Manual de Bioestadística

Grado de Enfermería

Universidad de Extremadura

Jesús Montanero Fernández

Universidad de Extremadura
DEPARTAMENTO DE MATEMÁTICAS

Introducción

Este volumen pretende ser un breve manual de iniciación a la Estadística. En principio, está concebido como apoyo a la docencia de la asignatura Bioestadística del Grado en Enfermería, aunque puede resultar también útil para alumnos que cursan estudios en cualquier titulación relacionada con las Ciencias de la Salud.

En lo que respecta a la materia en sí, es un hecho notorio que una gran parte de las investigaciones en las Ciencias Experimentales depende, en gran medida, de métodos estadísticos. La demanda de la Estadística viene motivada por distintas causas, según la especialidad en cuestión. En el caso de las Ciencias de la Salud, el problema estriba en la enorme variabilidad con que se presentan los fenómenos estudiados, variabilidad que, lejos de reducirse, se incrementa con frecuencia a medida que se profundiza en la investigación. Ello impide la formulación de leyes deterministas, propias de otras disciplinas, en favor de una descripción, lo más amplia y exhaustiva posible, de los distintos caracteres a estudiar.

Hemos de hacer hincapié en la trascendencia que tienen la recogida y tratamiento de datos, con la idea de extraer la mayor información posible acerca del fenómeno a estudiar. ¿Cómo recoger los datos y cómo tratarlos? La respuesta a esta pregunta es la Estadística. La siguiente definición de Estadística es debida a Barlett: *"La Estadística es la Ciencia que nos indica el proceso a seguir en el tratamiento de la información en aquellas circunstancias que envuelva la incertidumbre".* Estudiemos primeramente cuatro nociones elementales:

Población

Es el objeto del estudio. Es un concepto bastante abstracto, aunque en el caso de la Ciencias de la Salud, se identificará frecuentemente con un amplio conjunto de individuos, entendiendo como individuos personas, animales, células... En otras ocasiones, se entiende por población el conjunto de todos los posibles resultados en la medición de un fenómeno sometido a variabilidad como, por ejemplo, los diferentes pesos que ua misma báscula puede mostrar para una misma persona bajo las mismas condiciones, al menos aparentemente.

Carácter

Sobre la población se estudiará uno o varios caracteres. No podemos dar una definición de carácter. Lo entenderemos como una noción común. La expresión del mismo carácter da lugar a una función o aplicación lo que en el contexto estadístico se denomina variable.

Si estos pueden ser expresado numéricamente a partir de cierta unidad de medida se denominarán cuantitativos; de lo contrario se denominarán cualitativos.

Variable

Como hemos dicho anteriormente el estudio de un carácter sobre una población se asocia a una variable que hace corresponder a cada individuo la expresión de su carácter. Desde un primer punto de vista, las variables pueden clasificarse en dos categorías:

- Cualitativas: se dice que una variable es cualitativa cuando expresa un carácter de forma no numérica. Ejemplos: sexo (varón o hembra); color de os ojos (azul, negro, marrón ...).

- Cuantitativas: se dice que una variable es cuantitativa o numérica cuando expresa un carácter mediante un número real. En este apartado podemos distinguir a su vez dos variedades:

  - Discretas: aquéllas que sólo admiten una cantidad numerable de valores, es decir, los valores que pueden tomar pueden escribirse uno detrás de otro en una secuencia. Ejemplos: número de hijos de una familia (0,1,2...); edad en años (0,1,2...); precio en euros de un producto de venta en supermercados (0.01, 0.02, 0.03...)
  - Continuas: aquéllas que admiten cualquier valor dentro de un intervalo por lo que sus posibles valores no pueden enumerarse. Ejemplos: peso, tiempo de vida, glucemia...

  El alumno aventajado se habrá percatado sin duda de que la distinción puede ser en muchos casos meramente teórica pues nuestra percepción de la realidad es siempre discreta. De hecho, en la Estadística Descriptiva, las variables discretas y continuas se distinguirán atendiendo únicamente a criterios de carácter gráfico y, por lo tanto, estéticos.

Esta clasificación será la que regirá en la mayor parte de la práctica estadística. No obstante y atendiendo a otros criterios podemos considerar otra clasificación que casi podría considerarse como una versión refinada e la anterior:

- Nominal: es lo que entendemos exactamente por cualitativas.

- Ordinal: no se exige que la expresión del carácter sea numérica pero sí al menos que exista un orden natural establecido. Ejemplo: grado de satisfacción en una encuesta (muy bajo, bajo, medio, alto, muy alto). Téngase en cuenta que, con frecuencia, estos resultados se asocian a números (1,2,3,4 y 5).

- De intervalo: se trata de una variable cuantitativa que expresa la magnitud de un carácter en relación con una unidad de medida. Ejemplos: peso en kilogramos, temperatura en grados Celsius, etc.

- De razón: se corresponde con el concepto de cantidad. Se trata de una variable de intervalo en la cual el valor 0 expresa la ausencia del carácter que se mide. Ejemplos: la variable peso es de razón, mientras que la variable temperatura en grados Celsius no lo es. Sí lo sería la temperatura en grados Kelvin. Por ello podemos decir que, si la medición de un individuo es el doble que la de otro quiere decir que el carácter se muestra en doble cantidad, de ahí su nombre.

Muestra de tamaño n

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. De 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.

El proceso de investigación, desde un punto de vista estadístico, consta de tres fases:

1. Selección de muestras.
2. Descripción de los datos de la muestra.
3. Inferencia o generalización al total de la población.

Esto nos sugiere el siguiente enfoque de la asignatura: empezaremos con una primera parte denominada Estadística Descriptiva, dedicada a la descripción —esto es, clasificación, representación y síntesis— de una muestra. Seguiremos con una segunda parte denominada Estadística Inferencial, dedicada a la generalización de los resultados de la muestra. Para realizar dicha generalización, partiremos de la premisa de que la muestra estudiada haya sido seleccionada al azar. La especialidad matemática dedicada al estudio de tal fenómeno (azar) se denomina Teoría de la Probabilidad, y constituye el fundamento teórico de la Estadística Inferencial, por lo que también será estudidada.

La exposición pretende ser una introducción a los principales aspectos de la Estadística. Se ha pretendido conseguir que sirva, conjuntamente, de referencia para afrontar problemas reales en la investigación y de guía para la comprensión lógica de los principios que rigen la Estadística, con el consiguiente riesgo de fracasar en ambos intentos.

Algunas consideraciones de carácter didáctico

Puede llamar la atención a las personas versadas en la materia la heterodoxia con la que se trata en diversos aspectos. En primer lugar, no se ha seguido la secuenciación clásica Descriptiva-Probabilidad-Inferencia. Concretamente, el problema de relación entre dos variables o caracteres se aborda en una primera fase desde una punto de vista meramente descriptivo. Estamos dispuestos a asumir la inconsistencia que conlleva esta transgresión en aras de facilitar al alumno el estudio de la Estadística a nivel básico. Nuestra modesta experiencia nos deja patente la confusión que en el alumno genera el concepto de probabilidad. Por ello, nuestra estrategia se basa en postergar en lo posible la aparición del mismo.

En segundo lugar, tampoco es convencional el enfoque que se da al capítulo 3, dedicado a la probabilidad. Es consecuencia de una actitud crítica hacia enunciados del tipo: “la probabilidad de que un individuo extraído aleatoriamente de la población padezca tal enfermedad es...”. Nos preguntamos qué entendemos exactamente por azar y qué necesidad hay del concurso de este concepto para referirnos a lo que, en el caso que nos ocupa, no es más que una proporción, a secas. En la sección 3.1 se incluyen algunas disquisiciones sobre el azar que no pretendemos que

sean asumidas por el lector. El único objetivo de las mismas es suscitar una reflexión sobre el concepto de probabilidad. Además, este capítulo puede resultar excesivamente formal para el lector a quien la Estadística le interesa en tanto en cuanto le sea de utilidad en el análisis de datos propios de las ciencias de la salud o, por qué no decirlo, a quien sólo le interesa aprobar cierta asignatura. No obstante, una lectura superficial puede ser suficiente para abordar con bastante garantía posteriores capítulos.

En tercer lugar, todas las técnicas de Inferencia Estadística e incluso de Estadística Vital se estudian en un mismo capítulo, el quinto, donde se muestra mayor interés por clasificarlas que por describirlas de manera exhaustiva. Optamos por esta disposición en virtud del papel preponderante que desempeñan los programas estadísticos en el proceso al que se someten los datos. A día de hoy, saber qué técnica debemos aplicar y cómo se interpretan los resultados obtenidos priman sobre los detalles y variantes de los procedimientos utilizados. Es claro que lo ideal sería dominar ambos aspectos, pero el tiempo de alumno es limitado y nos hemos decantado por el primero.

Los capítulos de mayor interés práctico son el primero, el segundo y el quinto. Los capítulos tercero y cuarto son de carácter teórico y se precisan para la mejor comprensión del quinto. Cada capítulo lleva asignada una relación de cuestiones teóricas o prácticas. Volvemos a recalcar que el objetivo de esta asignatura no es que el alumno muestre su capacidad de cálculo, sino que sea capaz de determinar a qué tipo de tratamiento deben someterse los datos en un problema práctico sencillo y, sobre todo, que sea capaz de interpretar los consiguientes resultados. Todo lo referente a memorización de procedimientos y cálculos numéricos tiene un interés secundario, pues puede ser realizado sin dificultad mediante un ordenador, utilizando cualquiera de los diversos programas estadísticos. Hacemos mención en la bibliografía el programa SPSS, de extendido manejo. De hecho, se incluyen numerosas salidas obtenidas mediante ese programa. Se incluyen por último algunas tablas estadísticas de utilidad en el manejo de las distribuciones Binomial, Normal, t-Student y χ^2.

Índice general

Capítulo 1

Estadística Descriptiva para una variable

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

El presente capítulo se dedica a la descripción de una variable mientras que el segundo afronta el estudio correlativo de dos variables. En ambos casos 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, cuantitativas discretas y cuantitativas continuas.

1.1. Tablas de frecuencias

La construcción de tablas de frecuencias ha sido hasta hace bien poco la fase preliminar a 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 el concurso de un programa informático que facilita 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 sencillos ejemplo para distintos tipos de variables.

Ejemplo 1: [Variable cualitativa]

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

Grupo i	f_i
0	2892
A	2625
B	570
AB	226
Total	6313

Esta tabla puede completarse con una columna donde queden reflejadas las correspondientes proporciones:

Grupo i	f_i	$\hat{p}_i$
0	2892	0,458
A	2625	0,416
B	570	0,090
AB	226	0,036
Total	6313	1

Los términos f_i y $\hat{p}_i$ hacen referencia, respectivamente, a los conceptos de frecuencia y proporción y se denominan comúnmente frecuencia absoluta y frecuencia relativa. La frecuencia relativa se expresa en ocasiones mediante un porcentaje, de manera que en nuestro caso tendríamos 42,5 %, 41,6 %, 9,0 % y 3,6 %. El símbolo $\wedge$ 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 o probabilístico que abordaremos en capítulos posteriores.

Ejemplo 2: [Variable cuantitativa discreta]

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.

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 pequeños 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, el cual 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 pues, obtenemos la siguiente tabla de frecuencias absolutas a la que añadimos las frecuencias relativas:

$\mathtt{x}_i$	f_i	$\hat{p}_i$
18	6	0.24
19	5	0.20
20	3	0.12
21	3	0.12
22	3	0.12
23	3	0.12
24	2	0.08
Total	25	1

En total, tenemos pues $k = 7$ valores distintos. 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:

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

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

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

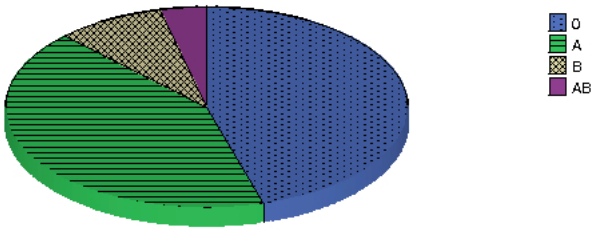
A esta columna puede añadírsele la de frecuencias relativas acumuladas que resulta de dividir las anteriores por el número total de datos

$$H_i = F_i/n$$

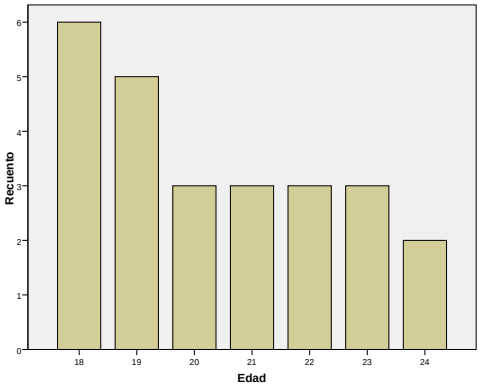
$\mathtt{x}_i$	f_i	$\hat{p}_i$	F_i	H_i
18	6	0.24	6	0.24
19	5	0.20	11	0.44
20	3	0.12	14	0.56
21	3	0.12	17	0.68
22	3	0.12	20	0.80
23	3	0.12	23	0.92
24	2	0.08	25	1
Total	25	1		

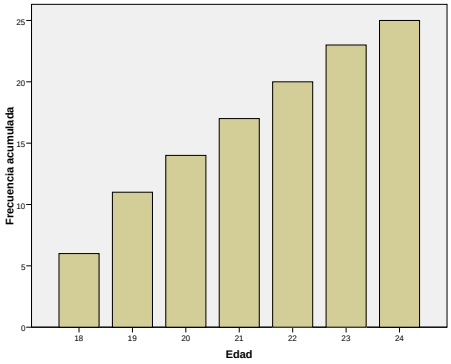
1.2. Representación gráfica

El segundo paso del proceso consiste en ilustrar mediante un gráfico lo obtenido en la tabla de frecuencia. Existen varios tipos de gráficos. El más simple es el conocido como diagrama de sectores. En el caso del ejemplo 1, la tabla de frecuencia quedaría plasmada de la siguiente forma:



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 este otro tipo de gráfico denominado diagrama de barras. Presentamos a continuación los digramas de barras para las frecuencias absolutas y las frecuencias absolutas acumuladas:





Los diagramas de barras para las frecuencias relativas y relativas acumuladas ofrecerían un aspecto idéntico al de los anteriores gráficos. Tan sólo cambiaría la escala del eje OY. Las líneas que unen las distintas barras se denominan polígonos de frecuencia. Seguramente el diagrama de barras para frecuencias acumuladas, a la derecha, resulte al lector menos intuitivo que el de la izquierda. No obstante, puede ser de gran interés para un estadístico.

