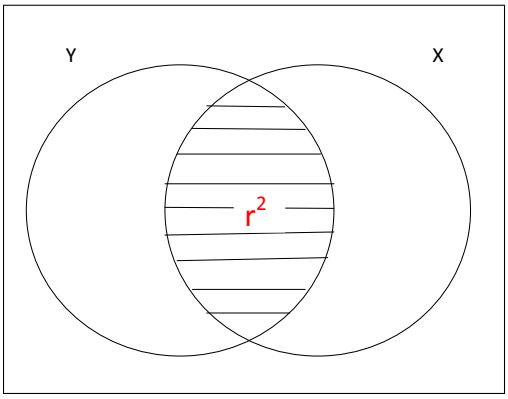


Figura 2.9: Interpretación esquemática de r^2.

Los casos extremos en el análisis de r^2 son $r^2 = 1$ y $r^2 = 0$. El primero se corresponde con una varianza residual nula, es decir, con el caso en que la recta de regresión lineal predice sin error los datos de Y a partir de X, y por tanto, se trata de una correlación lineal perfecta. El caso $r^2 = 0$ se corresponde con un varianza residual que iguala a la total, es decir, que la recta de regresión no ayuda en absoluto a reducir la incertidumbre inicial respecto a la variable Y y en consecuencia, corresponde con una recta de regresión de pendiente nula, es decir, constante. Concretamente, se trata de la constante $\overline{y}$, por ser la opción menos mala posible. Una situación similar ocurre en la Figura 2.10 cuando estudiamos la relación entre la talla y el IMC en 100 individuos adultos, a cuya muestra le corresponde $r = -0.035$.

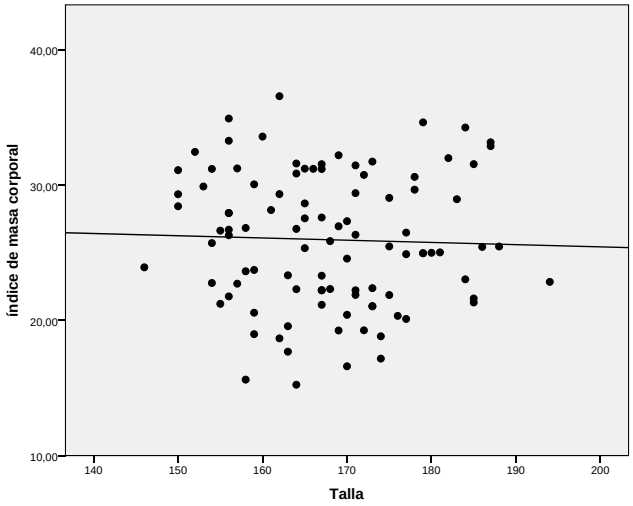


Figura 2.10: Diagrama de dispersión de las variables talla e IMC y recta de regresión.

Ejercicio 38. *¿Cómo interpretamos el valor de $r = -0.035$ en la Figura 2.10? ¿Te resulta paradójico? ¿Cómo será r si reemplazamos la talla por el peso: positivo, negativo o próximo a 0?*

Ejercicio 39. *En el ejemplo de relación entre el peso y la longitud del fémur del feto, ¿afectaría al valor de r^2 el hecho de expresar el peso en kg en lugar de en gr?*

Ejercicio 40. *En el mismo ejemplo, si reemplazamos la muestra de $n = 40$ fetos por otra diferente, de otros 40 fetos, por ejemplo, ¿obtendremos un mismo valor de r^2? ¿Obtendremos una misma ecuación de regresión? ¿Serán parecidas?*

2.3.1. Regresión lineal múltiple

Ya hemos visto que en lo que respecta a las variables peso y longitud de fémur (F), el grado de correlación observado en la muestra de $n = 40$ fetos es $r = 0.802$, por lo que la ecuación de regresión obtenida para dicha muestra, Peso=-29.1+13.1F permite explicar un 64.3 % (r^2) de la variabilidad del peso. Dependiendo del grado de fiabilidad que necesitemos en la predicción, la cual depende a su vez de n y r^2, la proporción anterior resultará grande o pequeña. Es decir, que si queremos mejorar la fiabilidad debemos incrementar el tamaño de la muestra o escoger otra variable con una correlación con el peso superior a la del fémur. Podríamos optar, en principio, por otras medidas del ecógrafo, como la circunferencia craneal (C) o la abdominal (A), pero ninguna de ellas presenta un grado de correlación con el peso superior al que presenta el fémur.

En situaciones como estas es más interesante añadir más variables independientes para predecir la variable dependiente Y a través de una ecuación lineal; en nuestro caso utilizaríamos las tres variables medidas directamente por el ecógrafo, F, C y A, como variables independientes X_1, X_2 y X_3 en una ecuación de tipo lineal cuya variable dependiente, Y, sea el peso. Es decir, se trata de construir, a partir de la muestra, una ecuación del tipo

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3.$$

En general, la ecuación concreta que buscamos, siguiendo de nuevo el criterio de Mínimos Cuadrados, es la que minimice la suma

$$\sum_{i=1}^{n}[y_i - (B_0 + B_1x_1 + B_2x_2 + B_3x_3)]^2.$$

La solución puede obtenerse mediante cualquier programa estadístico. En el problema del peso del feto, la ecuación de regresión múltiple obtenida para la muestra considerada es

$$\text{Peso} = -149.0 + 12.6 \cdot F + 9.8 \cdot C - 9.4 \cdot A. \tag{2.1}$$

Ejercicio 41. *Según eso, ¿qué peso cabría predecir a un feto con medidas F=43, C=172, A=167?*

Coeficiente R^2 múltiple: para valorar globalmente la fiabilidad de las predicciones que efectuemos mediante la ecuación anterior necesitamos un valor típico que generalice el coeficiente de correlación simple al cuadrado, r^2. Dicho coeficiente, que se obtiene mediante cálculos matriciales, se denomina coeficiente de correlación múltiple al cuadrado, y se denota por R^2. Expresa, por lo tanto, la proporción de variabilidad de Y explicada entre todas las variables independientes. Si sólo contamos con una variable independiente el valor de R^2 es igual al del correspondiente coeficiente de determinación. Nótese que programas estadísticos como SPSS ofrecen por defecto el valor de R^2 en un problema de regresión lineal porque se sobreentiende que la regresión debe ser múltiple. En la Figura 2.11 tenemos una visión esquemática del concepto.

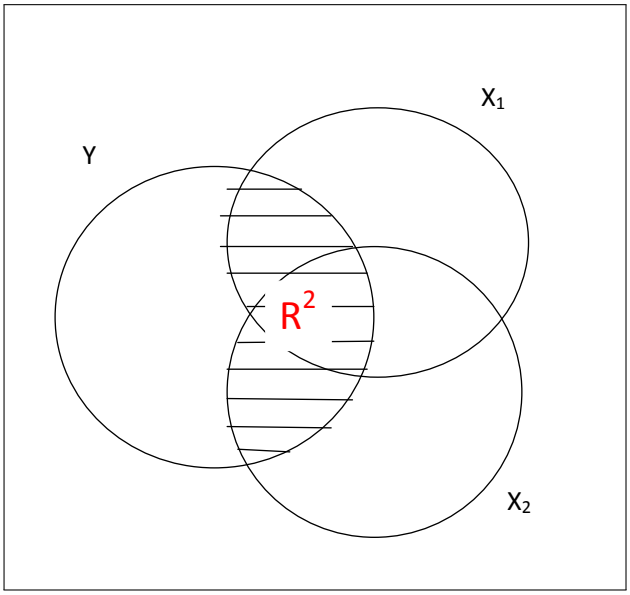


Figura 2.11: Interpretación intuitiva del coeficiente R^2.

Ejercicio 42. *¿Por qué crees que SPSS considera por defecto que la regresión debe ser múltiple en vez de simple?*

Ejercicio 43. *¿Puede disminuir R^2 si se introduce una nueva variable independiente en la ecuación, por ejemplo la longitud de la tibia?*

En el caso del peso del feto, obtenemos un valor $R^2 = 0.915$, lo cual justifica la inclusión de las dos nuevas variables dado que inicialmente teníamos $r^2 = 0.643$. Las predicciones efectuadas a partir de la ecuación (2.1) gozarán de mayor precisión que las correspondientes a la ecuación de regresión simple a partir del fémur, según se cuantifica mediante (5.4) en la segunda parte del manual.

Aunque no profundizaremos en los detalles, el coeficiente R^2 puede ser calculado e interpretado de forma idéntica (proporción de varianza explicada) aunque las variables independientes sean cualitativas o mezclas entre cualitativas y numéricas, como veremos más adelante.

Multicolinealidad: puede llegar a pensarse que el hecho de añadir variables independientes a la ecuación sólo conlleva ventajas, pero no es así. En primer lugar, estas variables hay que medirlas; en segundo lugar, nos impiden tener una visión gráfica sencilla de los datos; por último, pueden generar ciertas confusiones como consecuencia de la posible correlación lineal entre las distintas variables independientes, cosa que puede apreciarse incluso en la ecuación propuesta para el peso del feto. Este problema se denomina multicolinealidad. Lo más aconsejable es introducir una nueva variable en la ecuación solamente si su presencia incrementa sustancialmente el valor de R^2.

Ejercicio 44. *¿Qué aspecto de la ecuación* (2.1) *puede resultar paradójico?*

2.3.2. Regresión no lineal

Hasta ahora hemos afrontado únicamente el estudio de aquellas muestras en las que la relación entre las variables X e Y es de tipo claramente lineal, excluyendo situaciones dudosas como la de Figura 2.12, que corresponde al estudio de relación entre el marcador tumoral PSA y el volumen de un tumor prostático, estudiado en una muestra de $n = 97$ pacientes.

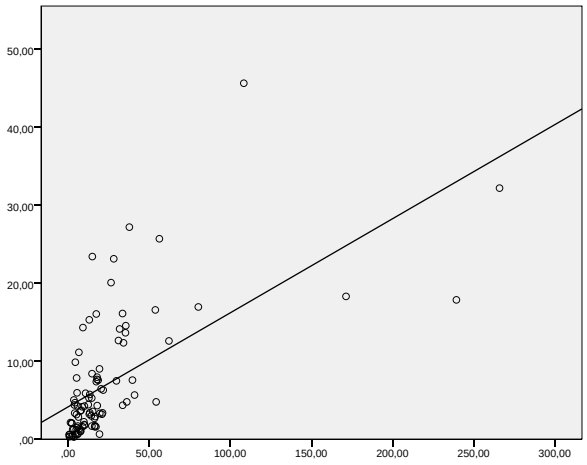


Figura 2.12: Diagrama de dispersión para las variables PSA y volumen tumor próstata, junto a la recta de regresión lineal.

La recta de regresión logra un aceptable ajuste a la nube de puntos, obteniéndose $r = 0.625$. No obstante, un estudio más profundo de ambas variables revela una relación lineal mucho más clara entre los logaritmos del volumen y del PSA, tal y como queda patente en el gráfico de la Figura 2.13, al que corresponde un coeficiente de correlación $r = 0.734$. No se trata de una casualidad, sino que ocurre porque la relación entre variables que se distribuyen según un modelo de campana de Gauss es de tipo lineal[1]. En las Figuras 1.6 y 1.7 apreciábamos que el volumen del tumor presentaba un fuerte sesgo positivo que quedaba anulado tras aplicar la transformación logarítmica. Algo similar ocurre con el

[1]Estrictamente hablando esto no ocurre necesariamente pero es lo más habitual.

PSA, de manera que la relación entre el logaritmo del PSA y el logaritmo del tumor sí es lineal, como se aprecia en la Figura 2.13.

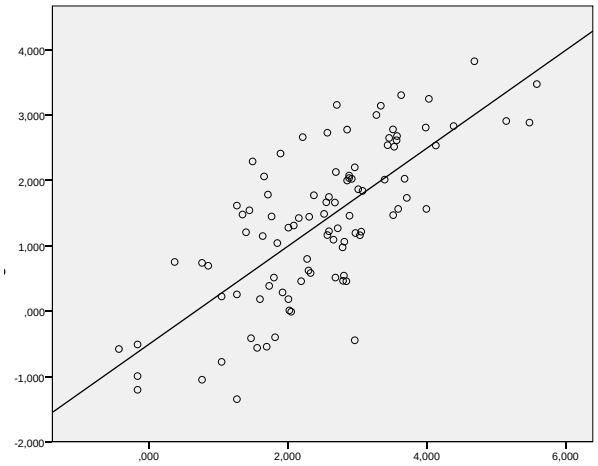


Figura 2.13: Diagrama de dispersión para las variables log(PSA) y log(volumen), junto a la recta de regresión lineal.

La ecuación de la recta de regresión representada en la figura anterior es $y = -0.590 + 0.750x$. Por lo tanto, las variables originales se relacionan aproximadamente según la ecuación:

$$\log \text{vol} = -0.509 + 0.750 \log \text{PSA},$$

luego, despejando, obtenemos $\text{vol} = 0.601 \cdot \text{PSA}^{0.750}$, que es la curva que se representa en la Figura 2.14.

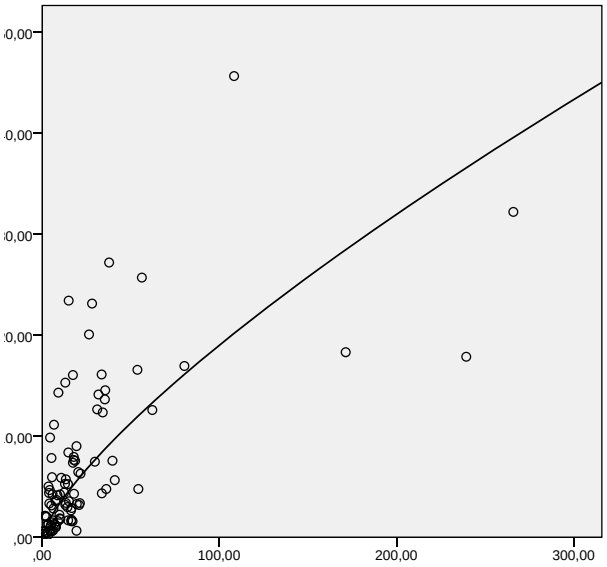


Figura 2.14: Diagrama de dispersión para las variables PSA y volumen del tumor, junto a la curva de regresión no lineal.

Este ejemplo ilustra cómo, en ciertas ocasiones, podemos lograr una mejor explicación de la variable dependiente si no nos restringimos a ecuaciones de tipo lineal, lo cual suele traducirse a grandes rasgos en considerar distintas transformaciones de las variables en juego, en especial la logarítmica. El programa estadístico SPSS ofrece la posibilidad de tantear con diferentes posibilidades. No obstante, debemos advertir que este tipo de estudios puede llegar a ser bastante complicado.

Ejercicio 45. *Si entre dos variables se da una relación de tipo exponencial* $y = a \cdot b^x$*, ¿qué transformaciones debemos aplicar a las variables* X *e* Y *para obtener una relación lineal?*

Ejercicio 46. *A izquierda y derecha de la Figura 2.15 se ilustran la relación entre la esperanza de vida global y la renta per cápita por un lado, y entre la esperanza de vida de los hombres y la de las mujeres por otro, calculadas todas ellas en 2009 para todos los países del mundo*[2]*. Comenta qué te sugiere cada gráfico y cómo crees que se ha llegado a la ecuación de la izquierda.*

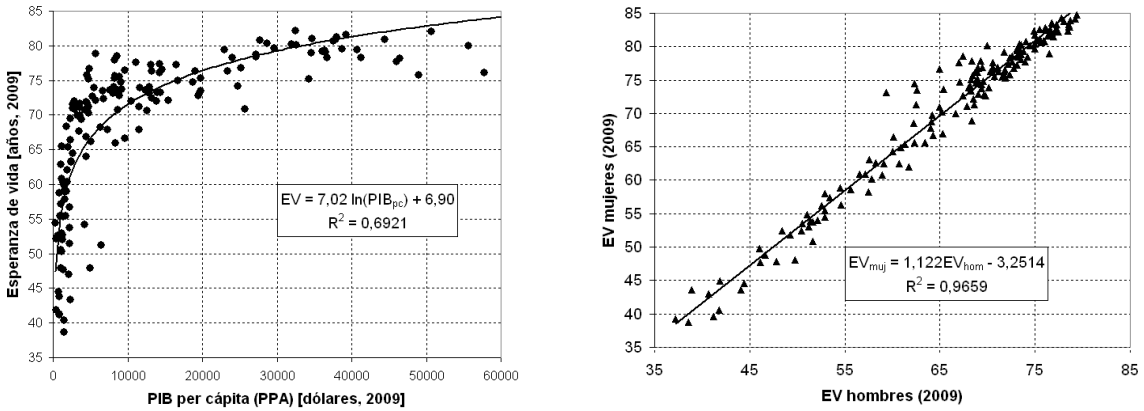


Figura 2.15: Esperanza de vida.

2.4. Relación entre una variable numérica y otra cualitativa

Como ya hemos comentado, este problema lo trataremos de manera más extensa en la segunda parte. El estudio a nivel meramente descriptivo es escueto y hemos optado por ubicarlo en este capítulo porque, desde un punto de vista teórico, el problema se formaliza mediante el mismo modelo que el de regresión.

Ejemplo 5. Se estudia la posible relación entre la acidosis en recién nacidos y la glucemia medida en el cordón umbilical. Para ello se toma una muestra de $n = 200$ recién nacidos distribuidos a partes iguales en cuatro grupos: sanos, enfermos con acidosis respiratoria, con acidosis metabólica y mixta. Los datos quedan representados mediante los diagramas de caja en la Figura 2.16.

[2]Gráficos obtenidos de Wikipedia.

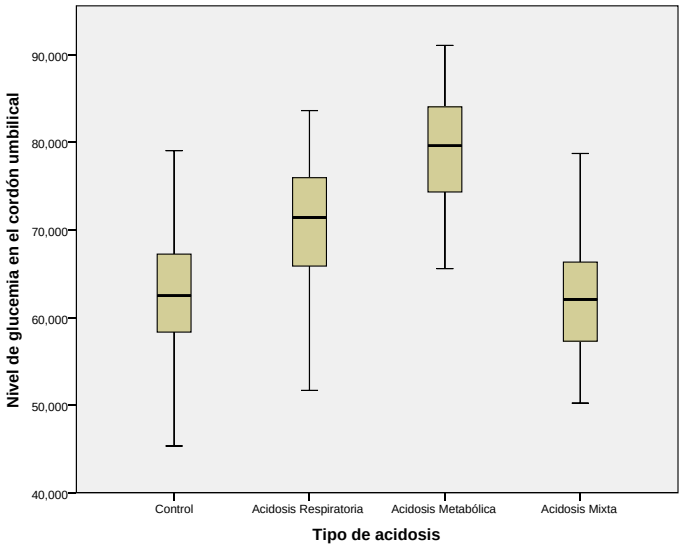


Figura 2.16: Diagramas de caja para la glucemia según el tipo de acidosis.

Podemos observar que los niveles de glucemia son mayores en los enfermos con acidosis respiratoria que en los sanos, al menos por término medio (mediano); que los niveles de glucemia en los enfermos de acidosis metabólica es aún mayor y que los enfermos de acidosis mixta poseen valores de glucemia similares al de los individuos sanos, al menos, insistimos, por término medio. En general, podemos afirmar que:

> La relación entre un variable cualitativa y otra numérica se traduce en la comparación de las medias que dicha variable numérica alcanza en las distintas categorías de la variable cualitativa.

Concretamente, entendemos las distancias entre las medias como una prueba de la relación entre ambas variables, que será más fuerte cuanto mayor sean dichas diferencias. La cuestión es algo más compleja, pues esta distancia debe evaluarse teniendo en cuenta el grado de variabilidad que presentan los datos, lo cual afecta a su vez a la variabilidad de las propias medias aritméticas calculadas. Es una situación análoga a la de regresión lineal, pues se trata en definitiva de medir la proporción de variabilidad explicada por la variable cualitativa, lo cual da lugar a un coeficiente R^2. Ya hemos dicho que no profundizaremos aquí en esa cuestión. En todo caso, el problema de comparación de medias presenta una casuística algo compleja que abordaremos en el contexto de la Inferencia Estadística (segunda parte), mientras que en esta primera parte realizaremos un primer análisis meramente intuitivo a partir del gráfico a partir de diagramas de cajas o de medias (ver tutorial).

2.5. Análisis de la covarianza

Recibe este nombre un tipo de estudio más complejo en el cual se relacionan entre sí al menos dos variables numéricas y una cualitativa. Mejor dicho, se estudia la posible

relación entre dos variables numéricas pero distinguiendo las diferentes categorías de otra variable cualitativa. En tal caso, se puede hablar de un coeficiente de correlación r^2 para cada categoría por separado y de un coeficiente R^2 múltiple, que expresa la proporción de varianza de la variable respuesta numérica explicada conjuntamente por la variable explicativa numérica y por la variable explicativa cualitativa.

Ejemplo 6. A partir de una muestra de $n = 403$ afroamericanos adultos se estudió la relación entre el perímetro de la cintura (Y) y el de la cadera (X) para hombres y mujeres por separado. Desde el punto de vista gráfico, los resultados se presentan en la Figura 2.17, que consiste en un diagrama de dispersión en el que se distinguen ambos sexos por colores y se calculan por separado ambas rectas de regresión lineal.

En este caso se obtuvo un valor $R^2 = 0.739$, es decir, que la variabilidad del perímetro de la cintura se explica en un 73.9 % a partir del de la cadera, con diferentes ecuaciones según el sexo. De hecho, podemos apreciar fundamentalmente que, para valores similares de cadera, los hombres tienden a presentar valores de cintura superiores a los de las mujeres. Eso explica que el índice cintura cadera tienda a ser superior en hombres que en mujeres.

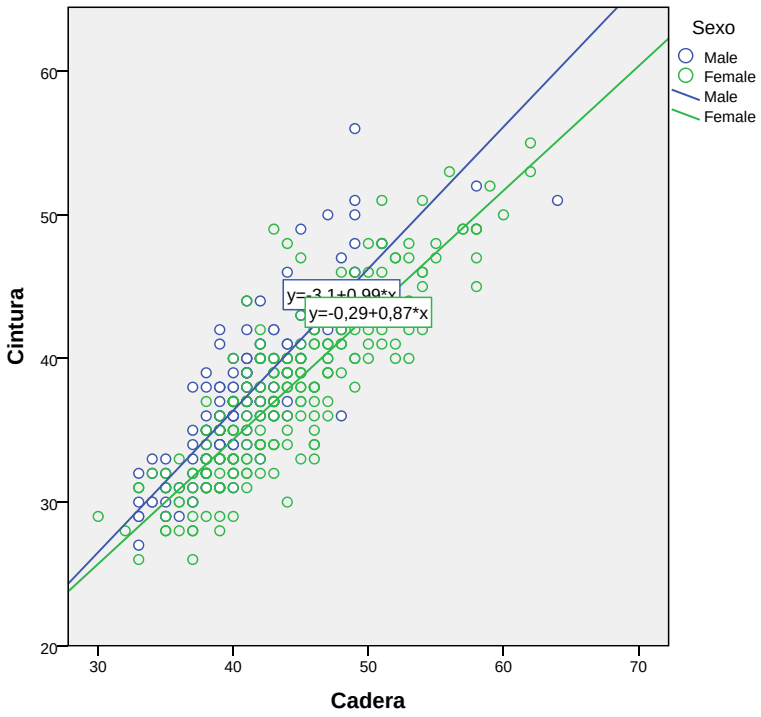


Figura 2.17: Diagrama de dispersión para el perímetro de cintura y el perímetro de cadera por sexos.

Otras cuestiones propuestas

Ejercicio 47. *Indica un ejemplo de 4 pares de datos que presenten un coeficiente de correlación lineal $r = -1$. Indica un ejemplo de 4 pares de datos que presenten un coeficiente*

de correlación lineal $r = 0$.

Ejercicio 48. *Supongamos que contamos con una muestra de tamaño n de una cierta variable X y que procedemos a tipificar los n datos, con lo cual obtenemos otros n valores de una nueva variable Z. Razona cuánto debe ser el valor el coeficiente de correlación lineal r entre X y Z.*

Ejercicio 49. *En un estudio de regresión lineal se obtuvo, a partir de una muestra de tamaño $n = 12$, una recta de regresión lineal $y = 3.2 - 4.1x$, y un coeficiente de correlación lineal $r = 0.93$. ¿Existe alguna contradicción entre estos resultados?*

Ejercicio 50. *Imaginemos que una variable bioquímica es muy interesante desde el punto de vista clínico aunque costosa de medir, pero que no obstante hemos observado, a partir de una muestra de n=341 individuos, una correlación lineal $r = -0.998$ con otra variable mucho más fácil de medir. Razona qué ventaja podemos extraer de este hecho y describe breve pero claramente cómo deberíamos proceder exactamente para sacarle partido a esta correlación. ¿Cómo afectaría al procedimiento el hecho de que la muestra estudiada hubiese sido de tamaño n=30 (suponiendo un coeficiente de correlación r similar)?*

Ejercicio 51. *Indica qué valor aproximado puede tener r en los siguientes ejemplos que se muestran en la Figura 2.18:*

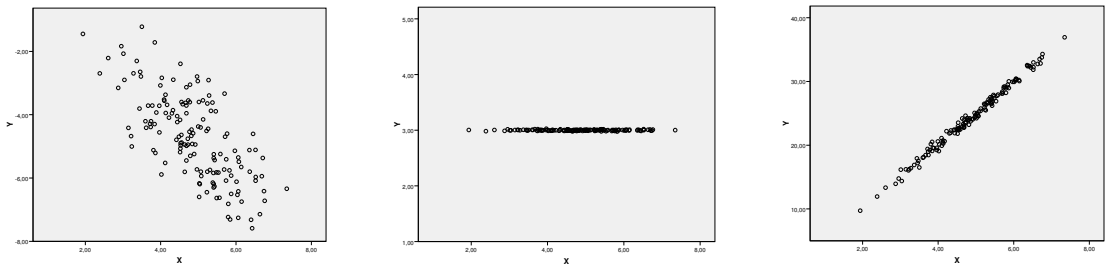


Figura 2.18: Algunos ejemplos de correlaciones.

Ejercicio 52. *Se midieron la presión sistólica (*mmHg*) y la concetración de colesterol LDL (*mg/l*) a $n = 462$ personas obteniéndose, entre otros resultados, los valores típicos que se muestran en la Tabla 2.3:*

	Presión (*mmHg*)	LDL (*mg/l*)
Media	138.33	57.40
Mediana	134	43.4
Desviación típica	20.50	20.71
Rango intercuartílico	24	25.2
Coeficiente de correlación	0.158	

Tabla 2.3: Valores típicos.

(a) Razona brevemente, a partir de estos resultados, cuál de las dos variables posee un mayor sesgo positivo.

(b) Razona cuál debe ser el valor del coeficiente de correlación lineal entre la presión arterial y el LDL si medimos este último en mg/dl.

(c) Se detecta posteriormente a la toma de datos que el medidor de tensión arterial comete un error sistemático consistente en indicar siempre 2 mmHg más de la cuenta. Sabido esto, ¿cuáles deben ser los verdaderos valores de la mediana y el rango intercuartílico de la presión arterial? ¿Cuál debe ser el verdadero valor el coeficiente de correlación lineal entre la presión arterial y el LDL (medido en mg/l*)?*

Ejercicio 53. *El diagrama de dispersión de la Figura 2.19 representa el área de la cabeza y la velocidad para una muestra de $n = 356$ espermatozoides con $r = 0.20$. ¿Qué proporción de variabilidad de la velocidad es explicada linealmente por el tamaño de la cabeza? ¿Qué proporción de variabilidad del tamaño de la cabeza es explicado linealmente por la velocidad? ¿Qué puedes extraer de este dato en términos prácticos?*

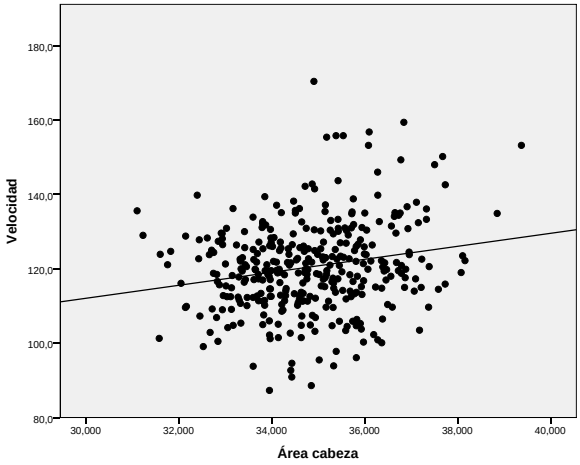


Figura 2.19: Diagrama de dispersión para el área de cabeza y la velocidad del espermatozoide junto a la recta de regresión.

Ejercicio 54. *El sustrato Inosina monofosfato reacciona produciendo Xantosina monofosfato ante la presencia de la enzima IMP de Hidrógeno. Se intenta explicar la velocidad de dicha reacción (medida en incremento de la densidad del producto por minuto) a partir de la concentración de sustrato (medido en* μmoles/l*). Tras medir ambas variables en $n = 7$ ocasiones, con las mismas condiciones ambientales, se obtuvo la Tabla 2.4:*

$[S]$	3.4	5.0	8.4	16.8	33.6	67.2	134.4
V	0.10	0.15	0.20	0.25	0.45	0.50	0.53

Tabla 2.4: Valores de concentración de sustrato ($[S]$) y velocidad (V).

(a) Representa la nube de puntos mediante un programa estadístico.

(b) Realiza el siguiente cambio de variables: $X = 1/[S]$, $Y = 1/V$. *Efectúa un estudio de correlación-regresión lineal entre las variables* X *e* Y *mediante un programa estadístico.*

(c) En general, en los procesos de reacción ante la presencia de una enzima, la velocidad de la reacción se relaciona con la concentración del sustrato según una ley del siguiente tipo:

$$V = \frac{V_{max} \times [S]}{K_m + [S]} ,$$

donde V_{max} *es la velocidad máxima posible en el proceso, que se corresponde con una concentración de sustrato muy grande, y donde* K_m *es una valor constante para condiciones ambientales fijas, denominado constante de Michaellis-Menten. Estima el valor de* K_m *y* V_{max} *en este proceso concreto.*

Ejercicio 55. *Se lleva a cabo un estudio con* $n = 100$ *individuos para determinar si el tipo de dieta influye en el IMC. Para ello, los individuos siguieron dos tipos de dieta, A y B; en concreto, 54 individuos siguen la dieta A y 46 siguen la B. En la Figura 2.20 se muestran los correspondientes diagramas de caja. Responde a la cuestión a un nivel puramente intuitivo.*

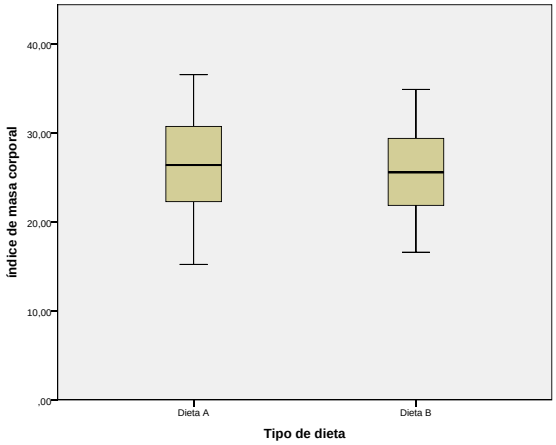


Figura 2.20: Diagramas de caja de IMC según el tipo de dieta.

Ejercicio 56. *En un estudio llevado a cabo en EE.UU. se efectuó un seguimiento de 16 años a una amplia muestra de individuos registrándose los casos en los que los sujetos sufrieron de infarto durante dicho periodo. En los diagramas de la Figuras 2.21, 2.22 y 2.23 se ilustran, respectivamente, las correlaciones observadas entre la edad y el nivel de colesterol sérico al comienzo del estudio, el registro o no de infarto durante el estudio y la edad del individuo al comienzo y, por último, entre el registro o no de infarto y el nivel de colesterol al comienzo. ¿Qué conclusiones te sugieren esos tres gráficos?*

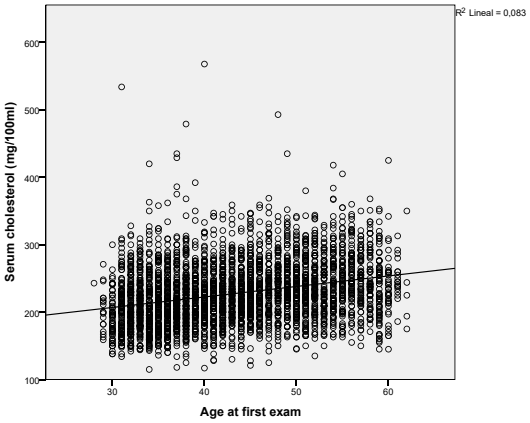


Figura 2.21: Diagrama de dispersión entre la edad y el nivel de colesterol al inicio del estudio, junto a la recta de regresión.

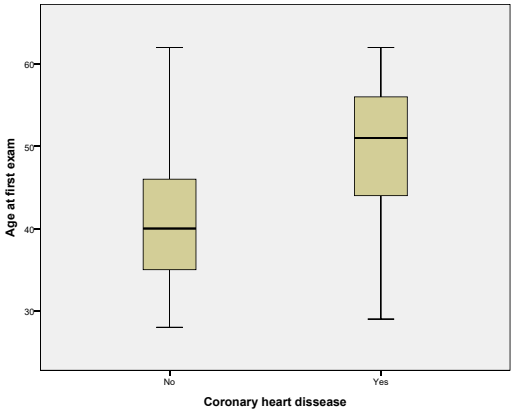


Figura 2.22: Diagramas de caja para la edad según el registro o no de infarto durante el estudio.

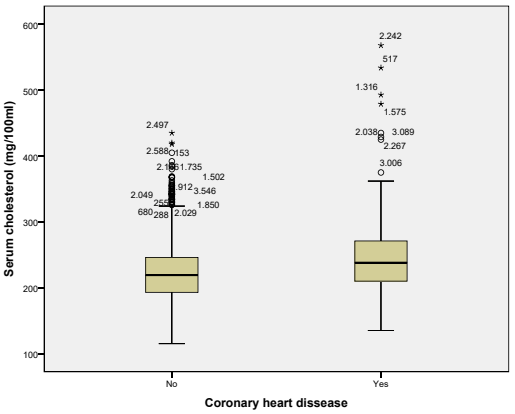


Figura 2.23: Diagramas de caja para el nivel de colesterol al inicio del estudio según el registro o no de infarto durante el estudio.

Ejercicio 57. *Siguiendo con los datos del Ejercicio 56, ¿qué te sugiere el diagrama de dispersión de la Figura 2.24, que relaciona la edad con el colesterol distinguiendo entre los individuos que sufrieron infarto y los que no?*

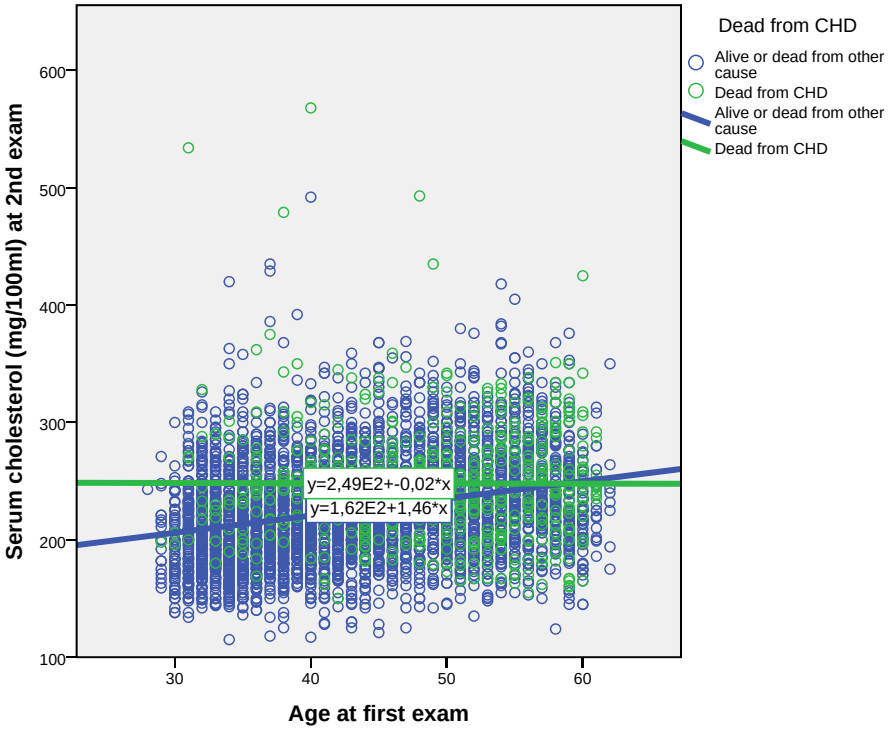


Figura 2.24: Diagrama de dispersión entre edad y nivel inicial de colesterol según el registro o no de infarto.

3. Relación entre variables cualitativas

En el capítulo anterior estudiamos la relación entre dos variables numéricas y entre una numérica y otra cualitativa. Para completar el esquema recogido en la Tabla 1 sólo queda estudiar la relación entre dos variables cualitativas. Entendemos que existe relación entre ambas cuando un cambio de categoría en una variable se asocia a un cambio de categoría en la otra y viceversa. El hecho de expresar un carácter de forma cualitativa puede resultar en principio más sencillo que medirla numéricamente, lo cual explica la abundancia de diseños de tipo cualitativos en la investigación experimental. Paradójicamente, desde un punto de vista meramente estadístico, el tratamiento de las variables cualitativas es mucho más engorroso que el de las numéricas, en especial a la hora de estudiarlas conjuntamente.

3.1. Estudio general de las tablas de contingencia

Comenzaremos con un estudio de carácter general para analizar posteriormente problemas más concretos en el contexto biomédico. En todo caso, repetiremos las mismas fases que en los capítulos anteriores pues nos situamos en un marco descriptivo, es decir: tabulación, representación gráfica y cálculo de los valores típicos correspondientes al estudio de la relación. Nótese que, a diferencia del estudio de variables numéricas, la tabulación de los datos tiene interés en nuestro caso porque, al tratarlos de manera categórica, se registrarán muchas repeticiones.

A diferencia del caso unidimensional, estudiado en el Capítulo 1, surgirán en este caso tres tipos diferentes de proporciones cuya relación y estimación se abordarán también en esta sección con vista a solucionar problemas de interés biomédico que aparecerán en las dos últimas secciones.

3.1.1. Tabla de contingencia

Partimos de una muestra compuesta por n individuos o unidades experimentales pertenecientes a una determinada población sobre los que se evalúan simultáneamente dos caracteres cualitativos, lo cual dará lugar a una tabla de frecuencias bidimensional o de

doble entrada denominada usualmente tabla de contingencia, en la que se indican las veces que se registra cada combinación de categorías. Veamos dos ejemplos.

Ejemplo 7. Según recientes investigaciones es posible que un índice cintura-cadera (ICC), definido como el cociente entre el perímetro de la cintura y el de la cadera, elevado se asocie a la aparición de ciertas patologías, como la diabetes y enfermedades cardiovasculares, de una manera más clara que el índice de masa corporal (IMC) elevado. Supongamos que, con el objeto de apoyar esa teoría, se analiza una muestra de $n = 252$ varones de más de 40 años que son clasificados en función de su ICC como normales, si ICC ≤ 0.94, o con cuerpo de manzana, si ICC > 0.94. Por otra parte, son también valorados médicamente distinguiendo entre sanos, diabéticos y enfermos cardiovasculares. Ambas clasificaciones se recogen de manera simultánea la siguiente tabla de contingencia:

	Estado de salud				
Tipo de ICC	2×3	Sano	Cardio	Diabetes	Total
	Normal	114	22	20	156
	Manzana	52	28	16	96
	Total	166	50	36	252

Tabla 3.1: Tabla de contingencia para las variables tipo de ICC y estado de salud.

Ejemplo 8. Se realiza un estudio a nivel cualitativo para considerar la posible asociación entre el nivel de SO_2 en la atmósfera (contaminación) y el estado de salud de cierta especie arbórea, en función del nivel de cloroplastos en las células de sus hojas. Se distinguen tres tipos de áreas según el nivel de SO_2: nivel alto, medio y bajo. Así mismo, se distinguen otros tres niveles de salud en los árboles: alto, medio y bajo. En cada zona se seleccionó una muestra de 20 árboles, así número total de árboles en la muestra final es $n = 60$. En cada caso se determina su nivel de cloroplastos. La tabla obtenida tras clasificar los 60 árboles fue la siguiente:

	Nivel de cloroplastos				
Nivel de SO_2	3×3	Alto	Medio	Bajo	Total
	Alto	3	4	13	20
	Medio	5	10	5	20
	Bajo	7	11	2	20
	Total	15	25	20	60

Tabla 3.2: Tabla de contingencia para las variables nivel de SO_2 (contaminación) y nivel de cloroplastos (salud de los árboles).

Comencemos con una breve descripción de la tabla correspondiente al Ejemplo 7. En este caso se distinguen $r = 2$ categorías (filas) diferentes en la variable tipo de ICC y $s = 3$ categorías (columnas) diferentes en la valoración médica, por lo que decimos que se trata de una tabla tipo 2×3. En los márgenes derechos e inferior de la tabla aparecen las frecuencias que denominaremos marginales, que corresponderían a un estudio por separado de las variables ICC y valoración, respectivamente, como ocurría en el Ejemplo 1. Las 6 frecuencias (2×3) que aparecen en el interior de la tabla pueden denominarse conjuntas o, también, observadas. Se denotan mediante O_{ij}, donde el subíndice i hace referencia a las filas y el j a las columnas. Así, por ejemplo, O_{12} se entiende como la frecuencia observada en la fila 1 y columna 2, es decir, con los datos del Ejemplo 7 estaríamos hablando del número de individuos con ICC normal y enfermedad cardiaca. Es obvio que la suma de frecuencias observadas de una misma fila es la frecuencia marginal que aparece en la columna derecha, y que la suma de frecuencias observadas en una misma columna es la frecuencia marginal que aparece en la fila de abajo. La suma total de las frecuencias conjuntas coincide con las de las marginales, tanto por filas como por columnas, y es el tamaño de muestra $n = 252$.

Una vez descrita dicha tabla, la cuestión esencial es en qué medida la tabla anterior corrobora la idea de que existe relación entre el estado de salud y el tipo de ICC, y en qué sentido. Es decir, nos preguntamos qué debe ocurrir para que podamos afirmar eso y cómo cuantificamos el grado de correlación observado. Para responder a estas preguntas introduciremos previamente los conceptos de proporción marginal, proporción condicionada y proporción conjunta.

Proporciones marginales: en primer lugar, podemos calcular las ya conocidas proporciones marginales o proporciones (simplemente). Por ejemplo, $\hat{P}(\text{Sano})$ denota la proporción de individuos de la muestra que están sanos. Así, para cada categoría se tiene:

$$\begin{aligned}
\hat{P}(\text{Sano}) &= \frac{166}{252} = 0.659, \\
\hat{P}(\text{Cardio}) &= \frac{50}{252} = 0.198, \\
\hat{P}(\text{Diabetes}) &= \frac{36}{252} = 0.143, \\
\hat{P}(\text{Normal}) &= \frac{156}{252} = 0.619, \\
\hat{P}(\text{Manzana}) &= \frac{96}{252} = 0.381.
\end{aligned}$$

Proporciones condicionadas: por otra parte, $\hat{P}(\text{Sano}|\text{Normal})$ se entiende como la proporción de individuos con ICC normal que están sanos según la valoración médica. Es lo que denominamos una proporción condicionada por fila, que se calculan, por ejemplo,

mediante los siguientes cocientes:

$$\begin{aligned}\hat{P}(\text{Sano}|\text{Normal}) &= \frac{114}{156} = 0.731,\\ \hat{P}(\text{Diabetes}|\text{Normal}) &= \frac{20}{156} = 0.128,\\ \hat{P}(\text{Diabetes}|\text{Manzana}) &= \frac{16}{96} = 0.167.\end{aligned}$$

De manera totalmente análoga pueden calcularse proporciones condicionadas por columnas:

$$\begin{aligned}\hat{P}(\text{Normal}|\text{Sano}) &= \frac{114}{166} = 0.659,\\ \hat{P}(\text{Normal}|\text{Diabetes}) &= \frac{20}{36} = 0.556,\\ \hat{P}(\text{Manzana}|\text{Diabetes}) &= \frac{16}{36} = 0.444.\end{aligned}$$

Proporciones conjuntas: por último, $\hat{P}(\text{Sano y Normal})$ denota la proporción de individuos de la muestra que son sanos según la valoración médica y, además, poseen un ICC normal. Es lo que denominamos proporción conjunta, que se calculan, por ejemplo, así:

$$\begin{aligned}\hat{P}(\text{Sano y Normal}) &= \frac{114}{252} = 0.452,\\ \hat{P}(\text{Diabetes y Normal}) &= \frac{20}{252} = 0.079,\\ \hat{P}(\text{Diabetes y Manzana}) &= \frac{16}{252} = 0.063.\end{aligned}$$

En definitiva, se trata siempre de calcular un cociente, aunque la composición del numerador y el denominador varía en función del tipo de proporción considerada.

Ejercicio 58. *Indica las siguientes proporciones relativas al Ejemplo 8 (puedes expresarlas si lo prefieres con porcentajes):*

(a) Proporción de árboles con alto nivel de cloroplastos entre aquellos que crecen en zonas poco contaminadas.

(b) Proporción de árboles que crecen en zonas poco contaminadas entre aquellos que cuentan con alto nivel de cloroplastos.

(c) Proporción de árboles de la muestra que crecen en zonas poco contaminadas y además cuentan con un alto nivel de cloroplastos.

(d) Proporción de árboles de la muestra que crecen en zonas poco contaminadas.

(e) Proporción de árboles de la muestra que cuentan con un alto nivel de cloroplastos.

Entre los distintos tipos de proporciones se verifica una relación muy clara, que es la que se indica en (3.1).

Ejercicio 59. *Razona por qué se verifica, con los datos del Ejemplo 7, que*

$$\hat{P}(Diabetes|Manzana) = \frac{\hat{P}(Diabetes\ y\ Manzana)}{\hat{P}(Manzana)} \tag{3.1}$$

Es muy común en Estadística denotar cada categoría de una variable cualitativa (en cierto contexto se denomina suceso al tal concepto) mediante una letra o signo, por ejemplo A; en ese caso y si la variable es binaria, se denotará la categoría opuesta mediante $\overline{A}$. También resulta útil identificar cada categoría con un subconjunto de un plano de área 1 y su proporción con el área de dicho subconjunto (véase Figura 3.1). Este tipo de gráfico, que suele denominarse diagrama de Venn, es el que podemos apreciar también a ambos lados de la Figura 3.4. En definitiva, podemos atribuir a las proporciones de sucesos las mismas propiedades que reconocemos al medir áreas de subconjuntos. De esta forma, la relación particular (3.1) puede expresarse, en general, en los siguientes términos:

$$\hat{P}(A \cap B) = \hat{P}(B|A) \cdot \hat{P}(A) \tag{3.2}$$

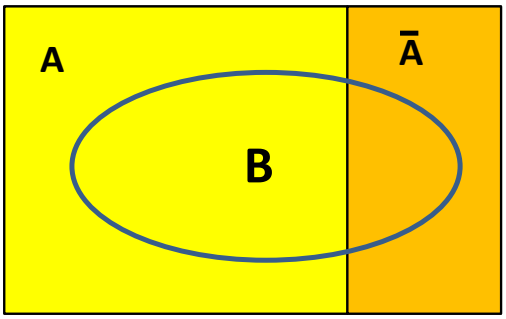


Figura 3.1: Esquema de la analogía entre proporciones y áreas.

Ejercicio 60. *Identifica en el plano de la Figura 3.1 las proporciones marginales de A y de su contrario, así como de B y su contrario, las proporciones conjuntas de A y B, así como del contrario de A y B, y, por último, la proporción de B condicionada a A.*

Ejercicio 61. *Según los datos del Ejemplo 7, la proporción de diabéticos en la muestra es del 14.9 %, mientras que la proporción de individuos con cuerpo de manzana entre los diabéticos es del 44.4 %. Utiliza la fórmula* (3.2) *para calcular directamente la proporción de individuos que son a la vez diabéticos y con cuerpo de manzana.*

3.1.2. Diagrama de barras agrupadas

Volviendo al estudio de proporciones muestrales, el diagrama de barras agrupadas resulta muy útil para ilustrar la asociación existente entre las dos variables cualitativas estudiadas. Dicho diagrama consiste en un diagrama de barras de las frecuencias absolutas de una variable cualitativa desglosadas en función de las categorías de otra. En el caso del Ejemplo 7 puede resultar más ilustrativo agruparlas en función del tipo de ICC.

También podemos agrupar las frecuencias del Ejemplo 8 en función del nivel de SO_2. Ambos diagramas se presentan en las Figuras 3.2 y 3.3, respectivamente.

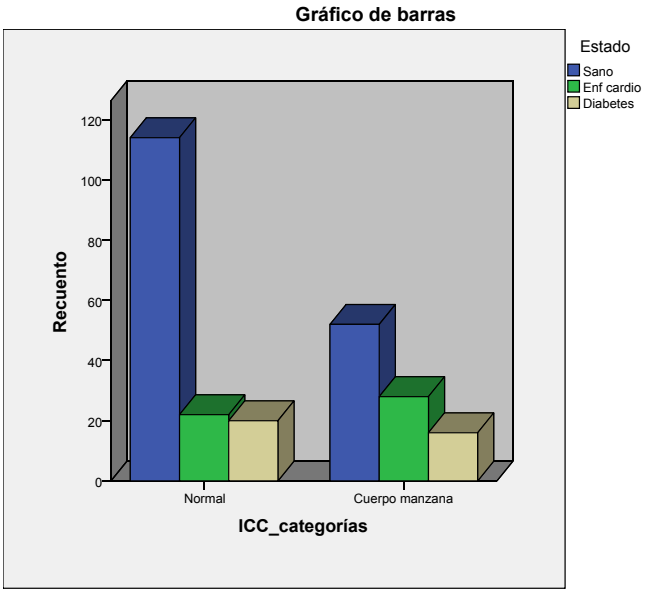


Figura 3.2: Diagrama de barras agrupadas para las variables tipo de ICC y estado de salud.