La variable estudiada en el ejemplo 2 admite 7 posibles valores, de ahí que el diagrama de barras resulte 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 frecuencia perdería su sentido último. Consistiría en una larga ordenación vertical de los valores obtenidos donde todos ellos presenta frecuencia absoluta 1. El diagrama de barra resultante se antojaría claramente mejorable en cuanto a su poder ilustrativo. Veamos otro ejemplo:

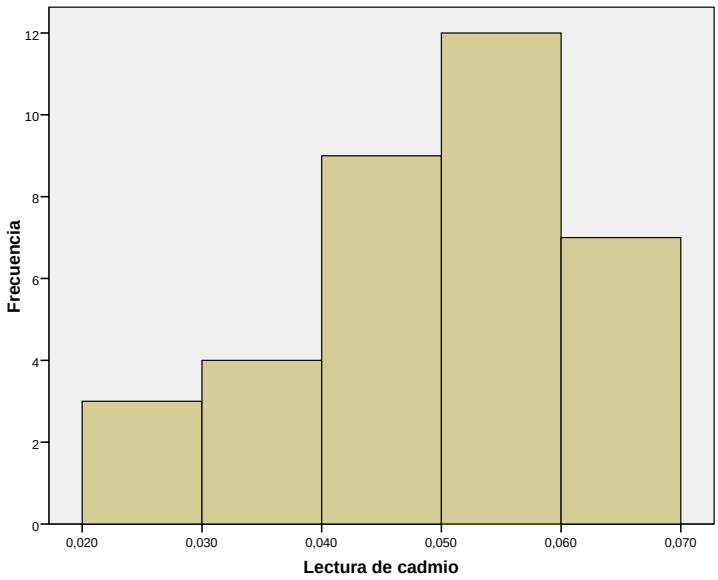
Ejemplo 3:[Variable cuantitativa continua]

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 de mide en miligramos de cadmio por metro cúbico de aire. Una muestra de 35 lecturas arroja estos datos: (Basado en un informe de Environmental Management, septiembre de 1981).

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

En este caso sucede también que la variedad de valores posibles es demasiado ampia 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. 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. 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 ésta como ya sabemos. Los gráficos correspondientes se denominan histogramas. En el caso del ejemplo 3 podemos obtener el siguiente histograma de frecuencias absolutas:



En definitiva, agrupar en clases significa simplificar, perder una parte de la información, en aras de una mejor ilustración de la misma. 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 Milton (2007) podemos encontrar un algoritmo perfectamente descrito. En la actualidad, todas las tareas gráficas se encomiendan a programas estadísticos que tiene implementados sus propios algoritmos. Por todo ello pasaremos de puntillas por esta cuestión indicando tan sólo unas normas básicas razonables:

1. Las clases serán intervalos contiguos, de ahí que en el histograma los rectángulos se peguen unos a otros.

2. Normalmente, los intervalos tendrán la misma amplitud. De no ser así, hemos de tener en cuenta que es el área del rectángulo y no su altura la que debe guardar proporción con la frecuencia del intervalo.

3. Todos los datos deben estar contenidos en los intervalos considerados y, a ser posible en su interior (no en la frontera).

4. El número de clases o intervalos a considerar debe guardar algún tipo de relación con el número total de datos. Con carácter orientativo, la ley de Sturges sugiere que, si disponemos de n datos, contruyamos el siguiente número de intervalos:

$$\operatorname{int}\big(1 + \log_2 n\big).$$

De esta forma, si hay entre 16 y 31 datos, se deberá tomar 5 clases, si hay entre 32 y 63, se tomarán 6, etc... Insistimos en que esta ley es meramente orientativa.

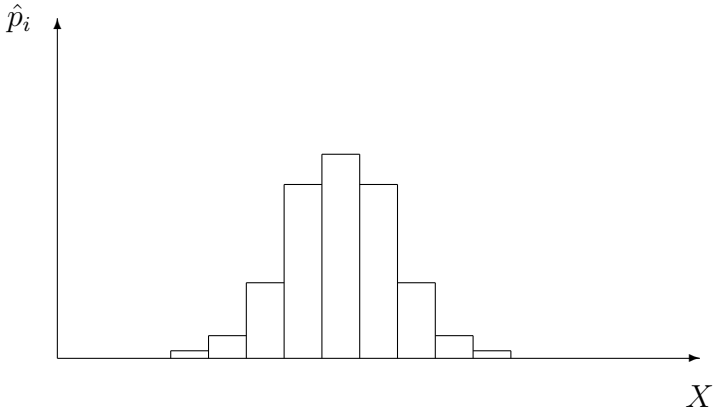
Otro tipo de gráfico de gran interés en estas situaciones y que guarda gran similitud con el histograma de frecuencias absolutas es el denominado diagrama tallo-hoja. A los datos de cadmio le corresponde el siguiente:

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

También hablaremos del denominado diagrama de caja o box-plot, pero eso será más adelante. Para acabar esta sección, destacamos que histogramas como lo que se observa en los datos del cadmio son bastante frecuentes en la naturaleza y desempeñan un papel central en la Estadística. Nos referimos concretamente a histogramas de este tipo:



Es lo que se conoce como curva normal o campana de Gauss y será objeto de un estudio más detallado en el capítulo 3. Fue estudiada inicialmente por Laplace y Gauss: ambos se ocupaban de problemas de astronomía y en ambos casos una distribución normal explicó 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, se distribuyen con frecuencia según curvas normales. Hechos de este tipo ya habían sido descritos por De Moivre con anterioridad a los trabajos de Gauss-Laplace. Desde un punto de vista teórico es el denominado Teorema Central del Límite, del que veremos una aproximación heurística en el tercer capítulo, el que confiere a la distribución normal un papel preponderante en la Estadística. Éste viene a decirnos, en términos intuitivos, lo siguiente: cuando los resultados de un experimento sean debidos a un conjunto muy grande de causas que actúan independientemente sumando sus efectos, siendo cada efecto individual de poca importancia respecto al conjunto, los resultados se distribuirán según una curva normal.

1.3. Valores típicos

El tercer paso del proceso descriptivo consiste en calcular una serie de parámetros, es decir, números, con la intención de recoger 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 de centralización

Las más importantes sin duda aunque por sí mismas no suelen bastar para resumir la información. La idea puede ser la siguiente: si pretendemos explicar la mayor parte posible de información con un único número, ¿cuál escogemos? Buscamos pues un número representativo, un valor central en algún sentido. Podemos pensar, por ejemplo, en el valor más frecuente, que se denomina moda o en otras opciones más o menos intuitivas como las denominadas media geométrica, media armónica o media truncada. Pero nos centraremos aquí en las dos opciones más naturales: la media aritmética y la mediana.

Media aritmética o muestral

Es el valor central en sentido aritmético. Se obtiene sumando los n datos de la muestra y dividiéndolos por el tamaño de ésta, 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.

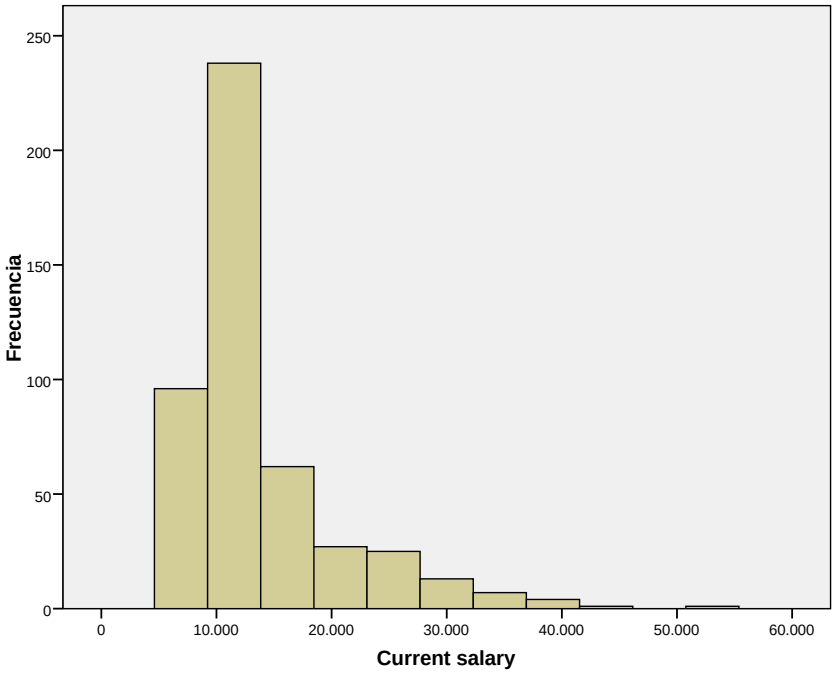
En el ejemplo 2 de las edades de 25 estudiantes tenemos $\overline{x} = 20{,}36$ años. La media se expresa, lógicamente, en las mismas unidades que los datos originales. Indicar dicha unidad es todo un detalle aunque no se considera preceptivo.

El hecho de que los datos estén agrupados en intervalos, como ocurre en el ejemplo 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 $\overline{x} = 0{,}0493$.

Mediana

Es el valor central $\tilde{x}$ en el sentido del orden, es decir, aquél 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. En el ejemplo 3 tenemos $\tilde{x} = 0{,}051$, que ocupa la posición 17.

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 puede 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 desplaza fuertemente la media restándole representatividad. Es lo que puede sucede en un caso como el siguiente, en el que se recogen los salarios de los empleados de cierto banco norteamericano:



1.3.2. Medidas de posición

Se trata de 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.

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 encomendarse a 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 $n/4$. Q_3 será el datos que ocupe esa posición pero contando desde el final.

Si dividimos la muestra en diez partes iguales obtenemos los denominados deciles $D_1,\ D_2, \ldots, D_9$. Obviamente, la mediana coincidirá con el el decil D_5. Si dividimos la muestra en 100 partes iguales, obtendremos los percentiles $p_1,\ p_2, \ldots, p_{99}$. 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.

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, el decil D_2 sería el cuantil 0.20, por ejemplo. 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 o, incluso de deciles.

1.3.3. 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. Uno de los más inmediatos es el denominado rango, que expresa la diferencia entre le valor mayor y el menor. En el ejemplo 2 sería igual $24 - 18$, es decir, 6. Esta medida es de utilidad en la Estadística no Paramétrica, de la cual hablaremos brevemente en el capítulo 5. Veamos cuáles son la más importantes en desarrollo de nuestra materia.

Varianza muestral

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 los datos que quedan a la derecha de la media se compensan con los que quedan a su izquierda. Se podría haber optado por considerar el valor absoluto de las diferencias, lo cual daría 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 y por razones que se comentarán en el capítulo 4, aparecerá en la mayoría de las ocasiones dividida por $n-1$ en vez de n:

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

Suele denominarse dicho parámetro varianza insesgada o cuasi-varianza. De ahora en adelante, si no se especifica lo contrario, cada vez que hablemos de varianza nos estaremos refiriendo a la insesgada $(n-1)$. El hecho de dividir por $n-1$ en lugar de n es apenas apreciable cuando n es grande, por o que no debe desviar nuestra atención de la esencia del parámetro. El cálculo de la varianza lo encomendamos el programa estadístico o, en su defecto, a la calculadora.

En el ejemplo de las edades en años de 25 alumnos se obtiene una varianza $s^2 = 4{,}157$ años^2. Podemos observar que 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 como desviación típica, que se denotará por s. Así pues,

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

En el ejemplo anterior, tendríamos $s = 2{,}039$ años.

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, la denominada Desigualdad de Chebichev establece que, para cualquier número k positivo, la proporción de datos de la muestra que se encuentran entre los valores $\overline{x} - k \cdot s$ y $\overline{x} + k \cdot s$ es al menos del

$$100 \times \left(1 - \frac{1}{k^2}\right) \%$$

De esta forma, tenemos por ejemplo $(k = 2)$ que, entre los valores $\overline{x} - 2 \cdot s$ y $\overline{x} + 2 \cdot s$, se encuentra, al menos, el 75 % de los datos.

Esta desigualdad no resulta demasiado esclarecedora. De hecho, en el caso $k = 1$ no dice absolutamente nada. No obstante, si nuestros datos se distribuyen según una curva normal ocurre que el mero conocimiento de $\overline{x}$ y s permite reproducir con exactitud el histograma y, por lo tanto, la distribución de los datos. 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 %. 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.

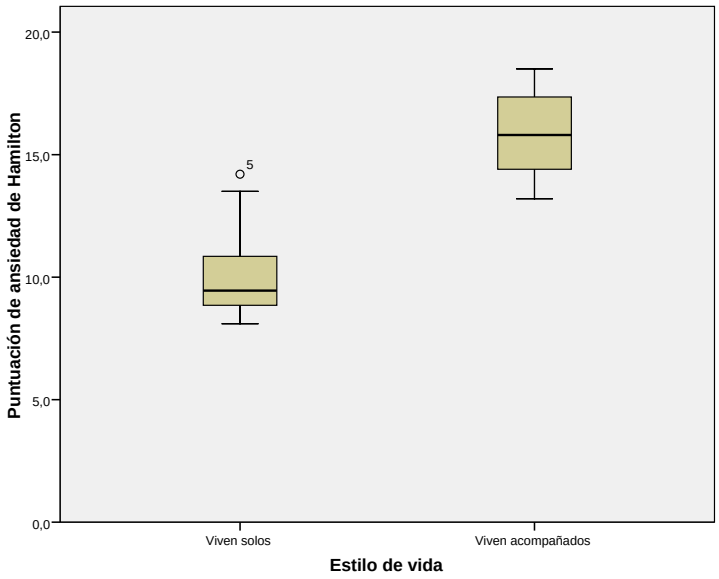
Rango intercuartílico o amplitud intercuartil

Pretende ser un complemento adecuado a la mediana. Está basado al igual que ésta en el orden de los datos y se define mediante $R_I = Q_3 - Q_1$. En el caso de los datos de edad, obtenemos $R_I = 2$.

En definitiva, si pretendemos resumir lo mejor posible la información contenida en la muestra debemos escoger al menor 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 la media adolece de falta de representatividad y el par anterior pierde su capacidad de resumen. La otra opción es el par $(\tilde{x}, R_I)$. Nos decantaremos por esta opción 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 cierto, existe un tipo de gráfico denominado diagrama de caja o box-plot especialmente útil a la hora de detectar ambas incidencias: asimetría y presencia de valores extremos. El gráfico debe ser elaborado por un programa estadístico, por lo que no nos extenderemos demasiado en su descripción: Consiete en dibujar una caja cuyos extremos coincidan con los cuartiles Q_1 y Q_3 y trazar dentro una linea donde se encuentre el cuartil Q_2. A continuación se calculan a ambos lados las vallas $Q_1 - 1{,}5 \cdot R_I$ y $Q_3 + 1{,}5 \cdot R_I$. El primer dato de la muestra por encima de la primera y el último por debajo de la segunda se de denominan valores adyacentes y se marcan mediante sendos segmentos que parten de la caja. Los valores que queden fuera del intervalo que determinan los valores adyacentes se consideran extremos. Se delimitan por último otras vallas más externas multiplicando por 3 el rango intercuartílico para distinguir los datos moderadamente extremos, si se nos permite la expresión, de los acusadamente extremos.

Veamos un ejemplo: se muestran los diagramas de caja para la puntuación de ansiedad de Hamilton en un grupo de 20 personas que viven solas y otro de 20 personas que viven acompañadas:



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

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

1.3.4. Medidas de forma

Por último, consideramos dos parámetros que pretenden dar cierta idea de la forma en la que se distribuyen los datos. Deben guardar pues una estrecha correspondencia con lo observado en los histogramas, diagramas tallo-hoja y diagramas de caja. Las dos medidas que definimos a continuación son muy difíciles de calcular si no se hace uso de un programa estadístico. Pero lo que nos interesa de ellas no es su cálculo sino su interpretación.

Coeficiente de asimetría

Es, como su propio nombre indica una medida del grado de asimetría o sesgo que se da en la distribución de los datos. Se define mediante

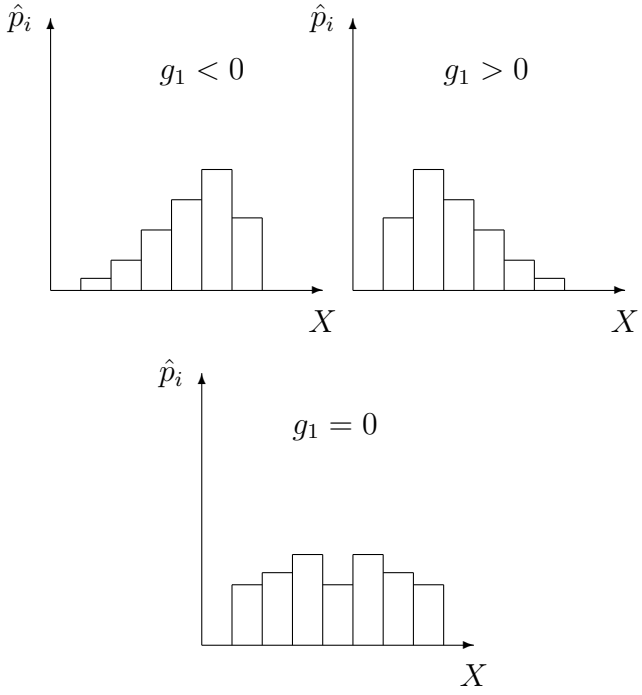
$$g_1 = \frac{m_3}{s^3}, \quad \text{siendo } m_k = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^k}{n}, \quad k = 1, 2, 3...$$

Distinguimos a grandes rasgos tres situaciones:

1. $g_1 < 0$: Distribución asimétrica de los datos con sesgo negativo.

2. $g_1 > 0$: Distribución asimétrica con sesgo positivo.

3. $g_1 = 0$: Distribución simétrica.



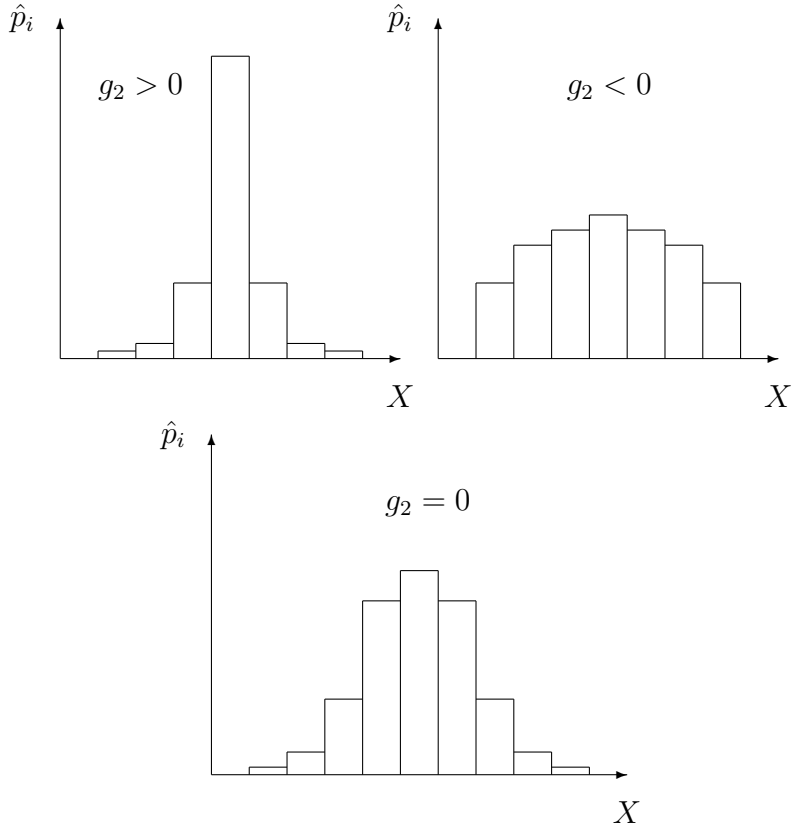
Coeficiente de aplastamiento o de Curtosis

El parámetro m^4/s^4 es una buena referencia acerca del grado de aplastamiento que presenta la gráfica de los datos cuando ésta es simétrica, de manera que cuanto mayor sea su valor tanto menor será su aplastamiento. En el caso de una campana de Gauss, se tendrá un valor 3. Entonces, el coeficiente de aplastamiento o Curtosis

$$g_2 = \frac{m_4}{s^4} - 3$$

expresa el grado de aplastamiento respecto a la curva normal, de la siguiente forma:

1. $g_2 > 0$: Distribución leptocúrtica (menos aplastada que la Campana de Gauss).
2. $g_2 < 0$: Distribución platicúrtica (más aplastada que la Campana de Gauss)
3. $g2 = 0$: Mesocúrtica (es decir, igual aplastamiento al de la Campana de Gauss).



1.4. Cuestiones propuestas

1. 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?

2. Considera los dos grupos de datos 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 en el capítulo?

3. 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.

   - Indica un intervalo que contenga aproximadamente al 68 % de los datos.
   - 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?

4. Se expresan a continuación las longitudes de 7 determinados objetos medidas en *mm* mediante un ecógrafo.

   7,0 7,4 8,9 9,6 10,5 11,7 12,5

   - Calcula (utilizando el modo estadístico de la calculadora) la media y desviación típica de los 7 datos.
   - Calcula (sin utilizar la calculadora) la media, desviación típica y varianza de los mismos datos expresados en *cm*.

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

   4 5 4,5 3,9 5,2 4 5,2 5,3 23 4,1

   Dar una medida de centralización y otra de dispersión adecuadas.

6. - Indica, si es que es posible, dos grupos, de 5 datos cada uno, que presenten la misma media pero distinta desviación típica.
   - Idem con misma desviación típica pero distinta media.
   - Idem con misma media y distinta mediana.
   - Idem con misma mediana y distinta media.
   - Idem con misma media y varianza pero distinto coeficiente de variación.

7. ¿Se puede dar una varianza negativa? ¿Y un rango intercuartílico negativo? Razónalo e ilústralo con un ejemplo, si es necesario.

8. 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 asigno a 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) estadísticico consideras más apropiado para efectuar la comparación.

9. Razona si son verdaderas o falsas cada una de las siguientes afirmaciones:

   - Si una muestra de datos presenta media 0, su desviación típica será pequeña.
   - Cuanto mayor es el tamaño de la muestra, mayor es su varianza.
   - Cuanto mayor es el tamaño de la muestra, mayor es su media.
   - Si $g_1 \simeq 0$ la media y la mediana deben ser parecidas.

10. La siguiente tabla representa el número de infartos de miocardio por día que se atendieron en un servicio especializado durante 30 días:

Infartos	0	1	2	3	4	5	6
f_i	2	3	8	11	2	3	1

    a) Representar el diagrama de barras para frecuencias absolutas y frecuencias absolutas acumuladas.

    b) Calcular la media, varianza, desviación típica y coeficiente de variación de los datos anteriores.

    c) Calcular la mediana y el rango intercuartílico.

11. Se ha desarrollado una nueva vacuna contra la difteria para aplicarla a niños. El nivel de protección estándar obtenido por antiguas vacunas es de 1 $\mu g/ml$ un mes después de la inmunización. Se han obtenido estos datos del nivel de protección de la nueva vacuna al transcurrir un mes: (Basado en un informe del Journal of Family Practice, enero 1990.)

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?

12. Considerar los datos del ejemplo 3.

    a) Obtener mediante la calculadora científica los valores de la media artimética, la desviación típica y el coeficiente de variación.

    b) Obtener, 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. ¿Y el del coeficiente de aplastamiento?

13. Se midió la altura en 200 individuos adultos de la ciudad de Cáceres. La información recogida en dicha muestra se ha agrupado en 6 clases de la misma amplitud, resultando la siguiente tabla:

Altura (cm)	f_i	F_i	$\hat{p}_i$	H_i
		2		
$(100, 120]$				0,06
	10			
	35			
			0,6	
			0,115	

Completar la tabla de frecuencias. Representar el histograma de frecuencias relativas acumuladas. Indica en qué intervalo se encuentra la mediana.

14. Los datos del siguiente diagrama tallo-hoja corresponden a la concentración de mercurio $[\mu gr/cm^3]$ en la sangre de 25 individuos de una zona contaminada. Se utiliza como unidad 1:

0	8
1	0 2
2	0 5 7
3	0 2 5 5 6 6 8
4	0 0 1 4 5 5
5	0 2 3
6	1 2
7	0

Calcula la moda, media, mediana, desviación típica y rango intercuartílico de estos 25 datos. ¿Qué par de valores consideras que resumen adecuadamente la información de toda la muestra? ¿Por qué? ¿Qué valores cabe esperar para los coeficientes de simetría y aplastamiento?

15. Considera el diagrama de caja de la sección 1.3 correspondiente a la puntuación de ansiedad de Hamilton sobre 20 individuos que viven solos (caja de la izquierda). Uno de los dos diagramas tallo-hoja corresponde a los datos mencionados. Razona cuál.

Frequencia	Tallo	Hoja
5	8 .	12367
7	9 .	0334456
3	10 .	137
2	11 .	06
1	12 .	9
1	13 .	5
1	14 .	2

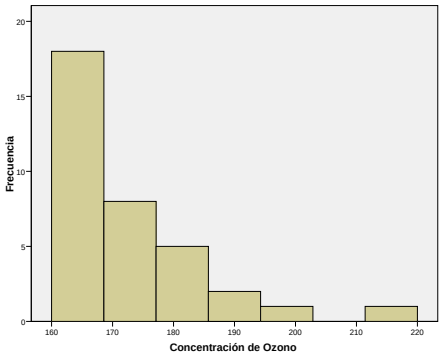
Unidad: 1,00

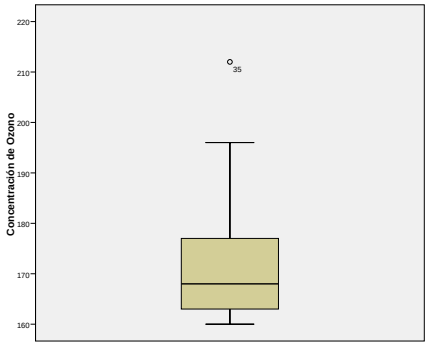
Frequencia	Tallo	Hoja
3	13 .	236
4	14 .	0267
3	15 .	246
3	16 .	019
4	17 .	2345
3	18 .	015

Unidad: 1,00

Indica un par de medidas que resuma lo mejor posible la información que aportan los 20 datos. ¿Qué podemos decir del coeficiente de asimetría?

16. En una zona boscosa cerca de Seattle se tomaron 35 medidas de concentraciones de ozono (partes por billón), obteniéndose los siguientes resultados:



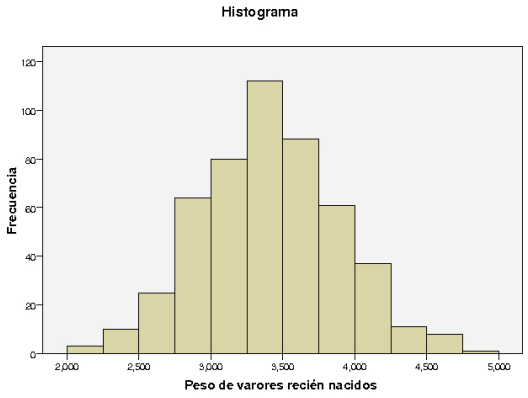


Descriptivos

		Estadístico
Media		171,66
Intervalo de confianza para la media al 95%	Límite inferior	167,62
	Límite superior	175,69
Media recortada al 5%		170,44
Mediana		168,00
Varianza		137,997
Desv. típ.		11,747
Mínimo		160
Máximo		212
Rango		52
Amplitud intercuartil		15
Asimetría		1,646
Curtosis		2,999

Comentar, a la luz de los gráficos y los coeficientes de forma, los aspectos más destacados de la distribución de los datos y seleccionar un par de parámetros que resuman lo mejor posible la información que contiene la muestra.