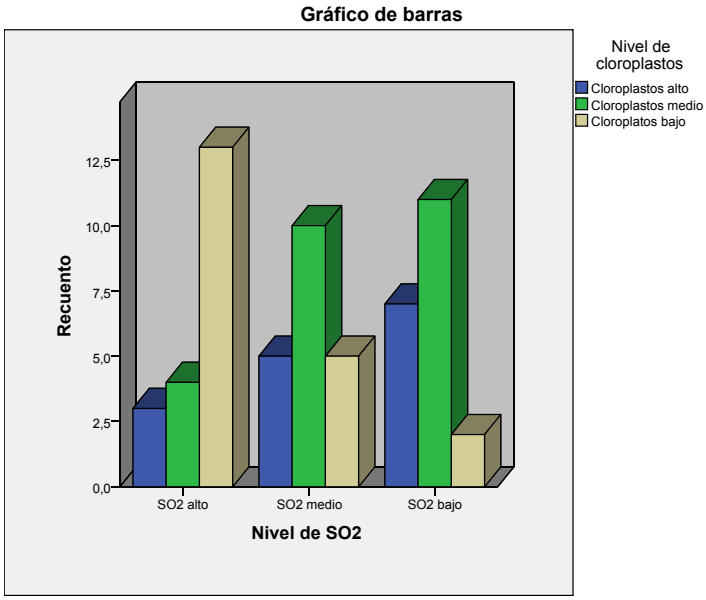


Figura 3.3: Diagrama de barras agrupadas para las variables nivel de SO_2 y nivel de cloroplastos.

Un diagrama de barras agrupado por filas nos da una información visual sobre los posibles cambios en las proporciones condicionadas por filas. Lo mismo sucede si agrupamos por columnas. Así, en el diagrama correspondiente al ICC observamos, por ejemplo, que la proporción de sanos (azules) es mayor entre los normales que entre los de cuerpo de manzana, lo cual se corresponde con una menor proporción de enfermos, sobre todo con enfermedad cardiaca, entre los primeros. Esas diferencias pueden resultar más acusadas en el caso del SO_2, donde apreciamos que la proporción de árboles con un nivel bajo de cloroplastos es mucho mayor en las zonas muy contaminadas (donde el nivel de SO_2 es alto). Es importante mencionar que podríamos haber llegado a conclusiones análogas si hubiéramos condicionado por columnas, es decir, condicionar por filas o columnas es indiferente desde el punto de vista teórico aunque no siempre lo es desde el punto de vista intuitivo. En general podríamos afirmar lo siguiente[1]:

> En términos estadísticos, entendemos que la correlación a nivel muestral entre las dos variables cualitativas observadas es más fuerte cuanto mayores sean las diferencias entre las proporciones condicionadas al pasar de una categoría a otra.

Así pues, en lo que se refiere a problemas de correlación entre dos variables, podemos distinguir tres situaciones:

Variable 1	Variable 2	Relación ↔ Cambio en la distribución
Numérica	Numérica	Los cambios **a lo largo** de la primera variable se asocian a cambios en los valores **medios** de la segunda.
Cualitativa	Numérica	Los cambios de **categoría** en la primera variable se asocian a cambios en los valores **medios** de la segunda.
Cualitativa	Cualitativa	Los cambios de **categoría** en la primera variable se asocian a cambios en las **proporciones** de la segunda.

Tabla 3.3: Tipos de relaciones estadísticas.

A la vista de las Figuras 3.2 y 3.3 podemos intuir pues que la correlación observada entre la valoración médica y el tipo de ICC es más débil que la correlación observada entre la salud de los árboles y la contaminación, pues en el segundo caso se aprecia una alteración drástica en el patrón de distribución (proporciones) cuando pasamos de una zona de contaminación baja o media a otra de contaminación alta. No obstante y al igual que sucediera con el coeficiente r en el caso numérico, necesitamos un coeficiente muestral que cuantifique de alguna forma el grado de correlación observado. En este caso será el denominado coeficiente de contingencia C de Pearson.

[1]Esta afirmación es válida sólo si estamos condicionando respecto a una variable con la suficiente heterogeneidad (es decir, tal que las frecuencias marginales de ambas categorías sean suficientemente grandes).

3.1.3. Coeficiente de contingencia C de Pearson

Para medir el grado de correlación muestral procederemos de manera similar a la forma de medir la variabilidad de un conjunto de datos numérico unidimensional: recordemos que no se trataba de evaluar las diferencias entre los datos, sino la distancia (al cuadrado) entre cada uno de ellos y una medida central de referencia, la media aritmética, que en ocasiones no es ni siquiera un valor posible[2], dando como resultado la varianza. En nuestro caso, dadas unas frecuencias marginales concretas, vamos a construir una tabla bidimensional de referencia cuyas sumas marginales se mantengan iguales a la tabla observada pero cuyos valores conjuntos, denominados valores esperados y denotados como E_{ij}, estén calculados de tal manera que las proporciones condicionadas permanezcan constantes al pasar de una fila (o columna) a otra. En ese caso deben ser necesariamente iguales a las proporciones marginales por filas (o columnas, respectivamente). La tabla de valores E_{ij} para el Ejemplo 7 resultante es la siguiente:

	Estado de salud				
Tipo de ICC	2×3	Sano	Cardio	Diabetes	Total
	Normal	102.8	31.0	22.3	156
	Manzana	63.2	19.0	13.7	96
	Total	166	50	36	252

Tabla 3.4: Tabla de valores esperados E_{ij} para las variables tipo de ICC y estado de salud.

Podemos comprobar que, efectivamente, con los datos de esta tabla ideal o esperada se verificaría:

$$\hat{P}(\text{Sano}) = \hat{P}(\text{Sano}|\text{Normal}) = \hat{P}(\text{Sano}|\text{Manzana}) = 0.659,$$
$$\hat{P}(\text{Cardio}) = \hat{P}(\text{Cardio}|\text{Normal}) = \hat{P}(\text{Cardio}|\text{Manzana}) = 0.198,$$
$$\hat{P}(\text{Diabetes}) = \hat{P}(\text{Diabetes}|\text{Normal}) = \hat{P}(\text{Diabetes}|\text{Manzana}) = 0.143,$$

y de igual forma,

$$\hat{P}(\text{Normal}) = \hat{P}(\text{Normal}|\text{Sano}) = \hat{P}(\text{Normal}|\text{Cardio}) = \hat{P}(\text{Normal}|\text{Diabetes}) = 0.619,$$
$$\hat{P}(\text{Manzana}) = \hat{P}(\text{Manzana}|\text{Sano}) = \hat{P}(\text{Manzana}|\text{Cardio}) = \hat{P}(\text{Manzana}|\text{Diabetes}) = 0.381.$$

Las diferentes proporciones conjuntas pueden entenderse desde un punto de vista gráfico como las respectivas áreas de los seis subconjuntos en los que se divide la muestra, a la que se le asigna un área total 1. De esta forma, la independencia o correlación nula se observaría si los diferentes subconjuntos mostraran la configuración de la izquierda en la Figura 3.4, mientras que lo realmente observado se ajusta a la configuración de la derecha. Obsérvese que en la primera las proporciones condicionadas no cambian al pasar de una categoría a otra y en ambas se mantienen las proporciones marginales .

[2]Como sucede, por ejemplo, cuando se dice que el número medio de hijos por mujer en España es 1.2.

PORTADA ÍNDICE

2 x 3	Sano	Cardio	Diab.	Total
Normal	102.8	31.0	22.2	156
Manzana	63.3	19.0	13.7	96
Total	166	50	36	252

Figura 3.4: Tabla de valores esperados E_{ij} (izquierda) y tabla de valores observados O_{ij} (derecha).

En el caso del Ejemplo 8, la tabla de valores E_{ij} es la siguiente:

	Nivel de cloroplastos				
Nivel de SO_2	3×3	Alto	Medio	Bajo	Total
	Alto	5	8.3	6.7	20
	Medio	5	8.3	6.7	20
	Bajo	5	8.3	6.7	20
	Total	15	25	20	60

Tabla 3.5: Tabla de valores esperados E_{ij} para las variables nivel de cloroplastos y nivel de SO_2.

Ejercicio 62. *Supongamos que se lleva a cabo un estudio para analizar la posible relación entre el factor Rh y el sexo. Se estudian un total de $n = 100$ personas con los siguientes resultados (parciales):*

	Rh			
Sexo	2×2	+	-	Total
	Masculino			40
	Femenino			60
	Total	75	25	100

Tabla 3.6: Tabla de contigencia para las variables sexo y Rh.

¿Qué cantidad de datos E_{ij} debería aparecer en cada una de las cuatro celdas interiores para que la proporción de Rh positivo fuera idéntica en hombres y mujeres. ¿Qué ocurrirá entonces con la proporción de Rh negativo?

Ejercicio 63. *En general, ¿serías capaz de determinar una fórmula general para calcular los valores E_{ij} a partir de las frecuencias marginales?*

Una vez construida esta matriz de referencia, entendemos que el grado de correlación correspondiente a nuestra muestra es más fuerte cuanto mayor sea la distancia (entendiendo en principio dicha distancia en sentido amplio) entre nuestra tabla de valores observados y la tabla de valores esperados. Así, en el ejemplo de la Figura 3.4 se trata de cuantificar de alguna manera la diferencia entre la configuración esperada de la izquierda y la observada de la derecha. La distancia que se utiliza para medir la diferencia entre ambas tablas es la siguiente:

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

Así, debe quedar pues claro que un valor χ^2_{exp} próximo a 0 debe entenderse como una correlación casi nula en la muestra, y que, cuanto mayor sea el valor de χ^2_{exp}, más fuerte será la dependencia o correlación observada en la muestra.

Coeficiente de contingencia C de Pearson: es útil normalizar la distancia χ^2 para obtener un valor con cotas universales. La normalización más popular es posiblemente el coeficiente de contingencia de Pearson, que pretende desempeñar un papel similar al coeficiente de correlación r introducido en el Capítulo 2, también denominado de Pearson. El coeficiente de contingencia de Pearson define mediante:

$$C = \sqrt{\frac{\chi^2_{exp}}{\chi^2_{exp} + n}}.$$

Este coeficiente debe estar comprendido, para toda tabla $r \times s$, entre 0 y $\sqrt{q^{-1}(q-1)}$, siendo $q = \text{mín}\{r, s\}$. La cota 0 corresponde a la ausencia total de correlación y la cota superior, que depende únicamente de las dimensiones de la tabla, a la máxima dependencia posible. En el Ejemplo 7, la cota máxima es, en general 0.707, por ser una tabla 2×3, y el valor obtenido en esta tabla concreta es $C = 0.201$; en el Ejemplo 8 la cota máxima es 0.816, al ser una tabla 3×3, y el valor concreto obtenido es $C = 0.444$. Es decir, en términos relativos se observa una mayor correlación en el segundo ejemplo en el sentido que indica el diagrama de barras de la Figura 3.2, es decir, zonas de poca contaminación se asocian a árboles sanos. En el Ejemplo 7 observamos una correlación débil y en el sentido que indica el diagrama de barras, es decir, un tipo normal de ICC está asociado a un estado sano.

Analicemos ahora cómo deberían ser los datos observados en el Ejemplo 8 para alcanzar el máximo grado de correlación, que se corresponde con $C = 0.816$. Una tabla que se ajusta a tal situación, que no es la observada en nuestro caso, es la siguiente siguiente:

	Nivel de cloroplastos				
Nivel de SO_2	3×3	Alto	Medio	Bajo	Total
	Alto	0	0	20	20
	Medio	0	20	0	20
	Bajo	20	0	0	20
	Total	20	20	20	60

Tabla 3.7: Ejemplo de máxima correlación entre el nivel de cloroplastos y el nivel de SO_2.

3.1.4. Tablas dos por dos

El caso particular en que se distinguen únicamente dos categorías en las dos variables consideradas, puede recibir además del tratamiento estudiado anteriormente, otro específico que destaca por su sencillez. La tabla de contingencia en esta situación tendrá la siguiente estructura:

	B			
A	2×2	B_1	B_2	Total
	A_1	a	b	a+b
	A_2	c	d	c+d
	Total	a+c	b+d	n

Tabla 3.8: Tabla de contingencia génerica de tipo 2×2.

Ejemplo 9. Se pretende averiguar en qué medida es efectiva una vacuna contra la hepatitis. Se estudió una muestra de $n = 1083$ individuos de los cuales algunos habían sido vacunados y otros no; transcurrido un largo periodo de tiempo, algunos habían llegado a contraer la hepatitis mientras que otros estaban sanos. La tabla de contingencia resultante es la siguiente:

	Vacunación			
Hepatitis	2×2	Sí	No	Total
	Sí	11	70	81
	No	538	464	1002
	Total	549	464	1083

Tabla 3.9: Tabla de contingencia para las variables hepatitis y vacunación.

Coeficiente ϕ: para un caso de este tipo y a la hora de medir el grado de asociación de las variables podemos utilizar, además del conocido coeficiente C, el denominado coeficiente ϕ, que se define mediante $\phi^2 = \chi^2_{exp}/n$, que es equivalente a la expresión:

$$\phi = \sqrt{\frac{(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}}.$$

Si analizamos detenidamente la última expresión, concluiremos que ϕ^2 es un parámetro completamente análogo al coeficiente de correlación lineal r^2; de hecho, si para ambas variables cualitativas asignamos sendos códigos numéricos a cada una de las posibles categorías, ϕ equivale al coeficiente de correlación r entre las variables numéricas resultantes. Concretamente, puede tomar cualquier valor entre 0 y 1. El valor 0 se corresponde con asociación nula y el valor 1 con una asociación máxima.

Ejercicio 64. *Comprueba que el valor de ϕ para los datos del Ejemplo 9 es 0.211.*

Por su parte, el coeficiente de contingencia, que en una tabla 2×2 debe estar comprendido entre 0 y 0.707, da como resultado en esta caso $C = 0.206$. Ambos valores coinciden en expresar un grado de relación medio-bajo en la muestra observada. El valor máximo $\phi = 1$ se corresponde con una tabla diagonal. Es lo que lo que habría ocurrido si los datos de la muestra hubieran sido los de la Tabla 3.10. Por contra, el valor $\phi = 0$ se corresponde con un grado nulo de relación, que se habría alcanzado si nuestros datos hubieran sido los de la Tabla 3.11. Efectivamente, si fuera este el caso podríamos observar que, tanto en el caso de vacunados como en el de no vacunados, la proporción condicionada de individuos afectados sería 1/3. Lo mismo ocurriría con la tabla resultante en el Ejercicio 62.

	Vacunación			
Hepatitis	2 × 2	Sí	No	Total
	Sí	0	81	81
	No	1002	0	1002
	Total	1002	81	1083

Tabla 3.10: Tabla de valores esperados para las variables hepatitis y vacunación en el caso $\phi = 1$.

	Vacunación			
Hepatitis	2 × 2	Sí	No	Total
	Sí	334	27	361
	No	668	54	722
	Total	1002	81	1083

Tabla 3.11: Tabla de valores esperados para las variables hepatitis y vacunación en el caso $\phi = 0$.

Con un propósito meramente didáctico y para hacer hincapié en la semejanza entre los parámetros r y ϕ, podemos convertir en cualitativas (categorizar) las variables numéricas X e Y del Ejemplo 4 ($r = 0.91$) que se representan en la Figura 2.4, asignándoles "+" cuando el valor queda por encima de su correspondiente media y "–" cuando queda por debajo. Así, obtendríamos la siguiente tabla 2×2, a la que corresponde un valor de $\phi = 0.86$.

		X		
Y	2×2	-	+	Total
	-	2	6	8
	+	4	0	4
	Total	6	6	12

Tabla 3.12: Tabla de contingencia para las variables peso y altura una vez categorizadas.

Ejercicio 65. *Compara el valor de ϕ que corresponde a esta tabla con el valor r obtenido para los datos numéricos originales. Confróntese esta tabla con las Figuras 2.4 y 2.22 para entender el concepto de relación estadística.*

Ejercicio 66. *Compara la tabla obtenida en el Ejercicio 62 con las Figuras 2.10 y 2.20 para entender el concepto de independencia.*

Recordemos que las conclusiones obtenidas hasta ahora se ciñen exclusivamente a la muestra considerada, es decir, no estamos aún en condiciones de extrapolarlas al conjunto de la población, entre otras cosas porque no sabemos en qué condiciones ha sido escogida esa muestra. Puede suceder que los individuos hayan sido seleccionados intencionadamente para obtener unos resultados concretos.

3.2. Estimando proporciones poblacionales

Este apartado constituye una primera incursión en la Inferencia Estadística, que se estudiará con mayor detalle la segunda parte del manual. Hemos de destacar que las proporciones se han denotado hasta ahora por $\hat{P}$ con la idea de resaltar que son parámetros descriptivos, es decir, que se refieren a la muestra estudiada, en contraposición con la proporción calculada a partir de toda la población, que se denotará por P y que, en la mayoría de los textos, se denomina probabilidad[3]. No obstante, podemos intuir que conocer proporciones a nivel poblacional puede quedar fuera de nuestro alcance en la mayoría de los casos. Precisamente, el objeto de este tipo de estudios suele ser calcular proporciones a partir de la tabla de frecuencias, es decir, a partir de la muestra, de manera que puedan considerarse estimaciones o aproximaciones a las proporciones correspondientes a la población.

[3]Intentaremos omitir dicho término para no inducir a confusión.





Sin embargo, que una proporción poblacional concreta pueda ser o no aceptablemente estimada a partir de la proporción muestral, calculada directamente a partir de la tabla de frecuencias, depende de cómo se haya obtenido la muestra. Efectivamente, parece obvio que, por ejemplo, si escogemos una muestra de una población con el requisito de que la cuarta parte sean hombres y el resto mujeres, esta no es válida para estimar la proporción de hombres y mujeres en dicha población. Por otra parte, si el hecho de ser o no diabético no se ha tenido en cuenta a la hora de seleccionar cada individuo, no está claro en principio si la muestra es adecuada para estimar la proporción de diabéticos puesto que no sabemos aún si este hecho guarda alguna relación con el sexo. Sin embargo, la muestra sí que puede ser adecuada en principio para estimar la proporción de diabéticos entre los hombres, por un lado, y la proporción de diabéticas entre las mujeres, por otro. También podría ser en principio adecuada para estimar la proporción de cualquier cualidad que no guarde relación con el sexo, como puede ser el Rh.

Por tanto, tiene sentido plantearse qué requisito debería cumplir una muestra para que fuera posible estimar cualquier proporción considerada. Como se explicará en el Capítulo 4, el procedimiento que justifica la estimación desde un punto de vista teórico es el denominado sorteo aleatorio, pero su aplicación estricta podría considerarse utópica en la mayoría de los estudios biomédicos. En ese sentido, podríamos enunciar de una forma algo imprecisa pero más realista la primera máxima de la Inferencia Estadística:

> A través de una muestra sólo podemos aspirar a estimar parámetros poblacionales relativos a variables que no hayan sido directa o indirectamente controladas durante el proceso de selección de la misma.

Este hecho tendrá bastante trascendencia cuando estudiemos los diversos tipos de estudios epidemiológicos así como los ensayos clínicos.

3.2.1. Diagramas de árbol y fórmula de Bayes

Sin embargo, a pesar de lo expuesto anteriormente, el tipo de relación expresado en (3.2) puede ser de utilidad para estimar indirectamente ciertas proporciones a partir de otras, algunas de las cuales pueden estimarse directamente a través de la tabla mientras que otras constan como datos ya conocidos por otros medios. El razonamiento aplicado se denomina fórmula de Bayes. La fórmula de Bayes es la respuesta a un problema muy común, un conflicto entre el razonamiento estadístico y el puramente intuitivo, tal y como se explica en [5]. Para ilustrar el problema del que hablamos, intentemos responder a la siguiente pregunta de manera rápida:

Ejercicio 67. *Es bien conocido que la proporción de lectores del New York Times es muy alta entre las personas que han obtenido un doctorado en Harvard, siendo bastante baja en el resto de norteamericanos. Si encontramos en el metro de Nueva York a una persona leyendo dicho periódico, ¿debemos inclinarnos a pensar que se trata de un doctor por la universidad de Harvard?*

En primer lugar, cabe plantearse la siguiente pregunta: ¿cómo hemos logrado saber que la proporción de lectores del periódico es mucho mayor entre los doctores por Harvard?

Pues, seguramente, mediante siguiente diseño: tomamos por un lado una muestra de doctores por Harvard y averiguamos el número de lectores del *New York Times* y, por otro, hacemos lo mismo con otra muestra de no doctores por Harvard. En conjunto habremos compuesto una tabla de contingencia tipo 2×2. Alguien podría plantear la posibilidad de escoger una muestra cualquiera de la población, sin más, y averiguar, por una parte, quiénes leían el periódico y, por otra, quiénes eran doctores por Harvard. Pero este segundo diseño conllevaría un serio problema.

Ejercicio 68. *¿En qué consiste el problema del segundo diseño?*

La tabla de contingencia del diseño primero permite estimar fácilmente las proporciones condicionadas de lectores entre los doctores y entre los no doctores, y es así como se llega a la primera afirmación. Sin embargo, la pregunta formulada hace referencia a la proporción condicionada contraria: la proporción de doctores por Harvard entre los lectores del *New York Times*. Esta no puede ser estimada adecuadamente a partir de la tabla porque el hecho de ser o no doctor por Harvard está controlado en el diseño, de tal manera que los doctores por Harvard están sobrerrepresentados en la muestra. Para poder estimarla indirectamente necesitamos un dato que no está en el enunciado: la proporción de doctores por Harvard en la población americana. El caso es que esta proporción es tan baja que hace casi imposible que el lector del periódico sea uno de ellos.

La fórmula de Bayes es la fórmula que debe utilizarse para llegar a dicha conclusión de manera precisa. Permite calcular $P(A|B)$ si se conocen, o al menos pueden aproximarse razonablemente, las proporciones $P(B|A)$, $P(B|\overline{A})$ y $P(A)$. Para deducir esta fórmula puede resultar de ayuda entender las proporciones cómo áreas en un diagrama como el de la Figura 3.1. En primer lugar, es claro que $P(B)$ puede descomponerse en dos sumandos según la ecuación siguiente: $P(B) = P(A \cap B) + P(\overline{A} \cap B)$. A su vez, aplicando la igualdad (3.2) a ambos sumandos obtenemos la siguiente ecuación:

$$P(B) = P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A}) \tag{3.3}$$

Nótese que la igualdad (3.3) justifica formalmente en el cálculo de proporciones el uso de los denominados diagramas de árbol, que resultarán familiares a muchos lectores. La Figura 3.5 intenta explicar esquemáticamente el proceso.

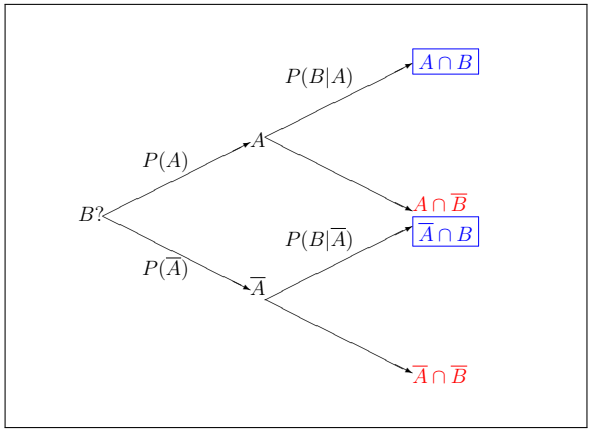


Figura 3.5: Diagrama de árbol para calcular $P(B)$.

Una vez obtenido $P(B)$ partiendo de proporciones que sí conocemos (aproximadamente) podemos obtener el valor de $P(A|B)$, utilizando de nuevo (3.2), para obtener la fórmula de Bayes:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\overline{A}) \cdot P(\overline{A})}. \tag{3.4}$$

Esta fórmula será de gran utilidad en las secciones siguientes. Podemos aplicarla a este otro ejemplo más concreto.

Ejercicio 69. *Supongamos que en una determinada población se conoce de antemano que el 5 % padecen diabetes tipo II. A través de una muestra, en la cual la mitad de los pacientes eran diabéticos y la otra mitad no, se estimó mediante la tabla de contingencia que la proporción de hipertensos era de un 60 % entre los diabéticos y de un 15 % entre los no diabéticos.*

(a) Estima la proporción de hipertensos en la población.

(b) Estima también la proporción de hipertensos que son diabéticos.

(c) Estima la proporción de diabéticos entre los hipertensos y compárala con la proporción de diabéticos entre los no hipertensos.

(d) Representa las cuatro posibilidades del estudio mediante un diagrama de Venn.

Ejercicio 70. *Plantea el Ejercicio 67 en estos mismos términos.*

3.3. Factores de riesgo

Nos centramos en esta ocasión en un tipo particular de tabla 2×2 de especial interés en Epidemiología. Supongamos que una de la variables cualitativas estudiadas es la ausencia o presencia de una enfermedad `E`, como puede ser un cáncer de pulmón, hepatitis, osteoporosis, etcétera, siendo la otra la ausencia o presencia de un posible factor de riesgo `FR` de cara a padecer dicha enfermedad, como, respectivamente, el hecho de fumar, el de no estar vacunado contra la hepatitis, el de no alimentarse correctamente, etc. El propósito de este tipo de estudios es determinar, a partir de una muestra, si ese supuesto factor de riesgo lo es efectivamente y en qué medida. Dado que en esta primera parte estamos en un contexto descriptivo, nos limitaremos por el momento a calcular una medida apropiada del riesgo que supone el factor en la muestra considerada. Los detalles sobre posibles inferencias o generalizaciones se exponen brevemente en la segunda parte del manual.

Ejercicio 71. *Indica 5 enfermedades y 5 respectivos posibles factores de riesgo. ¿Crees que están todos ellos confirmados estadísticamente o estamos hablando de meras suposiciones teóricas?*

En este tipo de estudios pueden considerarse diferentes parámetros de interés para una enfermedad concreta:

Prevalencia: es la proporción de individuos enfermos $P(\mathtt{E})$ en un instante dado en la población.

Incidencia: es la proporción de individuos que, estando sanos al inicio de un periodo de tiempo, enferman a lo largo del mismo. Se pueden distinguir distintos tipos de incidencias, por ejemplo, la incidencia entre los individuos que presentan un posible factor de riesgo o la incidencia entre los que no lo presentan. A partir de estas dos incidencias se calculan los riesgos relativo y atribuibles, que definiremos más adelante.

3.3.1. Tipos de diseños

En lo relativo al estudio de factores de riesgo, distinguiremos tres tipos de diseños:

Estudios transversales o de prevalencia: su objetivo principal es poder estimar la prevalencia, para lo cual se selecciona una gran muestra representativa de la población y se determina la cantidad de enfermos en un momento dado. La prevalencia $P(\mathtt{E})$ se estima entonces de manera obvia mediante la proporción de enfermos en la muestra, $\hat{P}(\mathtt{E})$.

Estudios de seguimiento o de cohortes: se selecciona una muestra de individuos sanos expuestos al factor de riesgo y otra de sanos no expuestos para estudiar su evolución durante un periodo de tiempo, que suele ser largo, anotándose cuántos llegan a contraer la enfermedad en cada caso. Este diseño permite estimar directamente las incidencias de la enfermedad para ambas cohortes mediante las proporciones condicionadas $\hat{P}(\mathtt{E}|\mathtt{FR})$ y $\hat{P}(\mathtt{E}|\overline{\mathtt{FR}})$, con el fin de compararlas entre sí[4].

Estudios retrospectivos o de casos-control: en un determinado momento se escoge una muestra de enfermos (casos) y otra de sanos (control), para a continuación averiguar qué individuos han estado expuestos al factor de riesgo. Suelen ser los menos costosos pues los de prevalencia requieren muestras muy grandes para que puedan registrarse suficientes enfermos, mientras que los de cohortes requieren de un seguimiento de las cohortes durante un largo intervalo de tiempo para que exista la posibilidad de que surja la enfermedad. Sin embargo, en los estudios tipo casos-control se seleccionan intencionadamente un grupo de enfermos que se comparan con otro de sanos, con lo que la presencia de la enfermedad en el estudio queda así garantizada. El inconveniente de este tipo de diseño consiste en que, al estar la enfermedad controlada en el estudio, no es posible dar a partir de la muestra una estimación válida de las diferentes incidencias ni prevalencias. Por contra, dado que la presencia del factor de riesgo no está controlada, sí podemos estimar las proporciones condicionadas $P(\mathtt{FR}|\mathtt{E})$, $P(\overline{\mathtt{FR}}|\mathtt{E})$, lo cual permitirá estimar adecuadamente el denominado Odds Ratio a través de la fórmula de Bayes, según indicaremos más adelante.

En todo caso, nuestros datos se recogerán en una tabla 2×2 donde se indicará, por un lado, si el individuo presenta el factor de riesgo y, por otro, si padece o desarrolla la enfermedad estudiada.

[4]Recordemos que, con la notación introducida, $\overline{FR}$ denota la categoría de las personas no expuestas al factor de riesgo.

	Factor			
Enfermedad	2×2	Sí	No	Total
	Enfermo	a	b	a+b
	Sano	c	d	c+d
	Total	a+c	b+d	n

Tabla 3.13: Tabla de contingencia para el estudio de factores de riesgo.

En el Ejemplo 9, la enfermedad estudiada es la hepatitis y el posible factor de riesgo la ausencia de vacunación. Se supone que estamos ante un estudio de cohortes pues se efectúa un seguimiento de individuos inicialmente sanos. Como hemos dicho anteriormente, en un estudio de cohortes tiene sentido estimar las incidencias de la enfermedad por grupos a través de la tabla. Concretamente:

$$\hat{P}(\mathtt{E}|\mathtt{FR}) = \frac{a}{a+c}, \qquad \hat{P}(\mathtt{E}|\overline{\mathtt{FR}}) = \frac{b}{b+d},$$

y se entenderán respectivamente como el riesgo observado en la muestra de contraer la enfermedad si se está expuesto al factor y el riesgo observado en la muestra de contraer la enfermedad si no se está expuesto al mismo. En un estudio de casos-control tiene sentido estimar a partir de la muestra la proporción de individuos enfermos que presentan el factor de riesgo y la proporción de individuos sanos que presentan el factor de riesgo. Concretamente, se calculan de la siguiente forma:

$$\hat{P}(\mathtt{FR}|\mathtt{E}) = \frac{a}{a+b}, \quad \hat{P}(\mathtt{FR}|\mathtt{S}) = \frac{c}{c+d}.$$

3.3.2. Medidas de riesgo

Veamos cuáles son las medidas más populares del riesgo que comporta un factor determinado. Aunque todas pueden en principio calcularse a partir de la tabla 2×2, estos valores podrán o no considerarse estimaciones razonables de los valores poblacionales en función del tipo de estudio del que se trate. Hemos de mencionar también que los propios coeficientes C y ϕ pueden entenderse como medidas de riesgo dado que expresan el grado de relación entre el factor y la enfermedad. No obstante, cuando la enfermedad estudiada no es muy frecuente estas medidas no suelen resultar intuitivas para explicar el grado de riesgo, de manera que se utilizan generalmente otras más específicas del contexto epidemiológico.

Riesgo atribuible: es la diferencia entre las incidencias de enfermos, es decir,

$$RA = P(\mathtt{E}|\mathtt{FR}) - P(\mathtt{E}|\overline{\mathtt{FR}}).$$

Este parámetro puede estimarse mediante estudios de cohortes. Un valor positivo indica que en la muestra se observa una mayor tendencia a la enfermedad en los que presentan

el factor de riesgo. Un valor aproximadamente nulo indica escasa relación entre el factor de riesgo y la enfermedad.

Con los datos del Ejemplo 9 y si consideramos como factor de riesgo el hecho de no estar vacunado, obtenemos una estimación del riesgo atribuible de

$$\widehat{RA} = 13.1\,\% - 2.0\,\% = 11.1\,\%.$$

El porcentaje de enfermos entre los no vacunados es 11.1 puntos superior al de lo vacunados. Esta medida adolece del mismo problema que el coeficiente ϕ pues, al restarse incidencias que suelen ser pequeñas aporta valores a su vez bajos.

Fracción atribuible a la exposición: se define como el cociente

$$FA = \frac{RA}{P(\mathtt{E}|\mathtt{FR})} = \frac{P(\mathtt{E}|\mathtt{FR}) - P(\mathtt{E}|\overline{\mathtt{FR}})}{P(\mathtt{E}|\mathtt{FR})}.$$

Se interpreta como la parte del riesgo de los expuestos que se debe propiamente al factor, entendiendo que una parte de los que están expuestos enferman por otras causas que comparten con los no expuestos. En el caso del ejemplo anterior es del 84 %. Lógicamente, este parámetro sólo puede estimarse en los estudios de cohortes.

Riesgo relativo: seguramente se trata de la medida de riesgo más intuitiva. Consiste de determinar en qué medida el factor de riesgo incrementa la incidencia de la enfermedad, es decir:

$$RR = \frac{P(\mathtt{E}|\mathtt{FR})}{P(\mathtt{E}|\overline{\mathtt{FR}})}.$$

Se puede estimar a partir de la tabla en un estudio de cohortes mediante:

$$\widehat{RR} = \frac{a}{a+c} : \frac{b}{b+d}.$$

Para los datos de la hepatitis tendríamos la siguiente estimación $\widehat{RR} = 13.1/2 = 6.55$. Es decir, en esta muestra se observa que el hecho de no estar vacunado aumenta 6.55 veces la proporción de enfermos.

Odds Ratio: constituye una alternativa muy socorrida al riesgo relativo que puede ser estimada razonablemente tanto en los estudios tipo cohortes como casos-control. Omitimos aquí la definición formal del parámetro que, a la postre y en virtud de la fórmula de Bayes, puede ser estimado directamente a partir de la tabla de contingencia de la siguiente forma

$$\widehat{OR} = \frac{a \cdot d}{b \cdot c}, \quad \text{o bien} \quad \widehat{OR} = \frac{b \cdot c}{a \cdot d}.$$

Se define de acuerdo con la expresión de la izquierda o de la derecha según cómo entendamos en principio el riesgo, que será mayor cuanto más grande sea el cociente. Un valor en torno a 1 se corresponde con una relación débil entre el posible factor y la enfermedad. Por

su expresión final se denomina también razón de productos cruzados. Así, en el Ejemplo 9 obtenemos:

	Vacunación			
Hepatitis	2×2	Sí	No	Total
	Sí	11	**70**	81
	No	**538**	464	1002
	Total	549	464	1083

Tabla 3.14: Tabla de contigencia para las variables hepatitis y vacunación.

$$\widehat{OR} = \frac{70 \cdot 538}{11 \cdot 464} = 7.10.$$

Esta medida no goza de una interpretación tan clara e intuitiva como el riesgo relativo. No obstante, en general si calculamos ambos a partir de una misma tabla y el Odds Ratio está por encima de 1, entonces aporta un valor superior al Riesgo Relativo; en el caso contrario, aporta un valor inferior. Por eso, es frecuente permitirse la licencia de interpretarlos de forma idéntica como medidas del incremento del riesgo, entendiendo que el Odds Ratio exagera ligeramente la percepción del mismo. Es decir, que en el Ejemplo 9 se entiende, exagerando, que el hecho de no vacunarse multiplica por 7 el riesgo de contraer hepatitis. Es de vital importancia entender bien la tabla para saber qué diagonal debe aparecer en el numerador y cuál en el denominador.

Ejercicio 72. *¿Qué diferencia existe entre $\widehat{RR}$ y RR?*

Ejercicio 73. *Razona lo mejor posible por qué en un estudio de tipo casos-control no podemos obtener una estimación razonable del riesgo relativo.*

Ejercicio 74. *¿Con qué valores de $\widehat{RA}$, $\widehat{FA}$, $\widehat{RR}$ y $\widehat{OR}$ se corresponde $\phi = 0$?*

Ejercicio 75. *¿Cómo se interpreta un valor $\widehat{RR} = 0.50$?*

Ejercicio 76. *Si se afirma que un hábito determinado incrementa en un 20 % el riesgo de padecer una enfermedad concreta, ¿qué podemos decir del riesgo relativo asociado?*

3.4. Diagnóstico Clínico

Otra cuestión de gran interés en Epidemiología que guarda una estrecha relación con las tablas 2×2 es el estudio de la eficacia de los diferentes procedimientos de diagnóstico de una patología o de detección de sustancias dopantes.

En primer lugar, hemos de mencionar que una gran cantidad de procedimientos de diagnóstico tienen una importante componente estadística. Efectivamente, nos referimos a aquellos métodos que consisten en medir una variable de tipo numérico que puede proceder de una analítica (concentración de leucocitos, marcador PSA, urea), de una ecografía

(anchura de un conducto, fracción de acortamiento entre sístole y diástole), etc. Si conocemos la distribución aproximada para los individuos sanos de una variable concreta, es decir, qué valores puede tomar y en qué proporciones, un valor anómalo respecto a dicha distribución puede ser considerado en principio patológico, lo cual supondrá un resultado positivo en el diagnóstico, que seguramente deberá ser corroborado mediante otra prueba más exhaustiva. Por contra, un valor dentro de los límites correspondientes a la población sana supondrá un resultado negativo, lo cual no tiene por qué excluir la posibilidad de que el individuo esté enfermo.

La forma de valorar la fiabilidad de un procedimiento de este tipo es aplicarlo a una muestra de individuos con un diagnóstico previo certero (sano o enfermo) y comprobar en qué medida los enfermos coinciden con los positivos. Se trata pues de un diseño tipo casos-control que dará lugar a una tabla 2×2 como la que aparece en el siguiente ejemplo:

Ejemplo 10. Se aplica un test diagnóstico a $n = 1000$ individuos, 200 de los cuales sabemos que están enfermos mientras que de los 800 restantes sabemos que están sanos. Los resultados se recogen en la Tabla 3.15.

	Diagnóstico			
Enfermedad	2×2	+	-	Total
	Enfermo	120	80	200
	Sano	90	710	800
	Total	210	790	1000

Tabla 3.15: Tabla de contingencia para valorar la validez de un diagnóstico.

3.4.1. Límites de normalidad

Antes de cuantificar la fiabilidad del procedimiento diagnóstico vamos a intentar detallar qué entendemos por valores anómalos. Por lo general, consideramos anómalos los valores extremos, ya sean demasiado grandes o demasiado pequeños, en relación con la distribución considerada, hasta completar un 5 % (aproximadamente). Si la variable se ajusta aproximadamente a un modelo de distribución de campana de Gauss, los límites a partir de los cuales los valores se consideran extremos son, según el Ejercicio 28,

$$\overline{x} \pm 2 \cdot s. \tag{3.5}$$

Un ejemplo interesante puede ser el estudio de concentración de hemoglobina glicosilada, cuya distribución en la población no diabética podemos apreciar en el histograma de la Figura 3.6, construido a partir de $n = 335$ individuos sanos que aportaron una media de 4.80 y una desviación típica de 0.60.

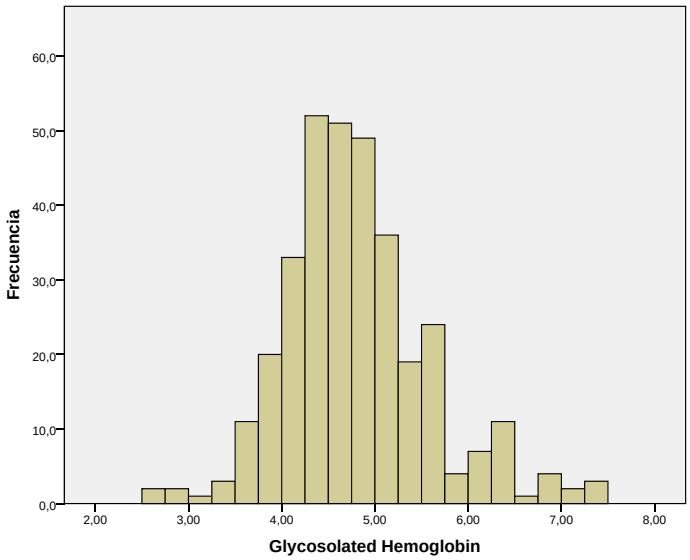


Figura 3.6: Hemoglobina glicosilada en individuos sanos.

Ejercicio 77. *¿Qué límites de normalidad podemos obtener aproximadamente a partir de estos datos? ¿A partir de qué valor puede pensarse en un diagnóstico de diabetes?*

En otros casos la variable en cuestión puede presentar un fuerte sesgo positivo, por lo que los límites de normalidad no deben calcularse según (3.5), pero que es corregido mediante una transformación logarítmica (como sucede, por ejemplo, el PSA) de manera que sí podemos determinar unos límites de tolerancia en función del logaritmo.

3.4.2. Fiabilidad de un procedimiento de diagnóstico

Una vez hemos entendido cómo puede diseñarse a grandes rasgos un procedimiento de diagnóstico, vamos intentar analizar la fiabilidad del mismo partiendo de una tabla de contingencia tipo 2×2, como la del Ejemplo 10, donde se confronta la enfermedad con el resultado del diagnóstico. Efectivamente, es posible, como se aprecia en la tabla, que un individuo sano sea diagnosticado erróneamente como enfermo (positivo), lo cual se denomina falso positivo. También es posible que un individuo enfermo sea diagnosticado como sano (negativo), lo cual sería un falso negativo. Por ello, definimos las siguientes medidas:

Sensibilidad: es la proporción de enfermos que son diagnosticados como positivos.

Especificidad: es la proporción de sanos diagnosticados como negativos.

Para el método diagnóstico del Ejemplo 10, obtendríamos las siguientes estimaciones

a partir de la tabla obtenida:

$$\texttt{Sensibilidad:} \qquad \hat{P}(+|E) = \frac{120}{200} = 0.600,$$
$$\texttt{Especificidad:} \qquad \hat{P}(-|S) = \frac{710}{800} = 0.887.$$

Es decir, la proporción de falsos negativos en la muestra es del 40.0 % y la de falsos positivos del 11.3 %. Nótese que estamos suponiendo que en el estudio la enfermedad está controlada, es decir, que hemos escogido un grupo de enfermos y otro de sanos, lo cual se conoce mediante un diagnóstico veraz previo. Sin embargo, desconocemos de antemano si estos individuos darán positivo o negativo con el nuevo procedimiento.

Ejercicio 78. *¿Qué sensibilidad y especificidad se espera de un procedimiento de diagnóstico completamente fiable?*

Curvas COR: ya hemos comentado que uno de los procedimientos más habituales de diagnóstico consiste en observar si una cierta variable, que correlaciona con la enfermedad estudiada, presenta un valor anómalo desde el punto de vista de la población sana, pero verosímil desde el punto de vista de la población enferma. Por ejemplo, es conocido que la enfermedad celiaca se asocia a concentraciones excesivamente elevadas del anticuerpo IgA en una analítica. Por lo tanto, un primer procedimiento para detectar la enfermedad puede consistir en establecer un umbral concreto de manera que un valor de IgA por encima del mismo se considere positivo en el test de diagnóstico. Si utilizamos uno de los dos límites de normalidad estudiados anteriormente podemos garantizar un procedimiento con una especificidad superior al 95 %, pero que puede ser poco sensible. Por contra, desplazar el umbral para aumentar la sensibilidad conduce necesariamente a una reducción de la especificidad.

Ejercicio 79. *Razona las dos afirmaciones anteriores.*

El problema estadístico se reduce pues a encontrar un umbral de la variable que permita obtener simultáneamente una sensibilidad y una especificidad razonables, lo cual se analiza gráficamente mediante la curva COR (característica receptiva del operador), como la que aparece en la Figura 3.7. En general, la variable analizada es tanto más válida cuanto más se aproxime a 1 el área subyacente a la curva, y el umbral ideal se corresponde con el punto de la curva más próximo al punto de coordenadas (0,1). En este caso particular, el área subyacente resulta ser 0.825, y el umbral que permite la mejor aproximación es IgA=33.8, para el cual se obtienen una especificidad del 80 % y una sensibilidad del 73 %, según indica el programa SPSS, aunque esta decisión es muy discutible.

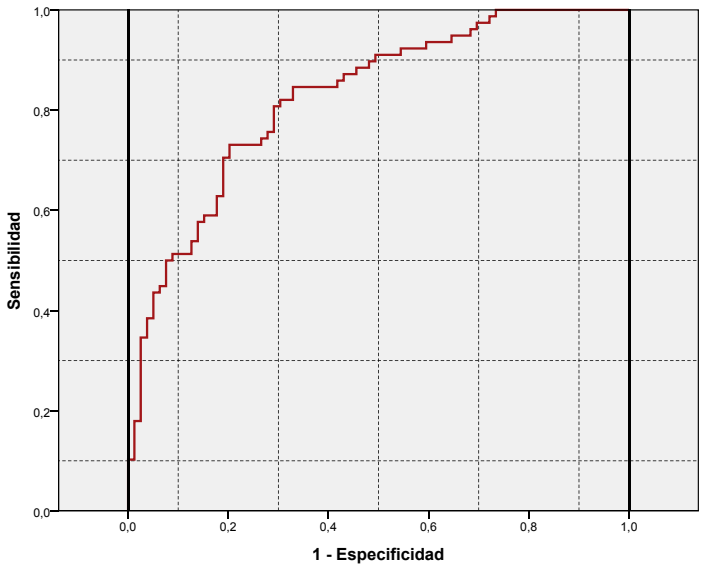


Figura 3.7: Curva COR para el diagnóstico de celiaquía a partir de IgA.

Valor predictivo positivo: se entiende como la probabilidad de estar enfermo si se ha dado positivo en el test[5].

Valor predictivo negativo: se entiende como la probabilidad de estar sano si se ha dado negativo en el test.

Ejercicio 80. *¿Qué valores predictivos positivo y negativo cabe esperar de un método de diagnóstico completamente certero?*

Ejercicio 81. *¿Cómo estimarías en principio los valores predictivos positivo y negativo directamente a través de la tabla? ¿Por qué el diseño habitual de casos-control utilizado en el Ejemplo 10 no permite unas estimaciones adecuadas según el procedimiento anterior?*

Dado que el diseño habitual de estos estudios no permite estimar los valores predictivos positivo y negativo directamente a través de las tablas, procederemos a estimarlos a partir de la sensibilidad (`sens`) y especificidad (`esp`), supuesta conocida de antemano (por otras vías) la prevalencia de la enfermedad. Para ello utilizaremos la fórmula de Bayes (3.4):

$$VP+ = \frac{\texttt{sens} \times \texttt{prev}}{\texttt{sens} \times \texttt{prev} + (1 - \texttt{esp}) \times (\texttt{1-prev})},$$

$$VP- = \frac{\texttt{esp} \times (1 - \texttt{prev})}{(1 - \texttt{sens}) \times \texttt{prev} + \texttt{esp} \times (\texttt{1-prev})}.$$

[5]Es la primera vez que mencionamos este concepto de probabilidad de manera explícita. Podemos interpretarlo de manera intuitiva o, también, entenderlo como la proporción de enfermos entre los individuos de la población que darían positivo en el test.

Así, si suponemos conocido que la enfermedad considerada en el Ejemplo 10 presenta una prevalencia del 2 %, tendremos las siguientes estimaciones:

$$\widehat{VP}+ = \frac{0.60 \times 0.02}{0.60 \times 0.02 + 0.113 \times 0.98} = 0.097,$$

$$\widehat{VP}- = \frac{0.887 \times 0.98}{0.40 \times 0.02 + 0.887 \times 0.98} = 0.990.$$

El procedimiento empleado parece ser pues mucho más útil para descartar la enfermedad que para detectarla. Otras veces ocurre lo contrario, por lo que la práctica habitual es combinar diferentes tests. Para más detalles al respecto consultar la bibliografía recomendada, en especial [1].

Otras cuestiones propuestas

Ejercicio 82. *Completa la siguiente tabla de contingencia de manera que podamos obtener un valor $\phi = 1$. ¿Cómo lo interpretarías en términos epidemiológicos?*

	Sexo			
Enfermedad	2×2	Hombre	Mujer	Total
	Enfermo			
	Sano			
	Total	6000	4000	10000

Tabla 3.16: Tabla de contingencia para las variables sexo y enfermedad.

Ejercicio 83. *Completa la siguiente tabla de contingencia de manera que podamos obtener un valor $\phi = 0$. ¿Cuál será entonces el correspondiente valor de C?*

	Sexo			
Rh	2×2	Hombre	Mujer	Total
	Rh+	40	60	
	Rh-			
	Total			120

Tabla 3.17: Tabla de contingencia para las variables sexo y Rh.

Ejercicio 84. *Si pretendemos probar la eficacia de una vacuna mediante una tabla 2×2 como en el caso del Ejemplo 9, ¿cómo debemos interpretar en términos clínicos un resultado $\phi = 0.02$?*

Ejercicio 85. *Supongamos que mediante un estudio de seguimiento se concluye que el riesgo de que un individuo con diabetes tipo II acabe desarrollando hipertensión triplica al de los individuos no diabéticos. ¿Qué parámetro estadístico se está manejando en el enunciado? ¿Cuál es su valor numérico en este caso? Suponiendo cierto el enunciado anterior, consideremos otro estudio estadístico en el que se selecciona una muestra de 1000, de las cuales 500 son diabéticas y otras 500 no. Teniendo en cuenta que estamos en condiciones de medir la presión arterial a los individuos de la muestra, razona lo mejor posible si la proporción de hipertensos de esta muestra constituye una estimación aceptable de la prevalencia de la hipertensión arterial.*

Ejercicio 86. *Supongamos que después de un largo estudio de seguimiento a fumadores habituales se estimó que el 15 % de los mismos acaba desarrollando cáncer de pulmón. Por otro lado, es conocido que aproximadamente el 90 % de los enfermos de cáncer de pulmón han sido fumadores habituales. Por último, otro estudio clínico diferente concluyó que el porcentaje de fumadores de la población se sitúa actualmente en torno al 30 %, con pocos cambios a lo largo de las últimas décadas. A partir de esta información, ¿serías capaz de estimar el riesgo relativo correspondiente al hábito de fumar? Interprétalo en términos intuitivos.*

Ejercicio 87. *Para estudiar la posible relación entre la exposición a un agente radioactivo se lleva a cabo un seguimiento durante 20 años de 5.000 individuos próximos a dicho agente y otros 95.000 lejanos, contabilizando en cada caso los tumores de tiroides que fueron diagnosticándose. Los resultados del estudio quedan recogidos en la siguiente tabla:*

	Exposición			
Tumor	2×2	Sí	No	Total
	Sí	25	30	55
	No	4975	94970	99945
	Total	5000	95000	100000

Tabla 3.18: Tabla de contingencia para las variables exposición y presencia del tumor.

(a) ¿De qué tipo de diseño se trata?

(b) Calcula dos medidas del riesgo que, según la muestra, supone la proximidad al agente radioactivo.

(c) ¿Cuál de ellas crees que es la más apropiada? Interprétala en términos clínicos.

(d) Calcula el coeficiente ϕ y compáralo con la medida anterior para entender por qué en epidemiología se utilizan parámetros de correlación específicos.

Ejercicio 88. *En las Figuras 3.8 y 3.9 se muestran sendos diagramas de barras agrupadas que ilustran la relación entre la agresividad de un tumor de próstata con la presencia de*

hiperplasia prostática, y con la presencia de penetración capsular, respectivamente. El estudio se realizó a partir de una muestra de 97 pacientes con tumor. Razona en cuál de los dos estudios se observa una mayor correlación y trata de proporcionar un valor aproximado para el coeficiente C en ambos casos.