17. Se midió el peso en kg de 500 varones recién nacidos después de la semana 38 de gestación. Los resultados son los siguietes:



Descriptivos

Media		3,40850
Intervalo de confianza para la media al 95%	Límite inferior	3,36673
	Límite superior	3,45028
Media recortada al 5%		3,40490
Mediana		¿?
Varianza		,226
Desv. típ.		,475467
Mínimo		2,138
Máximo		4,787
Rango		2,649
Amplitud intercuartil		,680
Asimetría		,106
Curtosis		-,052

Comentar los aspectos gráficos más destacados e indicar un par de medidas que resuman satisfactoriamente la información que aporta la muestra. Dar un valor aproximado para la mediana y para el percentil p_{84}. Razonar si deben aparecer valores extremos en el diagrama de caja.

Capítulo 2

Estadística Descriptiva para dos variables

Si en el capítulo anterior se afrontaba el estudio descriptivo de una variable cualitativa o cuantitativa, en el presente se aborda el estudio conjunto de dos, bien cualitativas o bien cuantitativas. Distinguimos también dos aspectos: la descripción de cada una de las variables por separado y el análisis de la relación existente entre ambas. Dado que el primero ha sido ya tratado en el anterior capítulo, nos centraremos en el segundo: la relación entre las dos variables en juego. Este estudio tiene carácter preliminar respecto a otra fase que se abordará en el capítulo 5. Empezaremos estudiando la relación entre variables cuantitativas para continuar después con la relación entre variables cualitativas. Indicar por último que la relación entre una variable cualitativa y otra cuantitativa se trata desde un punto de vista inferencial en la sección cuarta del capítulo 5.

2.1. Relación entre dos variables numéricas

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. Es preferible que las variables sean de intervalo o de razón, aunque en ocasiones pueden ser también ordinales. 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-ésimo 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:[Dos variables cuantitativas]

Se indica a continuación el peso (kg) y la estatura (cm) de 12 personas (no se especifica edad, sexo ni ningún otro aspecto):

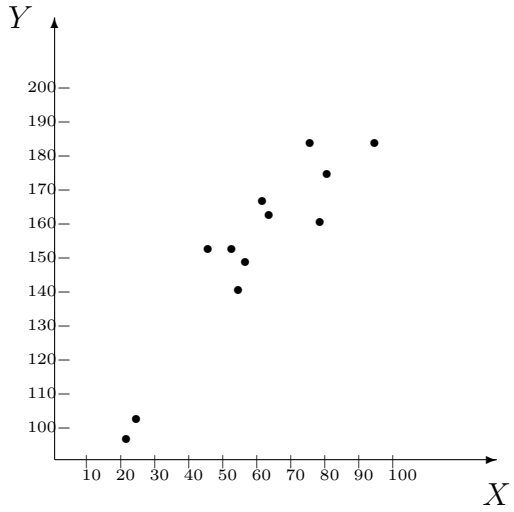
X =peso(kg)	80	45	63	94	24	75	56	52	61	34	21	78
Y =altura(cm)	174	152	160	183	102	183	148	152	166	140	98	160

El estudio debe empezar con una estadística descriptiva de cada variable por separado, cosa que se supone sabemos hacer. A continuación, nos dedicamos al estudio descriptivo de la relación

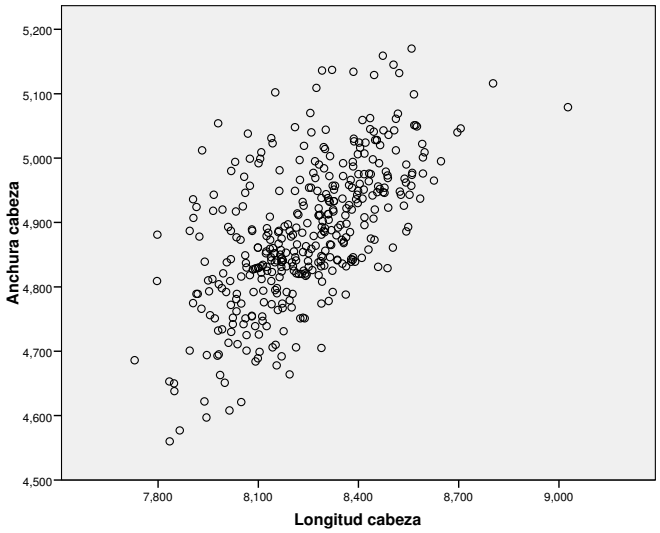
entre ambas. Podríamos empezar confeccionando una tabla de frecuencias donde se contabilice el número de ocasiones en el que aparece cada par, pero, salvo que se traten de variables con reducido número de valores posible, no tendrá utilidad alguna.

2.1.1. Diagrama de dispersión

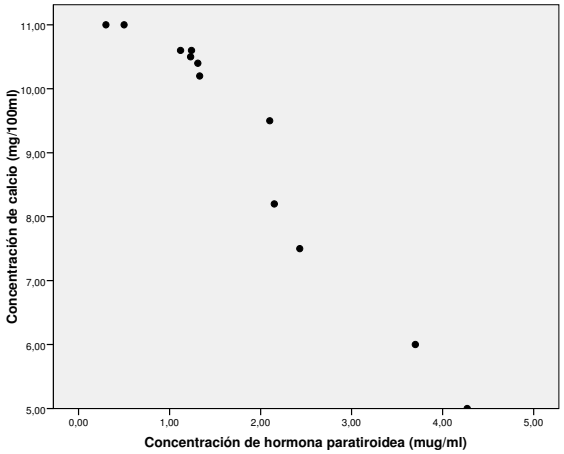
Así pues, lo primero que nos interesa realmente el la representación gráfica de la muestra. Esta tarea debe encomendarse a un programa estadístico aunque, en este caso y dado el escaso tamaño de la misma, podemos hacerlo nosotros mismos. El gráfico más adecuado para apreciar la relación entre dos variables 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 verían como sigue:



En este otro diagrama de dispersión se aprecia la relación entre la longitud y la anchura de la cabeza para $n = 391$ espermatozoides pertenecientes a cierta especie animal:



En ambos casos se observa en la muestra una relación positiva en el sentido de que el crecimiento de una variable suele venir emparejado al crecimiento de la otra. No será siempre el caso. Veamos, por ejemplo, el gráfico de dispersión correspondiente a $n = 12$ mediciones de las concentraciones de hormona paratiroidea (μg/ml) y calcio (mg/100ml) en sangre:



Como denominador común a los tres ejemplos considerados podemos resaltar que 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, la nube se forma en torno a una línea recta, que puede ser creciente o decreciente. Este tipo de relación se denomina lineal y es el objeto principal de estudio en este sección. Con ello no queremos decir que sea la única relación posible. Lo que sí es claro es que es la más sencilla. 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.

Una vez representados los datos pasamos al cálculos de los valores típicos. En primer lugar, necesitamos conocer la media y desviación típica de cada una de las variables por separado, es decir,

$$\overline{x} = \frac{\sum_i x_i}{n}, \qquad s_x = \sqrt{\frac{\sum_i (x_i - \overline{x})^2}{n-1}},$$
$$\overline{y} = \frac{\sum_i y_i}{n}, \qquad s_y = \sqrt{\frac{\sum_i (y_i - \overline{y})^2}{n-1}}$$

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

$$\overline{x} = 56{,}91kg, \quad s_x = 22{,}95kg, \quad \overline{y} = 151{,}5cm, \quad s_y = 27{,}45cm$$

2.1.2. Coeficiente de correlación

Uno de los principales objetivos de nuestro estudio es 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. Empezaremos definiendo la covarianza muestral como sigue:

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

La covarianza, que en el caso del ejemplo 4 se expresará en $kg \cdot cm$, puede ser tanto positiva como negativa, pero debe quedar necesariamenteacotada por los valores siguientes

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

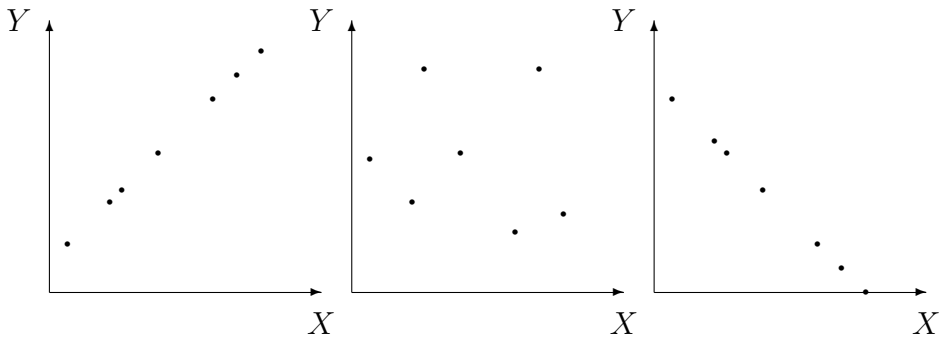
En el ejemplo 4, se tiene que s_{xy} debe estar comprendido entre $-630{,}71$ y $630{,}71$, siendo concretamente su valor $577{,}86\ kg \times cm$. La covarianza pretende expresar el grado de correlación lineal existente entre las variables X e Y de la siguiente forma:

- 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 $s_{xy} = +s_x \cdot s_y$ significa una correlación lineal perfecta, es decir, que la nube de puntos está incluida en una única recta, que será además creciente.

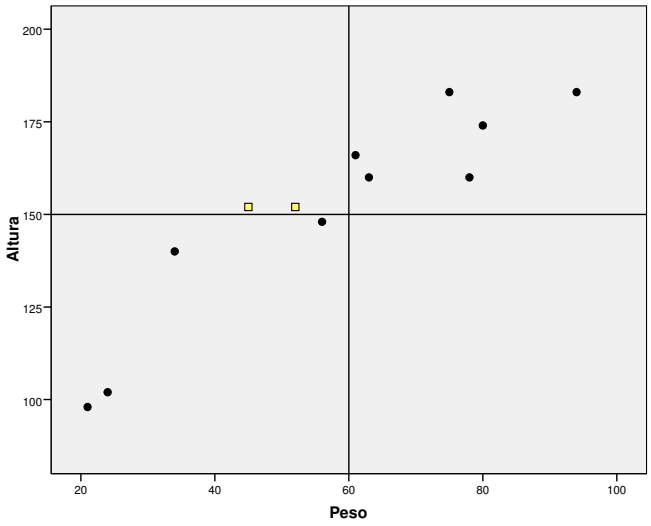
- Un valor negativo de s_{xy} significa una tendencia decreciente en la nube de puntos, es decir: si los valores de X crecen, los de Y decrecen. Existirá por tanto correlación (inversa) entre ambas variables, según la muestra. El caso extremo $s_{xy} = -s_x \cdot s_y$ significa una correlación lineal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

- $s_{xy} = 0$ se traduce, por contra, en la ausencia de relación lineal en los datos de la muestra.

Se ilustra lo dicho anteriormente mediante tres casos en los cuales se verifica, respectivamente, $s_{xy} = -s_x s_y$, $s_{xy} \simeq 0$ y $s_{xy} = s_x s_y$



Según lo dicho, en el ejemplo 4, se observa una alto grado de correlación lineal positiva. En el gráfico siguiente se aprecia el porqué:



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

Para evaluar qué entendemos por grande hemos de tener en cuenta la cota máxima que se puede alcanzar, que no es universal. Nos referimos a $s_x s_y$. De hecho, un cambio de unidades (pasar de centímetros a metros, por ejemplo), hace variar tanto las desviaciones típicas como la covarianza. Todo ello complica la interpretación del parámetro s_{xy}. Nos interesaría pues otro parámetro que se interprete de forma análoga pero cuyas cotas sean universales. La solución es fácil considerando

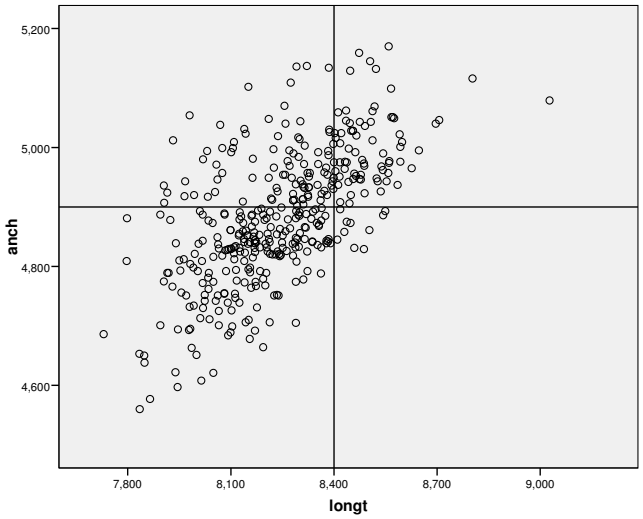
$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

Este parámetro, que se denomina coeficiente de correlación lineal muestral, se interpreta en los mismos términos con la salvedad de que es adimensional, encontrándose en todo caso entre -1 y 1 y alcanzando 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 $r = 0.9161$.

En la práctica será incluso de más utilidad el parámetro r_{xy}^2, denominado coeficiente de determinación muestral. Más adelante veremos su interpretación. En el caso del ejemplo 4 tenemos $r^2 = 0{,}8281$.

Existen algoritmos que tienen por objeto el cálculo del coeficiente r. No obstante, nosotros delegaremos esas tareas en el programa estadístico o, en su defecto, en la calculadora.

En el caso de la longitud y anchura de las cabezas de espermatozoides, se obtiene un coeficiente de corelación r =0.625, lo cual expresa una correlación positiva pero más débil que la observada anteriormente, cosa que debe quedar clara si en el diagrama de dispersión trazamos las lineal de referencia que pasan por las medias:



2.1.3. Recta de regresión muestral

En el caso de que se observe una considerable correlación lineal entre los datos de X y los de Y, puede ser interesante calcular la denominada recta de regresión muestral, que será la recta en torno a la cual se distribuyen los datos. Decimos recta y no curva pues estamos suponiendo, al menos por el momento, que la relación es de tipo lineal. Se trata pues de encontrar la recta que mejor se ajusta a nuestra nube de puntos. Pero, lógicamente, habrá que especificar primeramente que entendemos por "ajuste". En nuestro caso utilizaremos el criterio muy utilizado en Matemáticas conocido como el de Mínimos Cuadrados, cuya conveniencia fue argumentada hace casi dos siglos por el propio Gauss. Veamos en qué consiste.

Como ya sabemos, una recta en el plano puede expresarse de la forma $y = a + bx$, donde b es la pendiente y a el valor de corte con el eje de OY. Dado una unidad experimental de la muestra (x_i, y_i), al valor x_i correspondiente a las abcisas, es decir, a la variable X le corresponde, según la recta anterior, el valor $a + bx_i$ para las ordenadas. La diferencia entre dicho valor y el que realmente corresponde a la variable Y, es decir, y_i, se considera un error cometido por la recta anterior. El método de mínimos cuadrados propone cuantificar el error total mediante la suma de los cuadrados de los errores particulares, es decir,

$$\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$$

La recta que minimice dicho error será la solución deseada. Dicha solución puede encontrarse mediante argumentos geométricos o bien haciendo uso del cálculo diferencial. Obviando esos detalles, podemos afirmar que los parámetros de la recta de regresión buscada son los siguientes:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ a &= \overline{y} - b\overline{x}. \end{aligned}$$

Cabe realizar tres observaciones:

(i) El signo de b es el que le otorga la covarianza s_{xy}, que es a su vez el mismo de r. Es decir, que si la correlación es directa, la recta de regresión tiene pendiente positiva, y si es inversa, negativa, como cabía esperar.

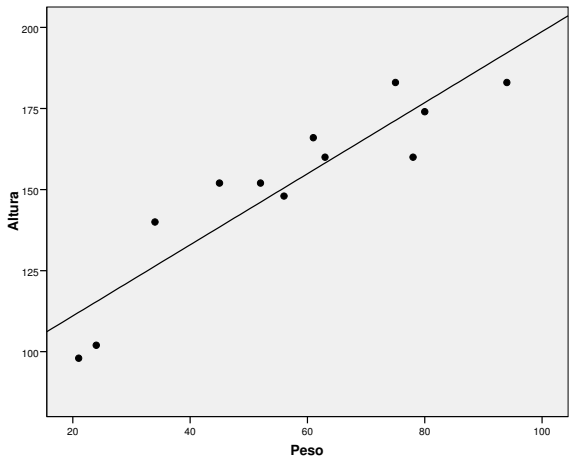
(ii) En todo caso, la recta pasará por el punto $(\overline{x}, \overline{y})$. Por decirlo de alguna forma, pasa por el centro de la nube de puntos.

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

En el caso del ejemplo 4, la recta de regresión lineal muestral es la siguiente:

$$y = 89{,}11 + 1{,}10x,$$

que se representa a continuación:



En la primera columna de la siguiente tabla se muestran los valores de X para los 12 datos; en la segunda, los correspondientes valores de Y; en la tercera, los valores de la ordenadas que se obtienen según la recta de regresión $y = 89{,}11 + 1{,}10x$; por último, en la cuarta columna tenemos las diferencias al cuadrado entre los segundos y los terceros, de manera que su suma cuantifica el error cometido por la recta de regresión.

x_i	y_i	$(a + bx_i)$	$[y_i - (a + bx_i)]^2$
80	174	176.80	7.86
45	152	138.44	183.94
63	160	158.17	3.36
94	183	192.15	83.70
24	102	115.42	180.05
75	183	171.32	136.37
56	148	150.50	6.23
52	152	146.11	34.69
61	166	155.98	100.48
34	140	126.38	185.51
21	98	112.12	199.66
78	160	174.61	213.47
			1335.32

Esa suma total, denominada error cuadrático, podrá resultarnos grande o pequeña, pero lo que es incuestionable es que cualquier otra recta que podamos considerar ofrecerá un error cuadrático mayor. También es claro que cuanto mas puntos tengamos mayor será el error cuadrático. Necesitamos pues una medida del grado de error relativa al tamaño de la muestra. Ese parámetro, que se denomina varianza residual o parcial, podría obtenerse dividiendo por n la suma anterior aunque, por detalles que obviaremos, la definimos dividiendo por $n-2$, es decir,

$$s^2_{y\leftarrow x} = \frac{1}{n-2}\sum_{i=1}^{n}[y_i - (a+bx_i)]^2$$

siendo $y = a + bx$ la recta de regresión. La varianza residual viene a expresar pues la parte de la variabilidad de los datos de Y no explicada por a variabilidad de los datos de X mediante la recta de regresión lineal. Por otra parte, se tiene lo siguiente:

$$\begin{aligned}(n-2)s^2_{y\leftarrow x} &= \sum_i [y_i - (a+bx_i)]^2 = \sum_i [(y_i-\overline{y}) - b(x_i-\overline{x})]^2 \\ &= \sum_i (y_i-\overline{y})^2 + b^2\sum_i (x_i-\overline{x})^2 - 2b\sum_i (y_i-\overline{y})(x_i-\overline{x}) \\ &= (n-1)(s_y^2 + b^2 s_x^2 - 2bs_{xy}) = (n-1)(s_y^2 - s_y^2 r_{xy}^2).\end{aligned}$$

Es decir,

$$\frac{n-2}{n-1}\cdot\frac{s^2_{y\leftarrow x}}{s_y^2} = 1 - r_{xy}^2$$

El primer factor de la ecuación no debe despistarnos pues su valor es prácticamente uno, en especial si la muestra es grande. Lo que tenemos en definitiva, haciendo caso omiso del mismo, es lo siguiente:

$$\frac{s^2_{y\leftarrow x}}{s_y^2} = 1 - r_{xy}^2$$

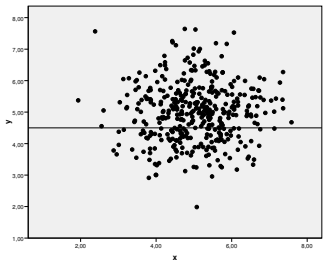
La interpretación de esta expresión es fundamental pues permite entender el significado exacto de r^2 y, en particular, de r. Concretamente, sabemos que s_y^2 expresa la variabilidad o dispersión de los datos de Y; por su parte, $s^2_{y\leftarrow x}$ expresa la parte de esa variabilidad que no es explicada por los datos de X mediante la recta de regresión lineal. Así pues, el cociente $s^2_{y\leftarrow x}/s_y^2$ puede interpretarse como la proporción de variabilidad de los datos de Y no explicada por la regresión, y es igual a $1 - r_{xy}^2$.

En consecuencia, r_{xy}^2 se entiende como la proporción de variabilidad de los datos de Y que es explicada por la regresión lineal respecto a los datos de X.

En el caso del ejemplo 4 teníamos $r^2 = 0.8281$, lo cual se traduce en que la recta de regresión explica un 82.81 % de la variabilidad de los datos de Y o, lo que es lo mismo, que conlleva un 17.19 % de error.

Los caso extremos serían $r^2 = 1$ y $r^2 = 0$. El primero se corresponde con $s^2_{y\leftarrow x} = 0$, es decir, la recta de regresión lineal predice sin error los datos de Y a partir de X. Se da por lo tanto una correlación lineal perfecta. El caso $r^2 = 0$ se corresponde con $s^2_{y\leftarrow x} = s_y^2$. Significa que toda la

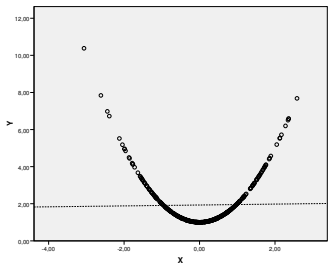
variabilidad de Y es error de regresión, es decir, que la recta de regresión no ayuda en absoluto a predecir los valores de Y. Este caso se corresponde con una recta de regresión de pendiente nula, es decir, constante. Concretamente, se trata de la constante $\overline{y}$, por ser la mejor opción posible. En definitiva, no aporta nada a la explicación de los datos de Y. Tal es. aproximadamente, el caso de la ilustración:



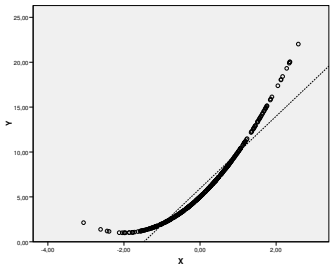
Para acabar este apartado llamamos la atención sobre el hecho de que los parámetros calculados se basan a su vez en otros parámetros descriptivos estudiados en el capítulo anterior, concretamente la media aritmética y la varianza. En dicho capítulo comentamos la necesidad de reemplazar la media aritmética por la mediana en determinados casos en los que la primera resultaba muy afectada por la asimetría y presencia de valores extremos. Nos preguntamos ahora si podemos llegar a hacer uso de la mediana en un problema de relación entre dos variables cuando aparezcan a su vez puntos extremos en el diagrama de dispersión. Efectivamente, la mediana se puede utilizar para construir lo que se denomina una recta de regresión resistente. No obstante, remitimos al lector a una bibliografía más avanzada para tratar estos problemas u otros que se nos quedan en el tintero.

2.1.4. 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, en mayor o menor grado, de tipo lineal. Hemos excluido pues aquellas situaciones en las que la función de X que mejor explica los datos de Y no es una recta sino una curva y este hecho debe tenerse muy en cuenta. De no ser así, interpretaríamos un coeficiente de correlación próximo a 0 como una ausencia de relación entre los datos de las variables, cosa que no es cierta tal y como se desprende del siguiente ejemplo, en el que observamos una dependencia absoluta de los datos de Y respecto a los de X. Sin embargo, se obtiene $r = 0.17$ y la recta de regresión es prácticamente plana, como podemos observar:



Tampoco es cierto que la presencia de un coeficiente de correlación lineal elevado implique que la relación entre las variables sea de tipo lineal, como ocurre en este ejemplo en el que $r = 0.97$:

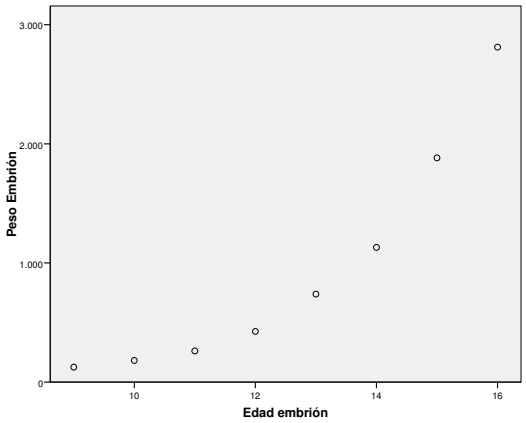


La mejor forma de determinar la conveniencia de un estudio de correlación-regresión lineal es echando un simple vistazo al diagrama de dispersión. Veamos un ejemplo.

Ejemplo 5:[Regresión no lineal]

Se pretende establecer la relación existente entre la edad en días (X) de un embrión y su peso en mg. (Y). La siguiente tabla presenta los pesos de 8 embriones con diferentes días de edad:

X (edad en días)	9	10	11	12	13	14	15	16
Y (peso en mg.)	125	181	261	425	738	1130	1882	2812



Observamos por un lado que existe una estrechísima relación entre la edad y el peso, de manera que la primera podría explicar perfectamente la segunda pero no mediante una recta sino mediante una sencilla curva. Lo más difícil del problema es determinar de qué tipo de función se trata: polinómica, exponencial, logística... Para responder a esa pregunta se precisa cierta experiencia tanto de carácter matemático como biológico en este caso, además del apoyo de un programa estadístico. Bajo estas premisas concluimos que la relación es de tipo exponencial, es decir, $Y = \mathtt{k} \cdot \mathtt{d}^X$. Necesitamos precisar los valores de los parámetros $\mathtt{k}$ y $\mathtt{d}$. Esto se consigue mediante reemplazando la variable Y original por $\tilde{Y} = \ln Y$. ¿Por qué? Pues porque si $y = \mathtt{k}\mathtt{d}^x$, entonces

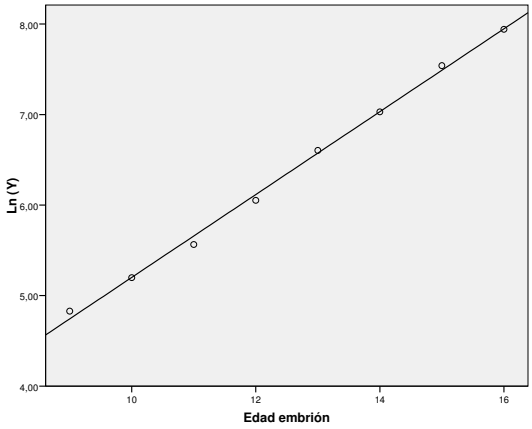
$$\begin{aligned} \ln y &= \ln(\mathtt{k}\mathtt{d}^x) \\ &= \ln \mathtt{k} + (\ln \mathtt{d})x \end{aligned}$$

En ese caso, si se denota $a = \ln \mathtt{k}$ y $b = \ln \mathtt{d}$, se tiene entonces que

$$\tilde{y} = a + bx$$

es decir, que la relación entre las variables transformadas sí es lineal. En nuestro caso, eso queda patente observando el diagrama de dispersión entre X y $\ln Y$ que, por cierto, aportan un coeficiente de correlación lineal

$$r_{X,\ \ln Y} = 0{,}99$$



Una vez hemos dado con el cambio de variables adecuado, calculamos los parámetros de la recta de regresión para los datos transformados como ya sabemos. En nuestro caso se obtiene

$$\ln y = 0{,}623 + 0{,}458x$$

Para deshacer el cambio basta con aplicar en ambos términos de la ecuación la función inversa del logaritmo neperiano, es decir la función exponencial, obteniendo

$$\begin{aligned} y &= \mathsf{e}^{\ln y} \\ &= \mathsf{e}^{0{,}623+0{,}458x} \\ &= \mathsf{e}^{0{,}623}\left[\mathsf{e}^{0{,}458}\right]^x \\ &= 1{,}86 \cdot 1{,}58^x \end{aligned}$$

Ya tenemos la función deseada:

$$Y = 1{,}86 \cdot 1{,}58^X$$

Esto significa que en un principio el embrión pesa 1.5mg, y que cada día transcurrido su peso se multiplica por 1.58, aproximadamente, claro está. Sabemos que esta aproximación es muy buena por el valor de r^2 obtenido.

Así pues, si damos con el cambio o los cambios de variables apropiados podemos resolver mediante la recta de regresión problemas con relaciones no lineales. En ese sentido decimos que el sencillo estudio de correlación-regresión lineal sirve de referencia para situaciones más complejas.

2.2. Relación entre dos caracteres cualitativos

La segunda parte del capítulo está dedicada al estudio de la relación entre dos caracteres cualitativos. Al igual que en los análisis anteriores, distinguiremos entre la tabulación de los datos, su representación gráfica y el cálculo de valores típicos.