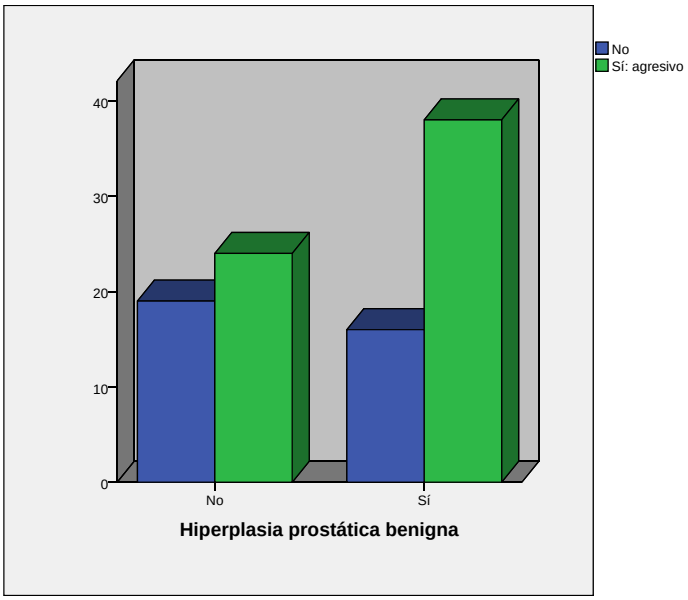


Figura 3.8: Diagrama de barras agrupadas para las variables agresividad del tumor e hiperplasia prostática.

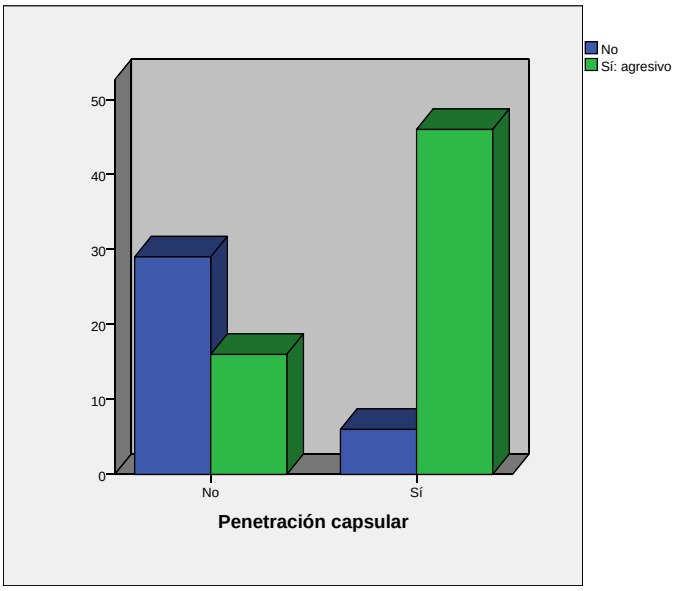


Figura 3.9: Diagrama de barras agrupadas para las variables agresividad del tumor e penetración capsular.

Ejercicio 89. *Se piensa que la presencia de cierta variedad de un gen puede predisponer a un cierto tipo de tumor. Para contrastarlo se seleccionaron 1000 individuos sanos y otros tantos afectados por el tumor. A continuación, se procedió a efectuar un análisis genético de todos los individuos de la muestra para determinar si presentaban o no la variedad del gen. Los resultados aparecen en la siguiente tabla:*

	Tumor			
Gen	2 × 2	Sí	No	Total
	Sí	610	360	970
	No	390	640	1030
	Total	1000	1000	2000

Tabla 3.19: Tabla de contingencia para las variables gen y presencia de tumor.

(a) ¿De qué tipo de diseño se trata?

(b) Calcula una medida de riesgo adecuada e interprétala en términos clínicos.

Ejercicio 90. *Calcula el valor de ϕ a partir de la Tabla 3.19.*

Ejercicio 91. *Considera una determinada enfermedad, un posible factor de riesgo asociado y diseña un hipotético estudio con vistas a medir el grado de riesgo de dicho factor.*

Ejercicio 92. *Si el peso medio medio de un bebé varón de nacido tras 40 semanas de gestación es de 3.5 kg con una desviación típica de 0.310 kg, determina los límites a partir de los cuales un bebé varón puede considerarse anormalmente pesado y anormalmente liviano.*

Ejercicio 93. *Se pretende valorar la efectividad de una prueba diagnóstica A para una enfermedad presente en el 2 % de la población. Para ello fue aplicada a una muestra constituida por 750 enfermos y 250 sanos con los siguientes resultados:*

	Diagnóstico A			
Enfermedad	2 × 2	+	-	Total
	Enfermo	730	20	750
	Sano	50	200	250
	Total	780	220	1000

Tabla 3.20: Tabla de contingencia para valorar la validez diagnóstico A.

(a) Estima la sensibilidad y especificidad de la prueba diagnóstico, así como las proporciones de falsos positivos y falsos negativos.

(b) Estima los valores predictivos positivos y negativos.

(c) Valora los resultados en términos prácticos.

Ejercicio 94. *Disponemos de otro procedimiento diagnóstico B para la misma enfermedad que en el Ejercicio 93. Sus resultados tras aplicarlo a los mismos individuos son los siguientes:*

	Diagnóstico B			
Enfermedad	2×2	+	-	Total
	Enfermo	610	140	750
	Sano	3	247	250
	Total	613	387	1000

Tabla 3.21: Tabla de contingencia para valorar la validez diagnóstico B.

(a) Estima nuevamente la sensibilidad, especificidad y los valores predictivos positivo y negativo.

(b) Valora los resultados y compararlos con los del procedimiento A.

Ejercicio 95. *Supongamos que el 50 % de los fetos de 8 semanas de gestación son varones y el resto, hembras. Se dispone de un método para detectar el sexo que reconoce correctamente como tales al 90 % de los varones y al 90 % de las hembras. Según estos datos, ¿qué proporción de fetos "diagnosticados" como varones lo son realmente?*

Ejercicio 96. *Supongamos que la sensibilidad de una prueba diagnóstica es del 100 %. ¿Cuánto vale entonces el valor predictivo negativo $VP-$?*

Ejercicio 97. *En un estudio llevado a cabo en EE.UU. se seleccionó una gran muestra de 3722 personas adultas que podemos asumir como aproximadamente aleatoria. Mediante cierto procedimiento basado en una serie de características físicas se diagnosticaron como positivos a los individuos que se consideraban candidatos a sufrir un infarto durante los siguientes 15 años, y como negativo a los que no. Después de un seguimiento de 15 años se registró qué individuos acabaron sufriendo realmente de infarto y se correlacionó con el diagnóstico previo mediante la siguiente tabla de contingencia:*

	Diagnóstico			
Infarto	2×2	+	-	Total
	Sí	190	668	858
	No	35	2829	2864
	Total	225	3497	3772

Tabla 3.22: Tabla de contingencia para la validez del diagnóstico de infarto.

Estima, a partir de la tabla, la sensibilidad, especificidad y valores predictivos positivo y negativo del método de diagnóstico previo del infarto.

PARTE

II

Inferencia Estadística

4. Conceptos básicos de Inferencia Estadística

Tal y como indicamos en la Introducción, el propósito final de la Bioestadística es explicar fenómenos biomédicos, que estarán en principio sujetos cierto nivel de incertidumbre, con el propósito de eliminarla en la medida de lo posible. Con esa intención se procede al análisis descriptivo de una muestra, en la que podemos observar un grado mayor o menor de correlación y en un sentido determinado. Recordemos algunos ejemplos:

- En el estudio de la longitud del fémur y el peso de 40 fetos ilustrado por la Figura 2.7 hemos observado una correlación lineal directa entre ambas variables ($r = 0.802$), que se mejora si añadimos al estudio las circunferencias de cabeza y abdomen, obteniendo entonces la ecuación (2.1) para predecir el peso del feto a partir de las medidas del ecógrafo.

- En el estudio de relación entre la acidosis y el nivel de glucemia en 200 recién nacidos del Ejemplo 5, que se ilustra en la Figura 2.16, observamos que la acidosis respiratoria y, en especial la metabólica, se asocian a un incremento del nivel medio de glucemia, hecho que no parece suceder con la acidosis mixta.

- En el estudio de eficacia de una vacuna contra la hepatitis expuesto en el Ejemplo 9, observamos que los individuos no vacunados de la muestra presentan un riesgo 6.5 veces mayor de padecer la hepatitis que los vacunados de la muestra.

Lo que resta es completar el esquema de la Figura 1 determinando en qué medida lo observado en la muestra puede generalizarse a la población de la que procede. Efectivamente, el hecho de que en una muestra concreta apreciemos cierto grado de correlación no debe hacernos descartar que, si la reemplazamos por otra diferente, nuestra conclusión sea otra. Esta variabilidad de las posibles muestras se debe a que el carácter que pretendemos explicar (peso, acidosis, hepatitis) se rige en buena parte por un conjunto de variables que no controlamos en el experimento y que por lo tanto, fluctúan de una muestra a otra. Es lo que se entiende comúnmente como azar. Debemos decidir pues si la correlación observada en la muestra es clara, es decir, significativa, o bien si puede ser explicada por el azar. Especialmente en el primer caso conviene determinar también un

margen de error para los diferentes valores típicos, dado que éstos varían de una posible muestra a otra. En definitiva, en Inferencia Estadística distinguimos dos tipos de problemas: los problemas de contraste de hipótesis y los problemas de estimación. Trataremos ambos en diferentes secciones haciendo especial hincapié en el cálculo del margen de error estadístico, el concepto de P-valor y la repercusión del tamaño de la muestra.

Obviamente, la Inferencia Estadística debe formularse en un lenguaje probabilístico. No obstante, haremos un uso intuitivo del concepto de probabilidad (que ya ha sido tratado, aunque no de forma explícita, en el capítulo anterior) que, en última instancia, se trata de una proporción. Es más, en el contexto de las Ciencias de la Salud podemos permitirnos la licencia de identificar probabilidad con proporción calculada respecto al total de una población. Así, por ejemplo, la probabilidad de medir más de 1.70 se entiende como la proporción de individuos de la población estudiada que verifica tal propiedad.

4.1. Parámetros poblacionales y muestrales

Todos los valores típicos estudiados en los Capítulos 1, 2 y 3 a partir de una muestra de tamaño n pueden definirse teóricamente en la población a partir de todos los valores de la población estudiada. Decimos teóricamente porque en la práctica no podrán ser calculados. Así por ejemplo, según vimos en (1.1), la media muestral viene definida por:

$$\overline{x} = \sum_{i=1}^{k} \mathtt{x}_i \hat{p}_i, \tag{4.1}$$

donde $\hat{p}_i$ denota la proporción de datos de la muestra que presenta el valor $\mathtt{x}_i$. Su homólogo poblacional, la media poblacional, que se denota como μ, se define entonces mediante

$$\mu = \sum_{i} \mathtt{x}_i p_i, \tag{4.2}$$

donde p_i denota la proporción de datos de la población que presenta el valor p_i, es decir, la probabilidad de $\mathtt{x}_i$. De la misma forma que definimos la media poblacional, podemos definir en la población todos los demás valores típicos. Como es usual, denotaremos por letras griegas los parámetros poblacionales para distinguirlos de sus homólogos muestrales o descriptivos, que se denotan por letras latinas. En otras ocasiones, los parámetros poblacionales se expresan directamente con letras latinas y los muestrales con la misma letra y, encima, el signo $\hat{}$.

Muestral	$\overline{x}$	s^2	r	B_j	$\widehat{RR}$	$\widehat{OR}$	...
Poblacional	μ	σ^2	ρ	β_j	RR	OR	...

Tabla 4.1: Parámetros muestrales y poblacionales.

Las conclusiones definitivas del estudio dependen de lo que sepamos acerca de los parámetros poblacionales. Por ejemplo, en el problema de relación entre el peso y la

longitud del fémur en fetos, que exista relación equivale a que el coeficiente de correlación lineal poblacional ρ no sea nulo; la relación es directa si es positivo y es más fuerte cuanto mayor sea ρ^2. La mejor ecuación para predecir el peso a partir de las medidas del ecógrafo viene dada por los valores β_0, β_1, β_2 y β_3 de la ecuación de regresión poblacional. Por otra parte, concluiríamos que la acidosis influye en el nivel de glucemia si encontramos diferencias entre las medias de glucemia de las cuatro categorías poblacionales, μ_1, μ_2, μ_3 y μ_4 (sanos, acidosis respiratoria, acidosis metabólica y acidosis mixta); en ese caso, el sentido de la relación vendría dado por el signo de las diferencias y el grado de relación, por la magnitud de las mismas. Por último, que el hecho de no estar vacunado incremente el riesgo de padecer hepatitis equivale a que el riesgo relativo poblacional RR sea mayor que 1, incrementándose más cuanto mayor sea RR.

En resumen, si pudiéramos calcular los parámetros poblacionales como calculamos los muestrales, el problema finalizaría aquí pues las conclusiones sería inapelables. La cuestión es que los parámetros poblacionales no pueden obtenerse en la práctica, sino que tenemos que conformarnos con sus homólogos muestrales, es decir, estimarlos a partir de unas muestras de las cuales nos fiamos parcialmente.

Ejercicio 98. *¿Por qué no podemos calcular en la práctica los parámetros poblacionales? De poder hacerlo, indica cómo probarías que se da una relación inversa entre la concentración en sangre de calcio y hormona paratiroidea. ¿Cómo determinarías una ecuación para explicar una variable a partir de la otra? ¿Serían exactas las predicciones?*

4.2. Muestreo

Dado que las posibles conclusiones de nuestro estudio pasan por el análisis previo de una muestra, deberíamos dar unas nociones mínimas de cómo deben seleccionarse. Si lo que pretendemos es extrapolar al global de la población la descripción de la muestra, la segunda debería ser representativa de la primera. La forma teórica de obtener una muestra representativa es mediante un muestreo aleatorio, que consiste básicamente en seleccionar a los individuos de la muestra mediante un proceso análogo a una lotería. Efectivamente, cualquiera de nosotros puede comprobar que si lanza un dado simétrico un número n suficientemente grande de ocasiones, las proporciones de unos, doses, treses, cuatros, cincos y seises obtenidas se aproximan a $1/6$. Es decir, que los resultados de n lanzamientos de un dado simétrico siguen los que se denomina Ley de azar, que constituye el fundamento de la Inferencia Estadística.

Ejercicio 99. *Relaciona en estos términos las ecuaciones* (4.1) *y* (4.2) *suponiendo que la muestra a partir de la cual se ha calculado $\overline{x}$ es aleatoria y grande, para así entender la aproximación de $\overline{x}$ a μ y, en general, de los valores típicos a sus respectivos homólogos poblaciones.*

En ocasiones, como en el problema de la acidosis en bebés, se precisa elegir una muestra aleatoria para cada categoría estudiada; ocurre lo mismo en los estudios de cohortes, donde se elige una muestra de expuestos y otra de no expuestos a un posible factor de riesgo, o en los de caso-control, donde se elige una muestra de enfermos y otra de sanos (el problema

de acidosis es una variante de este tipo). En el caso del estudio del fémur y el peso de los fetos, no deberíamos considerar ninguna estratificación a la hora de seleccionar la muestra, sino efectuar un sorteo simple.

Hay que advertir claramente que, salvo en estudios de gran calado, la obtención de la muestra mediante un sorteo en la población es casi utópica; que debemos conformarnos con analizar los datos de los que disponemos, siempre y cuando podamos descartar un claro sesgo o intencionalidad a la hora de incluirlos en el estudio. Si es así, la muestra puede considerarse, si no aleatoria, al menos arbitraria, lo cual puede ser suficiente si no sobrevaloramos los métodos que vamos a aplicar. Como ya comentamos en la Introducción, ello supone un primer error de partida que debemos estar dispuestos a arrastrar en el resto del estudio y al que se añadirán otros, aspecto que debemos tener muy presente en nuestras conclusiones, que deben relativizarse.

Punto de partida teórico: una buena parte de los prodecimientos que vamos a aplicar en lo sucesivo se basan de manera directa o indirecta en el resultado teórico que enunciaremos a continuación y que ya se introdujo intuitivamente en la Sección 1.1. Previamente, debemos tener presente que, si estamos estudiando una variable X definida sobre una población, con media μ y varianza σ^2, a partir de una muestra supuestamente aleatoria de tamaño n, tanto la media aritmética $\overline{x}$ como la varianza s^2 de la muestra pueden entenderse asimismo como variables numéricas, en el sentido de que pueden tomar diferentes valores en función de la muestra particular considerada. Como tales poseen, a su vez, una media y una varianza en relación al conjunto de las posibles muestras de tamaño n que pueden constituirse en la población.

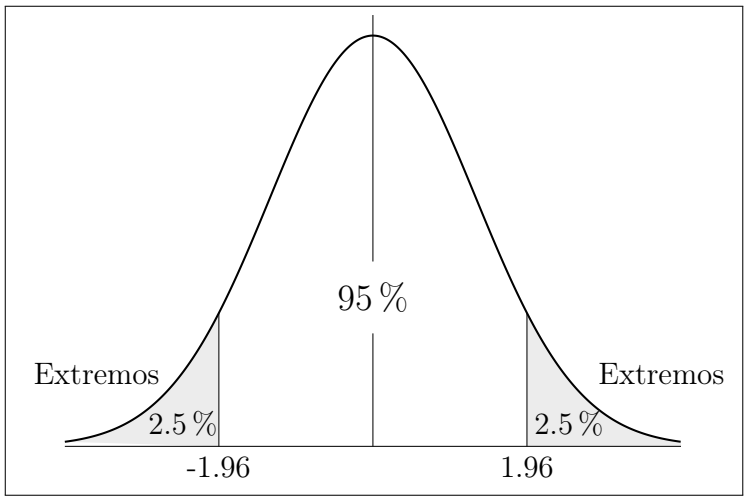


Figura 4.1: Distribución normal estándar $N(0,1)$.

Proposición 1. *En ese caso, se verifica que la media aritmética calculada a partir de una muestra de tamaño n tiene media μ y varianza σ^2/n, y se distribuye aproximadamente según un modelo de campana de Gauss si n es lo suficientemente grande. En consecuencia, si tipificamos la variable $\overline{x}$, se verifica, para n suficientemente grande, que $\sqrt{n}(\overline{x}-\mu)/\sigma$ sigue un modelo de distribución $N(0,1)$ (véase Figura 4.1). Si reemplazamos desviación típica poblacional σ por la desviación típica de la muestra obtenemos una distribución muy*

similar a la $N(0,1)$, que se denomina distribución t-Student. En definitiva, se verifica aproximadamente:

$$\frac{\overline{x}-\mu}{s/\sqrt{n}} \sim N(0,1), \tag{4.3}$$

En consecuencia, para un 95 % de las posibles muestras de tamaño n, se verifica que

$$\left|\frac{\overline{x}-\mu}{s/\sqrt{n}}\right| \leq 1.96. \tag{4.4}$$

Es decir, $|\overline{x}-\mu| \leq 1.96 \cdot s/\sqrt{n}$. El valor 1.96 ha aparecido ya en otras ocasiones pero redondeado como 2, por ejemplo, en la página 21 y en el Ejercicio 28. Se trata del valor que delimita dos colas con el 5 % de los datos más extremos en la distribución $N(0,1)$. De hecho, cuando en el Capítulo 1 afirmábamos que, en una campana de Gauss aproximadamente el 95 % de los datos quedan comprendidos en el intervalo $\overline{x} \pm 2s$, estábamos redondeando el valor 1.96.

4.3. Estimación

Ya sabemos que los valores típicos estudiados en la primera parte constituyen estimaciones o aproximaciones de los correspondientes parámetros poblacionales, que serán más certeros cuanto mayor sea la muestra. No obstante, suponiendo que la muestra sea aleatoria, estamos en condiciones de acotar el error con un cierto grado de confianza, es decir, de aportar un intervalo en el cual esperamos que se encuentre el valor desconocido del parámetro poblacional. Estas cotas se basan en cálculos probabilísticos más o menos básicos según el caso.

Intervalo de confianza para la media: el intervalo al 95 % de confianza para la media poblacional μ de una variable numérica a partir de una muestra de tamaño n con media $\overline{x}$ y desviación típica s es, según (4.4):

$$\overline{x} \pm 1.96\frac{s}{\sqrt{n}}. \tag{4.5}$$

Así pues, el margen máximo de error de la estimación $\overline{x}$ con una confianza del 95 % es:

$$E_{max} = 1.96 \cdot s/\sqrt{n}. \tag{4.6}$$

Ejemplo 11. Se pretende estimar la media, μ, de la estatura, que denotamos como X, de las mujeres de entre 16 y 50 años pertenecientes a una amplia población. Para ello se escogió una muestra (que supondremos aleatoria) de $n = 40$ mujeres, las cuales aportaron una media aritmética de 162.3 cm con una desviación típica de 5.2 cm.

En consecuencia, ya tenemos una estimación puntual de la media μ: la media aritmética $\overline{x} = 162.3$. El margen máximo de error al 95 % de confianza es:

$$E_{\text{máx}} = 1.96 \cdot \frac{5.2}{\sqrt{40}} = 1.6.$$


Por lo tanto, el intervalo de confianza al 95 % correspondiente es 162.3 ± 1.6. En definitiva, podemos afirmar con una confianza del 95 % que la altura media de la población se encuentra entre 160.7 cm y 163.9 cm.

La expresión (4.6) merece algunos comentarios aclaratorios:

- Cuanto mayor sea la desviación típica muestral s, es decir, cuanto más variabilidad se aprecie en la muestra, mayor será el margen de error. Efectivamente, una gran dispersión observada en la variable a través de la muestra se traduce a su vez en una variabilidad de la media aritmética muestral, en el sentido de que puede variar mucho de una muestra a otra y, por lo tanto, es poco fiable.

- Cuanto mayor sea el tamaño de muestra, n, menor es el margen de error. Efectivamente, es el tamaño de la muestra el que puede amortiguar la variabilidad cuantificada por s. De hecho, a medida que el tamaño tiende a infinito, el margen de error tiende a 0. En la práctica, podemos aprovechar la expresión (4.6) para determinar de manera aproximada el tamaño de muestra necesario, en función de un margen máximo de error establecido de antemano y con una confianza determinada (usualmente del 95 %), supuesta conocida una estimación inicial de la desviación típica mediante una pequeña muestra piloto.

- Cuando hablamos de 95 % de confianza no estamos expresando de forma vaga un grado de certeza psicológica sino que queremos decir lo siguiente: el procedimiento expresado en (4.6) proporciona un margen máximo de diferencia entre $\overline{x}$ y μ que se respetarían para el 95 % de las posibles muestras de tamaño n[1].

- En ocasiones se desea una confianza mayor, por ejemplo del 99 %. En ese caso, debemos reemplazar 1.96 por el valor que permite delimitar dos colas iguales con el 1 % del área en la curva anterior. Se trata concretamente de 2.58. Se denotan respectivamente por $z_{0.05}$ en el primer caso y $z_{0.01}$ en el segundo. En general, z_α es el valor que permite delimitar dos colas cuya suma de áreas sea α. Los distintos valores (cuantiles) pueden obtenerse a partir de una tabla numérica asociada a la distribución normal[2] $N(0,1)$.

Intervalo de confianza para una proporción: cuando estudiamos una variable cualitativa con dos categorías, como por ejemplo el hecho de padecer o no cierta dolencia, y pretendemos calcular un intervalo de confianza para la proporción global de enfermos p a partir de la proporción $\hat{p}$ en la muestra estudiada, se procede aplicando la proposición anterior a la variable numérica X que asigna el valor 1 al individuo que padece la enfermedad y 0 al que no la parece; este procedimiento está justificado por el hecho de que la media aritmética de dicha variable equivale a la proporción muestral de enfermos

[1] Es preciso entender dicha afirmación si se aspira a un comprensión más formal de la Inferencia Estadística.

[2] Existen otras tablas probabilísticas muy utilizadas en Inferencia Estadística y relacionadas con la $N(0,1)$ que también consideraremos, como la t-Student (ya mencionada), la χ^2 y la F-Snedecor. Todas ellas llevan asociados unos parámetros enteros denominados grados de libertad que las modulan. Para entender estos conceptos remitimos al lector textos más completos (consúltese la bibliografía).

y la varianza viene dada por $p(1-p)$, que es en todo caso inferior a $1/4$, y donde p es la proporción de enfermos en la población. En ese caso, para calcular un tamaño de muestra (conservador) que garantize una margen máximo de error $E_{\text{máx}}$ en la estimación de la proporción poblacional p, basta con despejar n en la fórmula siguiente:

$$E_{max} \leq \frac{1}{\sqrt{n}}.$$

Es una equivocación muy común asumir por defecto un margen máximo de error del 5 % en la estimación de la proporción p (es decir, confundirlo con la probabilidad de que el intervalo sea correcto), porque esa cantidad puede resultar o no aceptable en función del propio valor de p (desconocido). Por ejemplo, es un error considerar un margen de error del 5 % en la estimación de la prevalencia de una enfermedad rara.

En general, conocer de antemano el tamaño de muestra preciso para afrontar con garantías un estudio estadístico es uno de las grandes deseos del investigador experimental. Sin embargo y a pesar de las creencias que se propagan desde muchos ámbitos, es muy difícil satisfacer dicho deseo porque requiere determinar de antemano uno o varios parámetros[3] que pueden resultar más polémicos que el propio tamaño de muestra. No obstante, existen diversas fórmulas como podemos comprobar, por ejemplo, en [9, Capítulo 7], aunque hemos de ser muy cuidadosos en su aplicación y no hacer un mal uso de las mismas.

Ejercicio 100. *¿Estamos realmente en condiciones de determinar de manera aproximada un tamaño de muestra suficiente como para alcanzar el grado deseado de precisión en la estimación? ¿Cómo?*

4.4. Contraste de hipótesis

Como ya hemos comentado, en Inferencia Estadística distinguimos dos tipos de problemas: de estimación y de contraste de hipótesis. Este último consiste en contrastar o evaluar, a partir de la muestra considerada, si un modelo teórico dado (hipótesis) es o no aceptable. Se denomina test de hipótesis al algoritmo numérico al que se somete la muestra para tomar tal decisión. Por desgracia, la teoría de tests de hipótesis está lejos de poder ofrecer un algoritmo satisfactorio para cualquier posible hipótesis que podamos concebir. Más bien se limita al contraste de hipótesis iniciales muy concretas, en el sentido de que acaban asociándose a una única distribución de probabilidad. De esta forma, el estudio de relación entre variables se traduce en el contraste de un modelo inicial de independencia. Tanto la hipótesis inicial, que se denota por H_0, como la hipótesis opuesta, que se denota por H_1 y se denomina hipótesis alternativa, pueden expresarse con frecuencia en términos de parámetros poblacionales, como en los siguientes ejemplos que podemos encontrar en los Capítulos 2 y 3.

- Relación del peso de fetos con la longitud de su fémur: $H_0 : \rho = 0$, o, equivalentemente, $H_0 : \beta_1 = 0$.

[3] Como el margen máximo de error asumible en un intervalo de confianza, o la mínima potencia de un test dado un cierto tamaño del efecto en un problema de contraste de hipótesis.

- Relación del peso del feto con la longitud de su fémur y las circunferencias craneal y abdominal: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$.
- Relación de la acidosis en recién nacidos con el nivel de glucemia: $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$.
- Riesgo de no vacunarse de cara a padecer hepatitis: $H_0 : RR = 1$, o $H_0 : OR = 1$, según el diseño.

No podemos afirmar que todas las hipótesis iniciales sean de este tipo pero sí al menos las más importantes.

El test de hipótesis puede considerarse como una prueba de compatibilidad entre la hipótesis inicial y la muestra considerada. Por ejemplo, si contrastamos una igualdad de medias a partir de una muestra debemos evaluar la verosimilitud de la muestra suponiendo que la igualdad de medias se verificase. El criterio intuitivo que rige la posterior decisión se denomina Principio de Máxima Verosimilitud y podríamos formularlo así:

> En todo caso caso debemos optar por el modelo que haga más verosímil nuestra muestra. Es decir, si nuestra muestra es poco verosímil según un modelo teórico dado, debemos pensar que dicho modelo no explica correctamente la realidad y descartarlo.

En definitiva, partiremos de un modelo inicial (igualdad de medias, por ejemplo) y evaluaremos lo verosímil o compatible que es nuestra muestra según dicho modelo, de forma que, si resulta verosímil, lo aceptaremos y, en caso contrario, lo rechazaremos.

***P*-valor:** se trata de uno de los conceptos más importantes de la Estadística. Es una probabilidad y como tal se obtiene haciendo uso del Cálculo de Probabilidades, pero el aspecto más importante en relación al estudio que aquí realizamos es que:

> Debe entenderse como la medida de la verosimilitud de la muestra según el modelo teórico inicial H_0.

En consecuencia, un valor grande de P expresa que la muestra es verosímil (no extrema) según la hipótesis inicial, por lo que no estamos en condiciones de rechazarla. Por contra, un valor pequeño de P indica que la muestra es poco verosímil (extrema) según H_0, por lo que, siguiendo del Principio de Máxima Verosimilitud, debemos rechazar la hipótesis inicial H_0 en favor de su alternativa H_1.

Falta por determinar qué entendemos por grande o pequeño o, dicho de otra forma, que entendemos por verosímil y qué entendemos por extremo o raro. Como ya habremos comprobado, en Estadística se conviene, siguiendo una cierta tradición, que lo raro o extremo debe suponer a lo sumo un 5 % del total, de ahí que 0.05 sea el valor de referencia o nivel de significación habitual[4]. En definitiva:

[4]Esta elección está sujeta a una creciente controversia (véase [15]).

- $P > 0.05$: La muestra es compatible con la hipótesis inicial (resultado no significativo).
- $P < 0.05$: La muestra no es compatible con la hipótesis inicial (resultado significativo).

4.4.1. La importancia del tamaño muestral

En ningún caso debe confundirse un test de hipótesis con una demostración matemática, pues el resultado del primero es sólo una decisión razonable a partir de los datos que debe relativizarse. De hecho, hay que tener muy presente que los tests de hipótesis tienden a aportar resultados no significativos cuando se aplican a muestras de pequeño tamaño. Por contra, con muestras muy numerosas se pueden obtener resultados significativos por pequeñas evidencias contra H_0, afirmación que intentaremos razonar a continuación.

¿Cómo dar entonces sentido al uso de tests de hipótesis ante este hecho? Para los estudios más habituales (problemas de correlación en sentido amplio) podría valer la siguiente afirmación:

Si el resultado de un test es significativo entonces tenemos claro en qué sentido se da la correlación: en el que indica la muestra observada. Por contra, si el resultado no es significativo, el sentido de la correlación observada en la muestra no es extrapolable a la población.

Por ejemplo, si estamos estudiando la posible relación entre una variable cualitativa con dos categorías y una variable numérica cuyas medias poblacionales son μ_1 y μ_2, respectivamente, a partir de sendas muestras aleatorias, podemos inclinarnos a pensar que existe una cierta tendencia o correlación si, por ejemplo, la media aritmética de la primera muestra es superior a la de la segunda.

No obstante, cabría pensar también que una nueva muestra del mismo tamaño podría aportar una visión contraria como consecuencia del propio azar del muestreo. Si es eso lo que pensamos no estaremos en condiciones de saber si μ_1 es mayor que μ_2 o lo contrario, es decir, si la diferencia entre las medias poblacionales es positiva o negativa, siendo entonces el 0 un valor posible para dicha diferencia. Es decir, que las medias podrían incluso ser iguales. Ello justifica que en tal caso el test consista en medir la compatibilidad entre los datos obtenidos y la hipótesis inicial $H_0 : \mu_1 = \mu_2$. Dado que el procedimiento se limita a cuantificar en qué medida la muestra observada es compatible con dicha hipótesis mediante un P-valor, un resultado no significativo se interpretaría como una compatibilidad entre ambas, en cuyo caso la correlación observada quedaría en suspenso. Un resultado significativo indicaría que la muestra es extrema desde el punto de vista de la hipótesis inicial y, por lo tanto, poco compatible con la misma. En ese caso descartaríamos la hipótesis inicial de igualdad de medias en favor de la superioridad de μ_1 respecto a μ_2. En definitiva, estaríamos extrapolando el sentido de la correlación observada en la muestra a toda la población.

Ahora bien, las muestras pequeñas están sometidas a una gran variabilidad, es decir,

que en ellas la mayor parte de las circunstancias teóricamente posibles pueden ocurrir con una probabilidad aceptable, por lo que es difícil que una muestra pequeña pueda considerarse extrema desde el punto de vista de H_0. Por contra, las muestras grandes presentan un comportamiento muy regular, por lo que cualquier pequeña desviación respecto al patrón medio teórico correspondiente a H_0 puede considerarse una circunstancia extrema según H_0.

Es algo similar a lo que ocurriría con la máquina de Galton (véase Figura 1.5): dejando caer una bola (muestra) por un par de niveles no se puede manifestar un defecto de fabricación de la máquina. Sin embargo, al dejarla caer a través de muchos niveles (tamaño de muestra grande), si la máquina está efectivamente bien diseñada, es muy probable que la bola acabe aproximadamente en el centro. Luego, si acaba en un extremo ($P < 0.05$), el Principio de Máxima Verosimilitud nos moverá a pensar en un defecto de construcción (tendencia significativa).

Eso explica que, en problemas de correlación, se obtengan con frecuencia resultados no significativos con muestras pequeñas a pesar de observar en las mismas correlaciones moderadas; por contra, también podemos obtener con facilidad resultados significativos con muestras grandes a pesar de observar correlaciones[5] pequeñas. Llevando el razonamiento al extremo, con muestras enormes casi todos los contrastes de interés resultarán significativos. Dicho de otra forma más intuitiva, las tendencias observadas en la muestra enormes son casi automáticamente extrapoladas a la población.

En definitiva, los tests de hipótesis constituyen una herramienta fundamental de la Estadística, en especial para muestras de tamaño mediano, pero su uso es más cuestionable cuando las muestras son muy pequeñas o muy grandes (téngase en cuenta que este último caso es el más deseable desde el punto de vista estadístico).

Ejercicio 101. *¿Por qué afirmamos que las muestras pequeñas están sometidas a mayor variabilidad que las grandes?*

4.5. El test de Student como ejemplo

Veamos un ejemplo de cómo funciona un test de hipótesis. Hemos escogido el test posiblemente más utilizado en Bioestadística. Se utiliza para tratar de determinar si existe una relación significativa entre una variable cualitativa binaria (como, por ejemplo, estar sano o enfermo, ser tratado o no tratado) y una variable numérica (glucemia, presión arterial, etc). Según indicamos anteriormente, el problema de relación entre ambas variables se traduce en un problema de comparación de las medias poblacionales de la variable numérica, μ_1 y μ_2, correspondientes a cada una de las categorías consideradas. Es decir, la hipótesis inicial a contrastar es:

$$H_0 : \mu_1 = \mu_2.$$

Si seleccionamos de manera independiente sendas muestras aleatorias para cada categoría, el algoritmo al que se someten los datos se denomina test de Student para muestras

[5] En términos más genéricos hablaríamos de *tamaños del efecto*.





independientes.

Ejemplo 12. Se estudia la posible relación entre la edad de la primera menstruación (menarquia) y la enfermedad celiaca. Para ello se toma una muestra de $n_1 = 79$ mujeres sanas (no celiacas) y otra muestra de $n_2 = 78$ celiacas de edad similar. En cada caso se anotó la edad en años de la menarquia. Desde el punto de vista descriptivo, las sanas aportaron una media $\overline{x}_1 = 12.74$ y una desviación típica $s_1 = 1.48$, mientras que las celiacas aportaron una media $\overline{x}_2 = 13.33$ con una desviación típica $s_2 = 1.90$. En la Figura 4.2 se establece una comparativa de ambas muestras a través de los diagramas de caja[6].

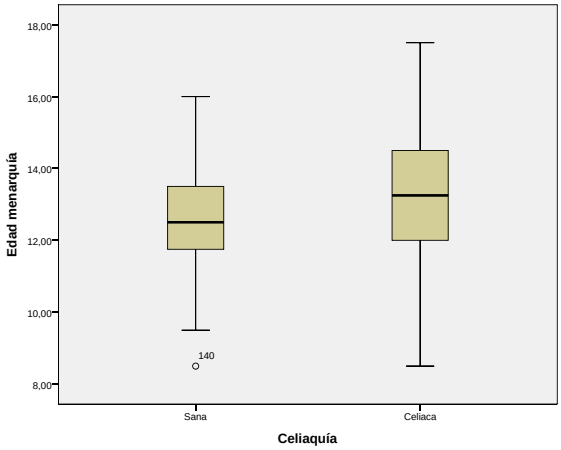


Figura 4.2: Diagramas de caja para la menarquia según la presencia de celiaquía.

Podemos observar que, al menos por término medio (y mediano), las mujeres celiacas de la muestra presentan una menarquia ligeramente más tardía que las sanas. Hemos de analizar si esa diferencia apreciada en esta muestra concreta es significativa. Sólo en ese caso podremos inferir que, en general, la celiaquía se asocia a una primera menstruación más tardía. Inicialmente, supondremos que ambas variables no guardan relación ($\mu_1 = \mu_2$) y evaluaremos si la muestra estudiada contradice claramente dicha suposición.

Según el modelo inicial las medias muestrales $\overline{x}_1$ y $\overline{x}_2$ deberían ser parecidas, es decir, la diferencia (en bruto) $\overline{x}_1 - \overline{x}_2$ debería ser próxima a 0. Obviamente, no podemos exigir que sea igual a 0 porque debemos asumir diferencias entre las muestras debidas exclusivamente al azar inherente al muestro. El problema es cuantificar qué estamos dispuestos a achacar al azar, lo cual es un problema de Cálculo de Probabilidades. Concretamente, según el modelo inicial, la diferencia de medias muestrales debería seguir un modelo de distribución normal de media 0, de manera que, al tipificarlo según (4.7), debería seguir una distribución $N(0,1)$ como la de la Figura 4.1.

$$t_{exp} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{4.7}$$

[6]Una comparación similar en función de las medias puede realizarse haciendo uso de los diagramas de medias cuya representación se indica en el Capítulo 6.

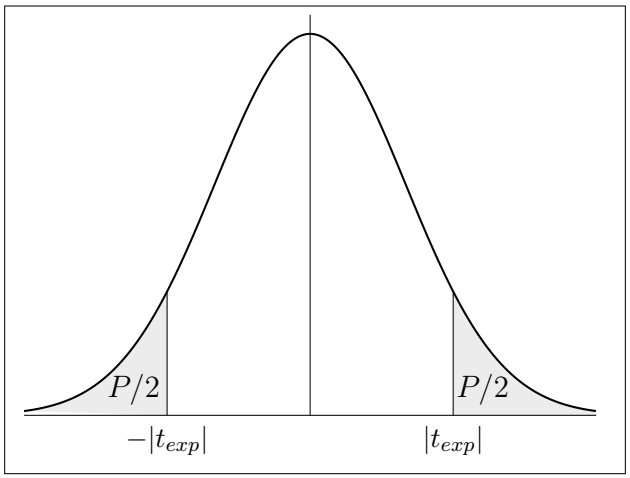


Figura 4.3: Distribución de t_{exp} según H_0.

El número t_{exp} resultante[7], denominado valor experimental, recoge toda la información que aporta la muestra estudiada en lo referente al contraste de la hipótesis $H_0 : \mu_1 = \mu_2$. De hecho, su valor absoluto se entiende como una distancia (tipificada) entre las dos medias muestrales que, bajo la hipótesis $H_0 : \mu_1 = \mu_2$, debería ser pequeña. Más concretamente, debería ajustarse a un modelo de distribución $N(0, 1)$ (véase Figura 4.3). El P-valor se define en este problema concreto como la probabilidad, según $N(0, 1)$, de obtener una distancia (tipificada) entre medias aritméticas al menos tan grande como la observada en la muestra. En otras palabras, el P-valor es el área de las colas que determinan $-|t_{exp}|$ y $|t_{exp}|$, como se indica en la Figura 4.3, lo cual expresa en qué medida es verosímil la muestra según H_0. En nuestro ejemplo, $t_{exp} = -2.18$, correspondiéndole entonces un valor $P = 0.031$. Según hemos convenido, el resultado es significativo, es decir, se opta por la hipótesis alternativa $H_1 : \mu_1 \neq \mu_2$, por lo que podemos concluir que la celiaquía se relaciona con la menarquia en el sentido indicado.

Por contra, obtener un valor t_{exp} próximo a 0, es decir, una escasa diferencia entre las medias muestrales, sería verosímil desde el punto de vista de la hipótesis inicial $H_0 : \mu_1 = \mu_2$, asociándose a un P-valor alto según la distribución $N(0, 1)$. Se entendería entonces que la muestra es compatible con la hipótesis inicial y, en definitiva, que el sentido de la tendencia observado en la muestra no es extrapolable a la población de la que procede. Ese no ha sido el caso, pues nosotros sí estamos en condiciones de generalizar la tendencia observada en la muestra: la celiaquía se asocia a una menarquia más tardía.

Intervalo de confianza para la diferencia de medias: los mismos cálculos probabilísticos que nos llevan a considerar (4.7) conducen también al siguiente intervalo[8] de

[7] En el test de Student propiamente dicho se reemplaza el denominador anterior por la expresión $s_c\sqrt{n_1^{-1} + n_2^{-1}}$, donde $s_c^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$.

[8] Al igual que en (4.7), se calcula en la práctica a través de s_c.

confianza al 95 % para la diferencia entre μ_1 y μ_2:

$$\overline{x}_1 - \overline{x}_2 \ \pm \ z_{0.05}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

En nuestro ejemplo, obtenemos que $\mu_1 - \mu_2$ debe encontrarse, con una confianza del 95 %, en el intervalo

$$-0.59 \pm 0.54 \ = \ (-1.13, -0.05),$$

lo cual indica que la media μ_1 (menarquia media para sanas) es en todo caso menor que μ_2 (menarquia media para celiacas), porque la diferencia observada en la muestra es mayor que el margen de error calculado. Ello concuerda con lo que ya sabíamos a través del P-valor. De hecho, puede comprobarse analizando la expresión (4.7) que $P < 0.05$ equivale a que el 0 quede fuera del intervalo al 95 % de confianza para $\mu_1 - \mu_2$. Pero el intervalo aporta algo que no expresa explícitamente el P-valor, pues cuantifica con un margen de error la diferencia entre las categorías, por lo que viene a dar una magnitud de la influencia de la variable cualitativa sobre la numérica. Esto es especialmente útil en el caso de muestras de gran tamaño, para las cuales los resultados suelen ser significativos.

Por último, advertimos que en este problema hemos precisado del conocimiento de la distribución $N(0,1)$, lo cual se debe en última instancia a que en la expresión (4.7) se están valorando sumas,[9] ya que las medias aritméticas se calculan sumando valores.

Ejercicio 102. *Existe la teoría de que el Bisfenol A, compuesto químico presente en muchos tipos de plástico y que nuestro organismo puede absorber, podría dar lugar a abortos tempranos en embriones masculinos, lo cual haría disminuir la proporción de nacimientos varones. Para contrastar dicha teoría, se efectuó un seguimiento de 6 embarazadas que, por su trabajo, estaban muy expuestas al Bisfenol A, resultando que todas ellas tuvieron finalmente niñas. ¿Corrobora eso la teoría? Responde directamente a través de un P-valor.*

4.6. Tests paramétricos y tests no paramétricos

Ya hemos comentado que en la mayoría de las ocasiones contrastaremos hipótesis iniciales expresadas en términos de parámetros poblacionales, como la media o el coeficiente de correlación. Este punto de vista está claramente vinculado a la distribución normal. Efectivamente, sabemos de la importancia que en general posee el parámetro media, y que este debe complementarse con alguna medida de dispersión para poder caracterizar la distribución de los datos. La desviación típica desempeña ese papel, al menos en el caso de la distribución normal. No obstante, cabe preguntarse, primeramente, qué utilidad tiene el estudio de estos parámetros cuando no podemos suponer la normalidad de la distribución (por ejemplo cuando se da un fuerte sesgo) y, segundo, si los tests de hipótesis que propondremos en el siguiente capítulo, o el propio test de Student, son válidos aunque no se satisfaga la normalidad de las variables numéricas consideradas. Esta problemática conduce a la fragmentación de la Inferencia Estadística en dos ramas. En la primera, la distribución normal desempeña un papel central, por lo que las inferencias se orientan a

[9] Conviene tener en cuenta aquí los comentarios acerca de la campana de Gauss del Capítulo 1.


conocer lo posible acerca de los parámetros asociados a dicha distribución. Esta rama se denomina por lo tanto Estadística Paramétrica. La otra corriente construye los distintos métodos partiendo de débiles supuestos sobre la distribución de las variables y no se busca por lo tanto el conocimiento de los parámetros que las caracterizan, de ahí que se denomine Estadística no Paramétrica. Podemos decir que los métodos no paramétricos clásicos se basan fundamentalmente en el orden de los datos, es decir, que de cada observación de la muestra importará sólo el rango o posición que ocupa respecto a los demás datos de la misma. Son, por lo tanto, métodos robustos ante la presencia de valores extremos (como sucede con el cálculo de la mediana) pero, por contra, bajo el supuesto de normalidad son menos potentes, es decir, tienen menor capacidad de detectar la violación de la hipótesis inicial a partir de los datos. Nosotros nos centraremos aquí en los métodos paramétricos, aunque indicaremos escuetamente en cada caso el procedimiento no paramétrico que podría reemplazar al método paramétrico propuesto en el caso de que este sea inviable.

Para decidir si la distribución original de los datos es o no normal contamos con los denominados tests de normalidad que introduciremos a continuación. No obstante y en virtud del Teorema Central el Límite, un tamaño de muestra suficientemente grande puede permitirnos en ciertos casos obviar el supuesto de normalidad y permitirnos aplicar en todo caso un método paramétrico. El esquema simplificado a seguir es el siguiente:

Distribución original normal o muchos datos $\longrightarrow$ Método paramétrico

Distribución original no normal y pocos datos $\longrightarrow$ Método no paramétrico

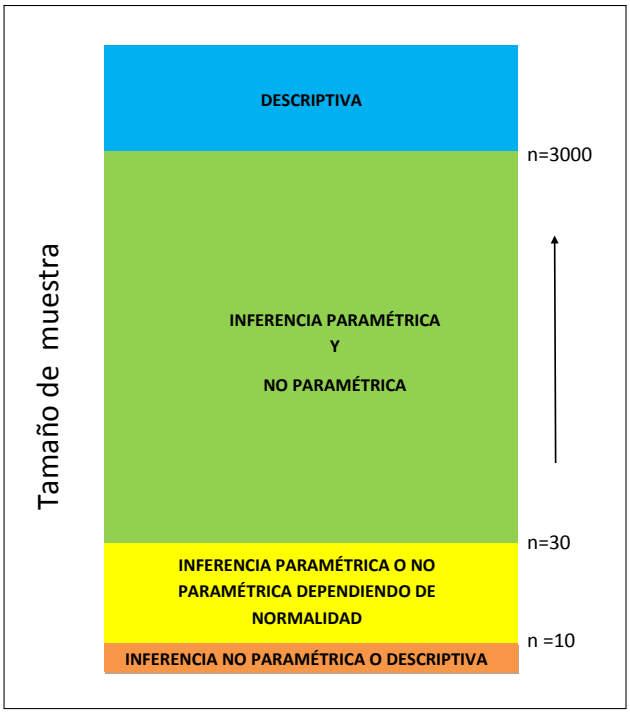


Figura 4.4: Métodos estadísticos y tamaño de muestra.

4.6.1. Pruebas de normalidad

Asumir el supuesto de normalidad significa aceptar que la distribución de frecuencias relativas de los datos de la población se adaptan aproximadamente a una curva normal. Esta situación ocurre con bastante frecuencia en Ciencias de la Salud, lo cual no quiere decir que se deba dar por descontado.

Precisamente, existen diversos métodos, como el test de Kolmogorov-Smirnov, el test χ^2, el test de Shapiro-Wilk o el test de D'Agostino, para contrastar la hipótesis inicial de que cierta variable sigue un modelo de distribución normal a partir de una muestra aleatoria de tamaño n. La mayoría de ellos están vinculados a aspectos gráficos. También existe un método basado directamente en los coeficientes de simetría y aplastamiento. Se trata en definitiva de contrastar la hipótesis inicial de normalidad de la variable numérica X estudiada:

$$H_0 : X \sim \text{Normal}.$$

De esta forma, se rechazará la normalidad cuando los datos observados la contradigan claramente. En este capítulo hemos afirmado que la mayoría de los contrastes se pretende probar si existe correlación entre variables, suponiendo como hipótesis inicial que ésta es nula. El contraste de normalidad puede considerarse una excepción en ese sentido, pues sólo entra en juego una variable numérica. Nótese además que la normalidad de la variable es la hipótesis inicial. En consecuencia, una muestra pequeña y, por lo tanto, con escasa información, difícilmente podrá conducir a rechazar la hipótesis de normalidad. Por contra, si la muestra es muy grande, los resultados serán significativos ante la menor violación del supuesto de Normalidad (Ejercicio 101). Por ello, debemos ser muy precavidos a la hora de interpretar los resultados si nos decidimos a aplicar un test de este tipo. No conviene perder de vista el tamaño de la muestra que se estudia y los aspectos gráficos (histogramas) de la misma.

Ejercicio 103. *Tras aplicar el test de normalidad de Shapiro-Wilk a los 30 datos de colesterolemia, obtenemos como resultado* $P = 0.973$. *Interprétalo en términos prácticos.*

5. Métodos de Inferencia Estadística

En este capítulo exponemos de manera muy esquemática las técnicas de Inferencia Estadística más utilizadas en los problemas de relación entre variables. Se trata pues de una continuación natural de los Capítulos 2 y 3. Por lo general, para cada problema estudiado indicaremos la alternativa no paramétrica al test paramétrico propuesto. La Tabla 5.1, que podríamos considerar como una evolución o mejora de la Tabla 1, puede servirnos como resumen de los métodos y como guion a seguir durante el capítulo. No obstante, en la última sección introduciremos algunas técnicas más avanzadas muy utilizadas por los investigadores en Ciencias de la Salud.

Este manual está ideado como guía para que un usuario de la Estadística sepa aplicar mediante el software adecuado las técnicas básicas, de ahí que los detalles teóricos queden relegados a la bibliografía recomendada. En definitiva, se pretende que, dado un problema concreto, el lector sea capaz de identificar el procedimiento estadístico a seguir e interpretar los resultados que se obtienen tras la aplicación del programa estadístico. Recordemos que en la tercera parte de este manual el lector cuenta con tutorial de SPSS que puede servirle de guía para ejecutar los diferentes métodos e interpretar los resultados.

Problema	Método Paramétrico	Alternativa no Paramétrica
Dos medias independientes	Student (2)	Mann-Whitney
Más de dos medias	Anova de un factor	Kruskal-Wallis
Dos medias apareadas	Student (1)	Wilcoxon
Correlación numérica	Test correlación r	Correlación de Spearman
Tabla de contingencia	Test χ^2	Test exacto de Fisher

Tabla 5.1: Métodos básicos en Inferencia Estadística.