2.2.1. Tabla de Contingencia. Coeficiente C de Pearson

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 A y B, en los que se distinguen r y s categorías, respectivamente. Es decir, la evaluación del carácter A puede dar lugar a r resultados posibles $A_1, A_2, ..., A_r$ y la del carácter B, a s resultados posibles $B_1, B_2, ..., B_s$. Reservaremos el subíndice i para denotar los niveles de A y el j para los de B.

Ejemplo 6:[Tabla de Contingencia 3×3]

Se realiza un estudio a nivel cualitativo para considerar la posible asociación entre el nivel de SO_2 en la atmósfera 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, con lo que el número total es $n = 60$. En cada caso se determina su nivel de cloroplastos. La tabla obtenida tras clasificar los 60 árboles, denominada de contingencia, fue la siguiente:

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

Empecemos con una breve descripción de la tabla. En este caso se distinguen $r = 3$ categorías o niveles para el carácter A fila (nivel de SO_2) y otras $s = 3$ categorías para el carácter B columna (nivel cloroplastos). De ahí que la tabla sea del tipo 3×3. Los valores que aparecen en las 9 casillas se denominan valores observados y se denotan mediante O_{ij}. Así, por ejemplo, tenemos $O_{11} = 3$, $O_{12} = 4$, $O_{23} = 5$, etc. A la derecha se expresan las sumas de las diferentes filas, que se denotan por $O_{i\cdot}$. Tenemos concretamente $O_{1\cdot} = 20$, $O_{2\cdot} = 20$ y $O_{3\cdot} = 20$. En este caso son todas iguales por el diseño utilizado, pero no tiene por qué ser así. De igual forma, se expresan abajo las sumas de las columnas, que se denotan por $O_{\cdot j}$. Así, $O_{\cdot 1} = 15$, $O_{\cdot 2} = 25$ y $O_{\cdot 20} = 20$. Por último, las suma de todas las observaciones es $n = 60$, que coincide tanto con la suma de las filas como con la suma de las columnas.

Todo nuestro estudio se basa en el análisis de las diferentes proporciones que se dan en la muestra, tanto brutas como condicionadas. Entre las primeras distinguimos las proporciones de las

diferentes categorías de A (SO_2). De esta forma, la proporción de árboles de la muestra que se encuentran en zonas con nivel alto de SO_2 es

$$\hat{P}(SO_2 \texttt{ alto}) = \frac{20}{60} = 0.33$$

En general se tiene que

$$\hat{P}(A_i) = \frac{O_{i\cdot}}{n}$$

Respecto a las distintas categorías de B (cloroplastos), la proporción de árboles de la muestra que presentan un nivel medio de cloroplastos es

$$\hat{P}(\texttt{Cloroplastos medio}) = \frac{25}{60} = 0.42$$

En general,

$$\hat{P}(B_j) = \frac{O_{\cdot j}}{n}$$

También, dadas sendas categorías de cada carácter, podemos calcular la proporción que supone respecto al total de la muestra una combinación o intersección de ambas. Por ejemplo,

$$\hat{P}(SO_2 \texttt{ alto y Cloroplastos medio }) = \frac{4}{60} = 0.067$$

En general,

$$\hat{P}(A_i \cap B_j) = \frac{O_{ij}}{n}$$

Hemos de destacar que las proporciones se denotan por $\hat{P}$ en lugar de P con la idea de resaltar que son parámetros descriptivos, es decir, que se refieren a la muestra estudiada, no al total de la población objeto del estudio, como veremos en el capítulo 5. También podemos hablar de las siguientes proporciones denominadas condicionadas, pues se calculan suponiendo que se verifique una categoría de las filas o de las columnas. Así, la proporción de árboles con SO_2 alto que presenta un nivel bajo de cloroplastos es

$$\hat{P}\big(\texttt{Cloroplastos bajo}\big|SO_2 \texttt{ alto}\big) = \frac{13}{20} = 0.65$$

Recíprocamente, la proporción de árboles con nivel medio de cloroplastos que presenta un nivel de SO_2 alto es

$$\hat{P}\big(SO_2 \texttt{ alto}\big|\texttt{Cloroplastos medio}\big) = \frac{4}{25} = 0.16$$

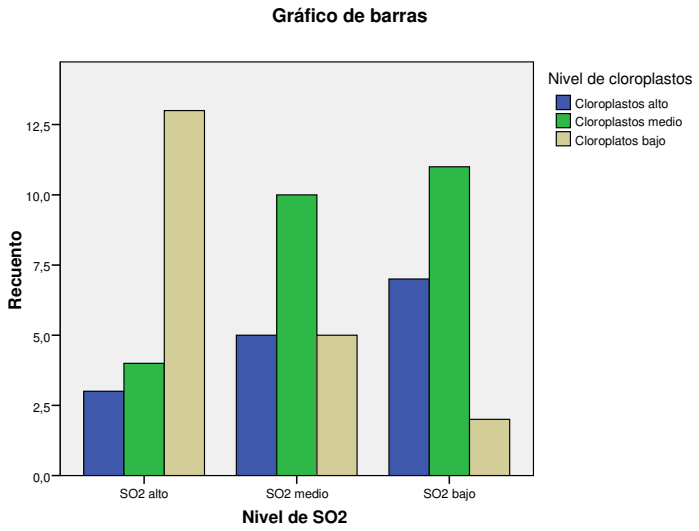
En general,

$$\hat{P}(A_i|B_j) = \frac{O_{ij}}{O_{\cdot j}}, \qquad \hat{P}(B_j|A_i) = \frac{O_{ij}}{O_{i\cdot}}$$

También pueden obtenerse las proporciones condicionales de esta otra forma:

$$\hat{P}(A_i|B_j) = \frac{\hat{P}(A_i \cap B_j)}{\hat{P}(B_j)} \qquad \hat{P}(B_j|A_i) = \frac{\hat{P}(A_i \cap B_j)}{\hat{P}(A_i)} \tag{2.1}$$

Un gráfico muy útil a la hora de ilustrar la asociación existente entre los dos caracteres es el denominado diagrama de barras agrupadas. En este caso, se muestra un diagrama de barras para cada categoría de SO_2:



Las marcadas diferencias entre los tres diagramas de barras hablan por sí solas de una considerable correlación o asociación entre los factores estudiados. No obstante, el análisis gráfico debe complementarse necesariamente con otro de tipo cuantitativo. Al igual que en el caso de variables numéricas, donde se define un valor típico, el coeficiente de correlación, que mide el grado de correlación lineal existente entre las variables, calcularemos aquí otro coeficiente que nos dará una medida del grado de dependencia existente entre los caracteres.

Debemos tener claro qué entendemos por dependencia entre dos caracteres: que las proporciones en las que se distribuyen las categorías de un carácter varíen en función de la categoría que corresponde al otro. Es lo que sucede con nuestros datos, pues observamos que la proporción de árboles muy sanos depende del grado de contaminación: de hecho es mucho más alta en las zonas poco contaminadas que en las muy contaminadas. Eso equivale a afirmar que la proporción bruta de árboles muy sanos varía al condicionar respecto al nivel de contaminación, de manera que en las zonas poco contaminadas aumenta y en las muy contaminadas disminuye. En general, se dice que los caracteres A y B presentan dependencia sobre la muestra cuando existen niveles i y j tales que

$$\hat{P}(B_j|A_i) \neq \hat{P}(B_j)$$

Para que no se apreciara el menor grado de dependencia en la muestra debería ocurrir pues que, para todas las categorías i y j de A y B respectivamente, se verificase que $\hat{P}(B_j|A_i) = \hat{P}(B_j)$, lo cual equivaldría afirmar que

$$\hat{P}(A_i \cap B_j) = \hat{P}(A_i) \times \hat{P}(B_j)$$

Para que eso sucediera debería verificarse

$$\frac{O_{ij}}{n} = \frac{O_{i\cdot}}{n} \times \frac{O_{\cdot j}}{n}$$

Es decir, en una muestra que presenta unos valores $O_{i\cdot}$ y $O_{\cdot j}$ determinados no se observaría grado alguno de dependencia si el valor observado para las categorías i y j de los caracteres A y B, respectivamente, fuera igual a

$$E_{ij} = \frac{O_{i\cdot} \times O_{\cdot j}}{n}$$

Estos valores, denominados esperados, son ideales no en el sentido de óptimos sino en el de irreales, pues pueden ser irrealizables en la práctica si poseen decimales. Deben entenderse como valores de referencia de manera que, cuanto más se alejen de ellos los valores realmente observados, mayor será el grado de dependencia. En nuestro ejemplo los valores esperados en el caso de dependencia nula, es decir, independencia, serían los siguientes:

		Nivel cloroplastos			
	(3×3)	Alto	Medio	Bajo	Total
Nivel SO_2	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

Debemos dar una medida de la diferencia o distancia entre la tabla de contingencia real y esta tabla ideal de valores esperados. Dicha distancia global nos dará idea del grado de asociación. Se basará lógicamente en las diferencias $O_{ij} - E_{ij}$, más concretamente y en la línea a la que debemos estar ya habituados, en la suma de sus cuadrados que, además, se ponderarán dividiendo por los respectivos valores esperados. De esta forma se obtiene la denominada distancia χ^2:

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

Debe quedar pues claro que un valor χ^2_{exp} nulo se correspondería con la independencia en los datos de la muestra y que, cuanto mayor sea su valor, más fuerte será la dependencia o correlación observada en la muestra.

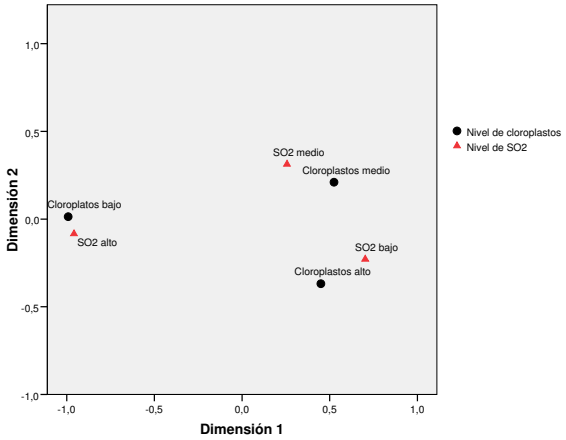
No obstante, no estamos aún en condiciones de valorar debidamente el resultado χ^2_{exp} obtenido. Sin embargo, existe otra medida del grado de asociación derivada del mismo que guarda cierto paralelismo con el coeficiente de correlación r y que sí podremos valorar fácilmente. Se trata del denominado coeficiente de contingencia de Pearson, y se define mediante

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

Este coeficiente debe estar comprendido 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 nuestro caso, la cota máxima es 0.816. Nos preguntamos ahora cómo deberían ser los datos para obtener dicha correlación máxima. Pues, por ejemplo, serviría la siguiente tabla de contingencias:

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

Si la cota máxima es 0.816 y el valor de C que aportan los datos es 0.444, lo cual nos da idea de que en la muestra se observa un grado de correlación medio. Existe un tipo de gráfico más sofisticado denominado biplot que permite evaluar de manera más pormenorizada las asociaciones entre las distintas categorías de los caracteres. En nuestro caso tendríamos lo siguiente:



2.2.2. Tablas 2×2. Coeficiente ϕ.

Este caso particular en el que se distinguen únicamente dos categorías en los dos caracteres considerados puede recibir, además del tratamiento estudiado en el apartado anterior, otro específico que destaca por su sencillez. En ese caso, la tabla de contingencia se reducirá a lo siguiente:

(2×2)	B_1	B_2	Total
A_1	$O_{1,1}$	$O_{1,2}$	$O_{1\cdot}$
A_2	$O_{2,1}$	$O_{2,2}$	$O_{2\cdot}$
Total	$O_{\cdot 1}$	$O_{\cdot 2}$	n

Ejemplo 7:[Tabla de Contingencia 2×2]

Se pretende averiguar en qué medida es efectiva una vacuna contra la hepatitis. Se estudió una muestra de 1083 individuos de los cuales algunos habían sido vacunados y otro no; por otro lado, algunos habían llegado a contraer la hepatitis mientras que otros no. La tabla de contingencia resultante es la siguiente:

		`Vacunación`		
	(2×2)	Sí	No	Total
`Hepatitis`	Sí	11	70	81
	No	538	464	1002
	Total	549	534	1083

Para un caso de este tipo, a la hora de medir el grado de asociación de los caracteres en la muestra, podemos utilizar, además del conocido coeficiente C, el denominado coeficiente ϕ, que

se define mediante $\phi^2 = \chi^2_{exp}/n$, o lo que es lo mismo,

$$\phi = \sqrt{\frac{(O_{1,1}O_{2,2} - O_{1,2}O_{2,1})^2}{O_{1.}O_{2.}O_{.1}O_{.2}}}.$$

Si analizamos detenidamente la última expresión, concluiremos que ϕ^2 es una parámetros completamente análogo al coeficiente de correlación lineal r^2. 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, que se obtiene cuando la tabla anterior es diagonal. Es lo que habría ocurrido si los datos de la muestra hubieran sido los siguientes:

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

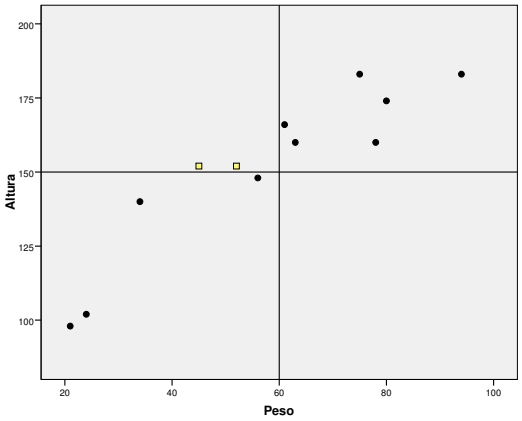
La asociación nula se daría, por ejemplo, si nuestros datos fueran los siguientes:

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

Efectivamente, podemos observar que, tanto en el caso de vacunados como en el de no vacunados, la proporción de individuos afectados es 1/3.

En nuestro ejemplo concreto se obtiene $\phi = 0.211$. 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 $C = 0.206$. Estos valores nos hablan del grado de asociación entre vacunación y hepatitis, es decir, de la eficacia de la vacuna, en la muestra considerada. Las conclusiones obtenidas se ciñen exclusivamente a dicha muestra, 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. Cabe incluso pensar que los individuos hayan sido seleccionados intencionadamente para obtener unos resultados que favorezcan la comercialización de la vacuna, o todo lo contrario. Nos planteamos pues por primera vez el problema de Inferencia Estadística, que intentaremos resolver a partir del próximo capítulo.