5.1. Tests de Student y Welch para muestras independientes

En el Capítulo 2 de la primera parte adelantamos que el estudio de la relación entre una variable cualitativa y otra numérica puede traducirse en una comparación entre las medias (o parámetros de centralización en general) que dicha variable numérica posee en cada categoría de la variable cualitativa. Ahora estamos en condiciones de abordar este estudio desde el punto de vista inferencial, lo cual dará pie a las técnicas más populares de la Bioestadística. El test de Student para muestras independientes es la primera de ellas. Ya ha sido introducido en el Capítulo 4 a raíz del Ejemplo 12, en el que se comparaban las edades medias de la menarquia de dos categorías de mujeres: celiacas y no celiacas. Para ello se procedió a seleccionar, de manera independiente, sendas muestras de tamaños n_1 y n_2 que fueron sometidas al test de Student(2), consistente en confrontar con la tabla t-Student(n_1+n_2-2), similar a la $N(0,1)$, el valor experimental

$$t_{exp}=\frac{\overline{x}_1-\overline{x}_2}{s_c\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}.$$

El resultado fue $P<0.001$. Además, se concluyó que la diferencia entre medias poblacionales debía encontrarse, con una confianza del 95 %, en el intervalo $(-1.13,-0.05)$.

La comparación de medias puede realizarse en todo caso con dos tests diferentes: el de Student, descrito anteriormente, y el test de Welch, que supone una ligera variación. Esto es así porque el test de Student –aquí lo denominamos también Student(2)– requiere en principio que las distribuciones de la variable numérica en las categorías consideradas sean de tipo normal y con idénticas varianzas. El test de Welch sólo requiere normalidad. La normalidad podría contrastarse mediante un test o método gráfico adecuado. Si la aceptamos en ambas categorías deberíamos, teóricamente, contrastar la hipótesis inicial de igualdad de varianzas $H_0:\sigma_1^2=\sigma_2^2$ mediante el denominado test de Levene. Si podemos aceptar también dicha hipótesis, el test más adecuado es el de Student y, en caso contrario, el de Welch. El esquema puede simplificarse teniendo en cuenta que, si las muestras son de tamaños similares y suficientemente grandes, el resultado del test de Student puede considerarse válido, aunque no se verifiquen la normalidad ni la igualdad de varianzas. Sin embargo, con muestras pequeñas no podemos proceder de esa forma. Es más, con muestras pequeñas puede ocurrir que ninguno de los dos tests sea válido porque no se verifique o no se pueda valorarse con garantías la hipótesis de normalidad.

5.1.1. Alternativa de Mann-Whitney

No obstante, existe una alternativa no paramétrica a ambos tests que no exige la normalidad de la variables estudiada y que es, por lo tanto, de especial utilidad con muestras pequeñas (ver Figura 4.4). Se denomina test de Mann-Whitney y consiste básicamente en una comparación de los rangos o posiciones promedios de la variable numérica en función de las categorías consideradas. En el ejemplo 12 el test de Mann-Whitney aporta también como resultado $P<0.001$. De hecho, es bastante habitual que los tres test propuestos (Student, Welch y Mann-Whitney) conduzcan a conclusiones similares para muestras

grandes. Además y desde un punto de vista global, el error que se asume al optar por un test que no es del todo apropiado para la situación es con frecuencia mucho menor que el que se asume de partida al considerar que la muestra es representativa y que los datos obtenidos son mediciones fiables de las variables estudiadas. Es decir, que vista desde un punto de vista global y realista, la discusión anterior peca de cierta falta de coherencia.

En este manual aconsejamos al usuario de la Estadística que no permita que un protocolo excesivamente complejo le impida entender el objetivo principal del análisis. Para que los métodos estadísticos sean aplicados de forma mínimamente consistente proponemos pues un procedimiento más sencillo que tiene en cuenta únicamente los tests de Student y Mann-Whitney para resolver el problema de relación planteado, tal y como queda esquematizado al final de la Subsección 5.2.2.

Ejercicio 104. *¿Qué ventaja puede reportar aplicar el test de Student en lugar del de Mann-Whitney si se dan las condiciones apropiadas para el primero?*

5.1.2. Problemas de comparación de proporciones

Un problema estadístico muy común consiste en decidir si dos proporciones son o no iguales. Dicho contraste podemos afrontarlo de dos formas diferentes con resultados muy similares. La primera de ellas fue ya introducida en el Capítulo 4 y consiste en entender la variable cualitativa a cuyas proporciones nos referimos como una variable numérica con valores 1, si la cualidad se da, y 0, en caso contrario, de manera que el problema planteado puede resolverse mediante el test de Student(2) para comparar dos medias con muestras independientes, siempre y cuando las muestras sean lo suficientemente grandes. Esta técnica posee la ventaja de que proporciona un intervalo de confianza para la diferencia de proporciones.

La segunda técnica consiste en entender el problema como un estudio de relación entre dos variables cualitativas y aplicar el test χ^2, que veremos más adelante. El método puede extenderse sin problemas a comparaciones de tres o más proporciones, siempre y cuando se verifiquen las condiciones de validez del test.

5.2. Anova de un factor

Este test es una generalización del test de Student(2) para dos muestras independientes que se aplica para un mismo tipo de estudio y de diseño, con la salvedad de que podemos distinguir un número de categorías y, por lo tanto, de medias, mayor de dos. Sería pues apropiado para los datos del Ejemplo 5, en el que se trata de contrastar si las medias de glucemia son idénticas en las cuatro categorías consideradas (control, acidosis respiratoria, acidosis metabólica y acidosis mixta):

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

El test que resuelve el contraste se denomina anova de una vía o factor y requiere en principio de las mismas condiciones de validez que el test de Student para dos muestras independientes.

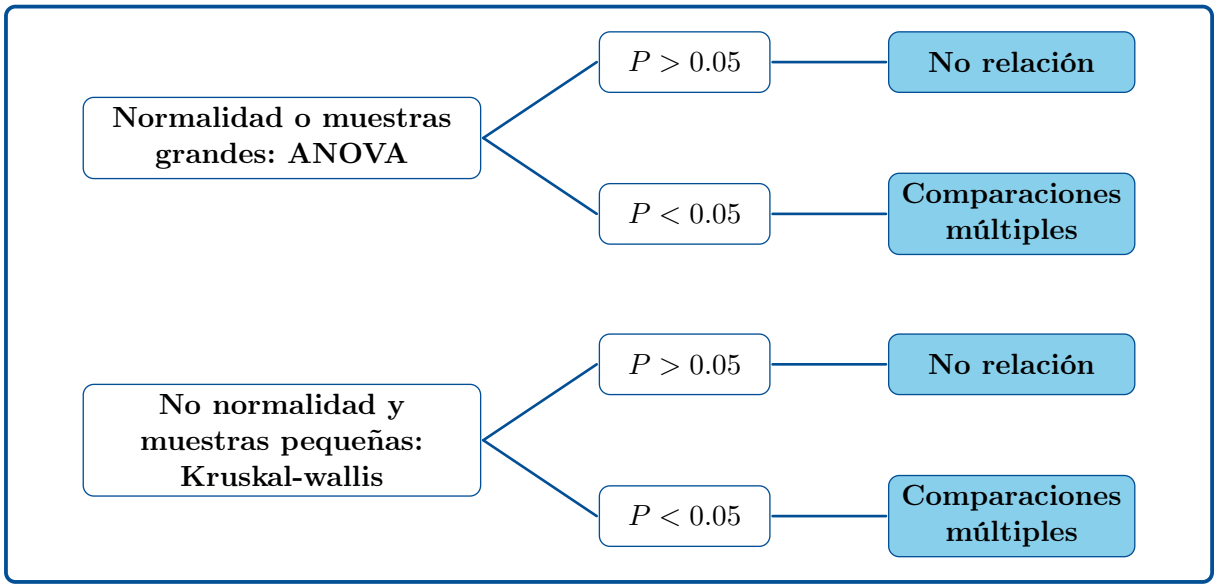
5.2.1. Alternativa de Kruskal-Wallis

Podemos efectuar, no obstante, las mismas consideraciones acerca de los tamaños muestrales que para el test de Student. Además, contamos con alternativas como el test de Brown-Forsyte y, especialmente, el test no paramétrico de Kruskall-Wallis, basado en rangos promedios, que a su vez generaliza el test de Mann-Whitney.

Ejercicio 105. *¿Qué sucederá si aplicamos el anova de una vía a un problema con dos medias?*

5.2.2. Método de Tukey

En el caso del Ejemplo 5, el P-valor obtenido es $P < 0.001$, es decir, las diferencias apreciadas a nivel muestral son realmente significativas, por lo que existe relación entre la acidosis y la glucemia. Para determinar de la manera más precisa en qué sentido se da dicha relación debemos proceder a comparar las medias por parejas de manera simultánea: se trata del denominado problema de comparaciones múltiples. Para ello tenemos a nuestra disposición diversos procedimientos aunque, para simplificar, podemos optar por el método de Tukey, que es ideal en el caso de que las muestras de las diferentes categorías sean de idéntico tamaño. Si hemos optado por aplicar el test de Kruskal-Wallis, podemos utilizar otros métodos de comparaciones múltiples. Un resumen de estos procedimientos se recoge en el siguiente esquema:



La Tabla 5.2 recoge los resultados de las comparaciones múltiples mediante el método de Tukey para los datos del Ejemplo 5. En dicha tabla, las categorías cuyas medias aparecen en columnas diferentes son las que se distinguen de manera significativa según el método Tukey:

Tipo de acidosis	1	2	3
Mixta	62.61		
Control	62.68		
Respiratoria		71.38	
Metabólica			78.80

Tabla 5.2: Método de Tukey aplicado a tipos de acidosis.

Podemos apreciar que, tal y como se intuía en la Figura 2.16, la acidosis mixta no se asocia a un cambio significativo de la glucemia mientras que la respiratoria y en especial la metabólica la aumentan significativamente.

5.3. Test de Student para muestras apareadas

Este test de Student es el apropiado para el diseño de muestras relacionadas o apareadas, que tiene como propósito controlar la variabilidad debida al individuo. Consiste en seleccionar una muestra aleatoria de n individuos a los que se les mide una variable numérica antes de iniciar un tratamiento para volver a medírsela después. En tal caso, no estaremos hablando de una variable sino de dos variables distintas, X_1 y X_2, medidas antes y después del tratamiento respectivamente, sobre una única población, sin distinguir categorías. Es decir, que mientras que el test de Student(2) de muestras independientes y el anova de un factor responden al problema de relación entre una variable cualitativa y otra numérica, el de Student(1) para muestras apareadas habría que encuadrarlo, en rigor, en el problema de relación entre dos variables numéricas.

Si el tratamiento es efectivo debe producirse una evolución, es decir, un cambio entre los valores de X_1 y X_2. No estamos en condiciones de exigir que ese cambio se dé en el mismo sentido para todos los individuos, pero sí al menos que se dé por término medio, de ahí que el problema se traduzca finalmente en una comparación entre las respectivas medias μ_1 y μ_2. Veamos un ejemplo.

Ejemplo 13. Se pretende probar los beneficios de la crioterapia en el tratamiento de la artrosis de rodillas en mujeres mayores. Para ello se seleccionó una muestra de $n = 30$ pacientes a las que se evalúo su nivel de dolor mediante la escala EVA antes de iniciar el tratamiento y tras 5 semanas de tratamiento. Los valores de dicha escala están comprendidos entre 0 y 10, donde 0 indica la ausencia de dolor y 10 indica dolor máximo. En resumen, obtenemos que la media muestral del dolor antes de iniciar el tratamiento es $\overline{x} = 5.37$, con una desviación típica $s_1 = 0.97$; el dolor medio muestral tras finalizar el tratamiento es $\overline{x}_2 = 5.59$, con una desviación típica $s_2 = 0.99$.

Podemos pues apreciar que, por término medio, en la muestra se ha producido un pequeño incremento del dolor. En consecuencia, esta muestra no supondrá en ningún caso una prueba significativa de la eficacia de la crioterapia para esta dolencia. Más bien

deberíamos preguntarnos si el tratamiento es contraproducente (o al menos incapaz de frenar un empeoramiento espontáneo), como en principio podría deducirse de la muestra. En todo caso, la hipótesis a contrastar es $H_0 : \mu_1 = \mu_2$.

El test de Student(1) para muestras relacionadas es especialmente sencillo, pues consiste en calcular la diferencia entre ambas variables, $d = X_1 - X_2$, cuya media media es $\mu_d = \mu_1 - \mu_2$, y contrastar la hipótesis inicial $H_0 : \mu_d = 0$. Para ello, considera la media aritmética $\overline{d}$ y desviación típica s_d de la diferencia[1] y confronta el valor

$$t_{exp} = \frac{\overline{d}}{s_d/\sqrt{n}},$$

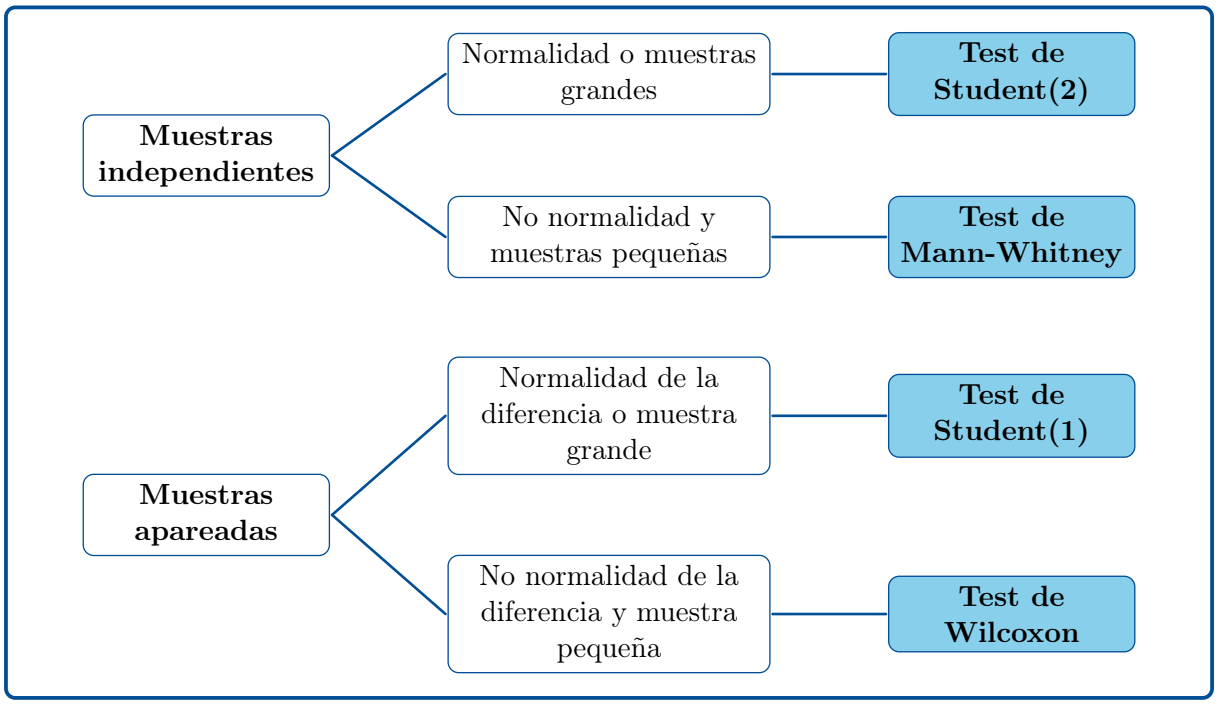
con la tabla t-Student$(n-1)$, similar a la $N(0,1)$. En nuestro caso se observa concretamente un incremento de 0.214 en el dolor medio que no resulta significativo ($P = 0.301$). De hecho, si analizamos el intervalo de confianza para la diferencia de medias podemos precisar que el nivel medio de evolución a nivel poblacional debe encontrarse entre un incremento de 0.630 puntos en dolor (empeoramiento) y un decremento de 0.201 (mejora). Es decir, no tenemos claro si se puede mejorar o empeorar.

El test de Student(1) para muestras relacionadas se plantea suponiendo que se verifica un requisito concreto: que la variable d distribuye según un modelo normal. Esto puede contrastarse mediante un test de normalidad, aunque hay que tener muy presente que, como en otros casos, el resultado del test puede considerarse válido aunque la distribución de la diferencia no sea normal, siempre y cuando la muestra sea lo suficientemente grande.

5.3.1. Alternativa de Wilcoxon

En todo caso, contamos con alternativas no paramétricas, especialmente útiles para muestras pequeñas. La más inmediata, denominada test de los signos, consiste en contrastar si la mediana de la diferencia es nula, lo cual se efectúa evaluando si hay diferencias significativas entre el número de diferencias positivas y el de diferencias negativos. No obstante, proponemos aquí como alternativa principal el test de la suma de rangos con signo de Wilcoxon, que combina la idea anterior con la que sustenta el test de Mann-Whitney. Concretamente, ordena los valores absolutos de las diferencias y les asigna rangos; a continuación a estos rangos se les asigna un signo + o - según sea la diferencia; por último, se compara la suma de los rangos positivos con la de los negativos que, bajo la hipótesis inicial, deberían ser similares. En nuestro caso aporta como resultado $P = 0.417$, por lo que la conclusión que se desprende del test de Wilcoxon es la misma que se desprende del de Student(1).

[1] Observemos que $\overline{d}$ puede calcularse directamente como $\overline{x}_1 - \overline{x}_2$ pero s_d no.

5.4. Test de correlación

Esta sección supone una continuación del Capítulo 2. El problema consiste en identificar una posible relación entre dos variables numéricas. En ocasiones, el objetivo es más ambicioso pues se busca explicar una variable numérica a partir de otras variables, a su vez numéricas, mediante una ecuación de regresión adecuada. En todo caso, utilizaremos la información de una muestra supuestamente aleatoria de tamaño n.

Empecemos por el caso más sencillo, el problema de correlación simple. Por ejemplo, consideremos el estudio de la relación entre del peso del feto y la longitud de su fémur, que se ilustra en la Figura 2.7. La muestra de tamaño $n = 40$ aportó un coeficiente de correlación lineal muestral $r = 0.802$ ($r^2 = 0.643$), es decir: en la muestra se aprecia un fuerte grado de correlación directa. La cuestión es si podemos extrapolarla al global de población para concluir que un fémur largo se asocia a un peso elevado. La respuesta parece obvia en este caso con sólo ver el gráfico, pero en otros casos no ocurrirá lo mismo.

En definitiva, estamos contrastando la hipótesis inicial de independencia entre peso y longitud de fémur, que puede expresarse a través del coeficiente de corrrelación lineal poblacional ρ mediante

$$H_0 : \rho = 0,$$

frente a la hipótesis alternativa $H_1 : \rho \neq 0$, que se corresponde con algún grado de relación lineal entre ambas. Por lo tanto, se trata de valorar si la muestra observada contradice significativamente la hipótesis inicial de independencia. De manera análoga a (4.7), la información que aporta la muestra queda resumida en el número

$$t_{exp} = \sqrt{(n-2)\frac{r^2}{1-r^2}}, \tag{5.1}$$

que se confrontará con la tabla de la distribución t-Student$(n-2)$ para obtener el P-valor correspondiente. Téngase en cuenta que para $m \geq 30$, la tabla de la t-Student(m) es prácticamente idéntica a la de la distribución $N(0,1)$. En nuestro caso obtenemos $t_{exp} = 8.27$, al que le corresponde un valor $P < 0.001$. Se dice entonces que la correlación observada es significativa, por lo que la tendencia observada es extrapolable. Por contra, un resultado no significativo en el test de correlación significaría que la posible relación observada en la muestra podría ser explicada exclusivamente por el azar, por lo que quedaría en suspenso, aunque esta situación no se ha dado en nuestro ejemplo.

Coeficiente de correlación de Spearman: cuando tenemos dudas acerca de la linealidad de la relación o advertimos la presencia de datos anómalos, podemos optar por la alternativa no paramétrica de Spearman, que consiste en calcular el coeficiente de correlación entre los rangos del mismo nombre y aplicarle un test específico. Trabajar con el coeficiente de correlación de Spearman puede ser un buen recurso cuando la relación observada no es lineal y encontramos una transformación adecuada para resolver el problema, situación que es muy común.

Ejercicio 106. *Tras aplicar el test de correlación a los datos correspondientes al Ejercicio 53 se obtiene $P < 0.001$. Interpreta el resultado en términos prácticos.*

Ejercicio 107. *Tras aplicar el test de correlación a los datos correspondientes a la Figura 2.10 se obtiene $P = 0.731$. Interpreta el resultado en términos prácticos.*

5.4.1. Regresión múltiple

Si nuestro objetivo es predecir una variable, como el peso del feto, de la mejor manera posible, debemos intentar explicarla a partir de varias variables que correlacionen con ella. Éstas serán incluidas en la regresión, dando lugar en un contexto poblacional, a una expresión a estimar del tipo

$$Y \simeq \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k. \tag{5.2}$$

Por ejemplo, en el caso del peso, podemos incluir, además de la longitud del fémur, las circunferencias del abdomen y cabeza, dado que son variables que también correlacionan con el peso (como puede comprobarse aplicando sendos tests de correlación) y porque entendemos que pueden explicar partes de la variabilidad del peso no explicada por la longitud del fémur, lo cual da lugar a $R^2 = 0.915$. La primera pregunta es si esta correlación es significativa. La respuesta es obvia y se obtiene mediante el test de correlación múltiple que es una generalización del anterior y cuyo resultado depende en este caso del valor

$$F_{exp} = \frac{n-(q+1)}{q} \frac{R^2}{1-R^2}, \tag{5.3}$$

siendo q el número de variables explicativas (en nuestro ejemplo $q = 3$). El valor F_{exp} se confronta con la tabla de la distribución $F_{q,n-(q+1)}$, que con n suficientemente grande es aproximadamente igual a la de la distribución denominada $\chi^2(q)$. El resultado es altamente significativo ($P < 0.001$), lo cual quiere decir simplemente que está claro que entre las tres variables logramos explicar algo del peso.

5.4.2. Intervalo de confianza para una predicción

Lo que realmente nos interesa en este caso es la ecuación (2.1), que tiene como objeto pronosticar el peso del feto a partir de las tres medidas proporcionadas por el ecógrafo. Por desgracia, no estamos en condiciones, ni mucho menos, de garantizar su exactitud aunque, en su defecto, podemos construir un intervalo de confianza al 95 % para cada predicción obtenida. Al margen máximo de error al 95 % de confianza para el pronóstico resulta ser proporcional a $1.96 s_y$, siendo s_y la desviación típica de la variables respuesta. Concretamente, si $d(x, \overline{x})$ denota la distancia tipificada entre el vector de valores explicativas y la media aritmética de la muestra, se verifica aproximadamente que

$$E_{max} = 1,96 \cdot s_y \cdot \sqrt{(1 - R^2)\left(1 + \frac{1}{n} + \frac{d^2(x, \overline{x})}{n}\right)}. \tag{5.4}$$

Es decir, que en términos relativos la precisión de la estimación dependerá de tres factores: el valor de R^2 obtenido, el tamaño de muestra n y la posición respecto a la muestra estudiada del individuo sobre el que se efectúa la predicción.

Ejercicio 108. *¿En qué sentido crees que influye en la precisión de la estimación cada uno de los factores anteriores?*

Ejercicio 109. *Mediante un programa estadístico construye un intervalo de confianza para la predicción efectuada en el Ejercicio 41.*

5.4.3. Contrastes parciales y selección de variables

Los coeficientes B_0, B_1, B_2 y B_3 de la ecuación son propios de la muestra estudiada y debemos pues interpretarlos como meras estimaciones de coeficientes β_0, β_1, β_2 y β_3 poblacionales. No obstante, estamos en condiciones de calcular intervalos de confianza para los mismos. Además, podemos aplicar los denominados tests parciales, que permiten contrastar hipótesis iniciales del tipo $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ o $H_0 : \beta_3 = 0$. El resultado de un test de parcial depende exclusivamente, de manera totalmente análoga la expresada en las ecuaciones (5.1) y (5.3), del tamaño de la muestra y del denominado coeficiente de correlación parcial, que expresa la capacidad de la variable explicativa en cuestión para predecir el valor de la respuesta en exclusiva, es decir, al margen de lo que ya predicen las demás variables explicativas. Finalmente, se confrontará un valor t_{exp} con la tabla de la distribución t-Student.

En el estudio del peso de los fetos los tres test parciales aportan resultados significativos, es decir, las tres variables explicativas son necesarias en la ecuación para explicar el peso. Cuando alguna variable aporta un resultado no significativo en su test parcial significa que no es esencial para explicar la variable respuesta pues su correlación parcial con la misma es débil. Eso no implica necesariamente que ambas variables no correlacionen. Podría deberse a que la variable explicativa no aporta nada que no explique ya el resto de las variables en la ecuación.

Multicolinealidad y selección de variables: nótese que, cuando las variables explicativas están fuertemente correlacionadas entre sí, se generan redundancias entre ellas que se traducen en una fuerte disminución de los coeficientes de correlación parcial y, por lo tanto, en una abundancia aparentemente sorprendente de resultados no significativos en los tests parciales. Dicha situación, que ya se mencionó en la Sección 2.2, se denomina multicolinealidad. Si queremos optimizar un modelo de regresión ante la presencia de multicolinealidad no debemos en ningún caso desechar simultáneamente todas las variables cuyos resultados en los tests parciales sean no significativos. Podemos optar por ir eliminando de una en una, recalculando el modelo en cada caso, hasta que obtengamos un modelo con resultados significativos en todos los tests parciales. Este algoritmo se conoce como método de selección hacia atrás o *backward*.

Por último, advertimos que, por motivos didácticos, no abordaremos en este manual técnicas de inferencia específicas para un problema de análisis de la covarianza (véase Sección 2.4), remitiendo en todo caso al lector interesado a una bibliografía más avanzada.

5.5. Test χ^2

Esta sección supone una continuación del Capítulo 3. Nuestro problema es determinar si una muestra dada supone una prueba significativa de la relación entre dos variables cualitativas. En esencia se trata de aplicar un test de correlación similar a (5.1) pero reemplazando r por una medida de asociación a nivel cualitativo: el coeficiente de contingencia C. De esta forma, el denominado test χ^2 se obtiene confrontando el valor

$$\chi^2_{exp} = n\frac{C^2}{1-C^2}, \tag{5.5}$$

con la tabla de la distribución $\chi^2(m)$, siendo $m = (r-1)(s-1)$, donde r denota el número de filas y s el número de columnas. Si nuestra tabla es del tipo 2×2, podemos calcular χ^2_{exp} a partir del coeficiente ϕ como $\chi^2_{exp} = \phi^2/n$.

Nótese la similitud[2] entre (5.5) y las expresiones análogas (5.1) y (5.3). En todo caso, el resultado del test se basa únicamente en el grado de correlación observado en la muestra, que se cuantifica mediante C^2, ϕ^2, r^2 o R^2 y el tamaño de la misma.

En el Ejemplo 8 relacionábamos la salud de los árboles, distinguiendo tres categorías según su nivel de cloroplastos, con la contaminación, distinguiendo a su vez tres categorías en función de la concentración de SO_2. En total contábamos con $n = 60$ árboles en el estudio que aportaron un valor $C = 0.444$. En consecuencia, obtenemos $\chi^2_{exp} = 14.74$ que se corresponde, según la tabla $\chi^2(4)$, con $P = 0.005$. Se trata pues de un resultado significativo. Por lo tanto, podemos concluir que, tal y como se aprecia en la muestra, las concentraciones elevadas de SO_2 se asocian a una peor salud de los árboles. Un P-valor similar se obtiene con los datos del Ejemplo 7, por lo que podemos concluir que la mejor valoración médica observadas en los individuos de la muestra con ICC de tipo normal podría extrapolarse al global de hombres de más de 40 años, suponiendo que esta muestra fuera representativa.

[2]Se trata de una similitud que resulta de forzar en cierta medida la teoría por razones didácticas.


5.5.1. Alternativa de Fisher

El test χ^2 precisa de una serie de condiciones de validez que, a grandes rasgos, se resumen en lo siguiente: debemos contar con una cantidad suficiente de datos, especialmente si pretendemos distinguir muchas categorías en las variables estudiadas. En caso contrario debemos agrupar categorías hasta llegar, si es preciso, a una tabla tipo 2×2. Si aun así el número de datos es demasiado pequeño, en concreto, si hay alguna casilla con un valor esperado E_{ij} menor que 5, debemos aplicar la alternativa no paramétrica conocida como test exacto de Fisher.

5.5.2. Inferencias para el Riesgo relativo y Odds Ratio

Como casos especiales de tablas tipo 2×2 tenemos los estudios epidemiológicos de factores de riesgo, que dan pie a las medidas conocidas como Riesgo Relativo y Odds Ratio. Ahora estamos en condiciones de entender también estos parámetros en términos poblacionales, en cuyo caso se denotan por RR y OR, respectivamente. Dado que un determinado factor comporte riesgo para una enfermedad concreta se traduce en $RR > 1$ o $OR > 1$, según la medida de riesgo considerada, esto nos conduce a contrastar las hipótesis iniciales $H_0 : RR = 1$, o bien $H_0 : OR = 1$. La primera, propia de un estudio de cohortes, se contrasta confrontando con la tabla $\chi^2(1)$ el valor experimental

$$\chi^2_{exp} = \frac{(\log \widehat{RR})^2}{s^2_{log\widehat{RR}}},$$

donde[3] $s^2_{log\widehat{RR}} = \frac{c}{a(a+c)} + \frac{d}{b(b+d)}$.

En el caso del Ejemplo 9, donde el posible riesgo es la ausencia de vacunación contra la hepatitis, obtenemos

$$s^2_{log\widehat{RR}} = 0.101, \quad \chi^2_{exp} = 34.97, \quad P < 0.001.$$

La hipótesis inicial $H_0 : OR = 1$ se contrastaría en un estudio tipo caso-control confrontando con la tabla $\chi^2(1)$ el valor experimental

$$\chi^2_{exp} = \frac{(\log \widehat{OR})^2}{s^2_{log\widehat{OR}}},$$

siendo $s^2_{log\widehat{OR}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$.
En nuestro caso,

$$s^2_{log\widehat{RR}} = 0.109, \quad \chi^2_{exp} = 35.24, \quad P < 0.001.$$

Queda pues claro que el hecho de no vacunarse contra la hepatitis implica un incremento en el riesgo de padecerla.

Ejercicio 110. *A partir de los datos del Ejercicio 87, contrasta si existe relación entre la exposición al agente radioactivo y el tumor de tiroides.*

[3] Siguiendo las notaciones de la tabla 3.13.


5.6. Algunas técnicas más avanzadas

Una vez completada la Tabla 5.1, procedemos a una breve y esquemática ampliación de la misma[4]. Sabemos que, en el problema de relación entre variables numéricas, es frecuente incrementar el número de variables explicativas para poder pronosticar mejor la variable respuesta, dando lugar a lo que conocemos como regresión múltiple. En general, en cualquier problema de relación entre variables podemos incrementar el número de variables explicativas dando lugar a diferentes tipos de estudios:

Explicativas	Respuesta	Problema estadístico
Varias numéricas	Numérica	**Regresión múltiple**
Numérica y cualitativa	Numérica	**Análisis de la covarianza**
Cualitativa y cualitativa	Numérica	**Anova de dos factores**
Numéricas y cualitativas	Cualitativa	**Regresión logística**

Tabla 5.3: Métodos avanzados.

Los dos primeros problemas de la Tabla 5.3 fueron ya estudiados desde un punto de vista descriptivo en el Capítulo 2; además, en este mismo capítulo, hemos estudiado también las inferencias relativas el problema de regresión múltiple. Para acabar, en esta sección abordaremos un estudio muy resumido de los dos últimos problemas.

5.6.1. Anova de dos factores

Es muy habitual en estudios de cierta envergadura, como los ensayos clínicos, intentar explicar una variable de tipo numérico a partir de dos variables cualitativas.

Ejemplo 14. Se desea probar la eficacia de cierto medicamento para reducir la presión arterial en personas hipertensas. Para ello se considera como variable respuesta la medida de la presión arterial y como variable explicativa la dosis de medicamento, distinguiendo categóricamente entre dosis nula (placebo), media y alta. De esta forma, el problema consistiría en comparar las medias de las tres dosis mediante un anova de una vía (en el caso de tener únicamente dos dosis quedaría reducido al test de Student para muestras independientes). Sin embargo, se consideró interesante introducir la dieta (distinguiendo entre dos posibilidades, A y B), como nuevo factor explicativo, lo que se traduce en una descomposición de la muestra total en 6 partes, en función de las diferentes combinaciones dieta-medicamento. De hecho, se procedió a distribuir aleatoriamente un total de 100 pacientes entre las dos posibles dietas, por un lado, los tres posibles tratamientos, por otro, resultando seis grupos de entre 14 y 19 individuos cada uno.

En este tipo de estudios resultaría muy ventajoso un diseño equilibrado, es decir, que los seis grupos tuvieran el mismo tamaño, o al menos similar. También sería conveniente que cada grupo tuviera un tamaño de muestra superior a 30, dado que el método que vamos

[4]Para un estudio más detallado consultar, por ejemplo, [12].

a aplicar en nuestro caso, denominado anova de dos factores, es de tipo paramétrico. Ello nos induce a intuir que este tipo de diseño exige un considerable tamaño muestral, máxime si pretendiéramos introducir un tercer factor.

La inclusión del segundo factor suele deberse a una de las siguientes causas, que no son mutuamente excluyentes:

- Porque se desea explicar con mayor precisión la variable respuesta. Para ello se introduce un factor de carácter secundario pero que reducirá el grado de azar en nuestro estudio dado que aumentará el porcentaje de variabilidad total explicada.
- Porque se desea estudiar si dos factores interaccionan entre sí a la hora de explicar la variable respuesta.
- Porque se desea determinar cuál de los dos factores tiene un efecto mayor en la variabilidad de las respuesta.

Ejercicio 111. *Analiza desde un punto de vista crítico el diseño utilizado en el Ejemplo 14.*

Como podemos intuir, y tal y como hemos adelantado en el Capítulo 2, el concepto de coeficiente de correlación múltiple R^2 puede extenderse perfectamente a métodos que, aparentemente, difieren de la regresión lineal múltiple, como es el caso del anova de dos factores y el análisis de la covarianza, interpretándose en todo caso como la proporción de variabilidad total de la respuesta numérica explicada por las variables en juego (ya sean numéricas, cualitativas o mezcla de ambas). En consecuencia, el contraste total de regresión (5.3), a partir del parámetro descriptivo R^2 y el tamaño de muestra n, puede extenderse igualmente a ambos problemas. Un resultado no significativo en este contraste (asociado a un R^2 bajo) indica que las variables explicativas no son apropiadas para la predicción de la respuesta numérica, lo cual pone fin al problema desde el punto de vista inferencial. La situación contraria nos habla de una relación entre las variables explicativas (cualitativas en ese caso) y la respuesta, que deberíamos analizar.

Aditividad - interacción: si es ese el caso, debemos examinar si ambos factores interactúan a la hora de explicar la variable respuesta o, por el contrario, sus posibles efectos se suman sin más. Así pues, un modelo aditivo consistiría en la descomposición de la media μ_{ij} de cada combinación de categorías en una suma tipo (5.6) que desglosamos en la Tabla 5.4

$$\mu_{ij} = \theta + \alpha_i + \beta_j, \qquad \sum_i \alpha_i = \sum_j \beta_j = 0. \tag{5.6}$$

2×3	`Dosis de medicamento`		
	Placebo	Media	Alta
`Tipo de dieta`			
Dieta A	$\theta + \alpha_1 + \beta_1$	$\theta + \alpha_1 + \beta_2$	$\theta + \alpha_1 + \beta_3$
Dieta B	$\theta + \alpha_2 + \beta_1$	$\theta + \alpha_2 + \beta_2$	$\theta + \alpha_2 + \beta_3$

Tabla 5.4: Modelo aditivo dosis-dieta.

El significado de cada parámetro es el siguiente:

- Parámetro común: θ se interpreta como la componente de la media común a todas las combinaciones.
- Factor dieta: α_1 y α_2 indican el aumento o disminución respecto a θ atribuible a la dieta en cuestión.
- Factor medicamento: β_1, β_2 y β_3 indican el aumento o disminución respecto a θ atribuible a la dosis en cuestión.

Nótese que, en este modelo, los efectos de la dieta y el medicamento se suman sin más entre sí, de ahí que se denomine modelo aditivo. En tal caso, un resultado significativo en el contrate de la hipótesis inicial $H_0 : \alpha_1 = \alpha_2 = 0$ se traduciría en una relación entre la dieta y la respuesta, que se dará en el sentido que indiquen las medias aritméticas. Igualmente, un resultado significativo en el contraste inicial $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ se traduciría en una relación entre la dosis del medicamento y la respuesta, que se daría, en este caso (al haber más de dos categorías) en el sentido que marcara el test múltiple de Tukey.

Tamaños del efecto: a nivel descriptivo, pueden estimarse los denominados tamaños del efecto o coeficientes η^2-parciales, que son análogos de los respectivos coeficientes de correlaciones parciales (mencionados en el Capítulo 2) y de los cuales dependen, en sentido aproximado, los resultados de los contrastes anteriores. En definitiva, vienen a indicarnos el peso de cada factor en la respuesta.

Medida de la interacción: sin embargo, la aditividad del modelo, entendida según (5.6), no sólo no debe darse por supuesta, sino que puede ser, precisamente, la hipótesis estadística más interesante a contrastar. En efecto, no deberíamos dar por hecho que los efectos de los factores se suman sin más, sino que cabe pensar que ambos puedan interaccionar. El modelo general de análisis de la varianza con interacción se expresa según (5.7) y se desglosa en la Tabla 5.5:

$$\mu_{ij} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \tag{5.7}$$

con la condición

$$\sum_i \alpha_i = 0, \quad \sum_j \beta_j = 0, \quad \sum_j (\alpha\beta)_{ij} = 0, \quad \sum_i (\alpha\beta)_{ij} = 0.$$

2×3	Placebo	Media	Alta
A	$\theta + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$	$\theta + \alpha_1 + \beta_2 + (\alpha\beta)_{12}$	$\theta + \alpha_1 + \beta_3 + (\alpha\beta)_{13}$
B	$\theta + \alpha_2 + \beta_1 + (\alpha\beta)_{21}$	$\theta + \alpha_2 + \beta_2 + (\alpha\beta)_{22}$	$\theta + \alpha_2 + \beta_3 + (\alpha\beta)_{23}$

Tabla 5.5: Modelo general dosis-dieta.

En este caso, el significado del nuevo parámetro es el siguiente:

- $(\alpha\beta)_{ij}$ se interpreta como el aumento o disminución respecto al modelo aditivo que se presenta en la combinación de la categoría i del primer factor y la categoría j del segundo.

Imaginemos que en el ejemplo 14 obtuviéramos que una dosis alta reduce en 5 puntos la presión arterial media respecto al placebo y que, por otra parte, la dieta A reduce 3 puntos la media la presión arterial respecto a la dieta B. En un modelo aditivo cabría esperar que la combinación de dieta A con dosis alta obtuviera una media 8 puntos más baja que la combinación de dieta B con placebo. Si el descenso fuera, por ejemplo, de 12 puntos, estaríamos apreciando una interacción, que en ese caso se denominaría sinergia.

En general, un resultado significativo en el contraste de la hipótesis inicial $H_0 : (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{23} = 0$ se interpreta como la presencia de interacción entre ambos factores, lo cual supondría pasar a un diseño tipo anova de un factor con seis categorías, una para cada posible combinación entre los dos factores iniciales. No obstante, un gráfico en el que se comparan nítidamente las medias aritméticas de todas las combinaciones resulta de enorme utilidad, al menos a nivel descriptivo. Por contra, un resultado no significativo se interpreta como la validez del modelo aditivo visto anteriormente, con todas sus consecuencias.

Ejercicio 112. *Dado que, bajo el supuesto de aditividad, los efectos de ambos factores se suman sin interaccionar entre sí, ¿por qué no estudiarlos en modelos separados y sumar posteriormente sus efectos? ¿Qué ventaja puede aportar el hecho de combinarlos en un modelo aditivo de dos factores?*

Nótese que el contraste total basado en R^2 es el primero que debemos efectuar, pues si éste no resulta significativo todo lo demás sobra. A continuación deberíamos aplicar el contraste de interacción y, según el resultado, decantarnos por un análisis de los gráficos de medias o por un modelo aditivo. En el caso del Ejemplo 14, si ejecutamos una anova de dos factores según se indica en el tutorial de SPSS obtenemos un valor $R^2 = 0.250$, que indica que sólo el 25 % de la variabilidad de la presión sistólica es explicado conjuntamente por la dieta y el fármaco.

Dado que el P-valor del contraste total es inferior a 0.001 concluimos que es significativo, por lo que, extrapolando, podemos hablar de una influencia de la combinación dieta-fármaco en la presión sistólica. El contraste de interacción da como resultado $P = 0.057$, por lo que no hemos detectado una violación significativa de la aditividad entre los factores. En ese caso, podemos estudiar cada uno por separado y sumar sus efectos.

En el caso de la dieta, hemos observado un tamaño del efecto de $\hat{\eta}^2 = 0.014$, asociado a un resultado $P = 0.249$ en el contraste para la dieta. Por lo tanto, no hemos logrado detectar un efecto significativo de la dieta en la presión arterial. Sin embargo, en el caso de la dosis del medicamento, se obtiene $\hat{\eta}^2 = 0.203$, con un $P < 0.001$ en el contraste.

Por lo tanto, sólo hemos detectado una influencia significativa del fármaco. Para saber en qué sentido se da y teniendo en cuenta que contamos con tres dosis distintas, aplicamos el método de Tukey que revela un descenso de la presión media significativamente superior en el caso de la dosis baja del medicamento. En la Figura 5.1 podemos apreciar gráficamente las diferencias entre las seis medias aritméticas. Por otra parte, en el esquema que sigue se resume el procedimiento anterior.

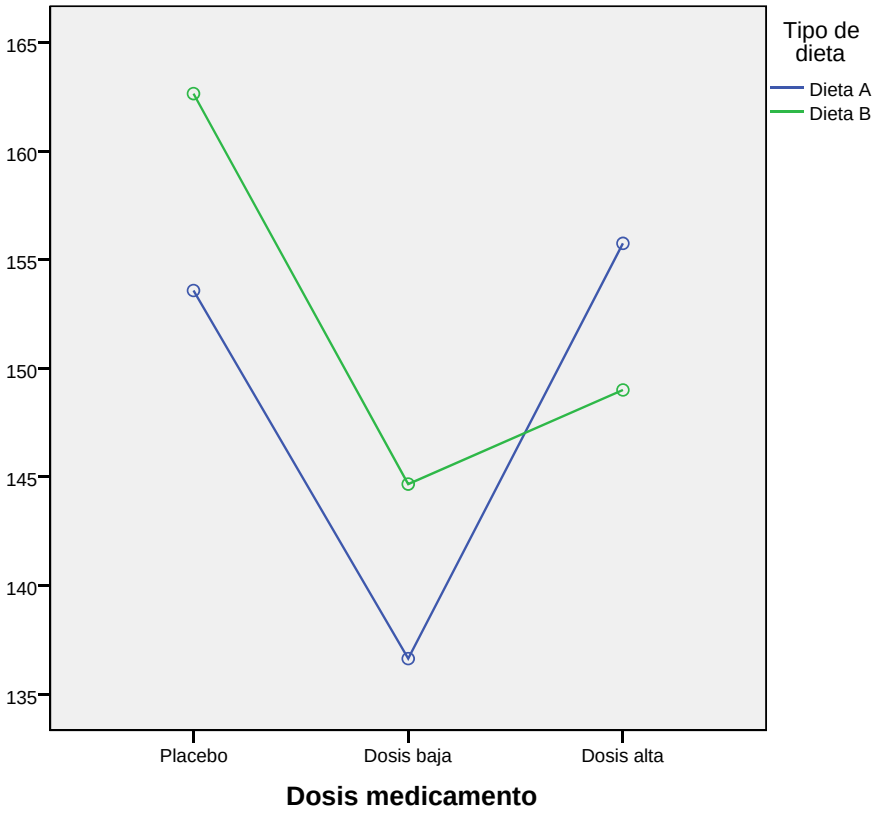
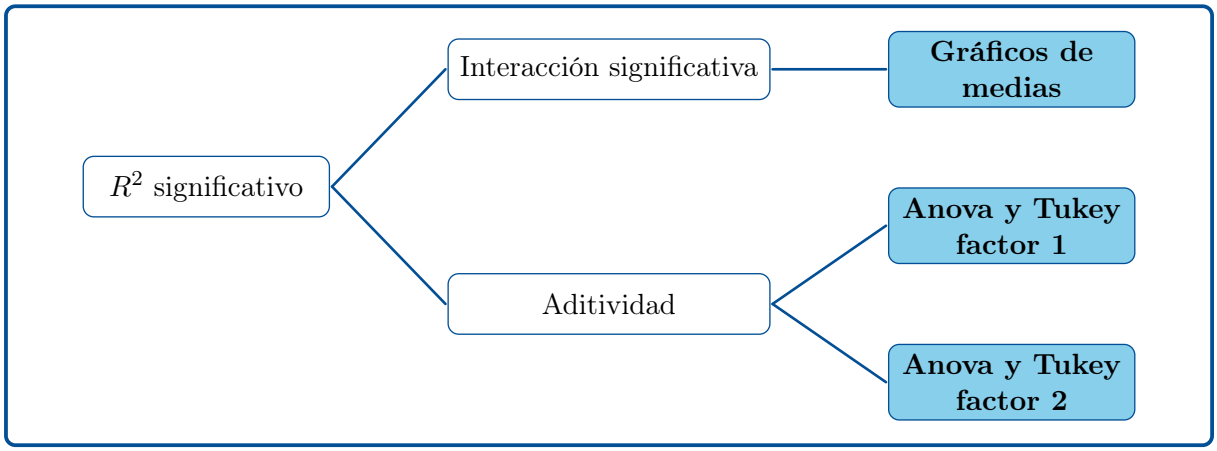


Figura 5.1: Presión sistólica media según dosis del medicamento y dieta.



Ejercicio 113. *A partir de los datos del archivo* `Ensayo clínico.sav` *aplica un anova de dos factores para llegar a las conclusiones anteriormente expuestas.*

5.6.2. Regresión logística binaria

El último problema a estudiar en este manual consiste en intentar determinar si cierto evento se produce o no en función de una serie de variables $X_1, \ldots X_k$. La variable res-

puesta, asociada a la ocurrencia del evento, es por lo tanto cualitativa y binaria. Conviene que esté codificada de manera que se asigne 1 a la ocurrencia del evento y 0 a lo contrario, aunque no es estrictamente necesario.

El modelo de regresión logística produce como respuesta un número que debe entenderse como la probabilidad de que el evento ocurra, dados los valores concretos de las variables explicativas. Así pues, teniendo en cuenta que dicha respuesta debe restringirse al intervalo $[0, 1]$, no cabe pensar en una ecuación del tipo (5.2). De hecho la ecuación que propone el modelo es una composición entre (5.2) y la función logística $f(x) = (1+\mathsf{e}^{-x})^{-1}$, es decir, una ecuación del tipo

$$P(\mathsf{evento}) \simeq \frac{1}{1+\mathsf{e}^{-(\beta_0+\beta_1 X_1+\beta_k X_k)}}. \tag{5.8}$$

Desde el punto de vista numérico el problema se reduce, al igual que en el caso de la regresión lineal múltiple, a encontrar estimadores $B_0, B_1, \ldots B_k$, de los coeficientes anteriores e insertarlos en la ecuación (5.8). Sin embargo, el método de cálculo en este caso es más aparatoso y no se basa en el criterio de Mínimos Cuadrados.

Ejercicio 114. *¿Por qué una variable respuesta Y en el intervalo $[0, 1]$ no puede obedecer a un modelo tipo (5.2)?*

Ejemplo 15. Tras un seguimiento de 15 años de $n = 462$ adultos sudafricanos intentamos determinar la ocurrencia o no de un infarto de miorcardio mediante un modelo de regresión logística, a partir de los valores de presión sistólica, consumo de tabaco, colesterol ldl, antecedentes familiares y nivel de obesidad al inicio del estudio.

Coeficientes del modelo: el modelo se ejecuta según se indica en el Capítulo 7. Las estimaciones de los coeficientes de regresión se encuentran en la segunda columna por la izquierda en la Figura 7.59. Podemos observar que la segunda columna por la derecha ofrece los resultados de los contrates parciales, como si se tratara de un problema de regresión múltiple. De hecho, estamos también en condiciones de aplicar diferentes algoritmos de selección de variables a partir de ellos. La ecuación resultante es la que se utiliza para estimar las probabilidades de que ocurra el evento (infarto en el caso del Ejemplo 15).

Odds Ratios: nótese que el modelo admite variables explicativas tanto numéricas como cualitativas. Cuando tratamos con una variable cualitativa binaria, como el caso de los antecedentes familiares, puede probarse que el valor que aporta la primera columna de la derecha, e^{B_j}, coincide con el Odds Ratio asociado a dicha variable, de manera que un valor en torno a 1 indica una escasa influencia de la variable en la respuesta. En el caso del Ejemplo 15 se ha obtenido un Odds Ratio de 2.884, con un P-valor menor que 0.001 en el contraste parcial, lo cual nos indica que los antecedentes familiares de infarto incrementan fuerte y significativamente el riesgo de infarto.

Coeficiente R^2 de Nagelkerke y tabla de clasificación: el modelo proporciona una medida de su fiabilidad denominado coeficiente de Nagelkerke, que es un sucedáneo del

coeficiente R^2 de regresión lineal múltiple y como tal debe interpretarse. En el Ejemplo 15 obtenemos como resultado $R^2 = 0.255$, lo cual nos habla de una escasa capacidad de explicar el infarto por parte de las variables consideradas. No obstante, existe otro procedimiento más claro para determinar la capacidad predictiva del modelo, que consiste en construir una tabla de contingencia, denominada tabla de clasificación, donde se indica, por un lado, qué individuos poseen una probabilidad de sufrir un infarto superior al 50 % según el modelo y, por otro, qué individuos sufrieron realmente un infarto durante el seguimiento. La conclusión es que el modelo reconoce correctamente al 87.1 % de los sanos y al 46.9 % de los enfermos. Estos dos datos pueden interpretarse pues en términos de especificidad y sensibilidad, respectivamente.

Otras cuestiones propuestas

En la siguiente lista de problemas se hace referencia a una serie de archivos que pueden encontrarse en diversos repositorios de datos, aunque también pueden descargarse directamente en formato SPSS desde la dirección `https://matematicas.unex.es/~jmf/`. Los problemas pueden resolverse en principio con cualquier programa estadístico. Si se decide hacer uso del SPSS puede resultar de utilidad el tutorial incluido en la tercera parte de este manual. En todos los problemas propuestos se supone que la muestra estudiada es representativa y se pretende generalizar las conclusiones a la población de la que procede.

Ejercicio 115. *A partir de los datos del archivo* `Tumor de próstata.sav`*:*

(a) Intenta explicar el volumen (log) del tumor a partir de la concentración de PSA (log) y la edad del paciente.

(b) Relaciona el PSA (log) con el pronóstico del tumor según la biopsia.

(c) Relaciona el volumen (log) del tumor con el porcentaje de Gleason 4-5.

(d) Relaciona el peso (log) del tumor con el porcentaje de Gleason 4-5. ¿Guarda más relación que el volumen?

Ejercicio 116. *A partir de los datos del archivo* `Diabetes Schorling.sav`*:*

(a) Relaciona la presencia de la diabetes con la presión sistólica (sbp).

(b) Relaciona la presencia de la diabetes con el nivel de colesterol HDL.

(c) Relaciona la presencia de la diabetes con el sexo.

(d) Relaciona la presencia de la diabetes con la complexión.

(e) Relaciona la concentración de hemoglobina glicosilada con la complexión.

(f) Relaciona la presión sistólica con la diastólica.

(g) Relaciona la glucemia con la hemoglobina glicosilada.

(h) *Selecciona los 30 primeros individuos del archivo y responde de nuevo a acada una de las preguntas anteriores.*

(i) *Intenta explicar el nivel de hemoglobina glicosilada a través de los factores sexo y complexión, considerados conjuntamente.*

Ejercicio 117. *A partir de los datos del archivo* `South Africa Heart Disease.sav`*:*

(a) *Relaciona la presencia de la enfermedad (chd) con la presión sistólica (sbp).*

(b) *Relaciona la presencia de la enfermedad con el nivel de colesterol (ldl).*

(c) *Relaciona la presencia de la enfermedad con el porcentaje de grasa corporal (adiposity).*

(d) *Relaciona la presencia de la enfermedad con el consumo de alcohol.*

(e) *Relaciona la presencia de la enfermedad con la edad. ¿Cuál de todas las variables mencionadas crees que guarda mayor relación con la enfermedad cardiaca?*

(f) *Relaciona la presencia de la enfermedad con los antecedentes familiares.*

(g) *Intenta explicar la presión sistólica a partir de la edad, el porcentaje de grasa corporal y el nivel de colesterol ldl.*

(h) *Selecciona los 30 primeros individuos del archivo y responde de nuevo a cada una de las preguntas anteriores.*

(i) *Intenta explicar la presencia de la enfermedad a partir del nivel de colesterol ldl, la edad, la presencia de antecedentes familiares, la presión sistólica, y puntuación en personalidad tipo A.*

Ejercicio 118. *En un estudio realizado en 68.183 mujeres adultas seguidas a lo largo de 16 años, aquellas que dormían 5 o menos horas no solo pesaban 2,5 kg más al inicio del estudio, sino que también ganaron una media de 4,3 kg más en comparación con las que dormían 7 o más horas. Además, las mujeres con 5 o menos horas de sueño tuvieron un 32 % más de posibilidades de ganar hasta 15 kg que las que dormían 7 o más horas a lo largo del estudio. Esta diferencia persistía tras ajustar los resultados según la ingesta calórica y la actividad física. Otros estudios muestran resultados similares también en los hombres. Se observó también que tanto el índice de masa corporal como el perímetro de cintura es significativamente mayor entre aquellos que duermen menos de 5 horas. En concreto, dormir menos se asocia con un aumento del perímetro de la cintura de 6,7 cm para los hombres y de 5,4 cm para las mujeres.*

¿Qué técnicas estadísticas (regresión lineal, test de Student, Wilcoxon, cálculos de medidas de riesgo, etc) crees que se han utilizado para llegar a estas conclusiones?

PARTE

III

Tutorial de SPSS

6. Estadística Descriptiva con SPSS

A continuación describiremos brevemente, mediante capturas de pantalla y resultados, cómo pueden ejecutarse la mayoría de los métodos explicados en las dos partes anteriores mediante un programa estadístico. Como ya comentamos en el Prólogo, nos hemos decantado en este caso por el programa SPSS, concretamente por la versión 22, el cual ha proporcionado todos los gráficos recogidos en la primera parte del manual. No existe realmente una razón de peso para elegir éste programa en lugar de otros, como por ejemplo R, a través de su paquete `Rcommander`, de similar manejo y disponible gratuitamente en la dirección `https://www.r-project.org/`. Recordamos que, además de en diversos repositorios de datos, podemos encontrar los archivos que usaremos en la dirección `https://matematicas.unex.es/~jmf/`.

6.1. Algunos aspectos generales

En la primera sección de este capítulo indicamos algunas funciones básicas del programa para pasar después, en el resto del capítulo y en el siguiente, al análisis de los datos desde un punto de vista descriptivo e inferencial, respectivamente. Para nuestro propósito serán de especial interés el menú Analizar y el menú Gráficos (Figura 6.1). Los menús Datos y Transformar contienen algunas opciones que serán de utilidad para la manipulación de datos (filas) y de variables (columnas), respectivamente, y se explican a continuación.

6.1.1. Datos y variables

El editor de datos dispone de dos tipos de vistas distintos: Vista de datos y Vista de variables. La Vista de datos está diseñada de manera que las variables se sitúan en las columnas y los elementos muestrales en las filas y es la vista que aparece por defecto en el editor de datos.

La opción Vista de variables muestra en la parte superior del área de datos propiedades predeterminadas por el programa, como son Nombre, Tipo, Anchura, Decimales, Etiqueta, Valores, Perdidos, Columna, Alineación, Medida y Rol. De esta información, serán de utilidad para el posterior análisis las siguientes:

- **Nombre**: nombre abreviado de la variable.

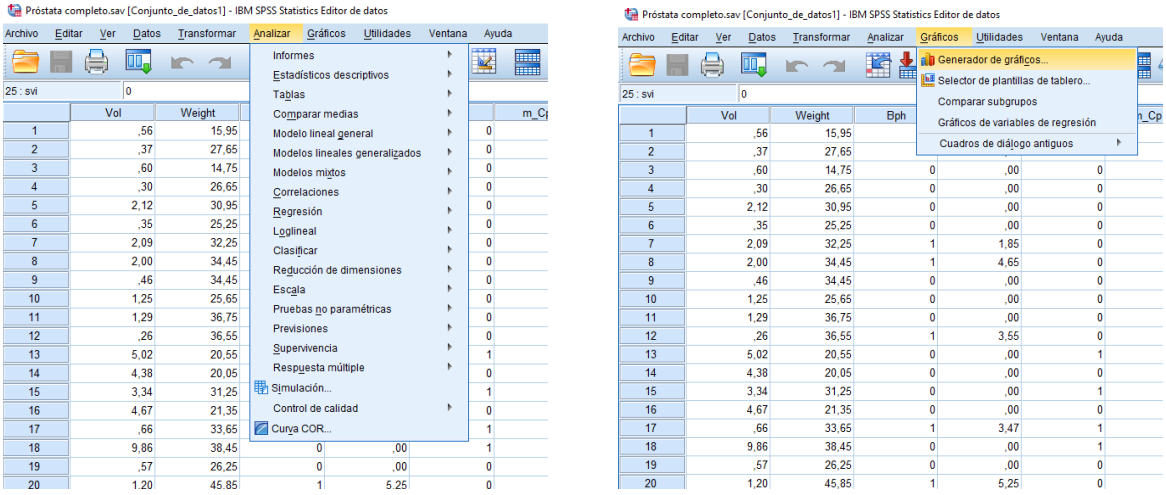
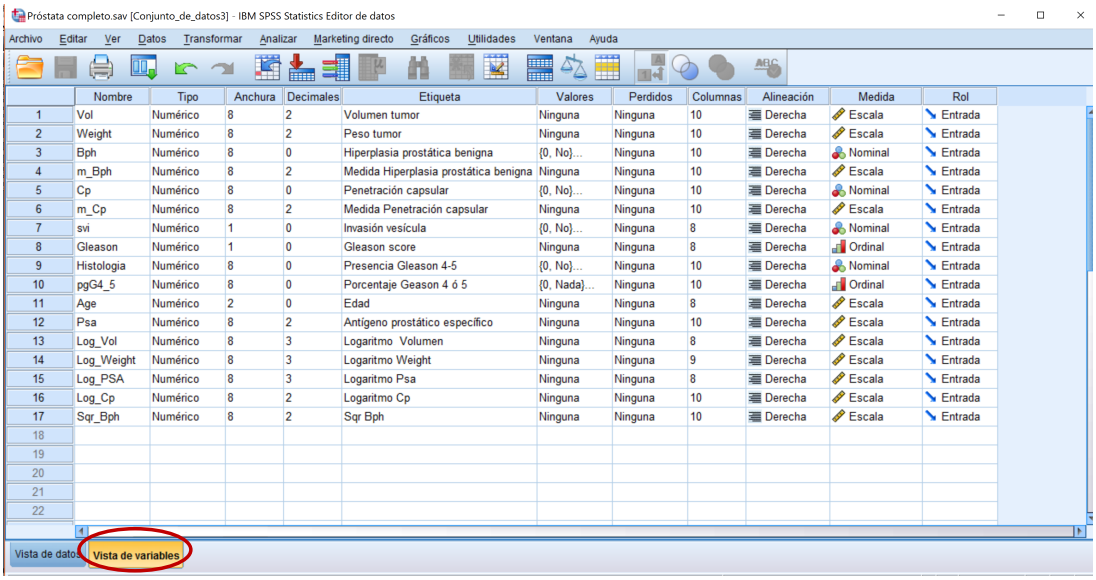


Figura 6.1: Menús Analizar y Gráficos.

- **Etiqueta**: descripción o nombre extendido de la variable.
- **Valores**: en el caso de variables cualitativas, es de interés para conocer a qué categoría corresponde cada valor.
- **Medida**: tipo de variable. El programa distingue tres tipos: escala, nominal y ordinal.



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Vol	Numérico	8	2	Volumen tumor	Ninguna	Ninguna	10	Derecha	Escala	Entrada
2	Weight	Numérico	8	2	Peso tumor	Ninguna	Ninguna	10	Derecha	Escala	Entrada
3	Bph	Numérico	8	0	Hiperplasia prostática benigna	{0, No}...	Ninguna	10	Derecha	Nominal	Entrada
4	m_Bph	Numérico	8	2	Medida Hiperplasia prostática benigna	Ninguna	Ninguna	10	Derecha	Escala	Entrada
5	Cp	Numérico	8	0	Penetración capsular	{0, No}...	Ninguna	10	Derecha	Nominal	Entrada
6	m_Cp	Numérico	8	2	Medida Penetración capsular	Ninguna	Ninguna	10	Derecha	Escala	Entrada
7	svi	Numérico	1	0	Invasión vesícula	{0, No}...	Ninguna	8	Derecha	Nominal	Entrada
8	Gleason	Numérico	1	0	Gleason score	Ninguna	Ninguna	8	Derecha	Ordinal	Entrada
9	Histologia	Numérico	8	0	Presencia Gleason 4-5	{0, No}...	Ninguna	10	Derecha	Nominal	Entrada
10	pgG4_5	Numérico	8	0	Porcentaje Geason 4 ó 5	{0, Nada}...	Ninguna	10	Derecha	Ordinal	Entrada
11	Age	Numérico	2	0	Edad	Ninguna	Ninguna	8	Derecha	Escala	Entrada
12	Psa	Numérico	8	2	Antígeno prostático específico	Ninguna	Ninguna	10	Derecha	Escala	Entrada
13	Log_Vol	Numérico	8	3	Logaritmo Volumen	Ninguna	Ninguna	8	Derecha	Escala	Entrada
14	Log_Weight	Numérico	8	3	Logaritmo Weight	Ninguna	Ninguna	9	Derecha	Escala	Entrada
15	Log_PSA	Numérico	8	3	Logaritmo Psa	Ninguna	Ninguna	8	Derecha	Escala	Entrada
16	Log_Cp	Numérico	8	2	Logaritmo Cp	Ninguna	Ninguna	10	Derecha	Escala	Entrada
17	Sqr_Bph	Numérico	8	2	Sqr Bph	Ninguna	Ninguna	10	Derecha	Escala	Entrada
18											
19											
20											
21											
22											

Figura 6.2: Vista de variables.

En la mayoría de las ocasiones, es más efectivo mostrar las etiquetas de valor de la variable en lugar del valor de datos (Figura 6.3).

Ver - Etiquetas de valor

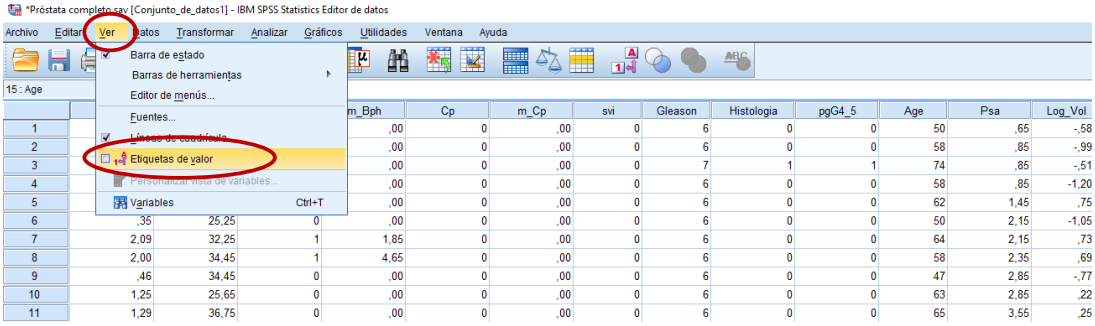


Figura 6.3: Mostrar las etiquetas de valor.

6.1.2. Cálculo de nuevas variables

Veamos cómo calcular una nueva variable a partir de otras variables ya definidas. Por ejemplo, en el archivo `Tumor de prostata.sav`, podemos calcular el logaritmo de la variable `PSA`:

- Abrimos el menú Calcular variable (Figura 6.4).

  Transformar - Calcular variable

- Escribimos el nombre de la variable que vamos a crear en el cuadro Variable de destino y la operación para calcular la nueva variable en Expresión numérica (Figura 6.5).

El menú también ofrece una lista de las funciones más comunes.

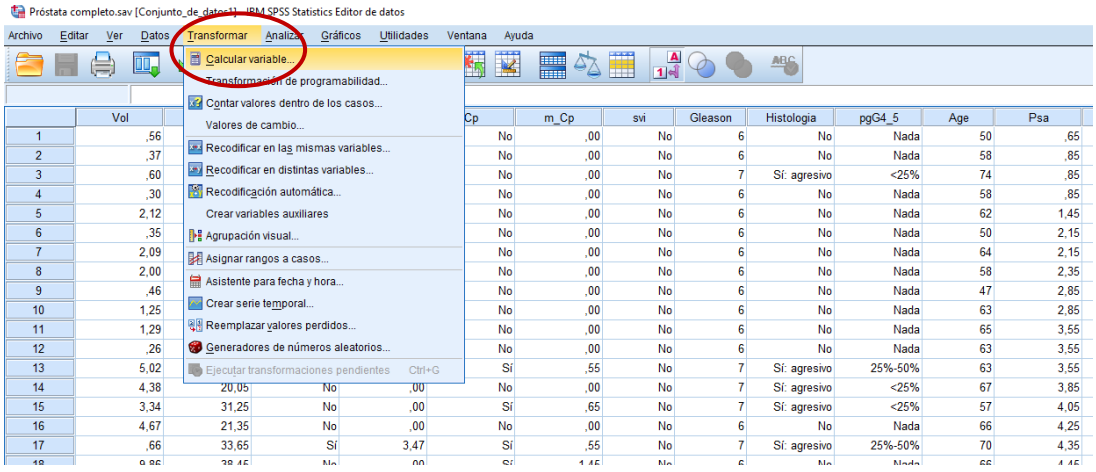


Figura 6.4: Cálculo de una nueva variable a partir de las ya registradas.

6.1.3. Selección de datos

Veamos cómo seleccionar un subconjunto específico de datos. Por ejemplo, en el archivo `Tumor de próstata.sav`, podemos seleccionar únicamente los pacientes con tumores agresivos:

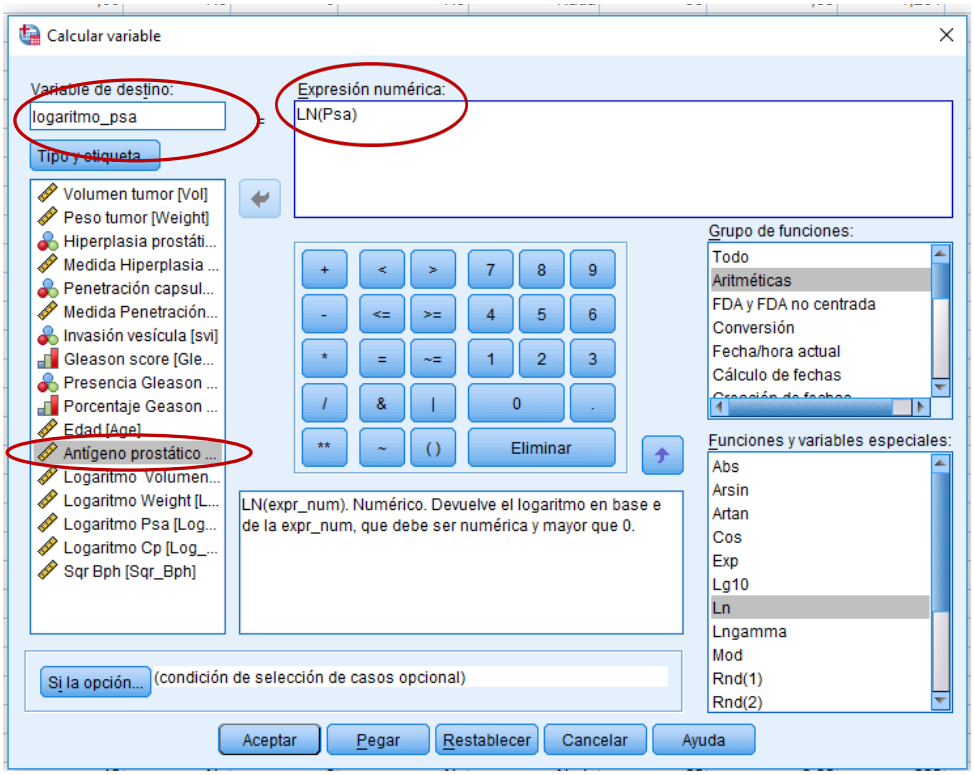


Figura 6.5: Cálculo de una nueva variable a partir de las ya registradas.

- Comprobamos qué valores de la variable `Histologia` corresponden a pacientes con tumores agresivos (Figura 6.6). En este caso es 1.

- Abrimos el menú Seleccionar casos (Figura 6.7).

Datos - Seleccionar casos

- Puesto que en este caso seleccionamos los pacientes que cumplen una determinada característica, marcamos la opción Si se satisface la condición e introducimos la condición en el cuadrado del menú Si la op... (Figura 6.8). En este caso sería `Histologia = 1`.

Como resultado obtendremos una nueva columna llamada `filter_$`, indicando los pacientes seleccionados.

Al igual que en el caso anterior, el menú también ofrece una lista de las variables del archivo y de las funciones más comunes.

Además de seleccionar datos utilizando una condición, existen otras opciones. Es importante recordar que, una vez finalizado el análisis con la selección de los datos, debemos borrar el filtro creado eliminando la nueva columna.

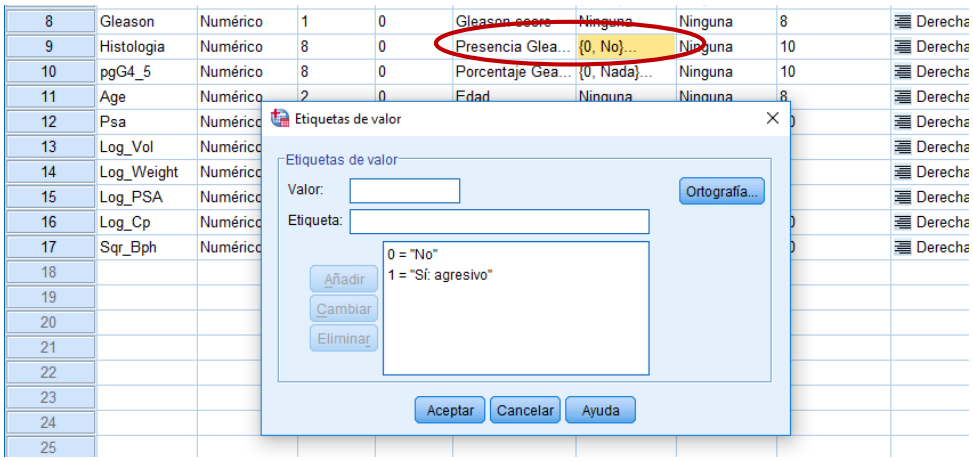


Figura 6.6: Selección de datos.

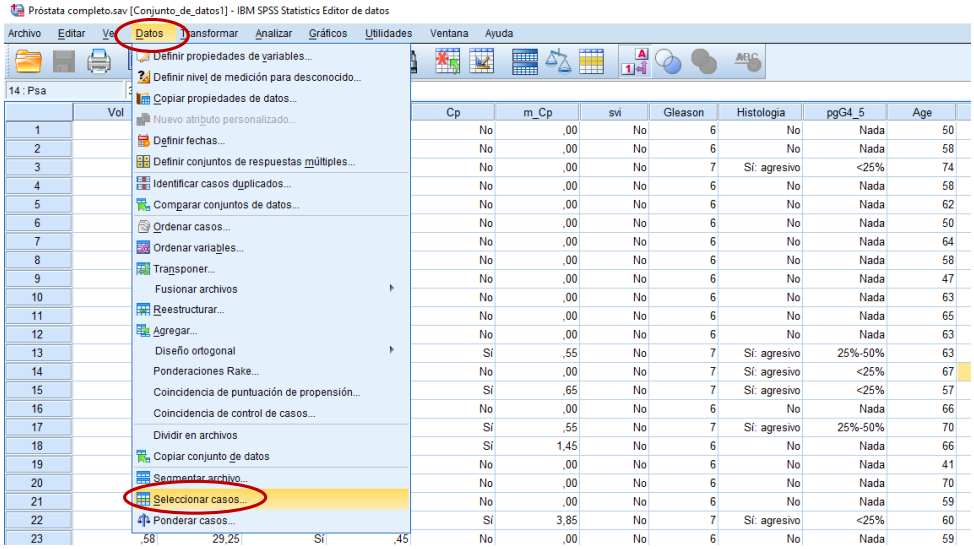


Figura 6.7: Selección de datos.

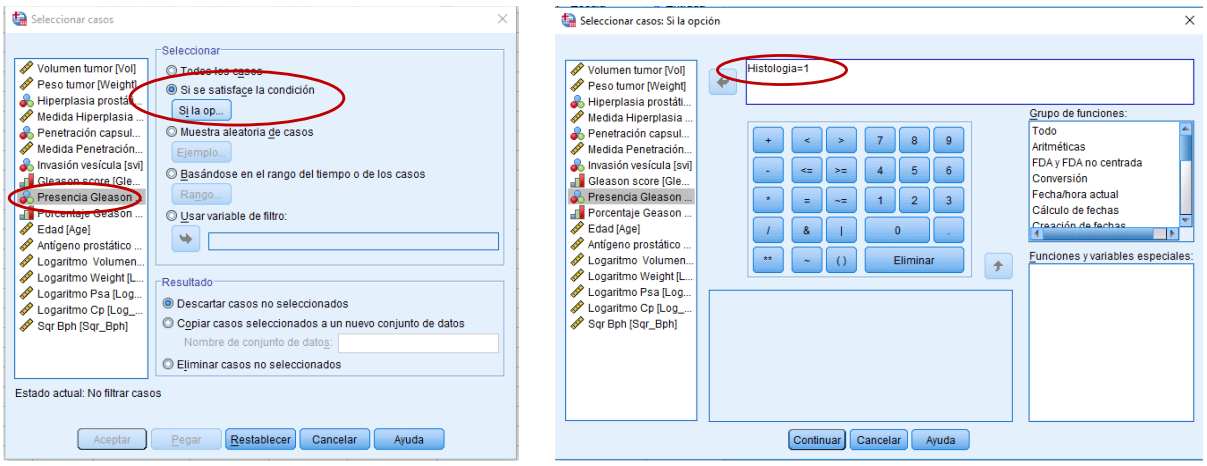


Figura 6.8: Selección de datos.

6.2. Análisis descriptivo de una variable

Es la primera fase del estudio estadístico y sus conclusiones se restringen a la muestra considerada. Empezaremos con el estudio descriptivo de variables de manera aislada. Las siguientes secciones están dedicadas al estudio descriptivo de relación entre diferentes dos variables variables.

6.2.1. Variable cualitativa

Las distintas herramientas para el estudio descriptivo de una variable cualitativa se encuentran en el menú Frecuencias de Estadísticos descriptivos (Figura 6.9).

Analizar - Estadísticos descriptivos - Frecuencias

Veamos cómo describir una variable cualitativa, por ejemplo, la variable `Estado` del archivo `ICC.sav`.

- **Tabla de frecuencias**: seleccionamos la variable `Estado` de la lista de variables y nos aseguramos de que tenemos marcada la opción Mostrar tabla de frecuencias (Figura 6.10).
- **Diagramas de barras o de sectores**: en la opción Gráficos, podemos elegir entre representar Gráficos circulares (sectores) o Gráficos de barras (Figura 6.10). Recordemos que este último gráfico es más conveniente si la variable presenta muchas categorías o si éstas pueden ordenarse de manera natural. La opción Valores del gráfico nos permite elegir si mostrar las frecuencias absolutas o las frecuencias porcentuales para el gráfico de barras.

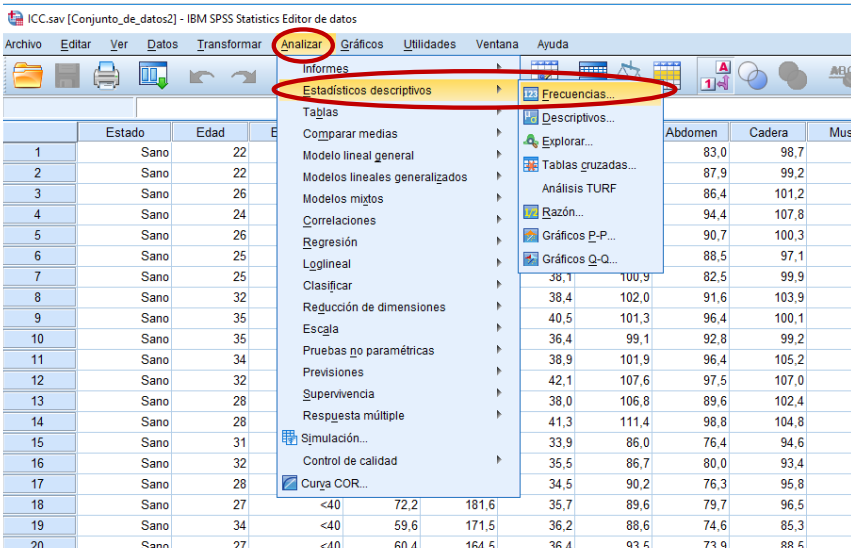


Figura 6.9: Análisis descriptivo de una variable cualitativa.

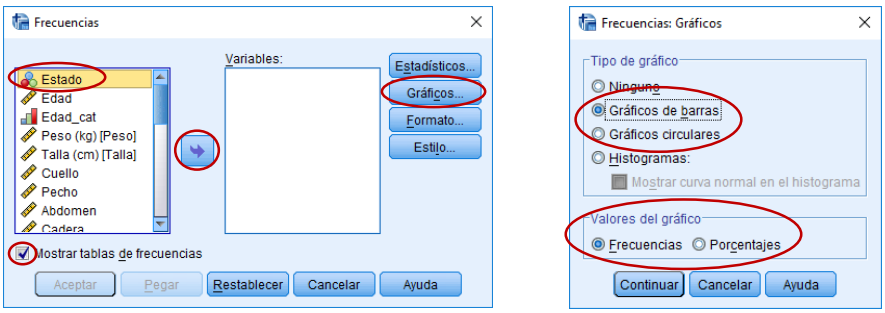


Figura 6.10: Análisis descriptivo de una variable cualitativa.

6.2.2. Variable cuantitativa

La mayoría de opciones disponibles para el estudio descriptivo de una variable cuantitativa se encuentran en el menú Explorar de Estadísticos descriptivos (Figura 6.11).

Analizar - Estadísticos descriptivos - Explorar

Veamos cómo describir una variable cuantitativa, por ejemplo, la variable `adiposidad` del archivo `Southafrica Heart Disease.sav`.

- **Valores típicos o medidas resumen**: seleccionamos la variable `adiposity` de la lista de variables y la introducimos en la Lista de dependientes (Figura 6.12). En el menú de Estadísticos nos aseguramos de que tenemos marcada la opción Descriptivos. Para calcular los cuantiles, marcamos la opción Percentiles (Figura 6.13).
- **Diagramas de caja**: se proporcionan por defecto.
- **Histogramas y diagramas de tallo-hoja**: en la opción Gráficos, podemos elegir representar Gráficos de tallo y hoja o Histogramas (Figura 6.13).

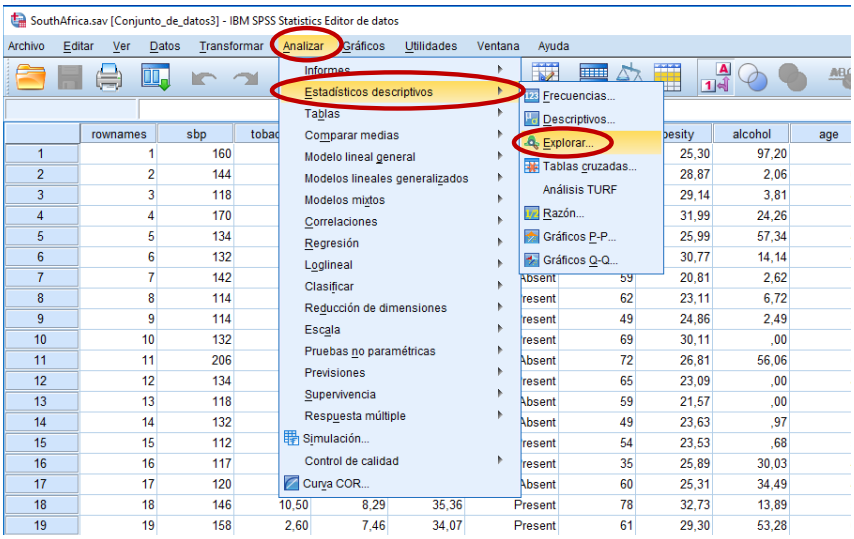


Figura 6.11: Análisis descriptivo de una variable cuantitativa.

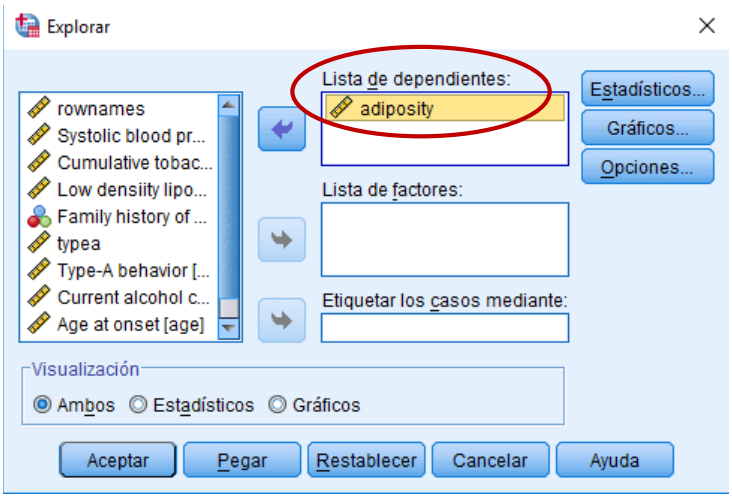


Figura 6.12: Análisis descriptivo de una variable cuantitativa.

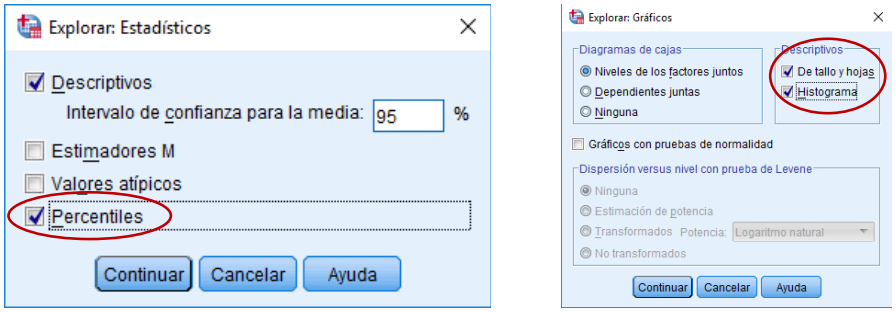


Figura 6.13: Análisis descriptivo de una variable cuantitativa.

A cualquiera de las opciones más básicas de este menú se puede acceder igualmente a través del menú Generador de gráficos de Gráficos. Por ejemplo, podemos solicitar directamente un histograma, arrastrando con el ratón la variable deseada al eje OX. Además, entre otras cosas, podemos representar la curva de una distribución normal que mejor se ajusta a nuestros datos sobre el histograma (Figuras 6.14 y 6.15).

Gráficos - Generador de gráficos - Histograma - Propiedades de elemento - Mostrar curva normal

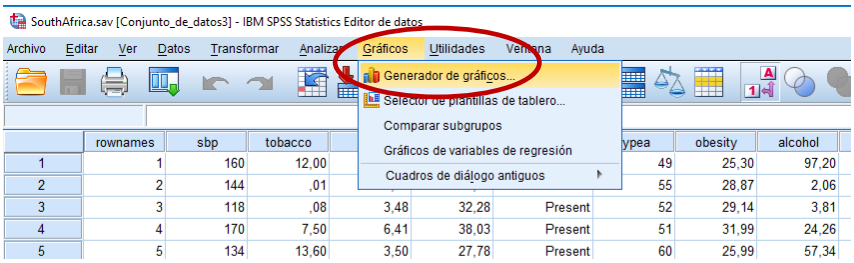


Figura 6.14: Mostrar una curva normal sobre el histograma.

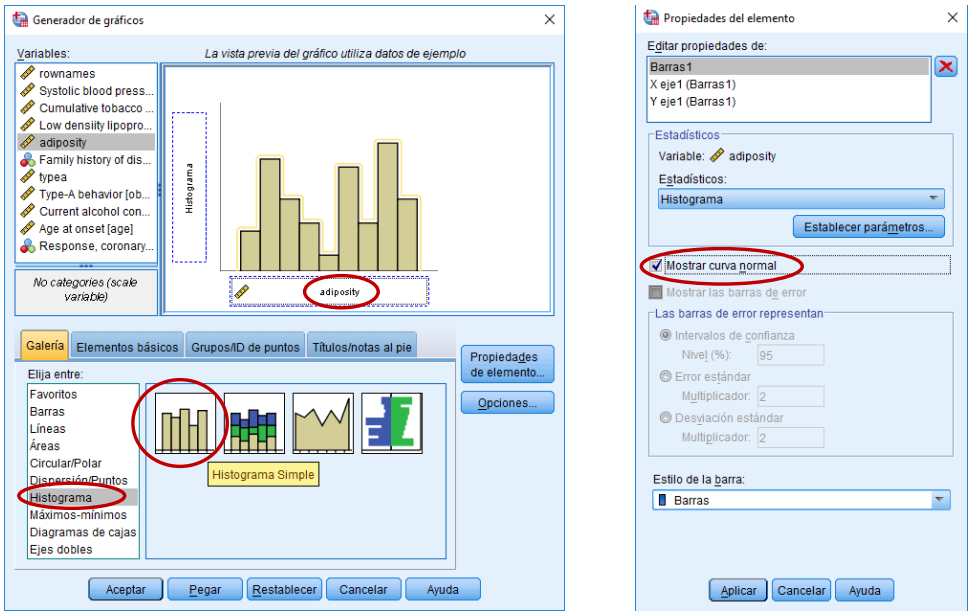


Figura 6.15: Mostrar una curva normal sobre el histograma.

El programa no respeta la fórmula de Sturges, aunque proporciona un número de intervalos adecuado a cada situación. Si deseásemos modificar el número de intervalos en el histograma, el proceso sería el siguiente (Figuras 6.14, 6.16 y 6.17)

Gráficos - Generador de gráficos - Histograma - Propiedades de elemento - Establecer parámetros - Tamaño de agrupaciones - Personalizado - Número de intervalos

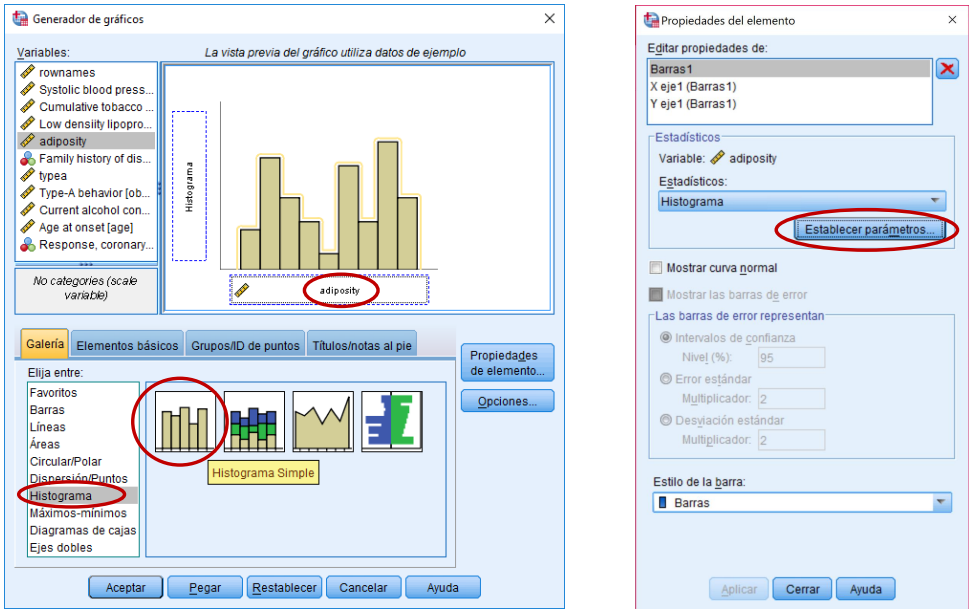


Figura 6.16: Modificar el número de intervalos en un histograma.

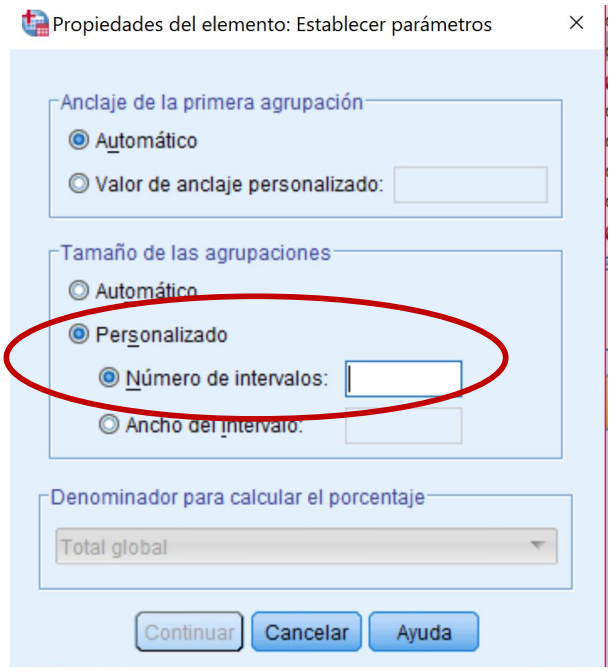


Figura 6.17: Modificar el número de intervalos en un histograma.

Además de utilizar el menú Explorar podemos obtener una tabla de valores típicos más elaborada a través de Tablas personalizadas del menú Tablas (Figura 6.18). Para ello, arrastramos la variable elegida (puede ser más de una) al rectángulo Filas (Figura 6.19) y en Estadísticos de resumen (Figura 6.20) elegimos los valores típicos que vamos a utilizar. Es importante recordar que entre las opciones no aparece el rango intercuartílico por lo que una alternativa es proporcionar el primer y el tercer cuartil.

Analizar - Tablas - Tablas personalizadas

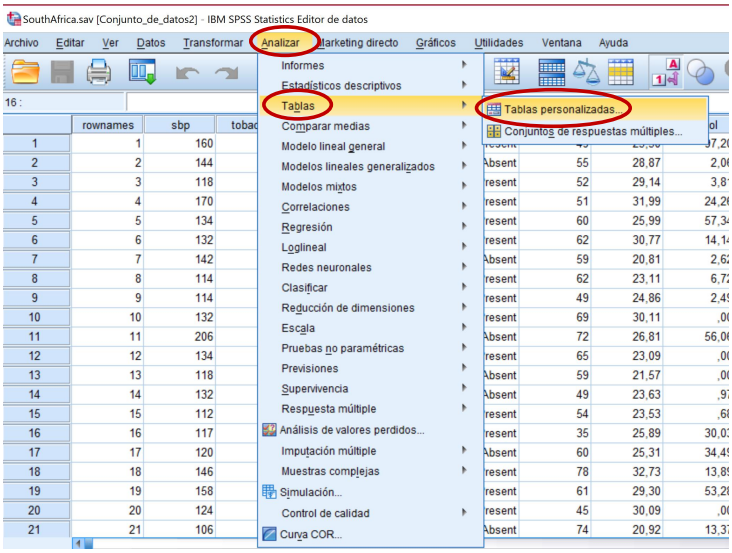


Figura 6.18: Resumir la información de una variable cuantitativa.

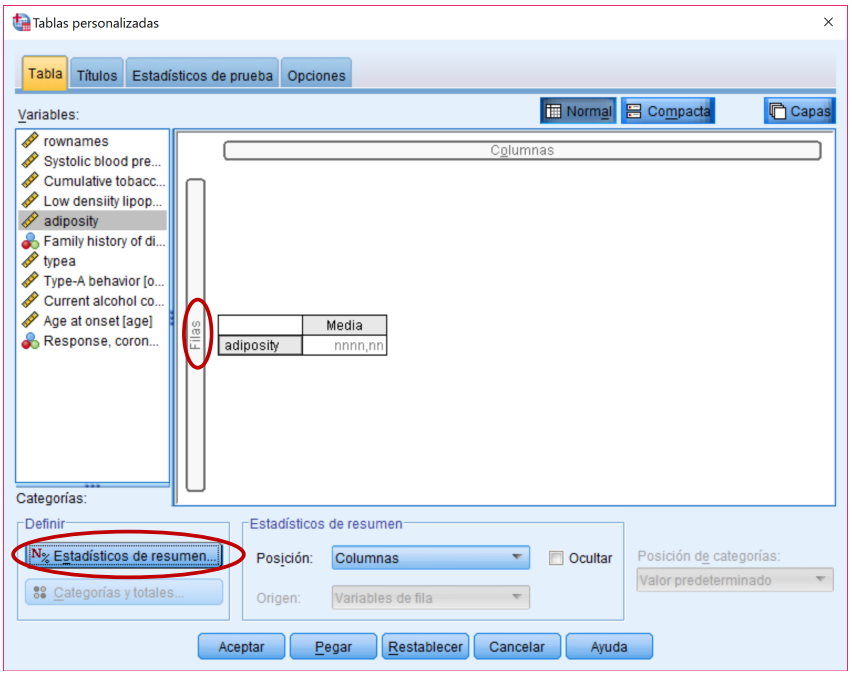


Figura 6.19: Resumir la información de una variable cuantitativa.

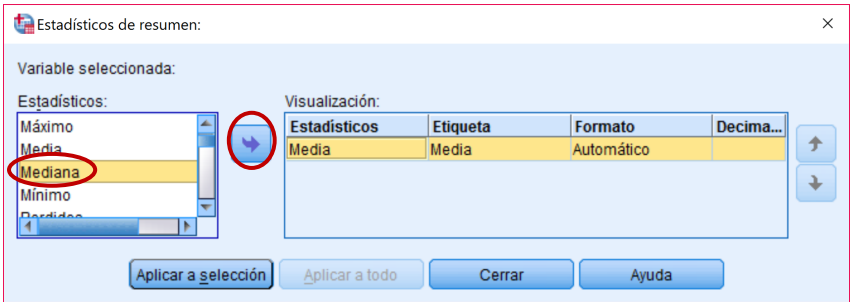


Figura 6.20: Resumir la información de una variable cuantitativa.

6.3. Relación entre dos variables cuantitativas

6.3.1. Problemas de correlación

Para analizar la relación entre dos variables cuantitativas utilizaremos los diagramas de dispersión y medidas para el grado de relación entre ambas variables.

Por ejemplo, en el archivo `Ecografia.sav`, analicemos la relación entre las variables `Peso` y `LF` (longitud del fémur).

- **Diagrama de dispersión**: a través del generador de gráficos.

  Gráficos - Generador de gráficos - Dispersión/Puntos

  Arrastramos la primera opción de la Galería e incorporamos cada variable cuantitativa a uno de los ejes del gráfico. En este caso, puesto que sólo analizamos la relación

entre dos variables, no importa cuál situamos en el eje OX y cuál situamos en el eje OY (Figuras 6.21 y 6.22).

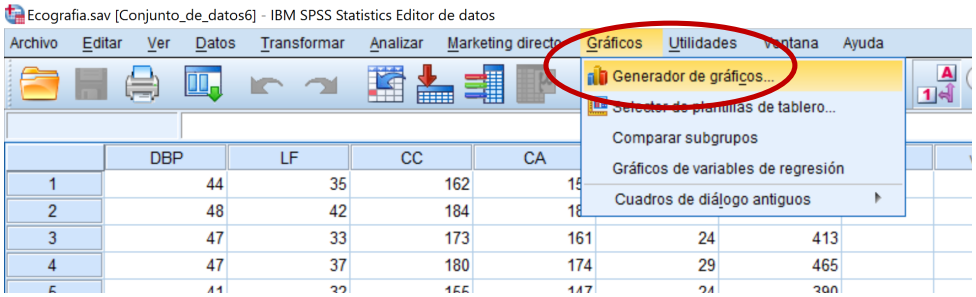


Figura 6.21: Análisis descriptivo de la relación entre variables cuantitativas.

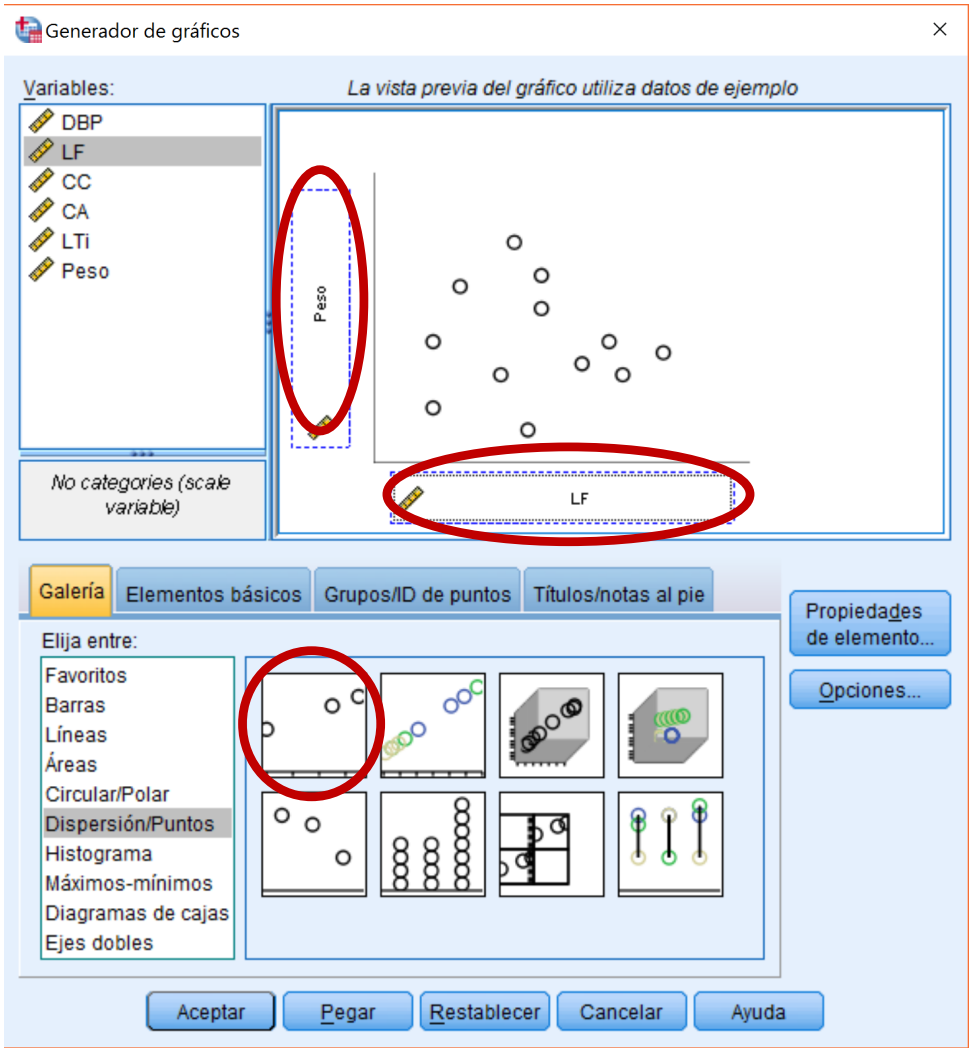


Figura 6.22: Análisis descriptivo de la relación entre variables cuantitativas.

Se puede incorporar la recta de regresión lineal haciendo doble click en el gráfico

resultante y a continuación en el icono que indica la Figura 6.23. Dicha recta aparecerá acompañada por la correspondiente ecuación para las versiones 22 o superiores del SPSS.

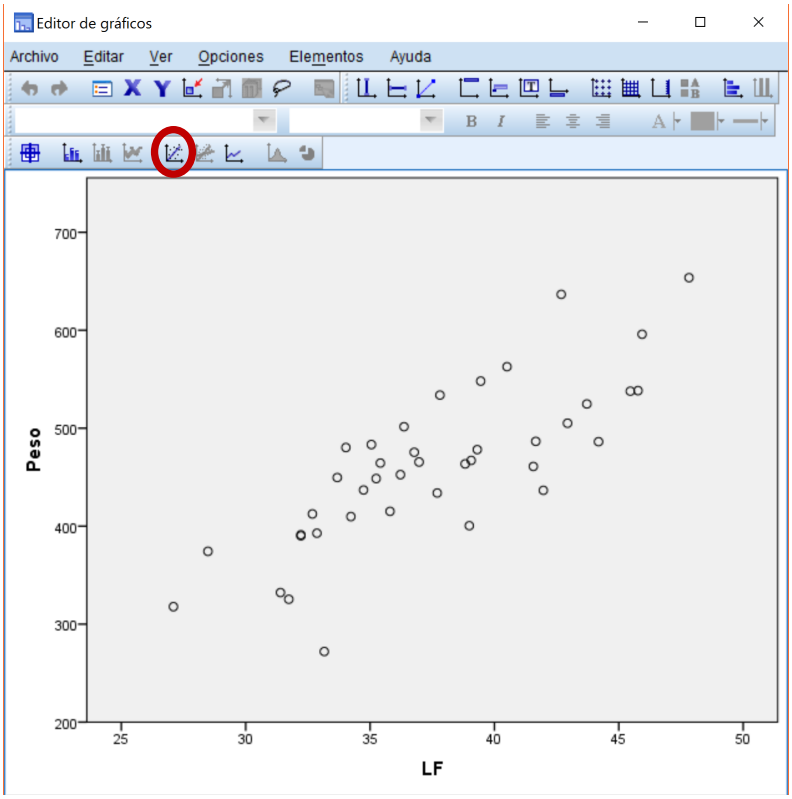


Figura 6.23: Análisis descriptivo de la relación entre variables cuantitativas.