Para hacer hincapié en la semejanza entre los parámetros r^2 y ϕ^2, podemos tratar de una forma cualitativa los datos correspondientes al ejemplo 4, que volvemos a representar:



Efectivamente, las medias aritméticas $\overline{x}$ y $\overline{y}$ dividen los ejes OX y OY, dando lugar a cuatro cuadrantes en los que se distribuyen los puntos 12 puntos. Los que se ubican en los cuadrantes superior izquierdo (dos) e inferior derecho (ninguno) rompen la tendencia que manifiestan los restantes puntos, haciendo disminuir así la correlación.

		X		
		$-$	$+$	Tot
Y	$+$	2	6	8
	$-$	4	0	4
	Tot	6	6	12

2.3. Cuestiones Propuestas

1. 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 $r = 0$.

2. 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?

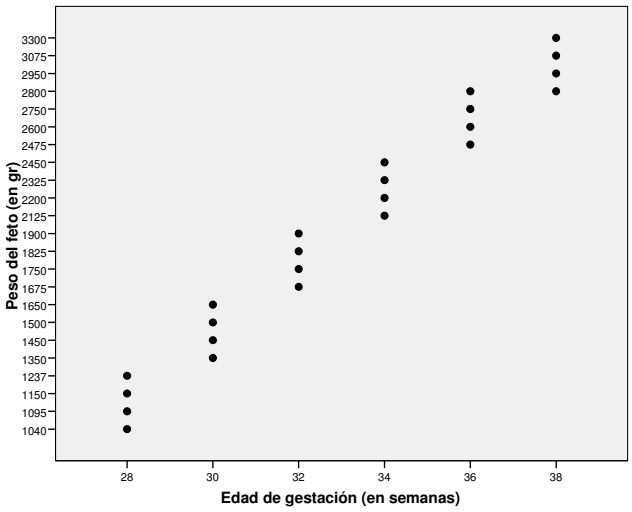
3. Si la dependencia de cierta variable Y respecto a otra variable X se expresa mediante una ley de la forma
$$y = \frac{K}{x^2},$$
¿qué cambio de variables *linealiza* la anterior relación?

4. Se estudian las variables X (masa en kg) e Y (longitud total en m) sobre una muestra de 1200 unidades experimentales. Se obtienen los siguientes parámetros descriptivos:
$\overline{x} = 2{,}00, \quad s_x = 0{,}10, \quad \tilde{x} = 1{,}99, \quad p_{16} = 1{,}91, \quad p_{84} = 2{,}12.$
$\overline{y} = 1{,}50, \quad s_y = 0{,}80, \quad \tilde{y} = 1{,}30, \quad g_1^y = 3{,}2, \quad r = -0{,}63.$

A Razona cuál de las dos variables se ajusta más satisfactoriamente a un modelo Normal. Esboza, basándote en los datos con los que contamos, las correspondientes representaciones gráficas.

B Si se expresaran la masa en gr y la longitud en mm, ¿cuáles serían los valores de $\overline{x}$, s_x^2 y r? Razónalo.

5. En el siguiente diagrama de dispersión se presentan 24 datos correspondientes a la medición del peso de un feto en función de su edad de gestación, comprendida en todo caso entre 28 y 38 semanas.



El valor del coeficiente de determinación es $r^2 = 0{,}964$ y la recta de regresión muestral es $y = -4301 + 192x$. Comentar los aspectos más relevantes, interpretando en términos muy prácticos el valor de r^2. ¿Qué utilidad puede tener la recta anterior?

6. En un estudio sobre la posible relación entre las concentraciones de calcio (en mg/100ml) y de hormona paratiroidea (en mug/ml) en plasma en individuos sanos, se obtuvieron los datos siguientes:

X pth	11.0	11.0	10.6	10.5	10.6	10.4	10.2	9.5	8.2	7.5	6.0	5.0
Y Ca	0.30	0.50	1.12	1.23	1.24	1.31	1.33	2.10	2.15	2.43	3.70	4.27

a) Representa la nube de puntos. ¿Qué tipo de relación se observa ente ambas variables?

b) Haciendo uso de un programa estadístico o, en su defecto, de una calculadora científica, obtener r y la recta de regresión muestral. Interpretar r^2 en términos muy prácticos.

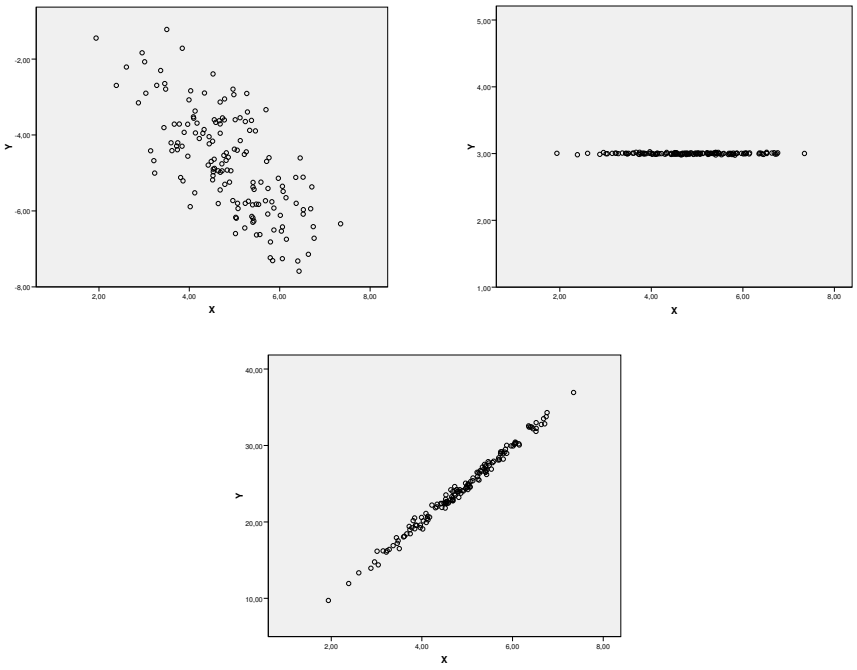
7. Se ha medido la presión sistólica (mm. Hg) en 12 individuos para relacionarla con la edad (años) de los mismos. Los resultados fueron los siguientes

X (edad)	30	50	60	30	70	60	60	40	40	50	70	40
Y (presión)	107	136	148	109	158	150	145	120	118	134	162	124

a) Representa la nube de puntos.

b) Haciendo uso de un programa estadístico, calcular r y la recta de regresión muestral. Interpretar r^2 en términos muy prácticos.

8. Indicar qué valor aproximado puede tener r en los siguientes ejemplos:



9. 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 variable en 7 ocasiones, con las mismas condiciones ambientales, se obtuvo:

$[S]$	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

a) Representa la nube de puntos.

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.

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.

10. Se estudia la posible relación entre la exposición a un agente radioactivo y la presencia de una determinada enfermedad en los individuos de una población. Se seleccionó una muestra compuesta por 620 individuos, distinguiendo en cada caso si el individuo estaba o no expuesto al agente y si padecía o no dicha enfermedad. Se obtuvo la siguiente tabla (2×2):

	Expuesto	No expuesto	Total
Enfermo	52	248	
No enfermo	48	272	
Total			620

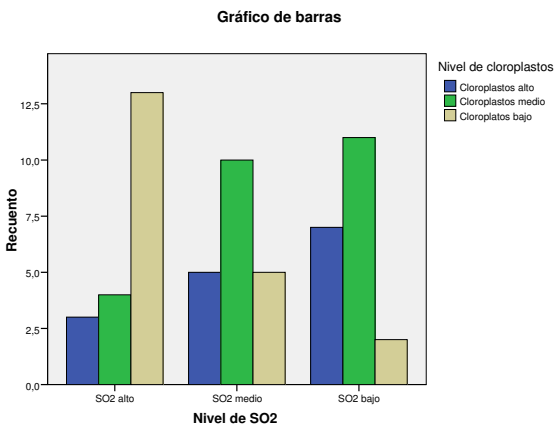
a) ¿Qué proporción de individuos de la muestra están expuestos al agente? ¿Qué proporción de individuos enfermos están expuestos al agente? ¿Qué proporción de individuos están expuestos y no están enfermos?

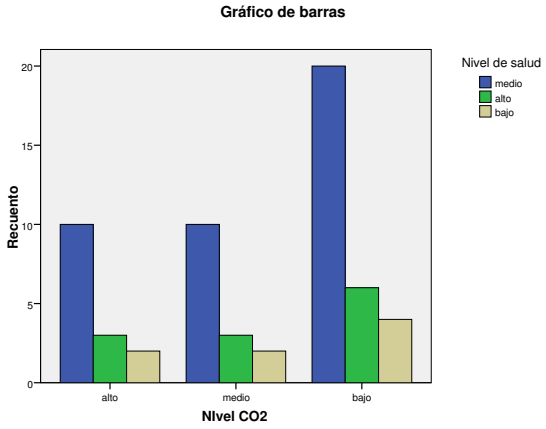
b) ¿Entre qué valores se encuentra, en todo caso, el coeficiente ϕ? Calcúlalo.

c) ¿Entre qué dos valores se encuentra el coeficiente C en una tabla 2×2? Cálculalo.

d) Valorar en términos muy prácticos el grado de correlación que refleja la muestra escogida.

11. Razonar en cuál de los dos casos obtendremos un coeficiente de contingencia mayor:





12. En un estudio sobre el efecto de tres técnicas diferentes utilizadas en reproducción asistida para el desarrollo *in vitro* de una muestra de óvulos fecundados, se obtuvieron los siguientes resultados:

		Desarrollo		
		Correcto	Defectuoso	Nulo
Técnica	Tratamiento A	23	9	6
	Tratamiento B	21	4	3
	Tratamiento C	34	24	17

¿Entre qué valores estará comprendido el coeficiente C? Calcúlalo, a ser posible con la ayuda de un programa estadístico, y valora el resultado.

Capítulo 3

Probabilidad

En contra de ciertas preconcepciones bastante extendidas, la Teoría de la Probabilidad, que introduciremos en el presente capítulo, constituye una disciplina con autonomía respecto a la Estadística. De hecho, los inicios y motivaciones de ambas materias fueron absolutamente dispares: mientras que la primera surge del estudio de los juegos de azar, la segunda emana de la necesidad de clasificación e interpretación de datos referentes a poblaciones. La fusión de ambas especialidades se produce avanzado el siglo `XIX`, como consecuencia de diversos estudios acerca de la evolución de las especies. Intentaremos ilustrar más adelante el porqué de la conexión entre ambas materias.

En cuanto a la Probabilidad hemos de decir que, si bien sus comienzos pueden presentar cierto tinte de frivolidad, su campo de aplicación se ha ido extendiendo paulatinamente al describirse multitud de fenómenos, a parte de los consabidos juegos de azar, que se ajustan a lo que entendemos por fenómenos aleatorios. No obstante, existen diversas opiniones respecto a este hecho, algunas ciertamente radicales, pues el concepto de azar es objeto de polémica. En la primera sección del capítulo intentaremos precisamente profundizar en dicho concepto. Ya advertimos en la introducción que la mayor parte del capítulo puede pecar de excesivo formalismo, de ahí que se recomiende el lector interesado en la Probabilidad y Estadística como mera herramienta para el análisis de datos una lectura rápida, que no obstante puede ser suficiente para afrontar los capítulos siguientes. En todo caso aconsejamos tener bien presente al menos el párrafo en el recuadro de la sección 3.1, que supone en cierta medida una desmitificación del concepto de probabilidad.

3.1. Fenómeno aleatorio

En esta sección intentaremos delimitar qué entendemos por fenómeno aleatorio y fabricaremos el modelo matemático que lo formaliza.

3.1.1. ¿Sabe alguien qué es el azar?

Solemos decir que un fenómeno es determinista cuando podemos predecir su resultado. Por contra, existen multitud de fenómenos cuyo desenlace no puede preverse pues ofrecen múltiples posibilidades. En ese caso, se denomina `suceso` a cualquiera de las posibles situaciones que en principio puedan acaecer tras la ejecución del experimento. Vamos a centrar nuestra atención en aquellos fenómenos no deterministas que verifica la siguiente propiedad:

(i) Pueden repetirse tantas veces como se quiera y aparentemente en idénticas circunstancias sin que el resultado de una ejecución pueda evidenciar una deformación o variación respecto a las mismas.

Tal podría ser el caso, por poner un ejemplo, de una serie de lanzamientos de una misma moneda. Efectivamente, no podemos predecir si el resultado de cada lanzamiento será cara o cruz, pero podemos aceptar que todos los lanzamientos se efectúan en igualdad de condiciones sin que el hecho de que un lanzamiento resulte cruz altere dicha perspectiva en los lanzamientos sucesivos.

En tal caso, pretendemos explicar este fenómeno observable mediante un modelo matemático en el que se asigna a cada suceso una medida precisa y cuantitativa de su grado de posibilidad. Podemos convenir que sea un número en el intervalo $[0, 1]$, de manera que un 0 significa que el suceso es imposible y un 1 significa que es seguro. De acuerdo con la condición (i), debe ser idéntica para toda la serie de ejecuciones. Además, la propia serie da lugar a sucesos compuestos (por ejemplo, tras dos lanzamientos de una moneda podemos hablar de los sucesos cara-cara, cara-cruz, cruz-cara o cruz-cruz). Teniendo en cuenta de nuevo la condición (i), la medida del grado de posibilidad de un suceso compuesto debe obtenerse de manera multiplicativa (es decir, la medida de la posibilidad de obtener cara-cruz se obtiene multiplicando la de cara por la de cruz). En este modelo matemático puede demostrarse que en una serie infinita de repeticiones es seguro que la proporción de resultados favorables a un suceso converja a la medida del grado de posibilidad que le hemos asignado. Dicha medida se denomina probabilidad.

Por lo dicho anteriormente concluimos que, para que en un fenómeno real pueda hablarse con propiedad de probabilidad, a la propiedad (i) debe añadírsele por coherencia esta otra:

(ii) Para cualquier suceso considerado, las proporciones de resultados favorables al mismo tienden a estabilizarse tras un gran número de repeticiones.