- **Diagrama de dispersión por categorías**: podemos obtener un diagrama de dispersión propio del análisis de la covarianza en el que aparezcan de diferentes colores los puntos correspondientes a distintas categorías de una variable cualitativa, como en el caso de la Figura 2.17. Para ello, elegimos el gráfico de colores a la derecha del gráfico de dispersión simple de la Figura 6.22, y especificamos en la opción Establecer color la variable cualitativa. En el gráfico obtenido se puede trazar tanto la recta de regresión lineal total, como ya sabemos, como las rectas correspondientes a cada categoría, para lo cual debemos hacer click en línea de ajuste por subgrupos.

- **Coeficiente de correlación r**: para calcular la matriz de correlaciones accedemos al menú de Correlaciones Bivariadas (Figura 6.24).

Analizar - Correlaciones - Bivariadas

Añadimos las variables (dos o más) en las que estamos interesados al cuadro Variables (Figura 6.25).

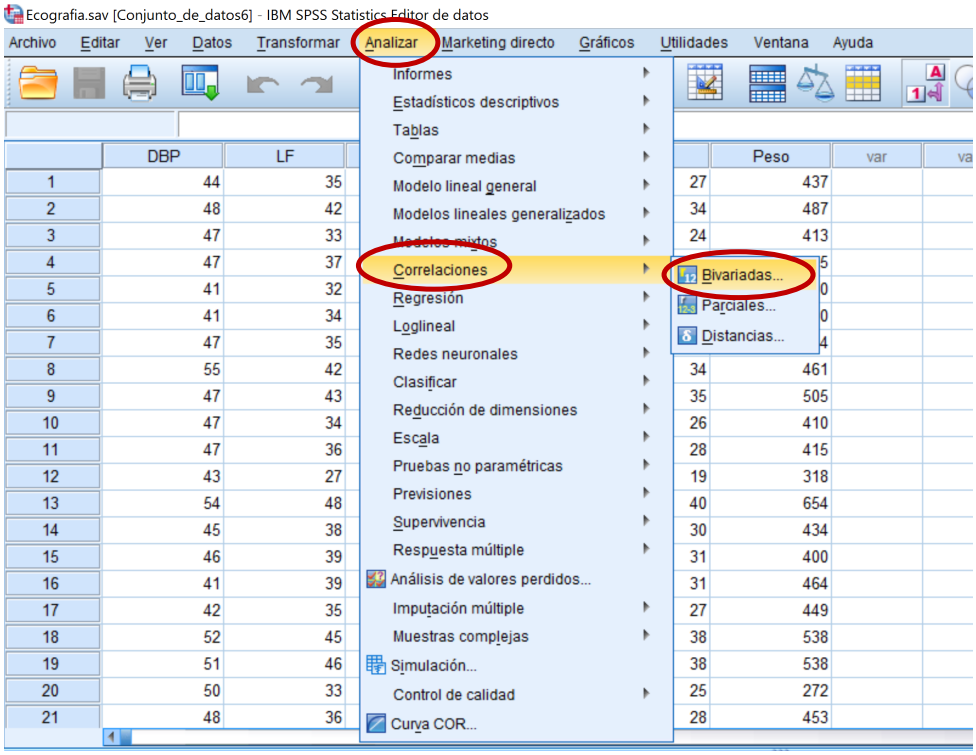


Figura 6.24: Análisis descriptivo de la relación entre variables cuantitativas.

Figura 6.25: Análisis descriptivo de la relación entre variables cuantitativas.

- **Coeficiente de determinación lineal muestral**: se puede calcular elevando al cuadrado el coeficiente de correlación lineal muestral. El cálculo directo se indica en el siguiente apartado.

6.3.2. Problemas de regresión

Regresión lineal simple: En este caso, estamos interesados en pronosticar el valor de una variable, que en general no se puede medir de manera sencilla, utilizando otra que es más fácil de medir.

Por ejemplo, veamos cómo predecir valores de la variable `Peso` a partir de la longitud del fémur (`LF`) en el archivo `Ecografia.sav`. Utilizaremos el menú de Regresión lineal (Figura 6.26).

Analizar - Regresión - Lineales

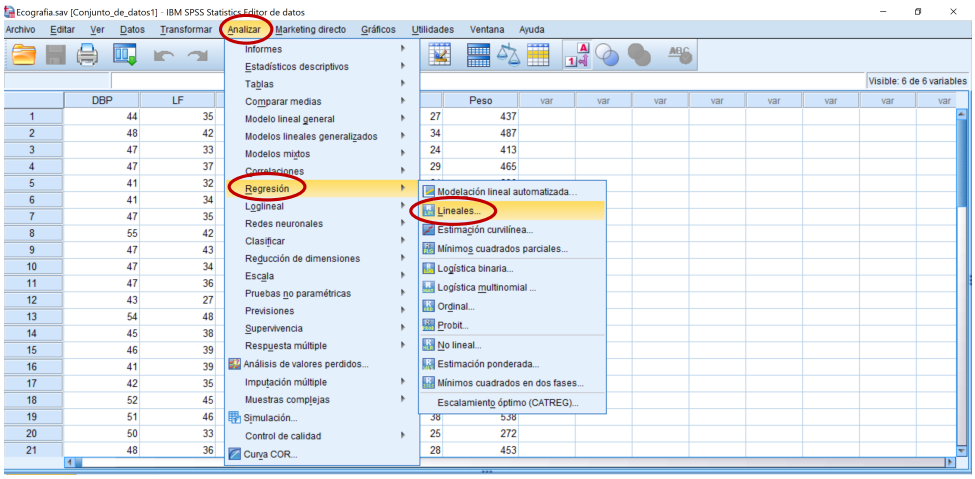


Figura 6.26: Regresión lineal simple.

Introducimos la variable que queremos predecir en el cuadro de Dependientes, en este caso el `Peso`, y en el cuadro de Independientes la variable que utilizaremos para ello, en este caso `LF` (Figura 6.27).

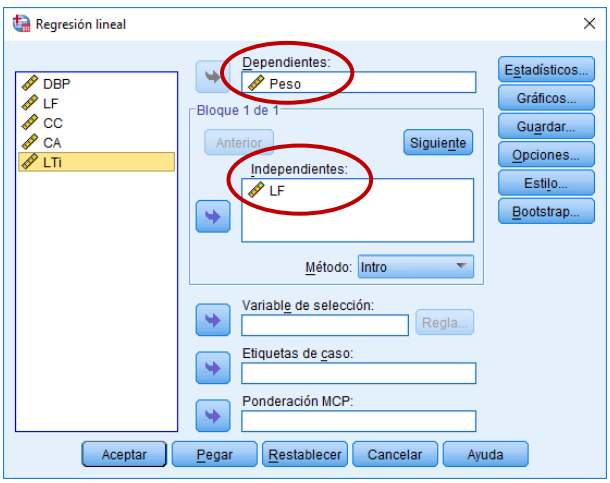


Figura 6.27: Regresión lineal simple.

El coeficiente de determinación lineal muestral se obtiene por defecto en la tabla de Resumen del modelo (Figura 6.28).

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,802[a]	,643	,633	49,233

a. Predictores: (Constante), LF

b. Variable dependiente: Peso

Figura 6.28: Regresión lineal simple.

La ecuación de regresión la proporciona la tabla de Coeficientes (Figura 6.29). En este caso la ecuación sería:

$$\text{Peso} = -29.188 + 13.058\ \text{LF}$$

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	-29,188	59,787		-,488	,628
	LF	13,058	1,579	,802	8,270	,000

a. Variable dependiente: Peso

Figura 6.29: Regresión lineal simple.

Regresión lineal múltiple: Podemos tratar de mejorar la predicción dada por un modelo de regresión lineal simple incorporando más variables predictoras al modelo.

Por ejemplo, veamos cómo pronosticar valores de la variable `Peso` a partir de las variables `LF`, `DBP`, `CC`, `CA` y `LTi` en el archivo `Ecografia.sav`.

Introducimos la variable que queremos predecir en el cuadro de Dependientes, en este caso el `Peso`, y en el cuadro de Independientes la variables predictoras, en este caso `LF`, `DBP`, `CC`, `CA` y `LTi` (Figura 6.30).

Al igual que en el caso anterior, el coeficiente de correlación múltiple R^2, se obtiene por defecto en la tabla de Resumen del modelo (Figura 6.31).

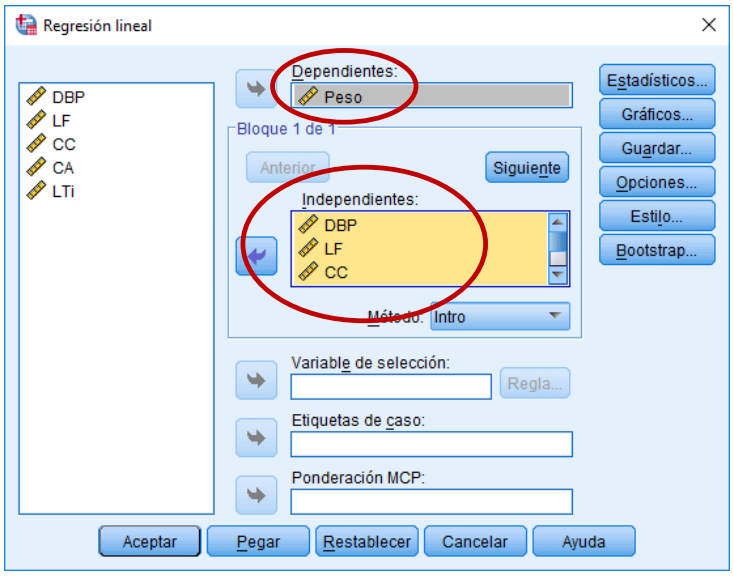


Figura 6.30: Regresión lineal múltiple.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,973[a]	,947	,939	20,045

a. Predictores: (Constante), LTi, CA, DBP, CC, LF

Figura 6.31: Regresión lineal múltiple.

La tabla de Coeficientes proporciona la ecuación de regresión múltiple (Figura 6.32), que en este caso sería:

$$\text{Peso} = -215.980 - 16.025\text{DBP} + 30.014\ \text{LF} + 13.541\ \text{CC} - 9.612\ \text{CA} - 16.114\text{LTi}$$

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	-215,980	248,953		-,868	,392
	DBP	-16,025	3,562	-,772	-4,499	,000
	LF	30,014	30,627	1,843	,980	,334
	CC	13,541	1,154	2,651	11,733	,000
	CA	-9,612	,730	-1,978	-13,170	,000
	LTi	-16,114	30,583	-,991	-,527	,602

a. Variable dependiente: Peso

Figura 6.32: Regresión lineal múltiple.

Regresión no lineal: En ciertas ocasiones se logra una mejor explicación de la variable dependiente si no nos restringimos a ecuaciones de tipo lineal. Por ejemplo, en el archivo `Tumor de próstata.sav` podemos utilizar un modelo de regresión en forma de potencia (es decir, correlacionamos linealmente los logaritmos de ambas variables y deshacemos el cambio, según vimos en la Figura 2.14) para predecir el volumen del tumor (`Vol`) a partir del antígeno prostático específico (`PSA`). Lo haremos a través del siguiente menú (Figura 6.33):

Analizar - Regresión - Estimación curvilínea

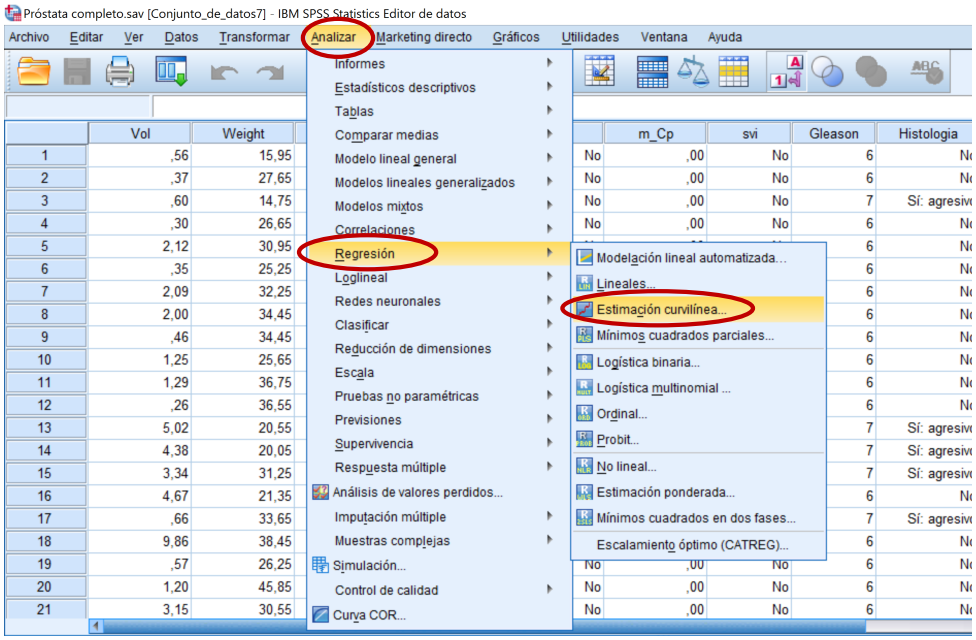


Figura 6.33: Regresión no lineal.

En el cuadro de Dependientes incluimos la variable a predecir, `Vol`, y en el cuadro de Independientes, la variable que utilizamos para predecir, `PSA`. En Modelos, elegimos el que deseemos, en este caso, Potencia (Figura 6.34).

Si queremos obtener la ecuación del modelo hacemos doble click en la curva del gráfico resultante, y en la opción Propiedades aparece la ecuación del modelo (Figura 6.35).

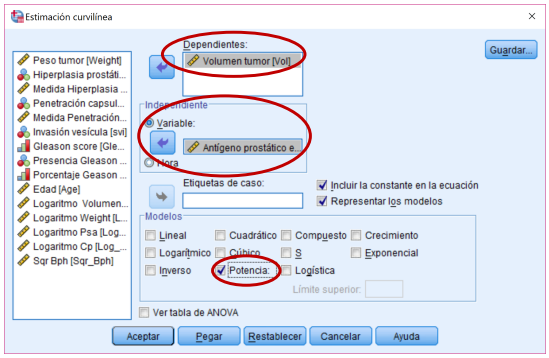


Figura 6.34: Regresión no lineal.

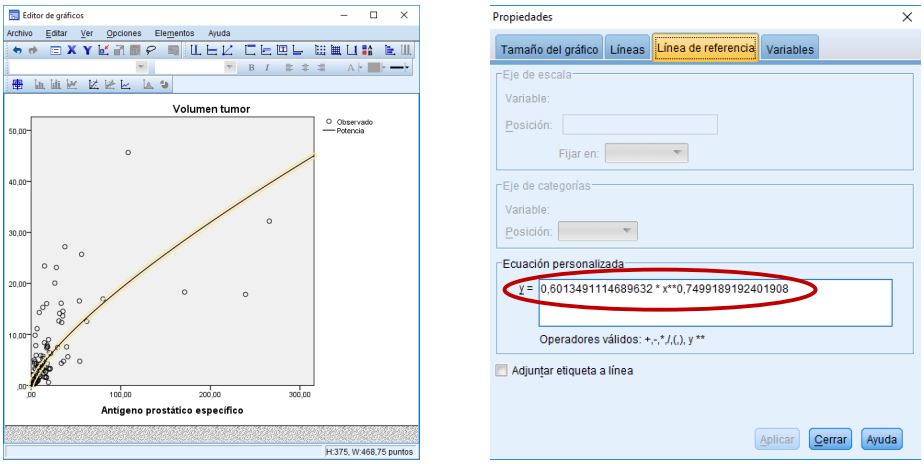


Figura 6.35: Regresión no lineal.

6.4. Relación entre una variable cuantitativa y una variable cualitativa

Fundamentalmente, haremos uso del menú Explorar de Estadísticos descriptivos (Figura 6.36) y trataremos la variable cualitativa como factor. En el siguiente capítulo ampliaremos este estudio mediante el uso de diferentes tests de hipótesis.

Analizar - Estadísticos descriptivos - Explorar

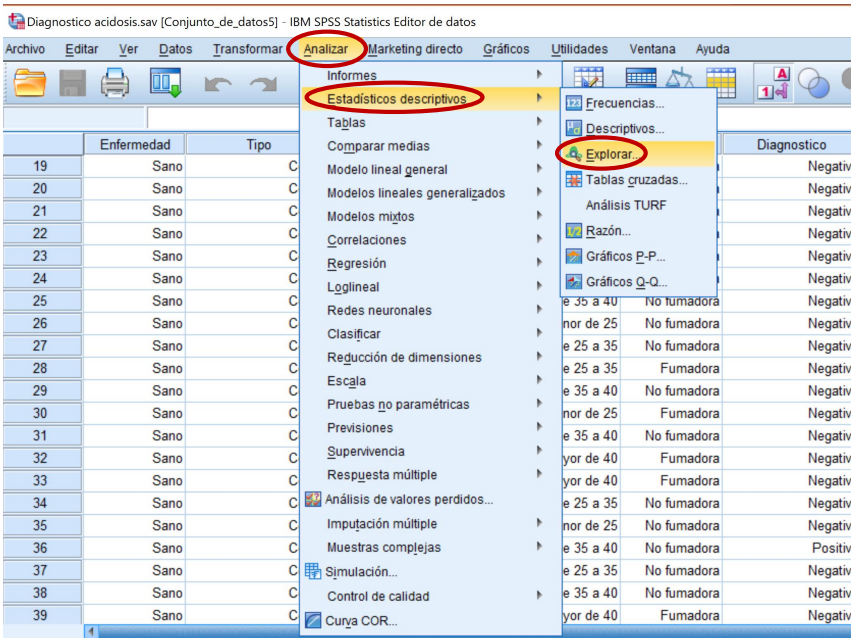


Figura 6.36: Análisis descriptivo de la relación entre una variable cualitativa y una cuantitativa.

Por ejemplo, veamos cómo estudiar en el archivo `Diagnostico acidosis.sav` la relación entre la acidosis en recién nacidos (`Tipo`) y la glucemia medida en el cordón umbilical (`Glucemia`).

- **Gráfico de puntos por grupos**: accedemos al Generador de gráficos.

  Gráficos - Generador de gráficos

  Arrastramos la primera opción de los Gráficos de puntos al cuadrado central e incluimos la variable cualitativa en el eje OX del gráfico y la variable cuantitativa en el eje OY (Figuras 6.37 y 6.38).

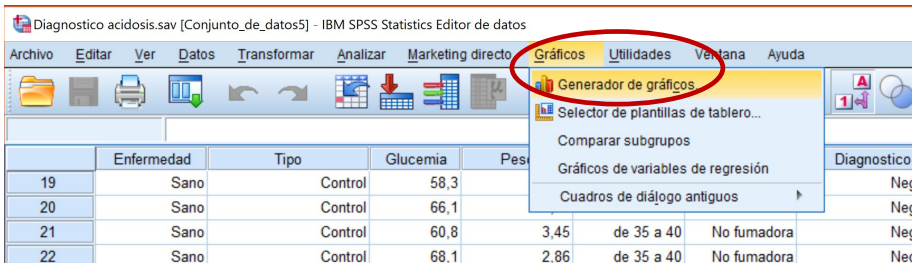


Figura 6.37: Análisis descriptivo de la relación entre una variable cualitativa y una cuantitativa.

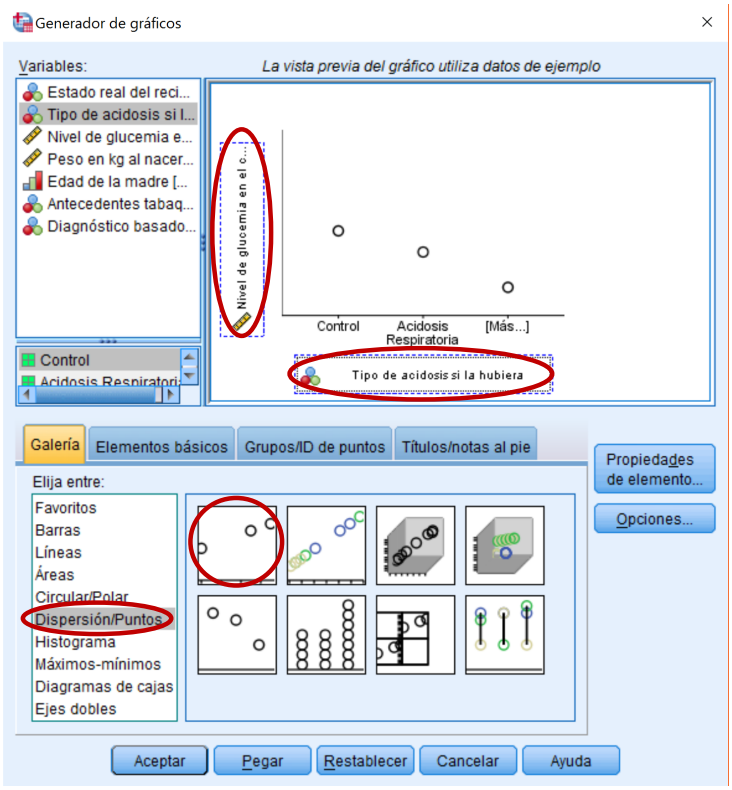


Figura 6.38: Análisis descriptivo de la relación entre una variable cualitativa y una cuantitativa.

- **Resúmenes por grupos**: en el menú Explorar, incluimos en la Lista de dependientes la variable cuantitativa y en la Lista de factores la variable cualitativa (Figura 6.39).
- **Diagramas de caja por grupos**: se obtiene por defecto al seleccionar las opciones anteriores.
- **Histograma por grupos**: si además queremos obtener un histograma por cada uno de los grupos, en la opción Gráficos seleccionamos Histograma (Figura 6.39).
- **Diagrama de medias**: en el Generador de gráficos del menú Gráficos, elegimos el Gráfico de medias de la opción Barras (Figura 6.40). Arrastramos la variable cualitativa al eje OX y la variable cuantitativa al eje OY.

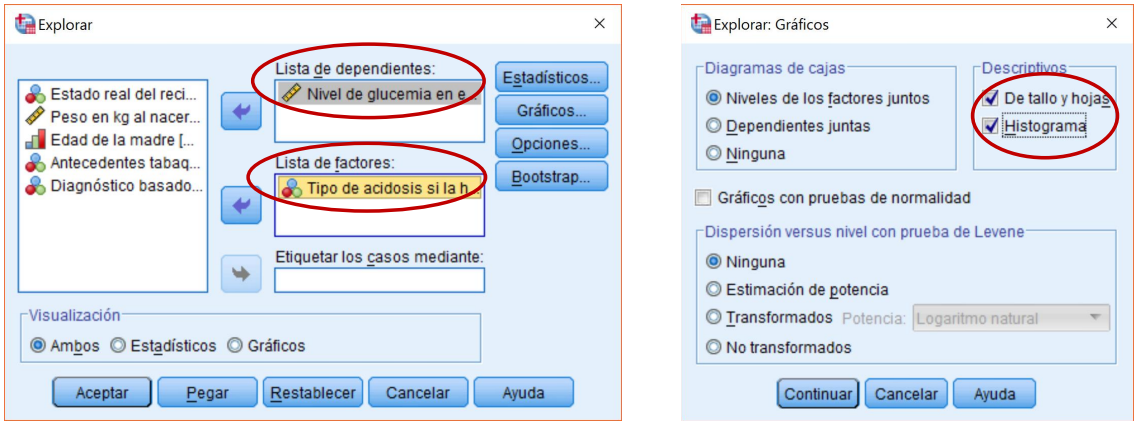


Figura 6.39: Análisis descriptivo de la relación entre una variable cualitativa y una cuantitativa.

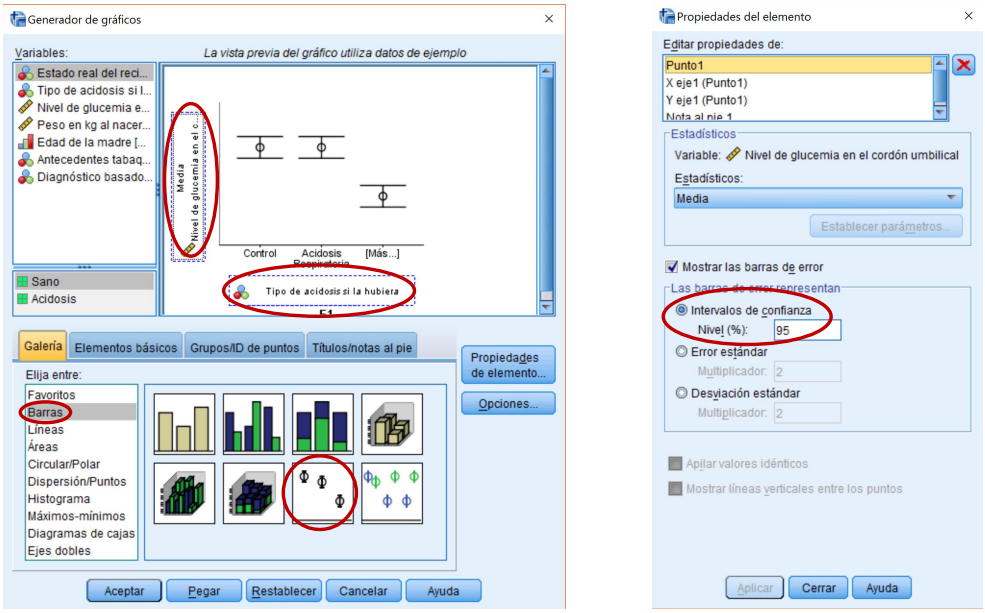


Figura 6.40: Análisis descriptivo de la relación entre una variable cualitativa y una cuantitativa.

6.5. Relación entre dos variables cualitativas

Utilizaremos en todo caso el menú Tablas cruzadas (o de contingencia) de Estadísticos descriptivos (Figura 6.41).

Analizar - Estadísticos descriptivos - Tablas cruzadas

Analicemos, por ejemplo, la relación entre las variables `Bph` e `Histologia` del archivo `Tumor de próstata.sav`.

- **Tablas de contingencia**: insertamos las variables elegidas en los cuadros Filas y Columnas (el orden sólo importa en la interpretación de los gráficos) (Figura 6.42).
- **Diagrama de barras agrupadas**: marcamos la opción Mostrar los gráficos de barras agrupados (Figura 6.42).
- **Coeficiente de contingencia C de Pearson**: para obtenerlo seleccionamos en el menú Estadísticos la opción Coeficiente de contingencia en el apartado Nominal (Figura 6.43).
- **Coeficiente ϕ**: si la tabla con la que trabajamos es 2×2, podemos seleccionar el coeficiente ϕ en la opción Phi y V de Cramer en el apartado Nominal del menú Estadísticos (Figura 6.43).
- **Tabla de valores esperados**: el programa también permite calcular la tabla con los valores que cabría esperar en ausencia de relación entre las variables en la opción Recuentos Esperados del menú Casillas (Figura 6.44).
- **Proporciones condicionadas**: podemos obtener una tabla con las proporciones marginales, las proporciones condicionadas y las proporciones conjuntas en el menú Casillas seleccionando las opciones Porcentajes Fila, Columna y Total (Figura 6.45).
- **Diagrama de barras apiladas**: se realiza a través del Generador de gráficos.

  Gráficos - Generador de gráficos

  Arrastramos la tercera opción de la Galería de Gráficos de Barras y añadimos las variables elegidas, una en el eje OX del gráfico y otra en la esquina superior derecha (Figuras 6.46 y 6.47).

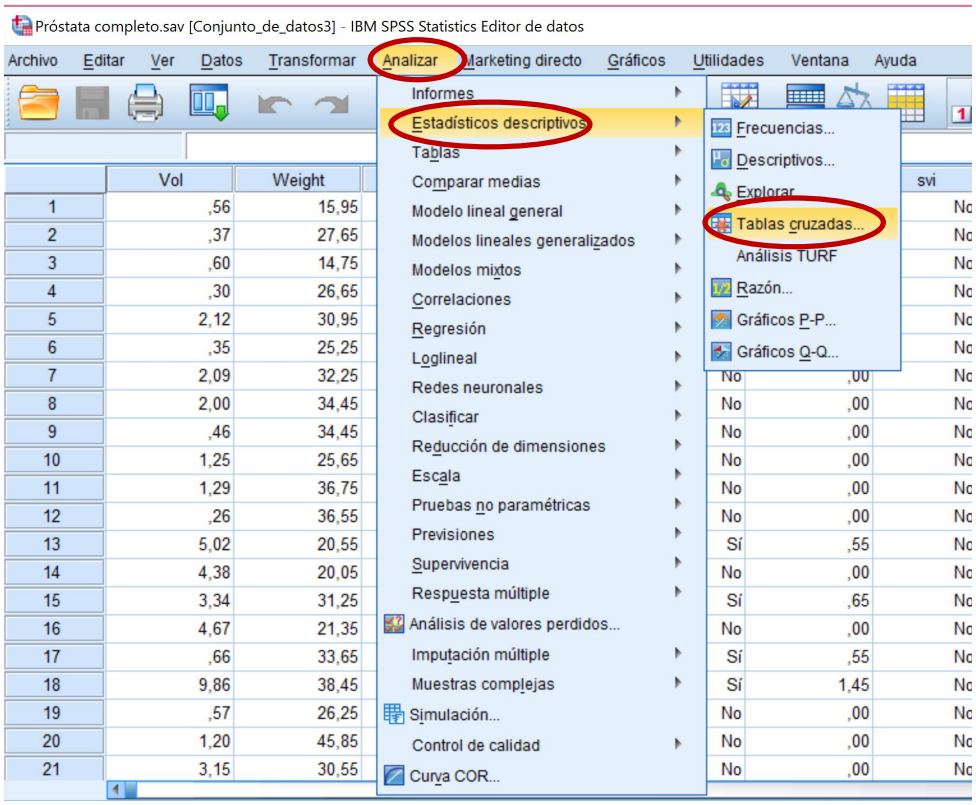


Figura 6.41: Análisis descriptivo de la relación entre variables cualitativas.

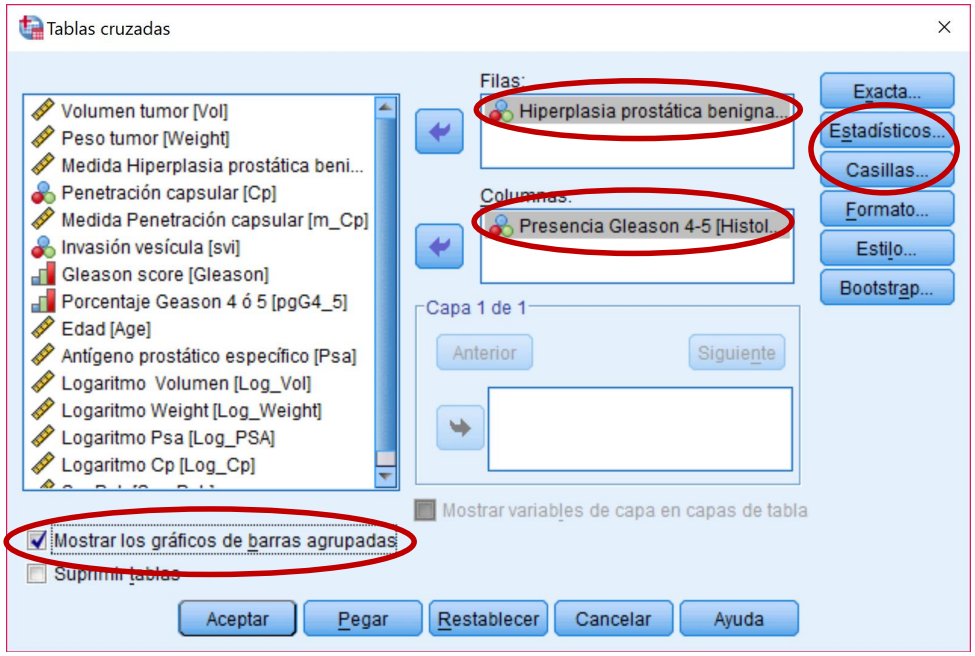


Figura 6.42: Análisis descriptivo de la relación entre variables cualitativas.

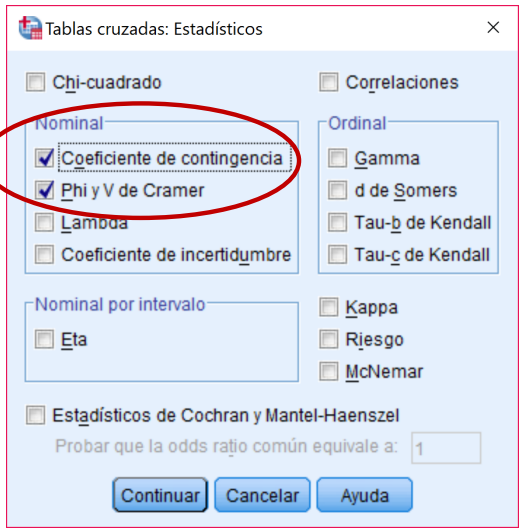


Figura 6.43: Análisis descriptivo de la relación entre variables cualitativas.

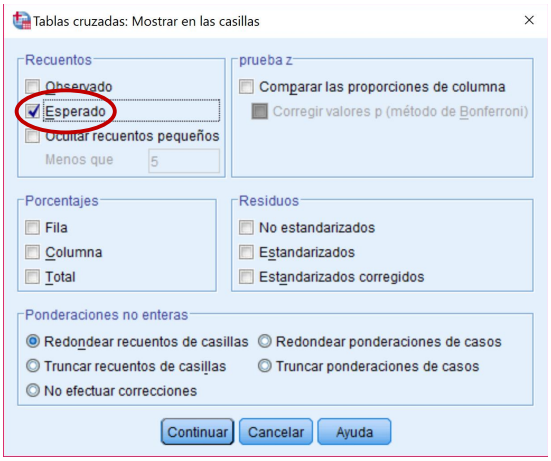


Figura 6.44: Análisis descriptivo de la relación entre variables cualitativas.

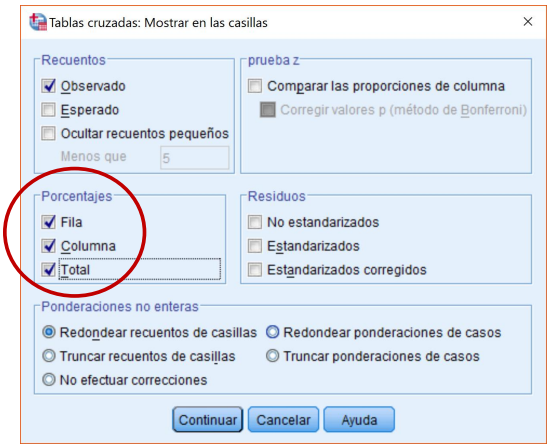


Figura 6.45: Análisis descriptivo de la relación entre variables cualitativas.

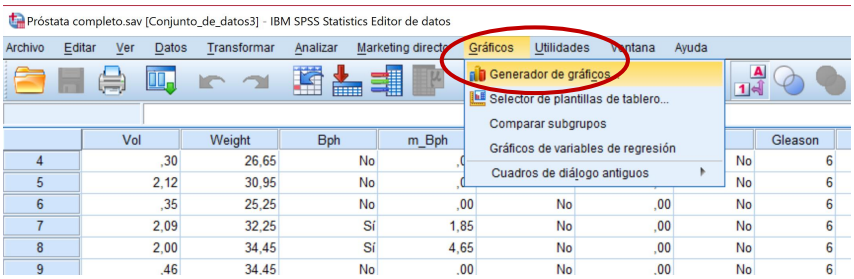


Figura 6.46: Análisis descriptivo de la relación entre variables cualitativas.

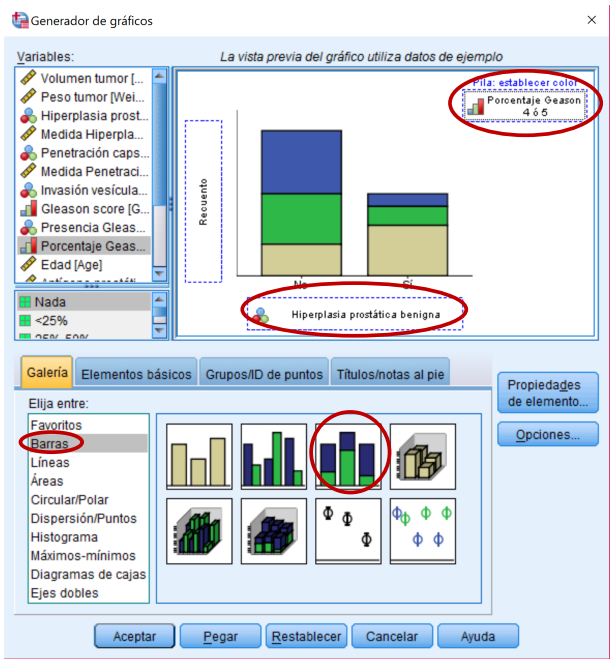


Figura 6.47: Análisis descriptivo de la relación entre variables cualitativas.

6.6. Medidas de riesgo y curvas COR

6.6.1. Medidas de riesgo

En el caso de tablas 2×2, cuando estudiamos la presencia o ausencia de una enfermedad y su relación con un posible factor de riesgo se suelen utilizar otras medidas más sensibles para cuantificar el grado de riesgo que comporta dicho factor (a parte del coeficiente de contingencia y del coeficiente ϕ). Dado que la ejecución del programa depende de cómo se hayan introducido los códigos en las dos variables cualitativas consideradas, aconsejamos que se efectúen los cálculos a través de la propia tabla de contingencia. No obstante, el programa aporta cálculos directos (y con intervalos de confianza) tanto del **riesgo relativo** como del **odds ratio** a través del menú Estadísticos en Tablas cruzadas.

Analizar - Estadísticos descriptivos - Tablas cruzadas - Estadísticos

Por ejemplo, en el archivo `Southafrica Heart Disease.sav`, estudiamos los antecedentes familiares (`famhist`) como posible factor de riesgo para presentar una enfermedad coronaria (`chd`).

Introducimos una de las variables en el cuadro Filas y la otra en Columnas (Figura 6.48). En la opción Estadísticos, marcamos Riesgo (Figura 6.48). El valor del odds ratio se recoge en la tabla de Estimación del riesgo (Figura 6.49).

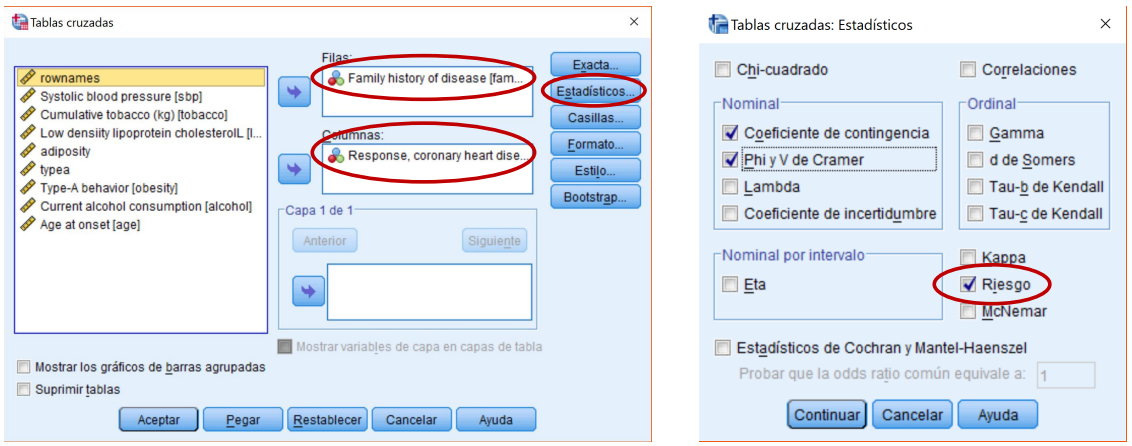


Figura 6.48: Cálculo del odds ratio.

Estimación de riesgo

	Valor	Intervalo de confianza de 95 %	
		Inferior	Superior
Odds ratio para Family history of disease (Absent / Present)	3,219	2,161	4,794
Para cohorte Response, coronary heart disease = No	1,526	1,305	1,784
Para cohorte Response, coronary heart disease = Yes	,474	,367	,613
N de casos válidos	462		

Figura 6.49: Cálculo del odds ratio.

Al igual que ocurre con el riesgo relativo y el odds ratio, la sensibilidad y especificidad de un procedimiento diagnóstico pueden calcularse a través de la correspondiente tabla de contingencia.

6.6.2. Curvas COR

En el menú Analizar se encuentra la opción de representar curvas COR (Figura 6.50).

Analizar - Curvas COR

Por ejemplo, en el archivo `Enfermedad celiaca.sav`, veamos cómo representar la curva COR y encontrar un umbral de la variable `Antigladina IgA` para determinar la presencia de enfermedad celiaca (`celiaquia`) obteniendo simultáneamente una sensibilidad y especificidad razonables.

Introducimos la variable cuantitativa en el cuadro Variable de prueba y la variable cualitativa indicando la enfermedad en el cuadro Variable de estado. Además, debemos indicar a qué categoría corresponde la presencia de la enfermedad en el cuadro Valor de la variable de estado. Por último, marcamos la opción Puntos de coordenadas de la curva COR (Figura 6.51).

Para elegir el valor umbral, nos fijamos en la tabla Coordenadas de la curva (Figura 6.51). La primera columna proporciona el umbral correspondiente al par Sensibilidad (segunda columna) y 1 - Especificidad (tercera columna). Por tanto, se busca un umbral correspondiente a una fila en la que los valores de la segunda y tercera columnas estén simultáneamente próximos a 1 y 0, respectivamente.

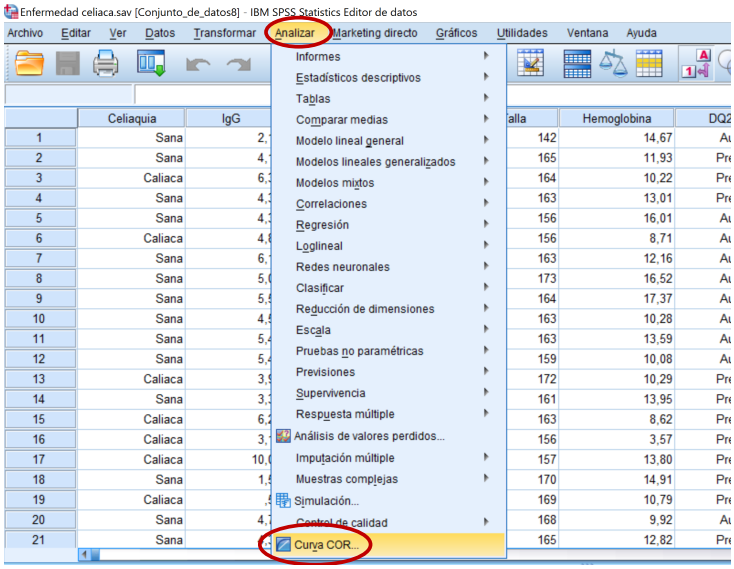


Figura 6.50: Curvas COR.

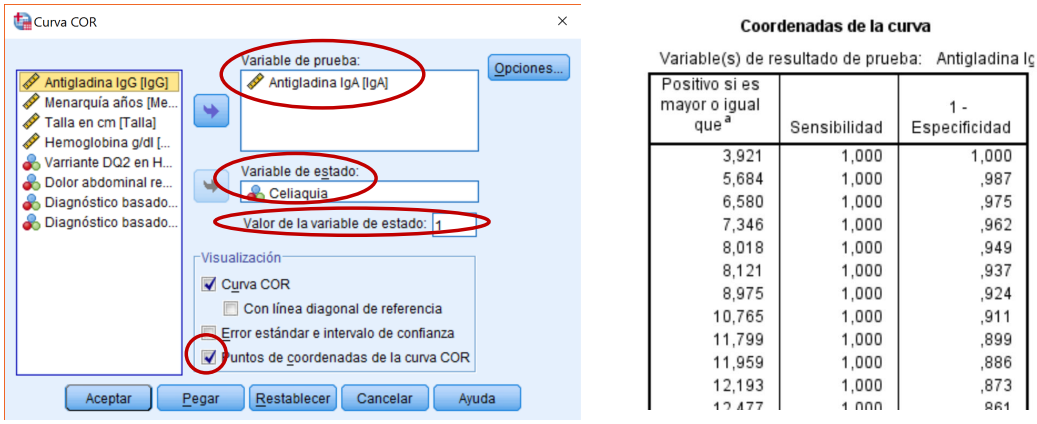


Figura 6.51: Curvas COR.

7. Inferencia Estadística con SPSS

En esta fase del estudio estadístico se pretende averiguar en qué medida son extrapolables los resultados obtenidos en la muestra a la población de la que procede, suponiendo que hubiera sido extraída aleatoriamente de la misma. Dedicaremos la primera sección al problema de Estimación y las siguientes a los problemas básicos de Contraste de Hipótesis.

7.1. Problemas de estimación

7.1.1. Intervalo de confianza para la media

A partir de los datos del archivo `Southafrica Heart Disease.sav` calculemos una estimación para la adiposidad (`adiposity`) media, un intervalo de confianza al 95 % y el error máximo cometido en dicha estimación. Para ello, utilizaremos el menú Explorar de Estadísticos descriptivos (Figura 7.1).

Analizar - Estadísticos descriptivos - Explorar

- Seleccionamos la variable `adiposity` de la lista de variables y la introducimos en la Lista de dependientes (Figura 7.2). En el menú de Estadísticos nos aseguramos de que tenemos marcada la opción Descriptivos (Figura 7.3).

- En la primera parte de la tabla de Descriptivos obtenida, nos fijamos únicamente en las dos primeras filas (Figura 7.2) que nos proporcionan la estimación de la media, en este caso 25.4067, y un intervalo de confianza al 95 %, que en nuestro caso es $[24.6954, 26.1181]$.

- El error máximo cometido en la estimación se obtiene de la forma usual, restando al límite superior del intervalo la estimación. En nuestro ejemplo, $26.1181 - 25.4067 = 0.7114$.

- La confianza para la cual calculamos el intervalo se puede modificar en la opción Estadísticos del menú Explorar (Figura 7.3).

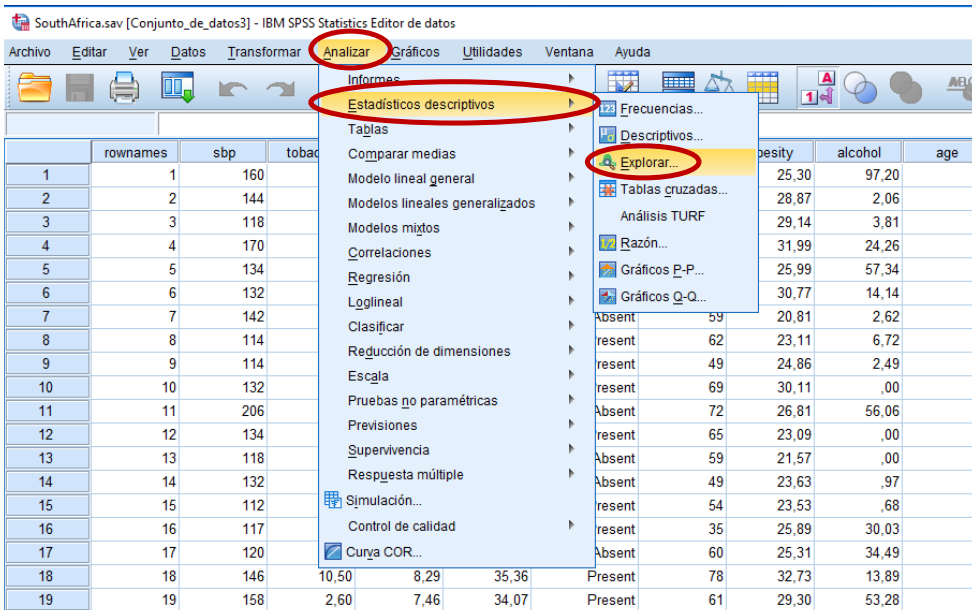


Figura 7.1: Intervalo de confianza para una media.

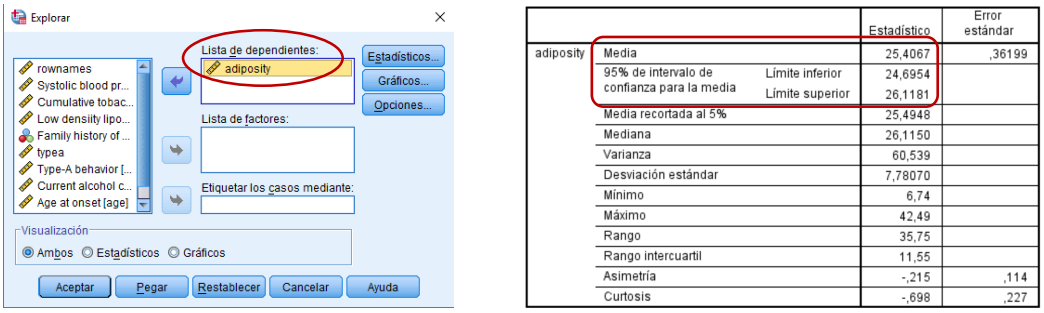


Figura 7.2: Intervalo de confianza para una media.

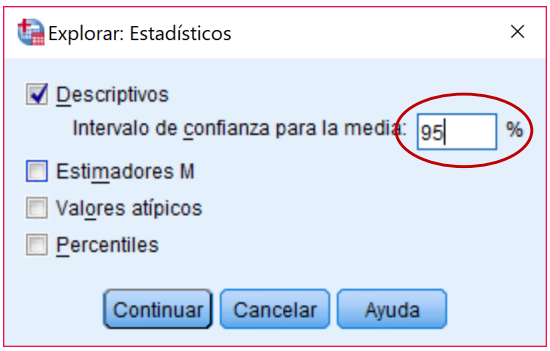


Figura 7.3: Intervalo de confianza para una media.

7.1.2. Intervalo de confianza para la proporción

A partir del archivo `Tumor de próstata.sav` calculemos una estimación para la proporción de individuos en la población que presentan `hiperplasia prostática benigna`

(Bph) y un intervalo de confianza al 95 % para dicha proporción. En primer lugar observemos que la variable Bph toma el valor 0 si el paciente no presenta hiperplasia prostática benigna y toma el valor 1 si el paciente sí la presenta.

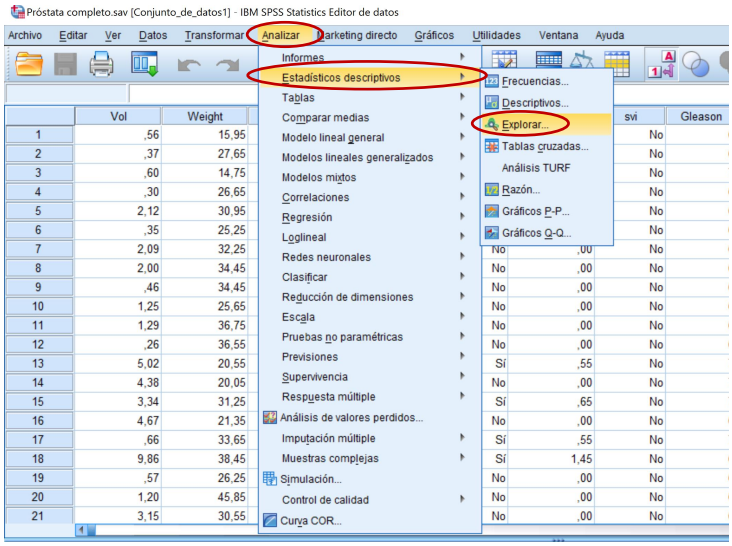


Figura 7.4: Intervalo de confianza para una proporción.

- Para estimar dicha proporción, tomamos la variable cualitativa, en este caso Bph, y calculamos su media e intervalo de confianza siguiendo los mismos pasos que en el apartado anterior, ya que la media aritmética de dicha variable equivale a la proporción muestral (Figura 7.4).

Analizar - Estadísticos descriptivos - Explorar

- Seleccionamos la variable Bph de la lista de variables, la introducimos en la Lista de dependientes y en el menú de Estadísticos nos aseguramos de que tenemos marcada la opción Descriptivos (Figura 7.5).

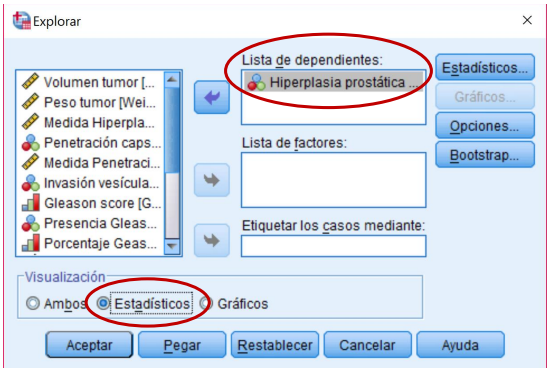


Figura 7.5: Intervalo de confianza para una proporción.

- En la primera parte de la tabla de Descriptivos obtenida, nos fijamos únicamente en las dos primeras filas que proporcionan la estimación de la proporción, en este

caso 0.56, el error máximo cometido es $0.66 - 0.56 = 0.1$ y el intervalo de confianza $[0.46, 0.66]$ (Figura 7.6).

- Teniendo en cuenta que la variable toma el valor 1 si el paciente presenta hiperplasia o comparando con el diagrama de barras (Figura 7.6) observamos que la estimación corresponde a la proporción de individuos con `hiperplasia prostática benigna`.

- Si queremos estimar la proporción de individuos que no la presentan y un intervalo de confianza, restamos a 1 los valores obtenidos. Así, en este caso estimamos la proporción de individuos que no presentan `hiperplasia prostática benigna` como $1-0.56 = 0.44$, y un intervalo de confianza al 95 % es $[1-0.66, 1-0.46] = [0.34, 0.54]$. El error máximo cometido es el mismo que para la proporción de individuos con `hiperplasia prostática benigna`.

- La confianza para la cual calculamos el intervalo se puede modificar de la misma manera que en el caso anterior.

Descriptivos

			Estadístico	Error estándar
Hiperplasia prostática benigna	Media		,56	,051
	95% de intervalo de confianza para la media	Límite inferior	,46	
		Límite superior	,66	
	Media recortada al 5%		,56	
	Mediana		1,00	
	Varianza		,249	
	Desviación estándar		,499	
	Mínimo		0	
	Máximo		1	
	Rango		1	
	Rango intercuartil		1	
	Asimetría		-,232	,245
	Curtosis		-1,988	,485

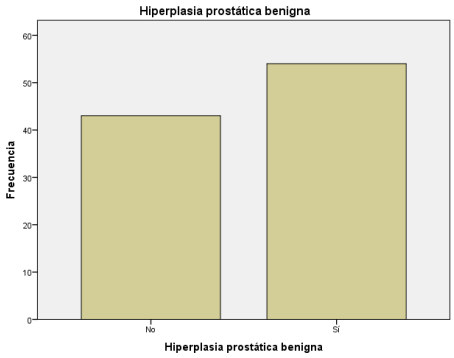


Figura 7.6: Intervalo de confianza para una proporción.

7.2. Tests de hipótesis en problemas de correlación y regresión

Esta sección constituye la continuación de la Sección 6.3 del capítulo anterior.

7.2.1. Problemas de correlación

En este apartado veremos cómo aplicar los test de hipótesis que nos permitirán concluir si la relación observada entre dos variables cuantitativas en la muestra puede extrapolarse a la población: el test de correlación de Pearson (paramétrico) y el test de Spearman (no paramétrico).

Por ejemplo, a partir del archivo `Ecografía.sav` veamos si la relación entre las variables `Peso` y `LF` en la población es significativa.

Test de correlación de Pearson: Para ello, accedemos al menú Correlaciones Bivariadas (Figura 7.7):

Analizar - Correlaciones - Bivariadas

- Añadimos las dos variables para las que estamos estudiando la relación al cuadro Variables y marcamos Pearson en Coeficiente de correlación (Figura 7.8).
- En la tabla Correlaciones se indica el P-valor del test de correlaciones (Figura 7.8).

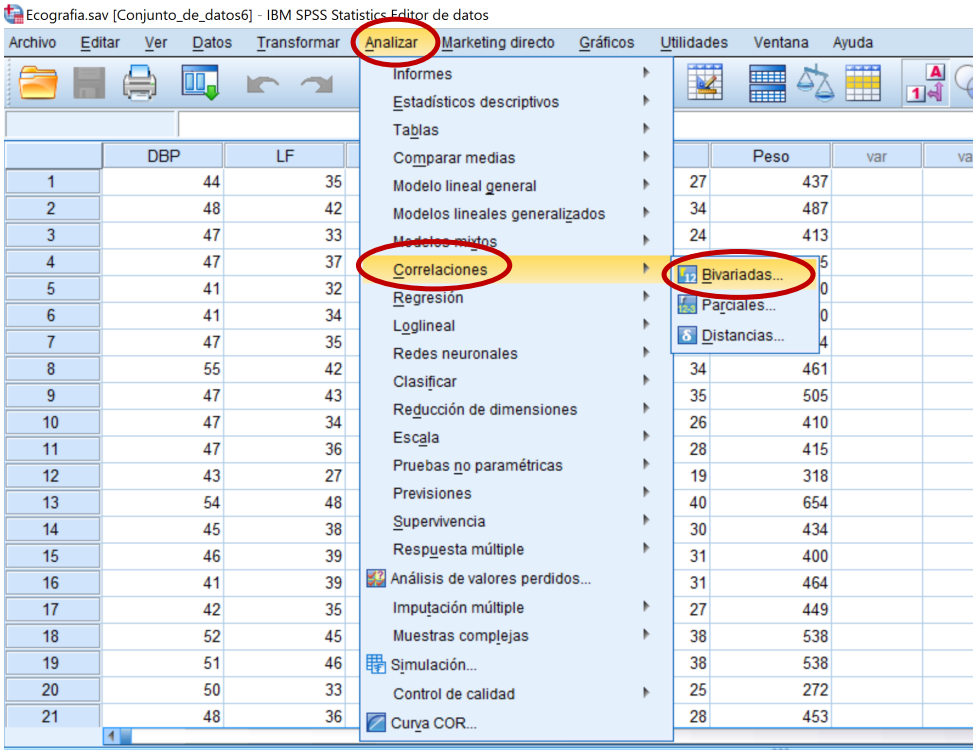
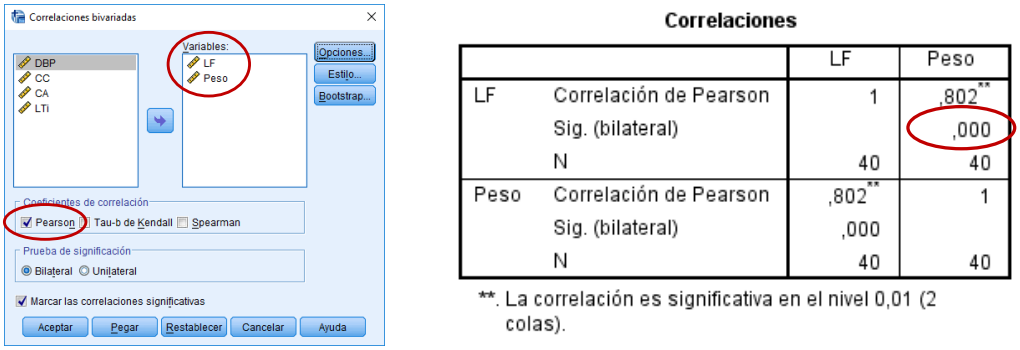


Figura 7.7: Problemas de correlación: tests de correlación.



Correlaciones

		LF	Peso
LF	Correlación de Pearson	1	,802**
	Sig. (bilateral)		,000
	N	40	40
Peso	Correlación de Pearson	,802**	1
	Sig. (bilateral)	,000	
	N	40	40

**. La correlación es significativa en el nivel 0,01 (2 colas).

Figura 7.8: Problemas de correlación: tests de correlación.

Test de Spearman: El procedimiento es completamente análogo al test anterior, accedemos al menú Correlaciones Bivariadas (Figura 7.7):

Analizar - Correlaciones - Bivariadas

- Añadimos las dos variables para las que estamos estudiando la relación al cuadro Variables y marcamos Spearman en el Coeficiente de correlación (Figura 7.9).
- En la tabla Correlaciones se indica el P-valor del test de Spearman (Figura 7.9).

Correlaciones

			LF	Peso
Rho de Spearman	LF	Coeficiente de correlación	1,000	,820**
		Sig. (bilateral)	.	,000
		N	40	40
	Peso	Coeficiente de correlación	,820**	1,000
		Sig. (bilateral)	,000	.
		N	40	40

**. La correlación es significativa en el nivel 0,01 (2 colas).

Figura 7.9: Problemas de correlación: test de Spearman.

7.2.2. Regresión múltiple

En este apartado continuamos con el estudio del modelo de regresión múltiple iniciado en la Subsección 6.3.2 del capítulo anterior con el objetivo de extrapolar las conclusiones obtenidas a toda la población. Recordemos que para ello, hacemos uso del menú Lineales de Regresión (Figura 7.10).

Analizar - Regresión - Lineales

Consideremos de nuevo el archivo `Ecografia.sav`, tomando como variable respuesta la variable `Peso` que queremos predecir utilizando las variables independientes `LF`, `DBP`, `CC`, `CA` y `LTi`.

- Incluimos la variable que queremos predecir en el cuadro de Dependientes y las variables predictoras en el cuadro de Independientes (Figura 7.11).
- Los coeficientes de la ecuación de regresión se obtienen por defecto.
- Los P-valores de los tests de correlación parciales también se obtienen por defecto; sin embargo, si también queremos calcular los coeficientes de correlación parcial es necesario marcar Correlaciones parciales y semiparciales en el menú Estadísticos (Figura 7.11).

- Si además queremos guardar las predicciones de la variable respuesta para cada individuo y un intervalo de confianza para las mismas, en la opción Guardar seleccionamos No estandarizados en los Valores pronosticados e Intervalos de predicción para los Individuos (Figura 7.12). Como resultado obtendremos tres nuevas columnas en el archivo original, correspondientes a las predicciones y a los límites inferior y superior del intervalo de confianza.

- El P-valor del test de correlación múltiple o total se proporciona en la tabla ANOVA (Figura 7.13).

- La estimación de los coeficientes de la ecuación de regresión se recoge en la columnna B de la tabla Coeficientes (Figura 7.14). Así, la recta de regresión en este caso será:

$$\text{Peso} = -215.980 - 16.025\ \text{DBP} + 30.014\ \text{LF} + 13.541\ \text{CC} - 9.612\ \text{CA} - 16.114\text{LTi}.$$

- Los P-valores de los test parciales, así como los coeficientes de correlación parciales se recogen también en la tabla Coeficientes, en las columnas Sig. y Parcial, respectivamente (Figura 7.15).

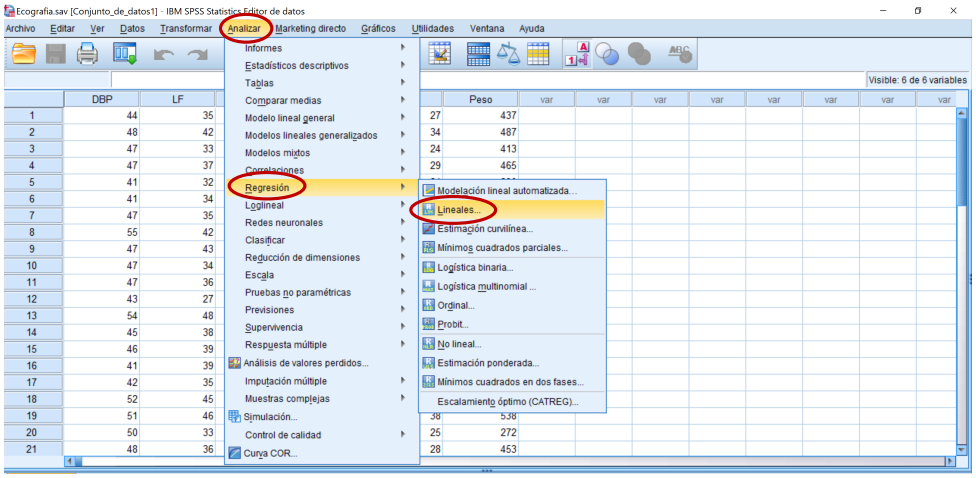


Figura 7.10: Regresión lineal múltiple.

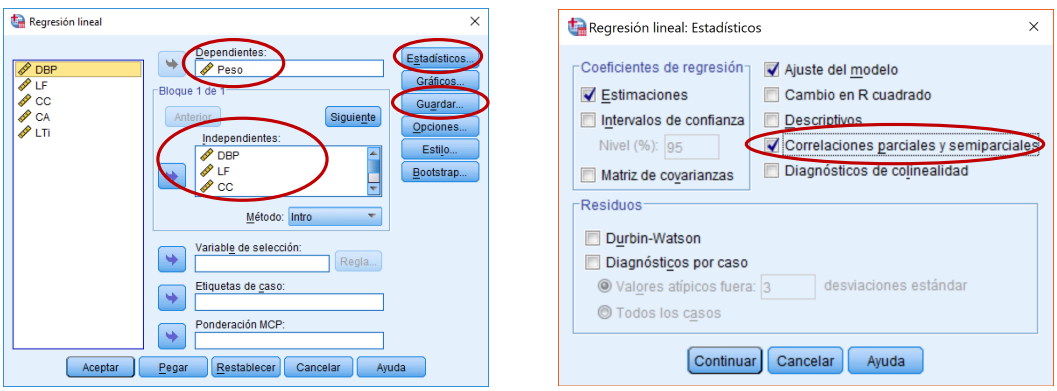


Figura 7.11: Regresión lineal múltiple.

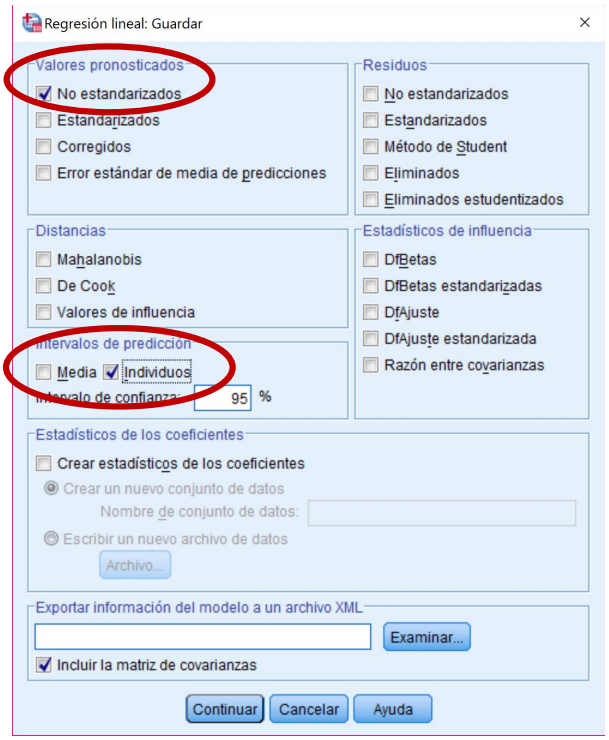


Figura 7.12: Regresión lineal múltiple.

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	244226,662	5	48845,332	121,560	,000[b]
	Residuo	13661,889	34	401,820		
	Total	257888,552	39			

a. Variable dependiente: Peso

b. Predictores: (Constante), LTi, CA, DBP, CC, LF

Figura 7.13: Regresión lineal múltiple.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Error estándar	Beta			Orden cero	Parcial	Parte
1	(Constante)	-215,980	248,953		-,868	,392			
	DBP	-16,025	3,562	-,772	-4,499	,000	,569	-,611	-,178
	LF	30,014	30,627	1,843	,980	,334	,802	,166	,039
	CC	13,541	1,154	2,651	11,733	,000	,577	,896	,463
	CA	-9,612	,730	-1,978	-13,170	,000	,420	-,914	-,520
	LTi	-16,114	30,583	-,991	-,527	,602	,799	-,090	-,021

a. Variable dependiente: Peso

Figura 7.14: Regresión lineal múltiple.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	95,0% intervalo de confianza para B		Correlaciones		
		B	Error estándar	Beta			Límite inferior	Límite superior	Orden cero	Parcial	Parte
1	(Constante)	-215,980	248,953		-,868	,392	-721,912	289,953			
	DBP	-16,025	3,562	-,772	-4,499	,000	-23,263	-8,786	,569	-,611	-,178
	LF	30,014	30,627	1,843	,980	,334	-32,228	92,255	,802	,166	,039
	CC	13,541	1,154	2,651	11,733	,000	11,196	15,887	,577	,896	,463
	CA	-9,612	,730	-1,978	-13,170	,000	-11,095	-8,129	,420	-,914	-,520
	LTi	-16,114	30,583	-,991	-,527	,602	-78,266	46,037	,799	-,090	-,021

a. Variable dependiente: Peso

Figura 7.15: Regresión lineal múltiple.

7.2.3. Selección de variables

Podemos aplicar el método de selección hacia atrás o *backward* para optimizar un modelo de regresión ante la presencia de un problema de **multicolinealidad**.

Volviendo al archivo `Ecografia.sav`, y con variable `Peso` como variable respuesta y `LF`, `DBP`, `CC`, `CA` y `LTi` como variables independientes, se observan resultados no significativos en los test parciales, lo cual nos indica que es posible optimizar el modelo.

El proceso es completamente análogo al de regresión lineal, utilizando el menú Lineales de Regresión (Figura 7.16).

Analizar - Regresión - Lineales

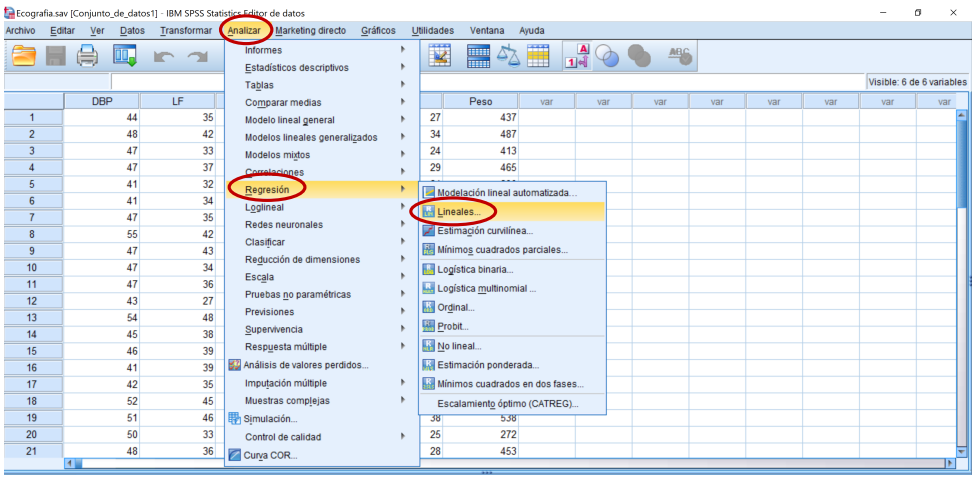


Figura 7.16: Problemas de multicolinealidad.

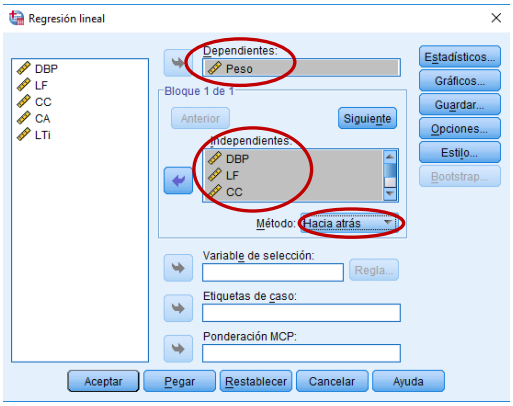
- Incluimos la variable que queremos predecir en el cuadro de Dependientes y las variables predictoras en el cuadro de Independientes (Figura 7.17).
- En el cuadro Método, seleccionamos Hacia atrás (Figura 7.17).
- La tabla Variables entradas/eliminadas muestra las variables introducidas inicialmente en el modelo y las variables eliminadas en cada paso, correspondiente a una fila de la tabla. En este caso, la tabla sólo tiene dos filas porque el método ha finalizado en dos pasos (Figura 7.17).

- Cada división horizontal de la tabla de Coeficientes corresponde al modelo en cada paso. Nos centraremos en el análisis de la parte final, que corresponde al modelo óptimo (Figura 7.18).

- En la tabla de Variables excluidas se indican las variables excluidas en cada paso (Figura 7.19).

Del análisis de las tablas anteriores se observa que se ha eliminado la variable `LTi` del modelo y la ecuación de regresión resultante es:

$$\text{Peso} = -86.496 - 16.058\ \text{DBP} + 13.883\ \text{LF} + 13.614\ \text{CC} - 9.680\ \text{CA}$$



Variables entradas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	LTi, CA, DBP, CC, LF[b]	.	Intro
2	.	LTi	Retroceder (criterio: Probabilidad de F-para-eliminar >= , 100).

a. Variable dependiente: Peso

b. Todas las variables solicitadas introducidas.

Figura 7.17: Problemas de multicolinealidad.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig.
1	(Constante)	-215,980	248,953		-,868	,392
	DBP	-16,025	3,562	-,772	-4,499	,000
	LF	30,014	30,627	1,843	,980	,334
	CC	13,541	1,154	2,651	11,733	,000
	CA	-9,612	,730	-1,978	-13,170	,000
	LTi	-16,114	30,583	-,991	-,527	,602
2	(Constante)	-86,496	39,462		-2,192	,035
	DBP	-16,058	3,524	-,773	-4,556	,000
	LF	13,883	,912	,852	15,221	,000
	CC	13,614	1,134	2,665	12,006	,000
	CA	-9,680	,711	-1,992	-13,619	,000

a. Variable dependiente: Peso

Figura 7.18: Problemas de multicolinealidad.

Variables excluidas[a]

Modelo		En beta	t	Sig.	Correlación parcial	Estadísticas de colinealidad: Tolerancia
2	LTi	-,991[b]	-,527	,602	-,090	,000

a. Variable dependiente: Peso

b. Predictores en el modelo: (Constante), CA, DBP, CC, LF

Figura 7.19: Problemas de multicolinealidad.

A continuación procedemos a completar el estudio iniciado en la Sección 6.4 del capítulo anterior.

7.3. Tests de comparación de medias para muestras independientes

En este apartado veremos los tests existentes para analizar a nivel poblacional la relación entre una variable cualitativa con dos posibles categorías y una variable cuantitativa: el test de Student y el test de Welch (tests paramétricos) y el test de Mann-Whitney (test no paramétrico).

Por ejemplo, analicemos en el archivo `ICC.sav` la relación entre el índice de cintura-cadera (`ICC`) y la hipertensión (`hip`).

7.3.1. Tests de Student y de Welch para muestras independientes

Haremos uso de la opción Prueba T para muestras independientes del menú Comparar medias (Figura 7.20).

Analizar - Comparar medias - Prueba T para muestras independientes

- Incluimos la variable cuantitativa en el cuadro Variables de prueba y la variable cualitativa en el cuadro Variable de agrupación (Figura 7.21). Es necesario indicar los grupos que deseamos comparar en la opción Definir grupos (Figura 7.21).

- La tabla Prueba de muestras independientes proporciona en la línea superior el P-valor y la estimación y un intervalo de confianza para la diferencia de las medias, según el método de Student asumiendo igualdad de varianzas, y los análogos para el método de Welch en la línea inferior en caso contrario (Figura 7.22).

- La primera columna de la tabla Prueba de muestras independientes recoge el P-valor para el contraste de igualdad de varianzas entre las dos muestras utilizando test de Levene. No obstante, en este punto aconsejamos tener presente las consideraciones efectuadas en el Capítulo 5, y más concretamente en el esquema de la Tabla 5.1.

- Cuando el resultado de aplicar el test de Student o el test de Welch sea significativo, podremos extrapolar a la población la relación entre la variable cuantitativa y la variable cualitativa en el sentido observado en la muestra a partir de los métodos de Estadística Descriptiva.

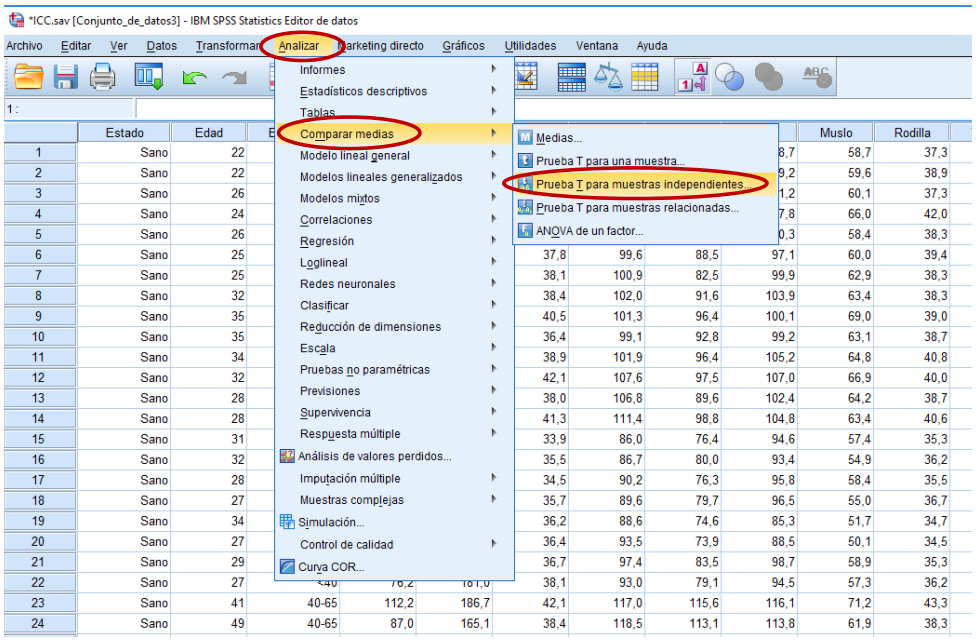


Figura 7.20: Test de Student y test de Welch para muestras independientes.

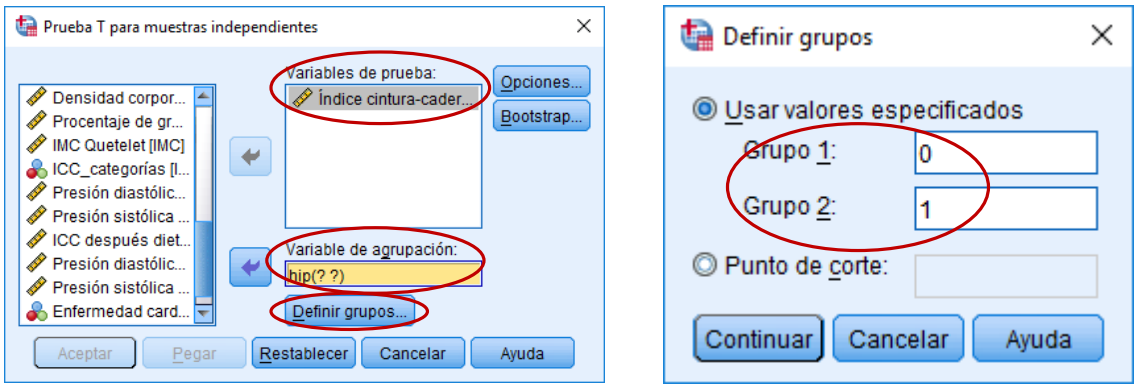


Figura 7.21: Test de Student y test de Welch para muestras independientes.

Prueba de muestras independientes

		Prueba de Levene de calidad de varianzas		prueba t para la igualdad de medias					
									95% de intervalo de confianza de la diferencia
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Índice cintura-cadera	Se asumen varianzas iguales	,018	,893	-9,333	250	,000	-,06181	,00662	-,07485	-,04877
	No se asumen varianzas iguales			-9,271	194,182	,000	-,06181	,00667	-,07496	-,04866

Figura 7.22: Test de Student y test de Welch para muestras independientes.

7.3.2. Test de Mann-Whitney

Constituye la alternativa no paramétrica a aplicar cuando no se verifican las condiciones de validez para los tests anteriores (véase Figura 4.4).

Para aplicar el test de Mann-Whitney utilizaremos la opción Muestras independientes del menú Pruebas no paramétricas (ver Figura 7.23).

Analizar - Pruebas no paramétricas - Muestras independientes

- En la pestaña Campos introducimos la variable cuantitativa en el cuadro Campos de prueba y la variable cualitativa en el cuadro Grupos (Figura 7.24).

- En la pestaña Configuración marcamos Personalizar pruebas y seleccionamos U de Mann-Whitney (2 muestras) (Figura 7.25).

- En la tabla resultante Resumen de contrastes de hipótesis se recoge el P-valor para el test de Mann-Whitney (Figura 7.26).

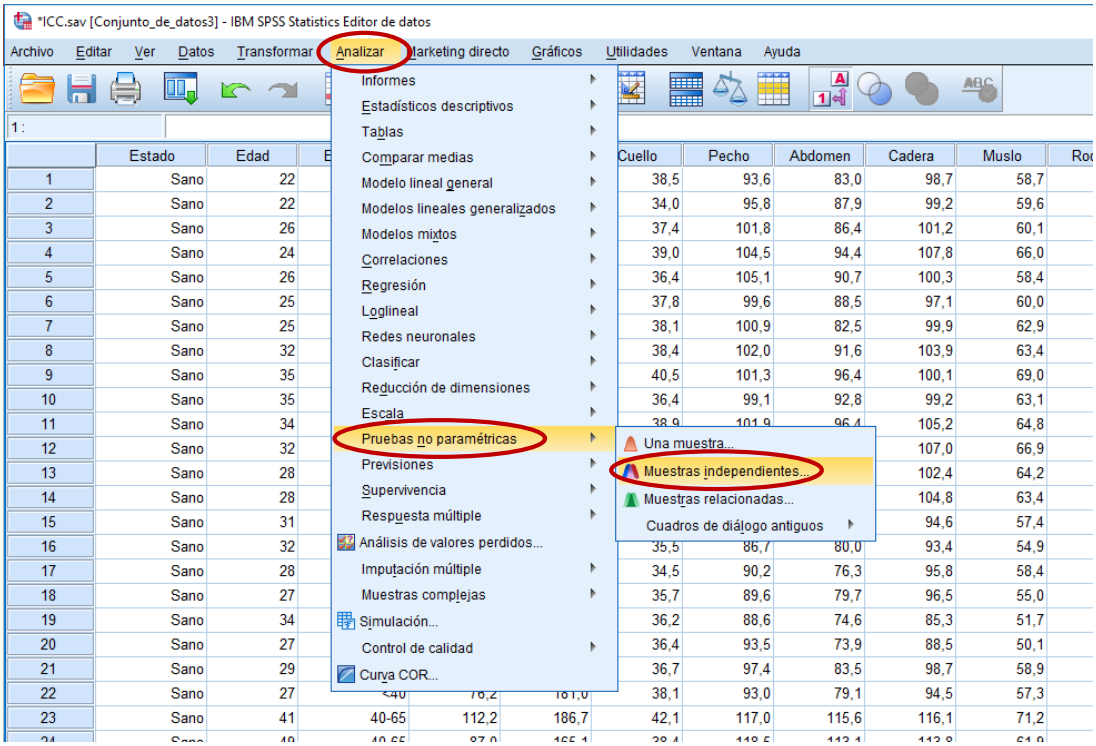


Figura 7.23: Test de Mann-Whitney.

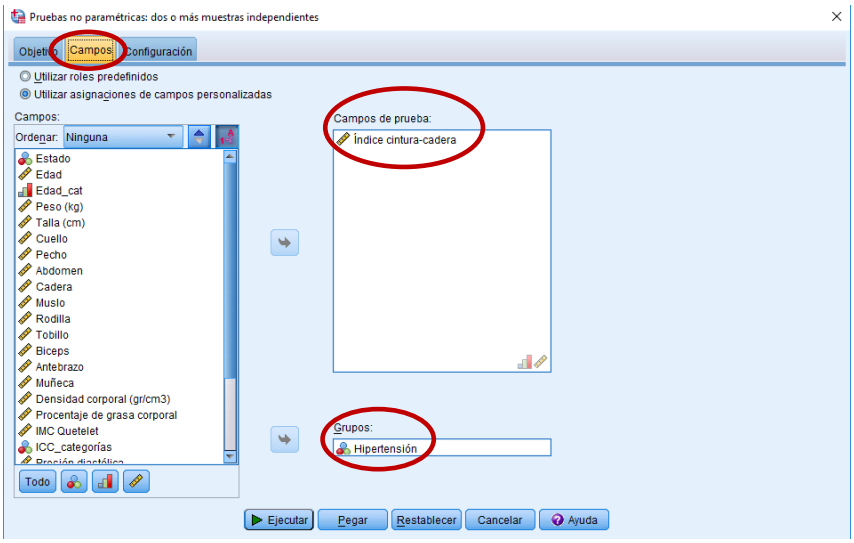


Figura 7.24: Test de Mann-Whitney.

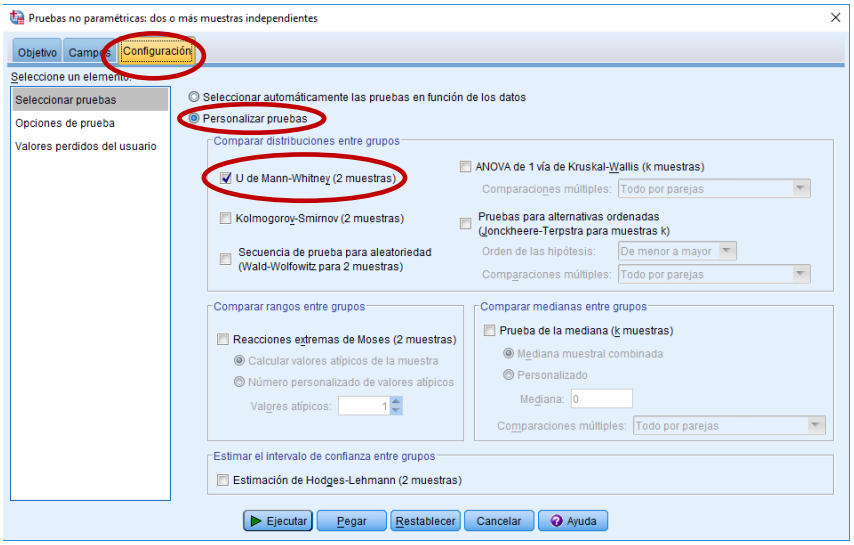


Figura 7.25: Test de Mann-Whitney.

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de Índice cintura-cadera es la misma entre las categorías de Hipertensión.	Prueba U de Mann-Whitney para muestras independientes	,000	Rechace la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significancia es ,05.

Figura 7.26: Test de Mann-Whitney.

7.4. Test de comparación de medias para muestras apareadas

En esta sección nos centraremos en comparar las medias de dos variables que resultan de sendas mediciones efectuadas sobre los mismos individuos (o individuos gemelos). Lo más habitual es que una de ellas sea la medición de un carácter antes de aplicar una técnica o tratamiento a cada sujeto y que la otra corresponda con la medición del mismo carácter después del mismo. Estudiaremos dos procedimientos: el test de Student para muestras relacionadas (test paramétrico) y el test de Wilcoxon (no paramétrico).

Por ejemplo, a partir del archivo `Ensayo clinico.sav` estudiemos si existen diferencias entre la presión sistólica antes (`pas_ini`) y después del tratamiento (`pas_fin`).

7.4.1. Test de Student para muestras relacionadas

Haremos uso de la opción Prueba T para muestras relacionadas del menú Comparar medias (Figura 7.27).

Analizar - Comparar medias - Prueba T para muestras relacionadas

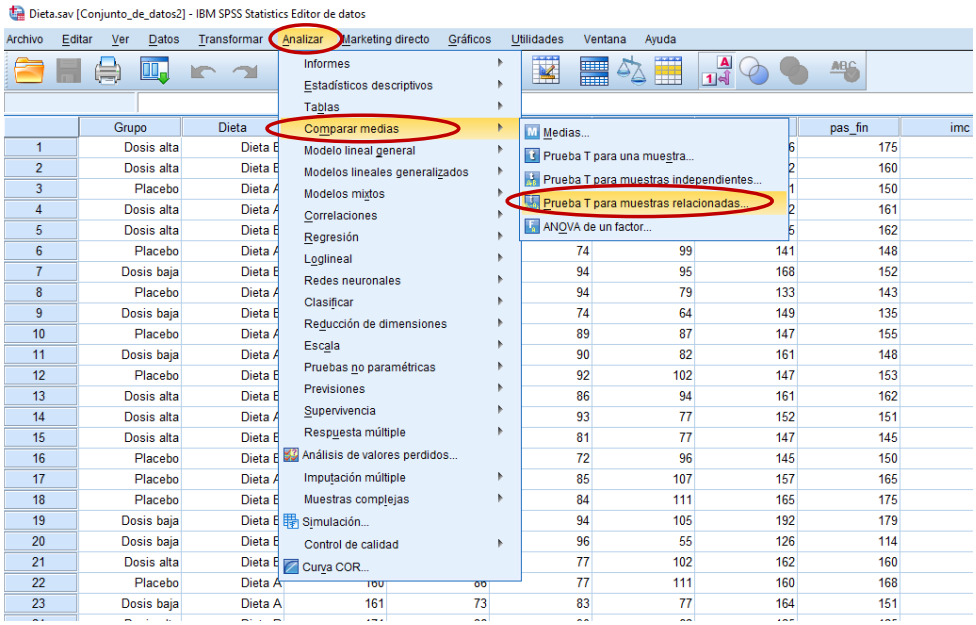


Figura 7.27: Test de Student para muestras relacionadas.

- Introducimos las variables correspondientes a las mediciones antes y después del tratamiento en el cuadro de Variables emparejadas (Figura 7.28).
- La tabla Correlaciones de muestras emparejadas proporciona el coeficiente de correlación lineal muestral entre los valores iniciales y finales, junto con el P-valor para el test de correlaciones de Pearson para las dos variables consideradas, que no es exactamente lo que queremos (Figura 7.29).

- La tabla Prueba de muestras emparejadas proporciona una estimación de la diferencia de medias y un intervalo de confianza al 95 % para la misma, así como el P-valor correspondiente al test de Student para muestras relacionadas (Figura 7.30).

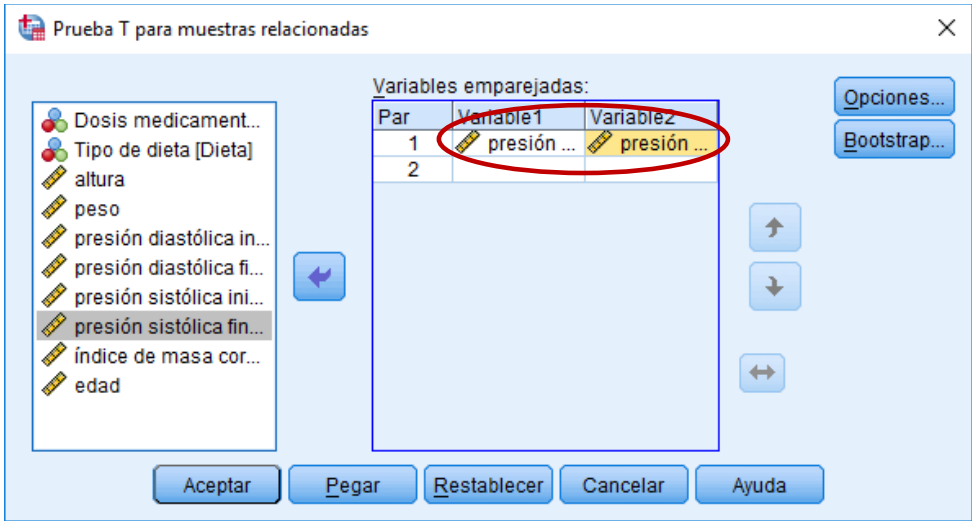


Figura 7.28: Test de Student para muestras relacionadas.

Correlaciones de muestras emparejadas

		N	Correlación	Sig.
Par 1	presión sistólica inicial & presión sistólica final	100	,843	,000

Figura 7.29: Test de Student para muestras relacionadas.

Prueba de muestras emparejadas

		Diferencias emparejadas					t	gl	Sig. (bilateral)
		Media	Desviación estándar	Media de error estándar	95% de intervalo de confianza de la diferencia Inferior	Superior			
Par 1	presión sistólica inicial - presión sistólica final	2,660	9,023	,902	,870	4,450	2,948	99	,004

Figura 7.30: Test de Student para muestras relacionadas.

7.4.2. Test de Wilcoxon

Lo emplearemos cuando no se verifiquen las condiciones de validez del test de Student para muestras relacionadas. Lo aplicaremos a través de la opción Muestras relacionadas del menú Pruebas no paramétricas (Figura 7.31).

Analizar - Pruebas no paramétricas - Muestras relacionadas

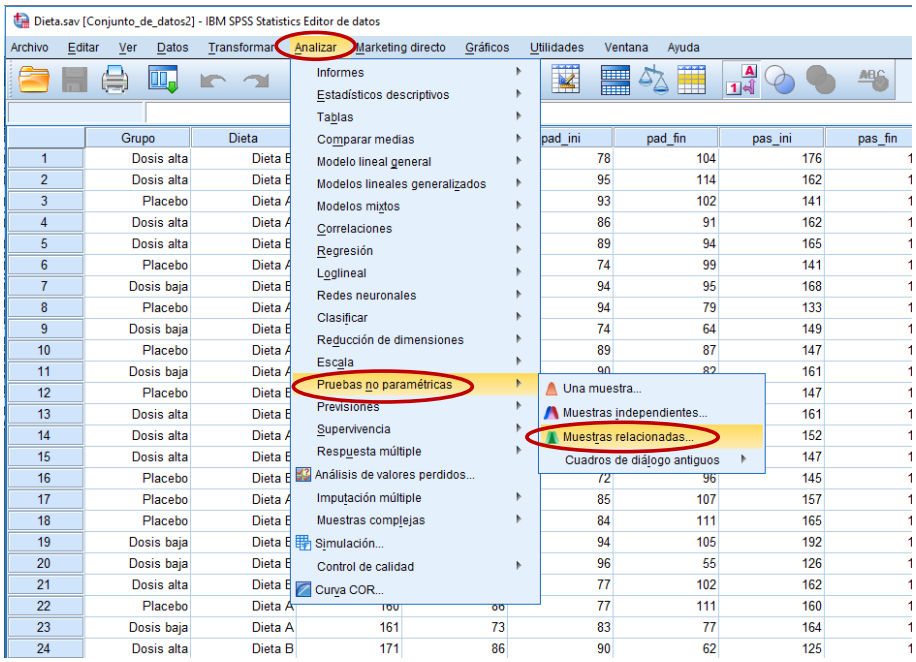


Figura 7.31: Test de Wilcoxon para muestras dependientes.

- En la pestaña Campos, en el cuadro Campos de prueba añadimos las dos variables cuantitativas que estamos comparando (Figura 7.32).
- En la pestaña Configuración marcamos Personalizar pruebas y Prueba de Wilcoxon de los rangos con signo para datos apareados (2 muestras) (Figura 7.33).
- En la tabla Resumen de contraste de hipótesis se indica el P-valor del test de Wilcoxon y la correspondiente decisión (Figura 7.34).

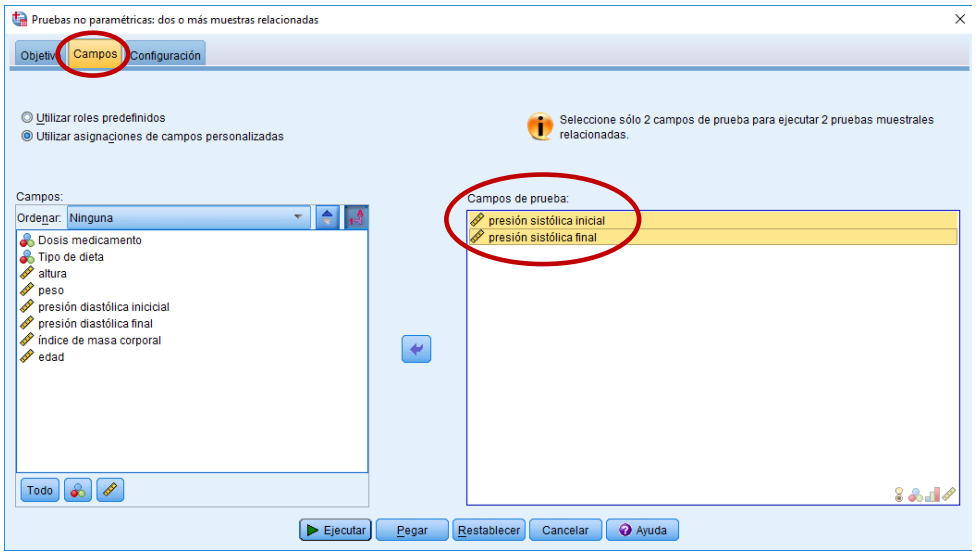


Figura 7.32: Test de Wilcoxon para muestras dependientes.

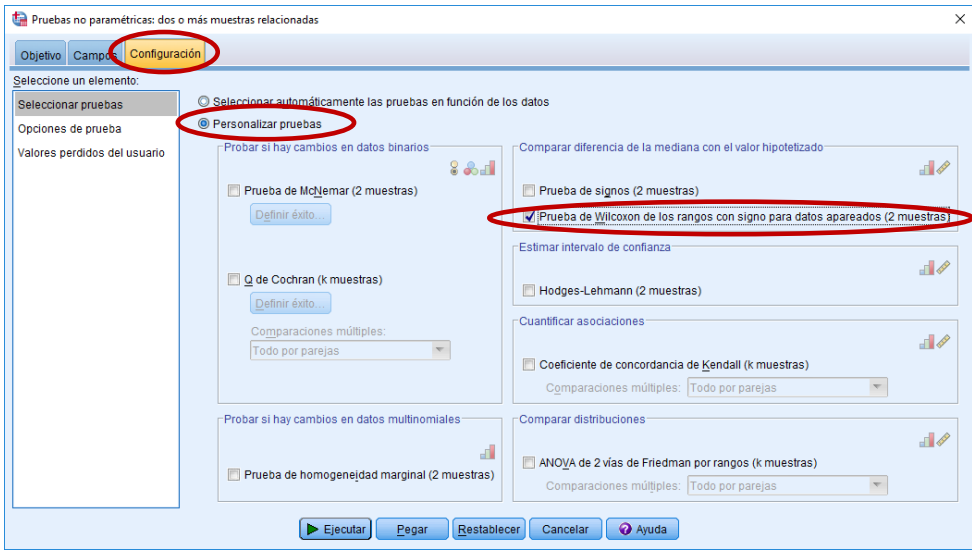


Figura 7.33: Test de Wilcoxon para muestras dependientes.

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La mediana de las diferencias entre presión sistólica inicial y presión sistólica final es igual a 0.	Prueba de Wilcoxon de los rangos con signo para muestras relacionadas	,005	Rechace la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significancia es ,05.

Figura 7.34: Test de Wilcoxon para muestras dependientes.

7.5. Anova de un factor y alternativa no paramétrica

En esta sección vemos cómo examinar la relación a nivel poblacional entre una variable cuantitativa y una variable cualitativa con más de dos categorías mediante el anova de una vía (test paramétrico) y el test de Kruskal-Wallis (test no paramétrico).

Por ejemplo, a partir del archivo `Diagnostico acidosis.sav` veamos si podemos extrapolar a la población la relación entre la acidosis en recién nacidos (`Tipo`) y la glucemia medida en el cordón umbilical (`Glucemia`).

7.5.1. Anova de una vía y comparaciones múltiples de Tukey

Utilizaremos la opción ANOVA de un factor del menú Comparar medias (Figura 7.35).

Analizar - Comparar medias - ANOVA de un factor

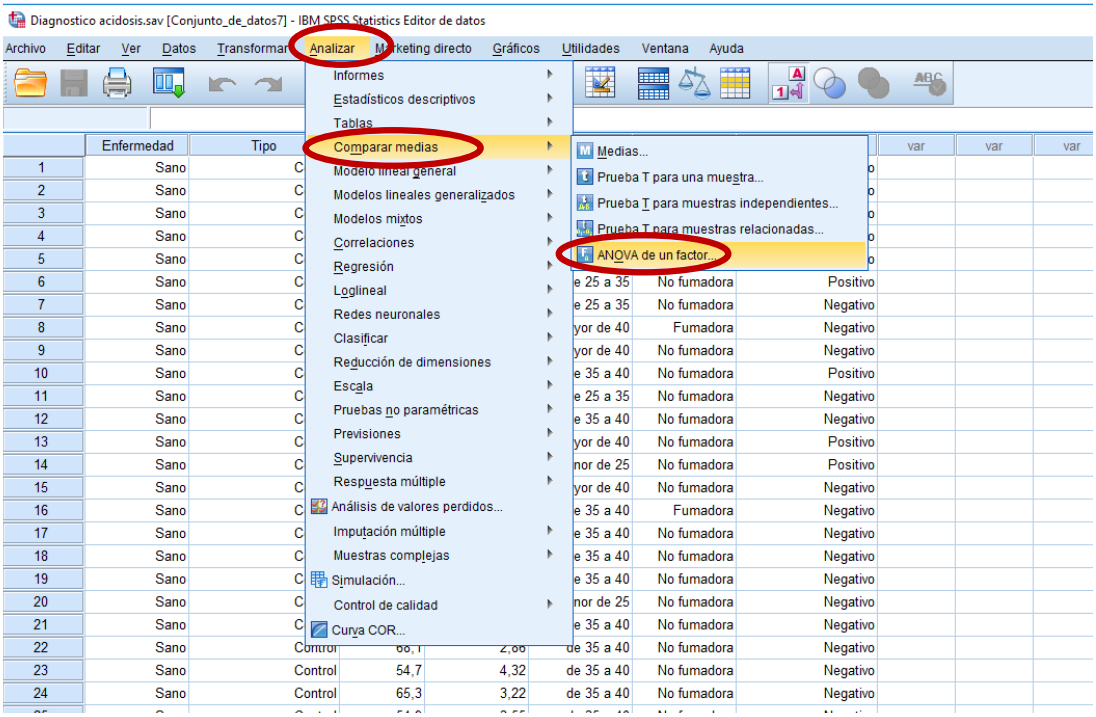


Figura 7.35: Anova de una vía.

- Introducimos la variable cualitativa en el cuadro Lista de dependientes y en el cuadro Factor introducimos la variable cualitativa (Figura 7.36).

- En la tabla ANOVA, obtenemos el P-valor para el test ANOVA de una vía (Figura 7.37). En el caso de que el resultado sea significativo, realizaremos las comparaciones múltiples de Tukey para examinar entre qué categorías (tipos de acidosis en este caso) existen diferencias significativas en cuanto a la variable cuantitativa (nivel de glucemia para este ejemplo).

- Para realizar las comparaciones múltiples, en el menú ANOVA de un factor, abrimos la opción Post hoc y marcamos Tukey (Figura 7.38).

- En la tabla de Comparaciones múltiples obtenemos los P-valores resultantes de la comparación de las medias por parejas (Figura 7.39).

- A continuación, aparece una tabla llamada Subconjuntos homogéneos, que en nuestro caso nos proporciona los distintos subconjuntos que se pueden distinguir para el nivel de glucemia en el cordón umbilical junto con las categorías del factor correspondientes (Figura 7.40).

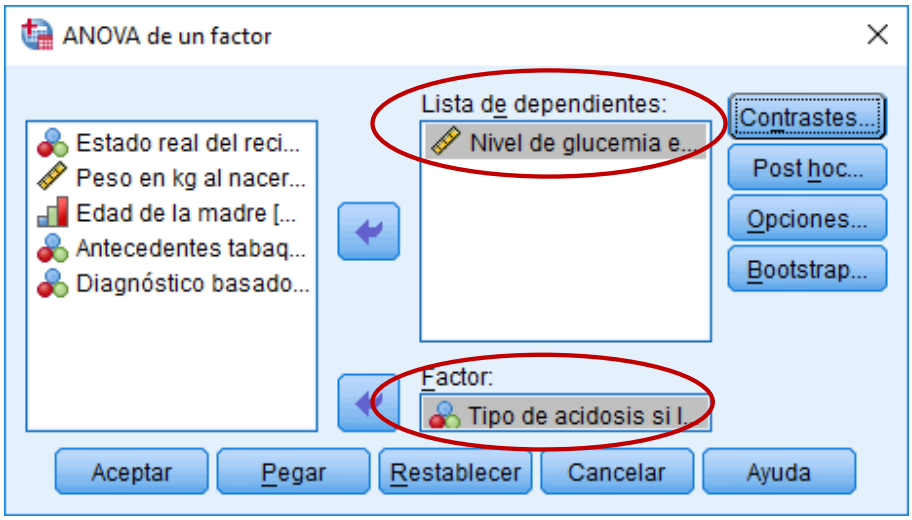


Figura 7.36: Anova de una vía.

ANOVA

Nivel de glucemia en el cordón umbilical

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	9124,624	3	3041,541	65,217	,000
Dentro de grupos	9140,844	196	46,637		
Total	18265,468	199			

Figura 7.37: Anova de una vía.

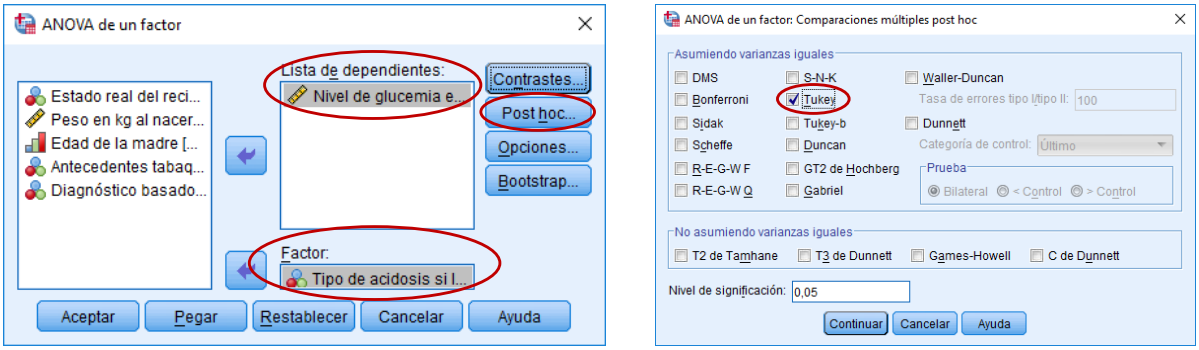


Figura 7.38: Anova de una vía. Comparaciones múltiples de Tukey.

Comparaciones múltiples

Variable dependiente: Nivel de glucemia en el cordón umbilical

HSD Tukey

(I) Tipo de acidosis si la hubiera	(J) Tipo de acidosis si la hubiera	Diferencia de medias (I-J)	Error estándar	Sig.	95% de intervalo de confianza	
					Límite inferior	Límite superior
Control	Acidosis Respiratoria	-8,7028*	1,3658	,000	-12,242	-5,164
	Acidosis Metabólica	-16,1243*	1,3658	,000	-19,663	-12,585
	Acidosis Mixta	,0687	1,3658	1,000	-3,470	3,608
Acidosis Respiratoria	Control	8,7028*	1,3658	,000	5,164	12,242
	Acidosis Metabólica	-7,4215*	1,3658	,000	-10,961	-3,882
	Acidosis Mixta	8,7715*	1,3658	,000	5,232	12,311
Acidosis Metabólica	Control	16,1243*	1,3658	,000	12,585	19,663
	Acidosis Respiratoria	7,4215*	1,3658	,000	3,882	10,961
	Acidosis Mixta	16,1930*	1,3658	,000	12,654	19,732
Acidosis Mixta	Control	-,0687	1,3658	1,000	-3,608	3,470
	Acidosis Respiratoria	-8,7715*	1,3658	,000	-12,311	-5,232
	Acidosis Metabólica	-16,1930*	1,3658	,000	-19,732	-12,654

*. La diferencia de medias es significativa en el nivel 0.05.

Figura 7.39: Anova de una vía. Comparaciones múltiples de Tukey.

Nivel de glucemia en el cordón umbilical

HSD Tukey[a]

Tipo de acidosis si la hubiera	N	Subconjunto para alfa = 0.05		
		1	2	3
Acidosis Mixta	50	62,611		
Control	50	62,679		
Acidosis Respiratoria	50		71,382	
Acidosis Metabólica	50			78,804
Sig.		1,000	1,000	1,000

Se visualizan las medias para los grupos en los subconjuntos homogéneos.

a. Utiliza el tamaño de la muestra de la media armónica = 50,000.

Figura 7.40: Anova de una vía. Comparaciones múltiples de Tukey.

7.5.2. Test de Kruskal-Wallis

Aplicaremos esta alternativa en el caso de que no se verifiquen las condiciones de validez para aplicar el anova. Consideraremos la opción Muestras independientes de Pruebas no paramétricas (Figura 7.41).

Analizar - Pruebas no paramétricas - Muestras independientes

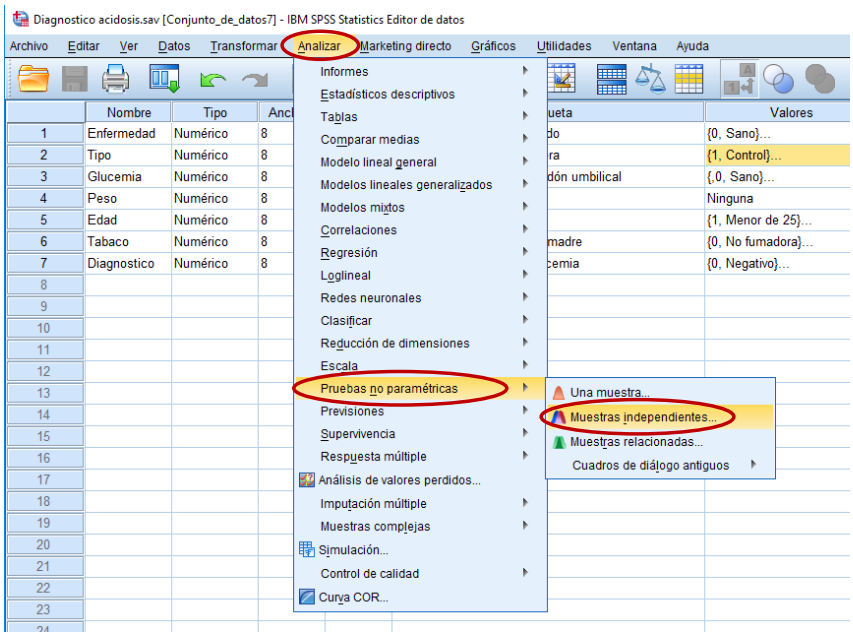


Figura 7.41: Test de Kruskal-Wallis.

- En la pestaña Campos, en el cuadro Campos de prueba añadimos la variable cuantitativa y en el cuadro Grupo añadimos la variable cualitativa (Figura 7.42).

- En la pestaña Configuración marcamos Personalizar pruebas y ANOVA de una vía de Kruskal-Wallis (k muestras) y en Comparaciones múltiples marcamos Todo por parejas (Figura 7.43).

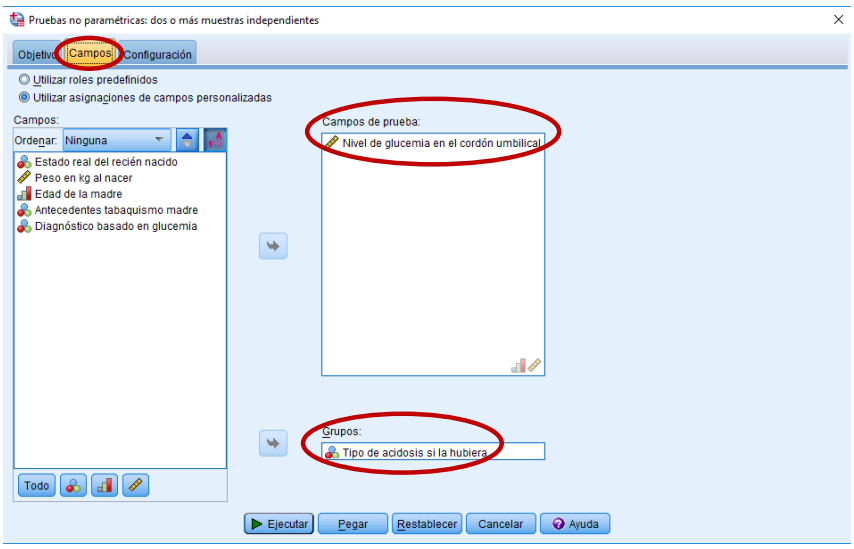


Figura 7.42: Test de Kruskal-Wallis.

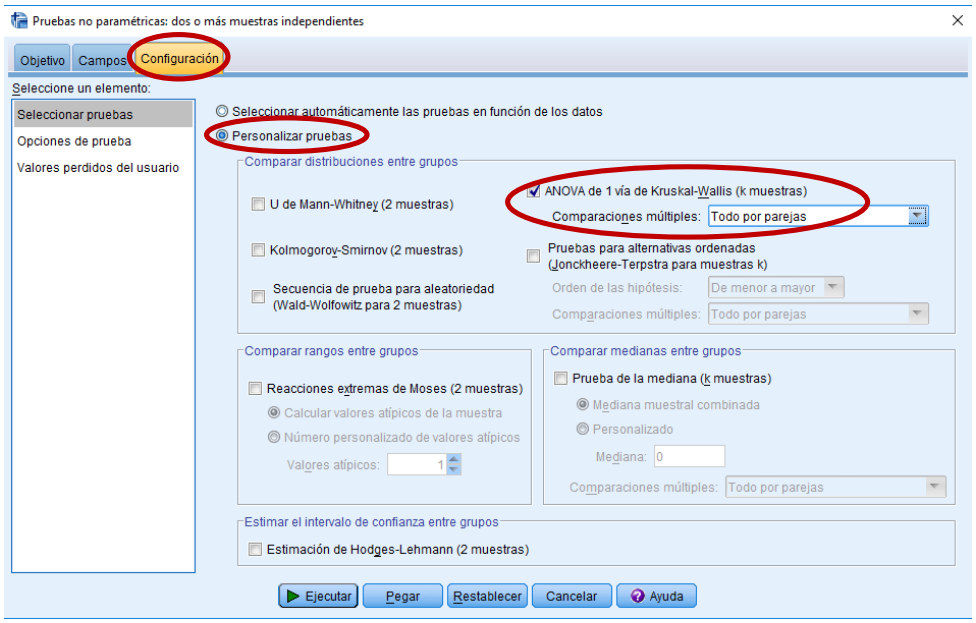


Figura 7.43: Test de Kruskal-Wallis.

- En la tabla Resumen de contraste de hipótesis se indica el P-valor del test de Kruskal-Wallis y la correspondiente decisión (Figura 7.44).

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de Nivel de glucemia en el cordón umbilical es la misma entre las categorías de Tipo de acidosis si la hubiera.	Prueba de Kruskal-Wallis para muestras independientes	,000	Rechace la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significancia es ,05.

Figura 7.44: Test de Kruskal-Wallis.

- Si el resultado de dicho test es significativo, analizaremos las comparaciones múltiples. Para ello hacemos doble click sobre la tabla de la Figura 7.44 y obtendremos un resumen más ampliado del test anterior. Para poder examinar las comparaciones múltiples, en el cuadro Vista seleccionaremos Comparaciones por parejas (Figura 7.45).

- Como resultado de esta acción, se muestra un poliedro cuyos vértices representan los distintos grupos (Figura 7.46). Los vértices que estén más próximos entre sí indicarán los grupos entre los que hay una menor diferencia en cuanto a la variable cuantitativa. La tabla obtenida muestra los P-valores obtenidos al realizar las comparaciones múltiples (Figura 7.46).

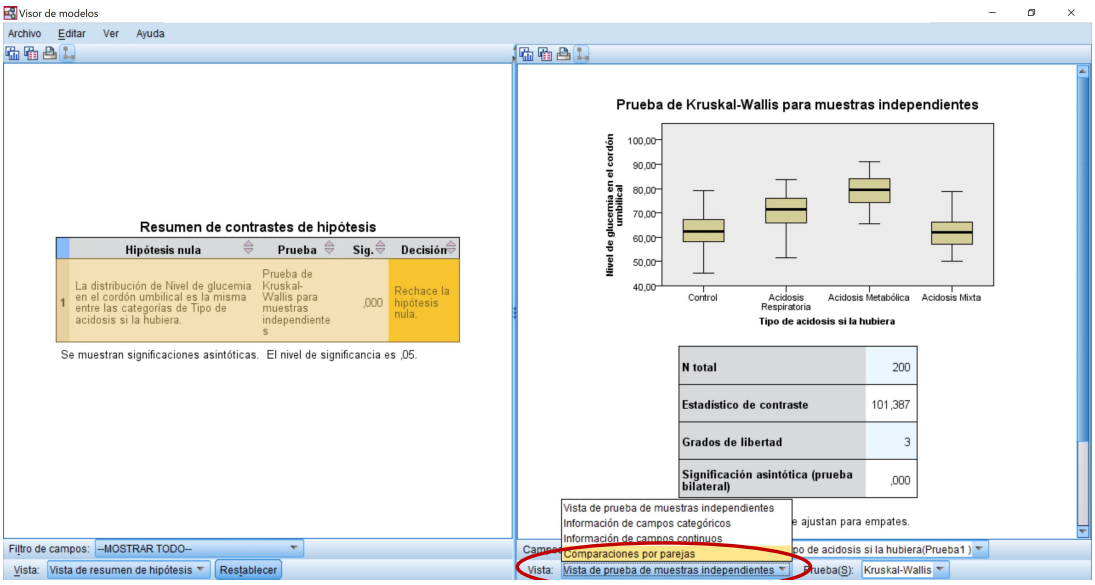


Figura 7.45: Test de Kruskal-Wallis. Comparaciones múltiples.

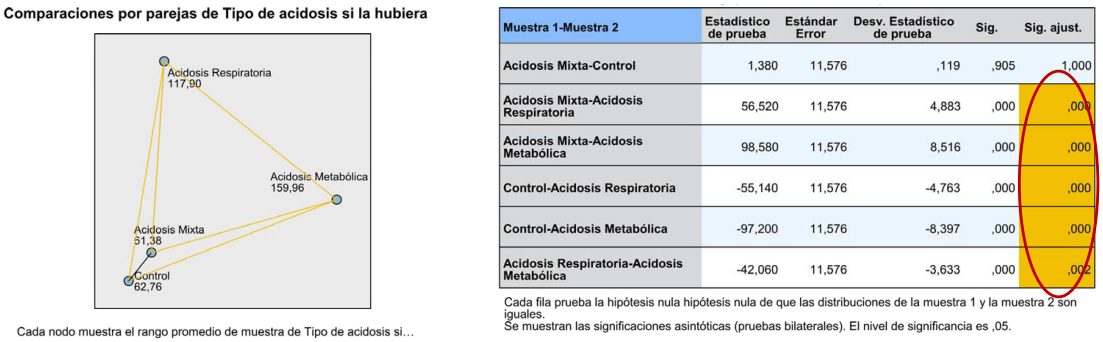


Figura 7.46: Test de Kruskal-Wallis. Comparaciones múltiples.

7.6. Relación entre dos variables cualitativas

La presente sección es la continuación de la Sección 6.5 del capítulo previo. Estudiaremos cómo aplicar los test de hipótesis que nos permitirán concluir si la relación observada entre dos variables cualitativas en la muestra puede extrapolarse a la población: el test χ^2 (paramétrico) y el test exacto de Fisher (no paramétrico). De nuevo, haremos uso del menú Tablas cruzadas de Estadísticos descriptivos (Figura 7.47).

Analizar - Estadísticos descriptivos - Tablas cruzadas

Por ejemplo, a partir del archivo `ICC.sav` veamos si la relación observada en la muestra entre la hipertensión (`hip`) y el tipo de ICC (`ICC_cat`) puede extrapolarse a la población.

7.6.1. Test χ^2

Para aplicar el test χ^2 el proceso es similar que para representar la tabla de contingencia.

- Introducimos una variable cualitativa en el cuadro Filas y otra en el cuadro Columnas (Figura 7.48).
- En la opción Estadísticos, marcamos Chi-cuadrado (Figura 7.48).
- El P-valor del test χ^2 aparece en la tabla Pruebas de chi-cuadrado. Además, al final de dicha tabla se indica si se verifican las condiciones de validez para dicho test, concretamente: que ningún valor esperado sea inferior a 1 y, a lo sumo un 20 % inferior a 5 (Figura 7.49).

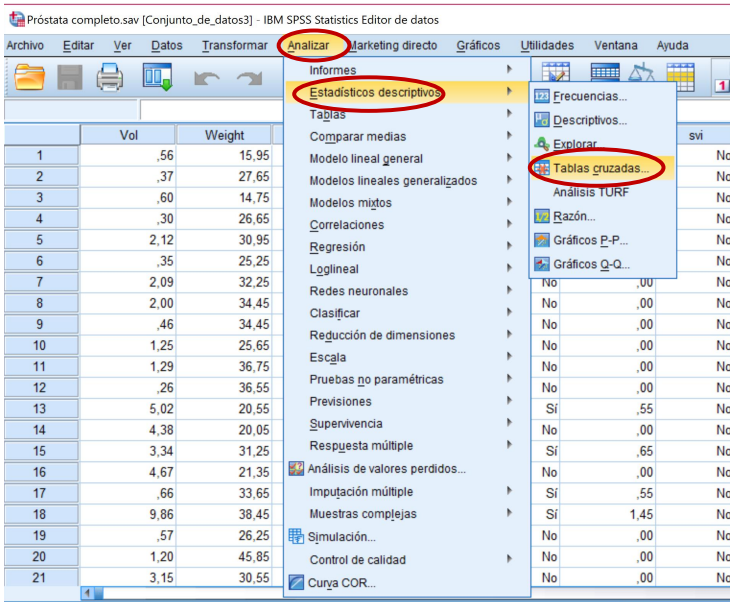


Figura 7.47: Relación entre dos variables cualitativas.

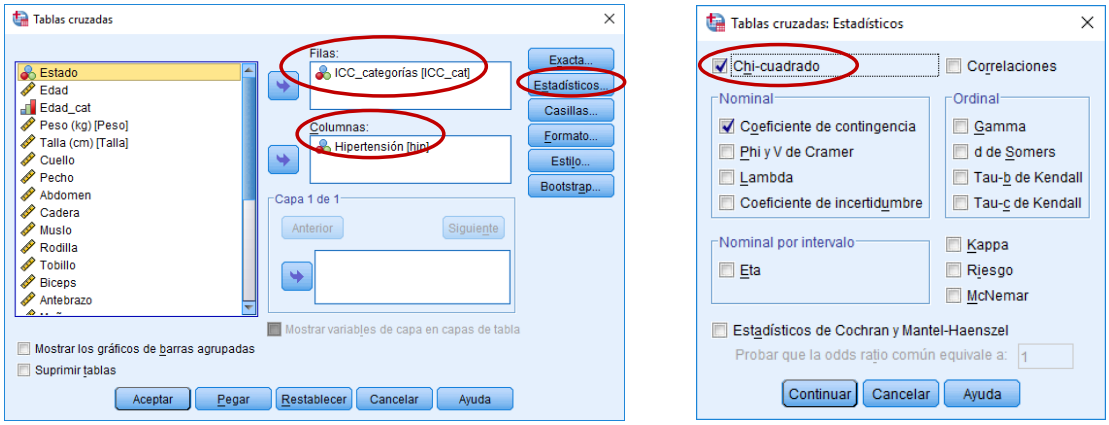


Figura 7.48: Relación entre dos variables cualitativas: test χ^2.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (2 caras)	Significación exacta (2 caras)	Significación exacta (1 cara)
Chi-cuadrado de Pearson	77,123[a]	1	,000		
Corrección de continuidad[b]	74,790	1	,000		
Razón de verosimilitud	79,286	1	,000		
Prueba exacta de Fisher				,000	,000
Asociación lineal por lineal	76,817	1	,000		
N de casos válidos	252				

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 36,19.

b. Sólo se ha calculado para una tabla 2x2

Figura 7.49: Relación entre dos variables cualitativas: test χ^2.

7.6.2. Test exacto de Fisher

El proceso para aplicar este test, que sólo se proporciona para tablas 2×2, es exactamente igual al anterior. En este caso, al marcar Chi-cuadrado en la opción Estadísticos se obtiene una nueva fila en la tabla Pruebas de chi-cuadrado, con el P-valor para el test exacto de Fisher. Además, al final de dicha tabla se indica si se verifican las condiciones de validez para el test χ^2 (Figura 7.50).

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (2 caras)	Significación exacta (2 caras)	Significación exacta (1 cara)
Chi-cuadrado de Pearson	15,812[a]	1	,000		
Corrección de continuidad[b]	11,049	1	,001		
Razón de verosimilitud	15,446	1	,000		
Prueba exacta de Fisher				,001	,001
Asociación lineal por lineal	15,059	1	,000		
N de casos válidos	21				

a. 3 casillas (75,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es ,95.

b. Sólo se ha calculado para una tabla 2x2

Figura 7.50: Relación entre dos variables cualitativas: test exacto de Fisher.

7.6.3. Problemas de comparación de proporciones

Para abordar el problema de decidir si dos o más proporciones son o no iguales, existen varias alternativas:

- Si la variable cualitativa es binaria, podemos tratarla como una variable cuantitativa tomando los valores 0 y 1, de forma análoga a como lo hicimos en la Subsección 7.1.2. En este caso, utilizaremos el test de Student para muestras independientes (siempre y cuando el tamaño muestral sea lo suficientemente grande).

- La segunda opción es entender el problema como un estudio de relación entre dos variables cualitativas y aplicar los métodos de la presente sección. Notemos que esta opción es válida tanto con dos como con más proporciones.

7.7. Anova de dos factores

Para aplicar un anova con dos (o más factores) haremos uso del menú Modelo lineal univariante (Figura 7.51)

Analizar - Modelo Lineal General - Univariante

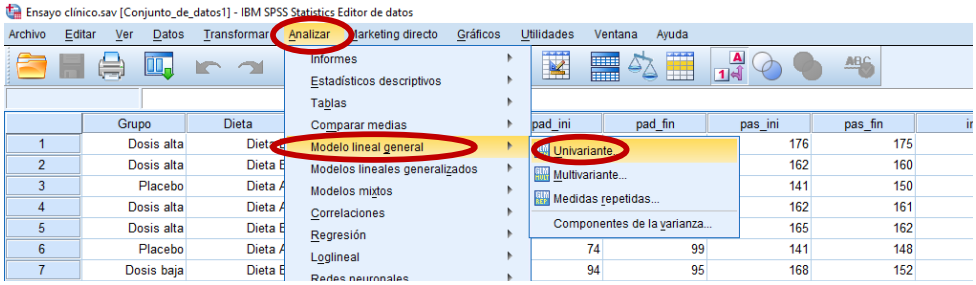


Figura 7.51: Anova de dos factores.

Realmente, este es un menú muy general que nos permitiría aplicar el test de Student para muestras independientes, el anova de uno y dos factores, las regresiones lineales simple y múltiple y el análisis de la covarianza. En este caso vamos a ejecutar el análisis descrito en el Ejemplo 14 para el archivo `Ensayo clinico.sav`.

- Tal y como se indica en la Figura 7.52, introducimos la variable cuantitativa, en este caso la `presión sistólica final`, como Variable dependiente, y las dos variables cualitativas, `Dieta` y `Grupo` (dosis), como Factores fijos.

- Tras hacer click en la pestaña Opciones conviene marcar Estadísticos descriptivos y Estimaciones del tamaño del efecto (Figura 7.52).

- En la pestaña Post hoc, desplazamos a la derecha los factores con más de dos categorías, si los hubiera (en nuestro caso `Grupo`) y marcamos la opción Tukey (Figura 7.53).

- Para obtener el gráfico de medias hacemos click en la opción de Gráficos, introducimos un factor en el Eje horizontal y otro como Líneas seperadas. Lo más aconsejable es que vaya como líneas separadas el de menos categorías, en nuestro caso `Dieta` (Figura 7.53). Es importante no olvidar pulsar en Añadir.

Las salidas que se muestran en las Figuras 7.54 y 7.55 son las que se comentan en el Ejemplo 14, y el gráfico solicitado es el que aparece en la Figura 5.1.

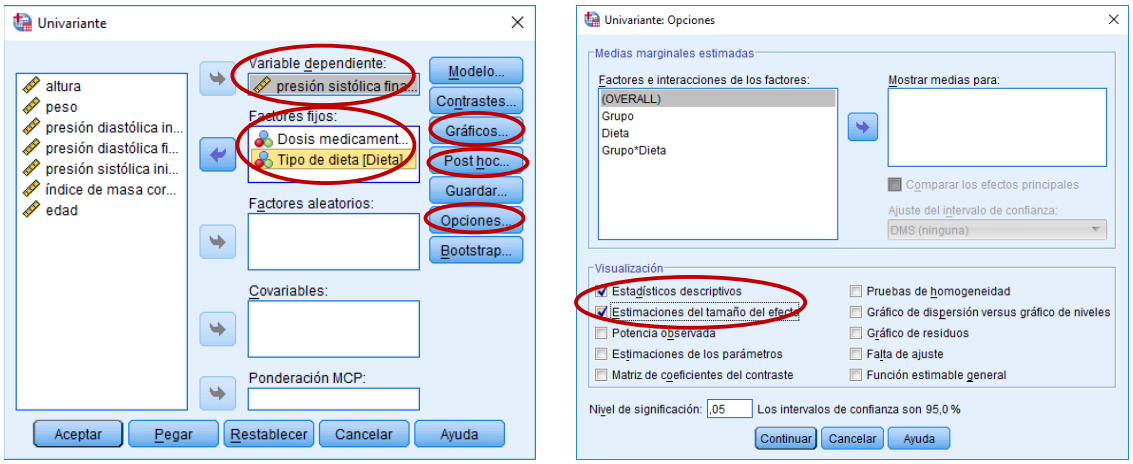


Figura 7.52: Anova de dos factores.

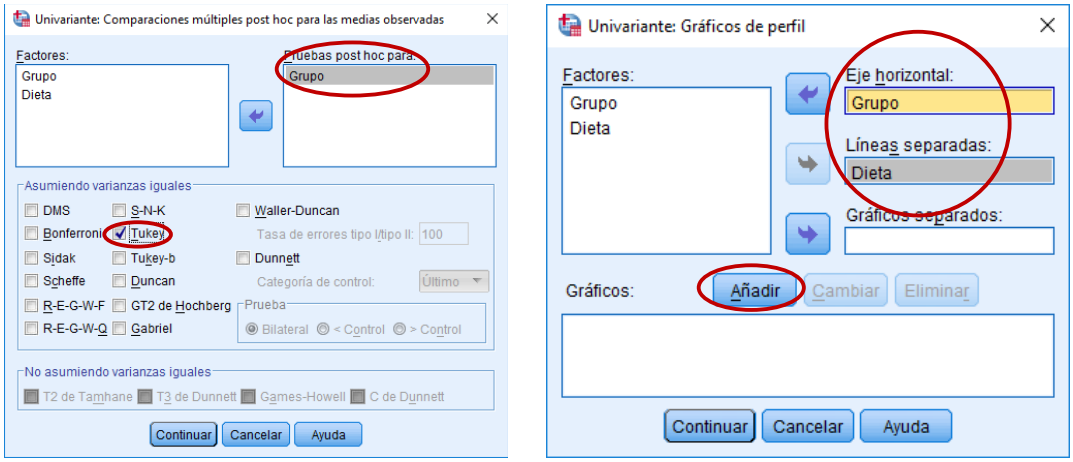


Figura 7.53: Anova de dos factores.

Pruebas de efectos inter-sujetos

Variable dependiente: presión sistólica final

Origen	Tipo III de suma de cuadrados	gl	Cuadrático promedio	F	Sig.	Eta parcial al cuadrado
Modelo corregido	6847,440[a]	5	1369,488	6,265	,000	,250
Interceptación	2232332,306	1	2232332,306	10212,845	,000	,991
Grupo	5231,869	2	2615,935	11,968	,000	,203
Dieta	293,685	1	293,685	1,344	,249	,014
Grupo * Dieta	1287,796	2	643,898	2,946	,057	,059
Error	20546,600	94	218,581			
Total	2273196,000	100				
Total corregido	27394,040	99				

a. R al cuadrado = ,250 (R al cuadrado ajustada = ,210)

Figura 7.54: Anova de dos factores.

presión sistólica final

HSD Tukey[a,b,c]

Dosis medicamento	N	Subconjunto 1	Subconjunto 2
Dosis baja	34	140,18	
Dosis alta	33		152,27
Placebo	33		157,42
Sig.		1,000	,334

Figura 7.55: Anova de dos factores.

7.8. Regresión logística binaria

Para aplicar un análisis de regresión logística, si tenemos la intención de explicar la probabilidad de que ocurra o no un determinado evento a partir los resultados de ciertas variables explicativas, debemos acceder al menú correspondiente, según la Figura 7.56, por la vía

Analizar - Regresión - Logística binaria

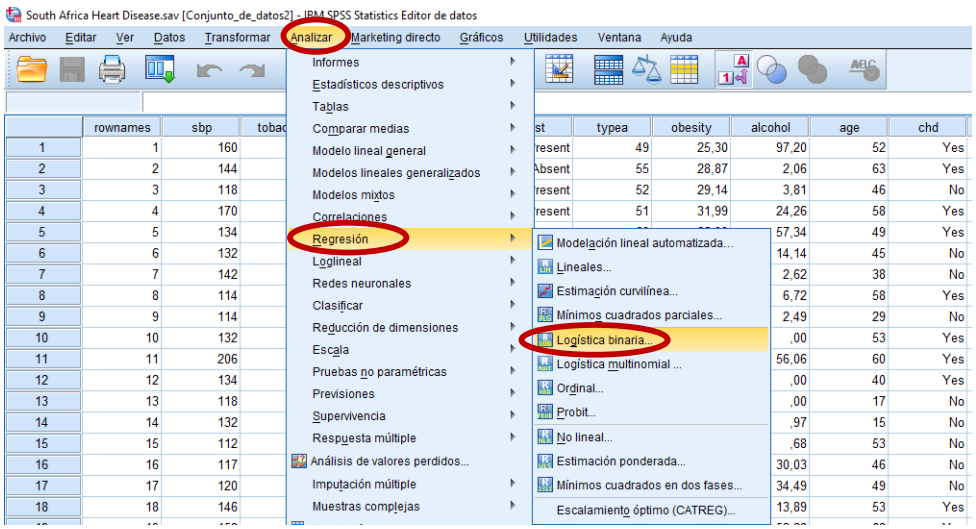


Figura 7.56: Regresión logística.

Lo aplicaremos a los datos de `Southafrica Heart Disease.sav` para explicar el infarto a partir de las variables `sbp`, `tobacco`, `ldl`, `famhist` y `obesity`, como se indica en el Ejemplo 15.

- Una vez dentro (Figura 7.57) se introduce el infarto en variable Dependiente y el resto como Covariables. La variable que se introduce como dependiente debe ser cualitativa binaria. Las covariables pueden ser tanto numéricas como cualitativas, pero en el último caso debemos especificarlo a través de la pestaña Categórica... incluyendo las variables cualitativas en el cuadro Covariables categóricas (Figura 7.58).

- Si deseamos obtener las probabilidades estimadas según el modelo, tanto para los datos ya estudiados como para otro individuo nuevo, debemos solicitarlo a través de la pestaña Guardar marcando Probabilidades y Grupo de pertenencia (Figura 7.58).

Figura 7.57: Regresión logística.

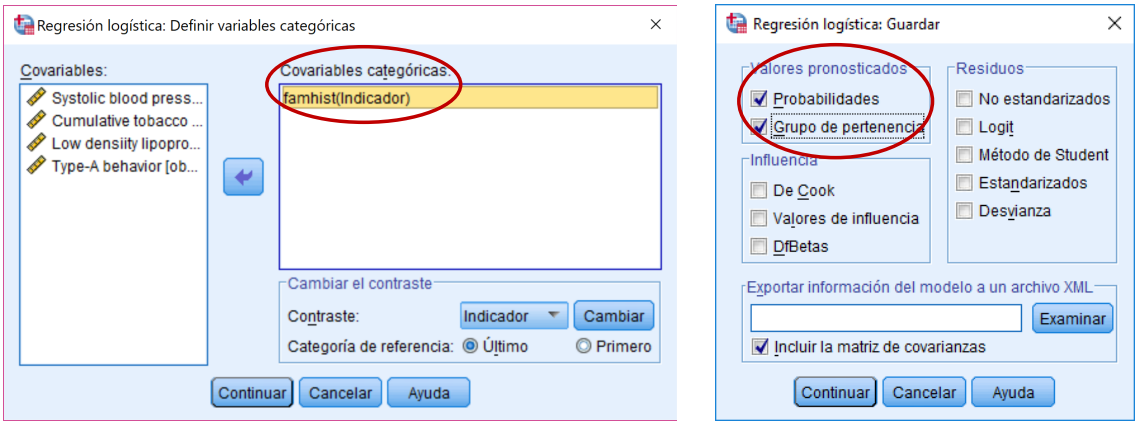


Figura 7.58: Regresión logística.

- En la tabla Resumen del modelo (Figura 7.59) podemos apreciar el valor obtenido para el coeficiente R^2 de Naglekerke, concretamente 0.255. No obstante, la Tabla de clasificación de la Figura 7.61 expresa con mayor claridad la capacidad del modelo para entender el comportamiento los propios datos de la muestra.

- En la tabla Variables en la ecuación (Figura 7.60) podemos apreciar las estimaciones de los coeficientes del modelo para cada una de las variables explicativas (segunda columna por la izquierda), los resultados de sus respectivos contrastes parciales (segunda columna por la derecha) y e elevado a los coeficientes (primera columna

por la derecha), que bajo ciertas condiciones puede interpretase como Odds Ratios. Así, por ejemplo, podemos concluir que la variable obesity no es esencial a la hora de explicar el infarto, justo lo contrario que la variable cualitativa famhist. Según los datos, tener antecedentes familiares incrementa fuertemente el riesgo de infarto.

Resumen del modelo

Escalón	Logaritmo de la verosimilitud -2	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	501,545[a]	,185	,255

a. La estimación ha terminado en el número de iteración 4 porque las estimaciones de parámetro han cambiado en menos de ,001.

Figura 7.59: Regresión logística.

Variables en la ecuación

		B	Error estándar	Wald	gl	Sig.	Exp(B)
Paso 1[a]	famhist(1)	-1,059	,219	23,351	1	,000	,347
	sbp	,013	,005	5,508	1	,019	1,013
	tobacco	,120	,025	23,218	1	,000	1,127
	ldl	,217	,057	14,475	1	,000	1,242
	obesity	-,022	,028	,615	1	,433	,978
	Constante	-2,784	,943	8,720	1	,003	,062

a. Variables especificadas en el paso 1: famhist, sbp, tobacco, ldl, obesity.

Figura 7.60: Regresión logística.

Tabla de clasificación[a]

			Pronosticado		
			Response, coronary heart disease		Corrección de porcentaje
	Observado		No	Yes	
Paso 1	Response, coronary heart disease	No	263	39	87,1
		Yes	85	75	46,9
	Porcentaje global				73,2

a. El valor de corte es ,500

Figura 7.61: Regresión logística.

7.9. Test de Kolmogorov-Smirnov

Dejamos intencionadamente para el final los tests de normalidad porque, al contrario de los expuestos hasta ahora, no responden a un problema de relación entre variables, sino que vienen a decidir si una variable concreta se ajusta o no aproximadamente a un modelo de distribución normal. Es bastante usual aplicar este tipo de test antes de resolver otros problemas, como los de comparación de medias. No obstante, aconsejamos tener siempre presente cuál es el objetivo del estudio, cuál es el tamaño de la muestra y qué aspecto gráfico (histograma) tienen las variables numéricas a estudiar (véase Tabla 5.1). En todo caso, existen diversos métodos para contrastar la hipótesis inicial de normalidad, aunque en este manual nos centraremos simplemente en cómo aplicar el test de Kolmogorov-Smirnov.

Por ejemplo, a partir del archivo `Southafrica Heart Disease.sav` veamos si la variable `adiposity` sigue una distribución normal.

La manera más sencilla es utilizar el menú Una muestra de Pruebas no paramétricas (Figura 7.62):

Analizar - Pruebas no paramétricas - Una muestra

- En el cuadro Campos de prueba de la pestaña Campos, extraemos todas las variables y dejamos únicamente la variable para la que queremos contrastar la normalidad, en nuestro caso, `adiposity` (Figura 7.63).

- A continuación, en la pestaña Configuración, marcamos Probar la distribución observada con el valor hipotetizado (prueba de Kolmogorov-Smirnov) (Figura 7.64).

- En la tabla Resumen de contrastes de hipótesis aparece el resumen del test aplicado, conteniendo el P-valor asociado al test y la correspondiente decisión (Figura 7.65).

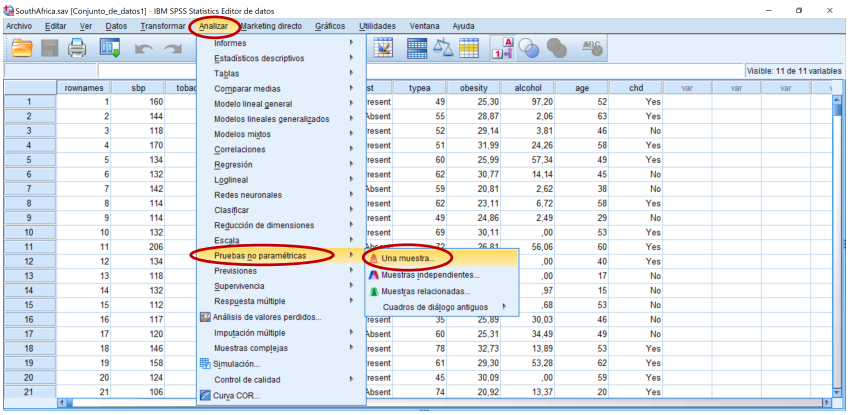


Figura 7.62: Contrastes de normalidad: test de Kolmogorov-Smirnov.

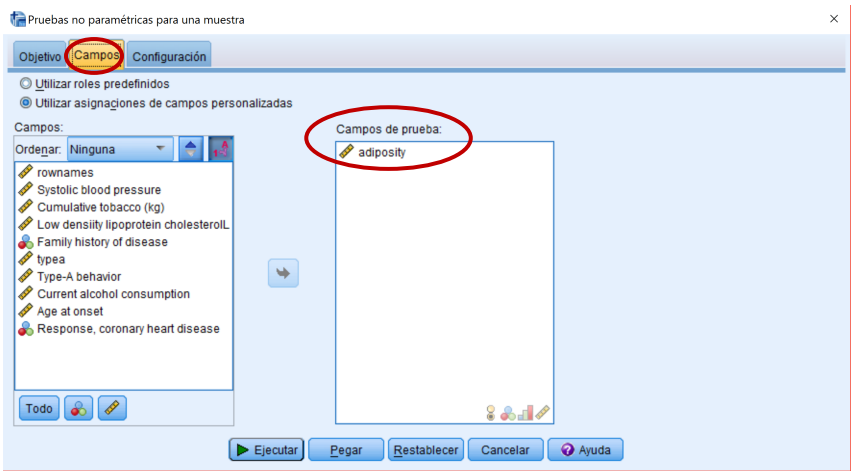


Figura 7.63: Contrastes de normalidad: test de Kolmogorov-Smirnov.

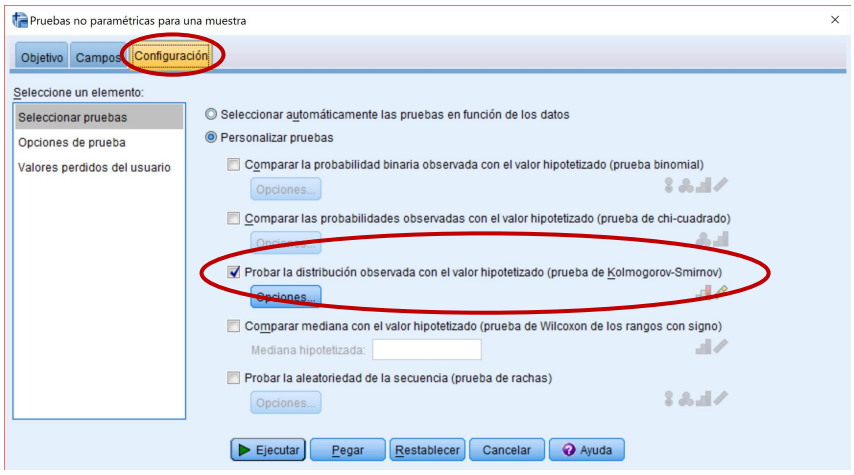


Figura 7.64: Contrastes de normalidad: test de Kolmogorov-Smirnov.

Resumen de contrastes de hipótesis

	Hipótesis nula	Prueba	Sig.	Decisión
1	La distribución de adiposity es normal con la media 25,407 y la desviación estándar 7,78.	Prueba de Kolmogorov-Smirnov para una muestra	,013[1]	Rechace la hipótesis nula.

Se muestran significaciones asintóticas. El nivel de significancia es ,05.

[1] Lilliefors corregido

Figura 7.65: Contrastes de normalidad: test de Kolmogorov-Smirnov.

El programa SPSS ofrece una forma alternativa para aplicar el test de Kolmogorov-Smirnov y el test Shapiro-Wilk a través del menú de Explorar (Figura 7.66).

Analizar - Estadísticos descriptivos - Explorar - Gráficos

- Seleccionamos la variable `adiposity` de la lista de variables y la introducimos en la Lista de dependientes (Figura 7.67).
- En la opción Gráficos marcamos Gráficos con pruebas de normalidad (Figura 7.67).
- Se obtiene la tabla de Pruebas de normalidad, donde aparecen el **valor experimental** y el **P-valor del test de Kolmogorov-Smirnov** y del **test de Shapiro-Wilk** (Figura 7.68).

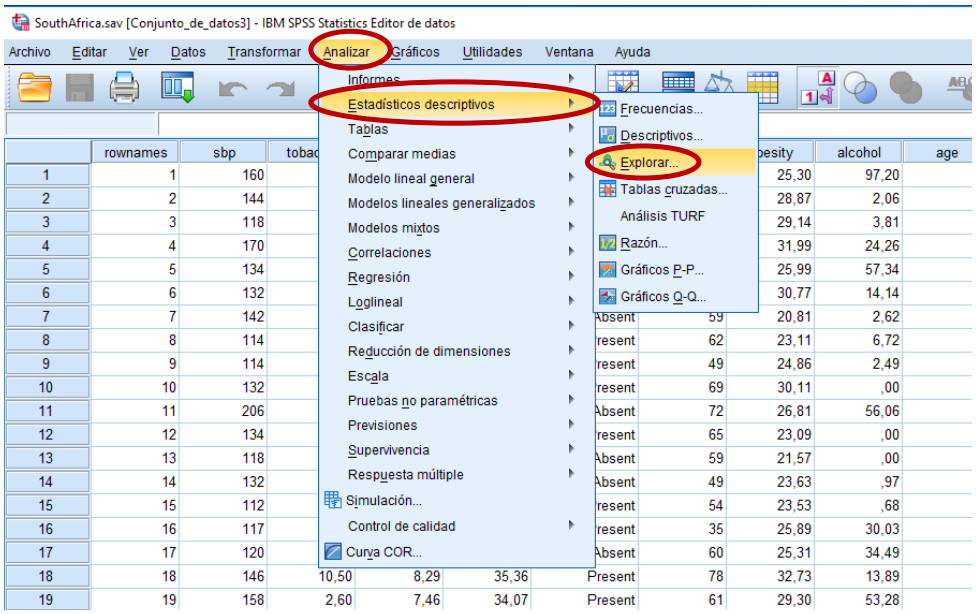


Figura 7.66: Contrastes de normalidad: tests de Kolmogorov-Smirnov y Shapiro-Wilk.

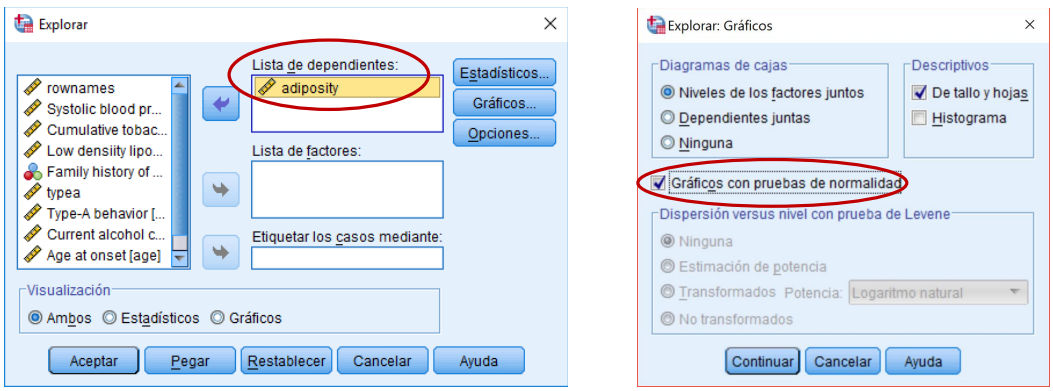


Figura 7.67: Contrastes de normalidad: tests de Kolmogorov-Smirnov y Shapiro-Wilk.

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
adiposity	,048	462	,013	,984	462	,000

a. Corrección de significación de Lilliefors

Figura 7.68: Contrastes de normalidad: tests de Kolmogorov-Smirnov y Shapiro-Wilk.

Es importante mencionar que esta última forma de contrastar la normalidad es por la que nos debemos decantar si deseamos contrastar la normalidad de una variable cuantitativa para cada una de las categorías de una variable cualitativa. En tal caso, añadiríamos la variable cualitativa a la Lista de factores y la tabla de Pruebas de normalidad proporcionaría los resultados de los tests mencionados para la variable cuantitativa en cada una de las categorías de la variable cualitativa.

Bibliografía

[1] Cobo, E., Muñoz, P., González, J.A., Bigorra, J., Corchero, C., Miras, F., Selva, A., y Videla, S. *Bioestadística para no estadísticos*. Elsevier Doyma, 2007.

[2] García Nogales, A. *Bioestadística Básica*. Abecedario, 2004.

[3] González-Ramírez, C., Montanero-Fernández, J. y Peral-Pacheco, D. A multifactorial study on duration of temporary disabilities in spain. *Archives of Environmental & Occupational Health*, pages 1–8, 2016. PMID: 27775491.

[4] Hospital Ramón y Cajal. Material docente de la unidad de bioestadística clínica. `http://www.hrc.es/bioest/M_docente.html#tema3`.

[5] Khaneman, D. *Pensar rápido, pensar despacio*. Debate, 2012.

[6] Macía Antón, A., Lubin, P., y Rubio de Lemus, P. *Psicología Matemática II*. UNED, 1997.

[7] Martín Andrés, A., y Luna del Castillo, J.D. 50 ± 10 *horas de Bioestadística*. Norma, 1995.

[8] Martín Andrés, A. y Luna del Castillo, J.D. *Bioestadística para Ciencias de la Salud*. Norma-Capitel, 2004.

[9] Martín González, M.A., Sánchez Villegas, A., Toledo Atucha, E.A., y Faulin Fajardo, J. (Eds.). *Bioestadística amigable*. Elsevier, 2014.

[10] Milton, J.S. *Estadística para Biología y Ciencias de la Salud*. McGraw-Hill/Interamericana de España, 2007.

[11] Montanero Fernández, J. Material docente sobre probabilidad e inferencia estadística. `http://matematicas.unex.es/~jmf/htm/material_enfermeria_medicina.html`.

[12] Montanero Fernández, J. *Modelos Lineales*. Sevicio de Publicaciones Universidad de Extremadura, 2008.

[13] Norman, G.R. y Streiner, D.L. *Bioestadística*. Mosby/Doyma Libros, 1996.

[14] Visauta Vinacua, B. *Análisis estadístico con SPSS para Windows: estadística multivariante*. McGraw-Hill, 1998.

[15] Wasserstein, R.L. y Lazar, N.A. The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

Índice alfabético