De hecho, la probabilidad del suceso coincidirá con el número hacia el que se aproxima dicha proporción. Decimos entonces que el fenómeno sigue la Ley del Azar y es por lo tanto aleatorio. La propiedad (ii) puede expresarse así: si A denota un suceso y $\hat{P}_n(A)$ la proporción de resultados favorables al mismo tras n repeticiones del experimento, es decir, la frecuencia relativa, existe un número $P(A)$ tal que

$$\text{(ii)} \qquad \boxed{\lim_{n\to\infty} \hat{P}_n(A) = P(A)}$$

A continuación nos planteamos la siguiente pregunta: ¿existen realmente fenómenos aleatorios? Pues parece ser que sí. ¿La respuesta dada se basa en premisas racionales o es de carácter empírico? Pues un poco de todo. De hecho, podemos establecer dos categorías de fenómenos aleatorios, más otra de propina:

Fenómenos a priori aleatorios

Nos referimos a aquéllos que responden a una clara simetría. Es el caso de una ruleta (círculo), una lotería (esferas), el lanzamiento de una moneda, de un dado convencional, es decir, un cubo, o de cualquier otro sólido platónico: tetraedro, octaedro, dodecaedro o icosaedro regulares. Cabe incluso conjeturar si en el fondo de todo fenómeno aleatorio existe una razón relacionada con la simetría, es decir, que lo que comúnmente denominamos azar no sea sino la consecuencia de una simetría más o menos compleja y más o menos evidente.

Dodecaedro e Icosaedro

En todo caso, en fenómenos de este tipo no parece que haya mucho inconveniente en asumir que pueden repetirse tantas veces como se quiera en igualdad de condiciones sin que el resultado de una ejecución condicione el de las restantes. Podemos aceptar pues la propiedad (i). En estas circunstancias, la propia geometría nos conduce al concepto de equiprobabilidad, pues no parece tampoco difícil convencerse de que, por ejemplo, los 6 lados de un cubo perfecto (simétrico) tienen un mismo grado de posibilidad de quedar arriba una vez lanzado el cubo (dado).

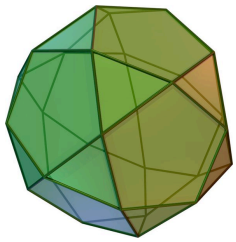
Nótese que es la propia simetría la que nos ha permitido asumir la propiedad (i). Efectivamente, si consideramos que la simetría en el fenómeno es extensible a la repetición del experimento, nada hace pensar que, en 10 ejecuciones del lanzamiento del dado, alguna de las 6^{10} posibles series resultantes tenga mayor grado de posibilidad de ocurrir que el resto. En particular, el experimento puede repetirse sin que el resultado de un lanzamiento condicione el de otro distinto.

Parece pues claro que estamos en disposición de formular un modelo matemático en el que asignamos a cada suceso una medida a priori de su grado verosimilitud. En este caso, a cada lado del cubo se le asigna probabilidad 1/6. La comprobación empírica de que, tras una larga serie de lanzamientos de un dado, las proporciones de resultados favorables a cada puntuación sea próxima a 1/6 debe considerarse un claro signo de que el modelo matemático propuesto se adecua satisfactoriamente al fenómeno real estudiado. Y, efectivamente, si el lector está lo suficientemente aburrido podrá comprobar cómo tras lanzar 100, o mejor 1000 veces un dado, la proporción de cincos obtenidos es muy próxima a 1/6. No descartamos que dicha proporción no converja realmente a 1/6, o que ni siquiera converja, pero estamos predispuestos a interpretar ese hecho como un defecto de construcción del dado. Es lo que denominaríamos un dado trucado

Fenómenos aleatorios a posteriori

Podemos pensar en otro tipo de fenómeno aleatorio que, al menos en apariencia, no se explica por un argumento de pura simetría. Son comúnmente admitidos como fenómenos aleatorios las variaciones accidentales en un proceso de medición o fabricación. Decimos que son fenómenos aleatorios en tanto en cuanto se dan la propiedades (i) y (ii). Puede que la primera pueda asumirse en virtud de la propia naturaleza del experimento, como ocurre con el lanzamiento de una moneda; sin embargo, se antoja imprescindible contrastar empíricamente la segunda propiedad (ley de azar), pues su violación dejaría patente la ineptitud del modelo matemático basado en el concepto de probabilidad a la hora de formalizar el fenómeno real.

Nos preguntamos, por ejemplo, si el lanzamiento de un sólido arquimediano, como el icosidodecaedro (sus caras forman 20 triángulos equiláteros y 12 pentágonos regulares) es un fenómeno aleatorio.



Icosidodecaedro

Posiblemente lo sea. No obstante, aunque podamos asumir la condición (i), respecto a la condición (ii) convendría contabilizar el número de veces en las que el poliedro cae sobre un pentágono y comprobar que la proporción de resultados favorables tiende a estabilizarse a medida que repetimos el experimento. Sólo entonces, es decir *a posteriori*, podremos aceptar que el fenómeno es aleatorio y la probabilidad de caer en pentágono será parecida a la frecuencia relativa de la serie.

Aunque sucediera eso no podemos pensar en una probabilidad universal para todos los icosidodecaedros, pues no se puede descartar que las frecuencias relativas converjan a distintos números dependiendo de si el poliedro utilizado es hueco o macizo, o incluso de su volumen, densidad, etc. En todo caso, en fenómenos de este tipo las probabilidades correspondientes a cada suceso no pueden calcularse *a priori*. Sólo podemos obtener una aproximación empírica a las mismas tras una larga serie de repeticiones.

Fenómenos inciertos

Son los más abundantes, quizás los únicos. Nos referimos a fenómenos como un partido de fútbol, la presencia de una enfermedad, la talla de un recién nacido, etc. No pueden considerarse aleatorios según hemos convenido dado que no verifican la condición (i), pues ni siquiera pueden volver a repetirse en idénticas circunstancias. Por lo tanto y en rigor, no deberíamos hablar de probabilidad en estos casos. Este abuso del término, muy frecuente en el lenguaje habitual, se debe a la idea bastante extendida de que los fenómenos se dividen en dos clases: deterministas y aleatorios. Deberíamos decir quizás que todo fenómeno descompone en una componente determinista y otra aleatoria.

Efectivamente, este tipo de fenómenos inciertos no pueden repetirse en idénticas condiciones porque existen unas causas o factores concretos que influyen en el resultado y fluctúan en las diversas repeticiones del experimento. La conjunción de dichos factores da lugar a una componente determinista a la que posiblemente se sume otra componente aleatoria en sentido estricto que sí verifica las condiciones (i) y (ii). De hecho, desde el punto de vista estadístico el diseño de un experimento tiene como objetivo aislar lo mejor posible esa componente aleatoria pura.

Por poner un ejemplo, no podemos afirmar que el lanzamiento de icosidodecaedros sea un fenómeno aleatorio porque es muy posible que las tendencias en los lanzamientos dependan de factores como el volumen o densidad del objeto utilizado. Sin embargo, si controlamos estos factores, lo cual puede conseguirse utilizando el mismo objeto en toda la serie lanzamientos, tal vez podría considerarse aleatorio a *posteriori*. En ese caso, podríamos diseñar varios experimentos paralelos con icosidodecaedros de distinto volumen o composición, para determinar si estos factores influyen realmente en las probabilidades obtenidas en cada caso.

La descomposición de los fenómenos en componentes deterministas y aleatorias viene a ser una solución ecléctica entre dos posturas radicalmente enfrentadas: por un lado, una que entiende que se habla de simetría o equiprobabilidad en aquellas circunstancias en las que renunciamos por completo a controlar las causas del resultado; es decir, que la clasificación de los fenómenos en deterministas y no deterministas no obedece a la naturaleza de los mismos sino a nuestra capacidad o actitud a la hora de explicarlos. Desde ese punto de vista, el azar no sería más que una especie de saco donde se refugian las causas que no podemos o no queremos controlar. Cabría entonces esperar que el progreso científico fuera menoscabando los dominios del azar. No obstante, parece haber sucedido lo contrario. En ese sentido son paradigmáticos los casos de la Física Cuántica que introduce el concepto de azar para explicar el comportamiento de lo pequeño, o la Teoría de la Evolución de Darwin, según la cual el azar es en última instancia el motor de los cambios biológicos. ¿No podría incluso pensarse que el azar o simetría es la explicación última de todo fenómeno, de manera que incluso aquello que damos por seguro sea sólo muy probable? Ésa sería la postura contraria.

3.1.2. El modelo de probabilidad

Tras esta delicada discusión y aunque han sido ya esbozados anteriormente, pasamos a determinar con claridad los elementos que intervienen en el modelo probabilístico asociado a una fenómeno o experimento aleatorio. Primeramente, debemos distinguir el modelo que corresponde a una única ejecución del experimento del que corresponde a una serie de n repeticiones verificando (i). Al primero lo denominaremos modelo de probabilidad original y al segundo, modelo de probabilidad producto. Advertimos ahora y volveremos a hacerlo al final de la sección que estas distinciones no responden en esencia a aspectos formales sino sólo didácticos, y que en la práctica podremos hablar de un único modelo de probabilidad, a secas.

Modelo original

Pensemos como ejemplo en el lanzamiento de un dado simétrico. Lo primero que debemos tener en cuenta es el conjunto pormenorizado de los posibles resultados en que puede desembocar el experimento. Dicho conjunto, que se denota por la letra Ω, se denominará espacio original. En el caso del dado distinguiremos seis posibilidades, tantas como caras tiene el cubo, es decir:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$

Si entendemos por suceso cualquier circunstancia susceptible de ocurrir tras la ejecución del experimento, debemos definirlo formalmente como cualquier subconjunto de Ω. Por ejemplo, que ocurra el suceso $\{2, 4, 6\}$ significa que el resultado del lanzamiento sea `par`. En general, decimos que se verifica un suceso cuando el resultado del experimento es cualquiera de los elementos que lo componen. El propio espacio Ω es un suceso que, por lo tanto, ocurre siempre, de ahí que se denomine suceso seguro. Por contra, el espacio vacío $\emptyset$ no ocurre nunca pues suponemos que el experimento aporta siempre un resultado, de ahí que se denomine suceso imposible. Es un elemento necesario en el álgebra de sucesos. Los elementos de Ω son a su vez sucesos, con la particularidad de que no pueden descomponerse en otros más simples. Se denominan sucesos elementales.

El conjunto de los sucesos está dotado de un algebra que nos permite unirlos, intersecarlos y complementarlos. Concretamente, dados dos sucesos A y B, se verificará la unión $A \cup B$ cuando se

verifique A o B (o ambos); se verificará la intersección $A \cap B$ cuando se verifiquen simultáneamente A y B, y el complementario $\overline{A}$ cuando no se verifique A. Decimos que dos sucesos son incompatibles o disjuntos cuando $A \cap B = \emptyset$.

Una vez configurado el espacio inicial y, en consecuencia, el conjunto de los posibles sucesos, debemos asignar a cada uno de ellos su **probabilidad**, que será un número en el intervalo $[0, 1]$ que asigne un 1 al suceso seguro y con las característica propia de una medida, es decir, que si A y B son incompatibles entonces

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

La probabilidad de cualquier suceso es igual por lo tanto a la suma de las probabilidades de los sucesos elementales que lo componen. En el caso de que la aleatoriedad del fenómeno responda a una simetría perfecta, como es el caso del dado, los sucesos elementales serán equiprobables. Por lo tanto, cuando se da una simetría perfecta, la probabilidad de un suceso cualquiera será igual al número de sucesos elementales que lo componen dividido por el número total de sucesos elementales, es decir, será el cociente entre el número de casos favorables al suceso y el número de casos posibles. Así, por ejemplo, la probabilidad de que el resultado de un lanzamiento sea par es 3/6.

Hemos visto que existe compatibilidad entre la unión disjunta de sucesos y la suma de probabilidades. ¿Es también cierto que la intersección de sucesos se traduce en el producto de sus probabilidades? En general no. Por ejemplo, en el caso del lanzamiento de un dado, la intersección de los sucesos `par` e `impar` es el conjunto vacío, luego su probabilidad es nula. Sin embargo, la probabilidad de `par` multiplicada por la probabilidad de `impar` es igual a 1/4. Decimos que dos sucesos A y B son independientes cuando sí se verifica que $P(A \cap B) = P(A) \times P(B)$. En caso contrario se dice que son dependientes.

Por ejemplo, son independientes los sucesos `múltiplo de 2` y `múltiplo de 3`. Efectivamente, el primero está compuesto por $\{2, 4, 6\}$ siendo su probabilidad 1/2; el segundo está compuesto por $\{3, 6\}$ siendo su probabilidad 1/3; la intersección de ambo sucesos es el suceso elemental $\{6\}$, cuyas probabilidad puede obtenerse multiplicando 1/2 por 1/3. Un ejemplo más ilustrativo del concepto de independencia podemos encontrarlo en el lanzamiento de dos dados que veremos a continuación.

Modelo producto

El modelo producto de orden n pretende explicar globalmente el fenómeno aleatorio, pues viene a formalizar n ejecuciones del experimento aleatorio. Un ejemplo muy sencillo puede ser dos lanzamientos consecutivos de un dado o, equivalentemente, el lanzamiento simultáneo de dos dados. El espacio Ω^n de las posibles series de resultados se denomina espacio muestral. En nuestro ejemplo tendríamos el espacio $\Omega^2 = \{(1, 1), (1, 2), \ldots, (6, 6)\}$ con un total de 36 elementos denominados series o muestras aleatorias. El hecho de que las repeticiones verifiquen (i) se formaliza construyendo la probabilidad P^n sobre este espacio como producto n veces de la probabilidad original. Efectivamente, si, por ejemplo, lanzamos dos veces un dado, podemos obtener un total de 36 series o muestras aleatoria de tamaño 2 diferentes, y por pura simetría hemos de asignar a cada cual idéntica probabilidad, es decir, 1/36. Nótese entonces que la probabilidad P^2 en el espacio muestral se obtiene de forma multiplicativa a partir de la probabilidad P en el espacio

original. Por ejemplo:

$$\begin{array}{ccccc} P^2(\texttt{dado[1]=5},\texttt{dado[2]=3}) & = & P(\texttt{dado[1]=5}) & \times & P(\texttt{dado[2]=3}) \\ \frac{1}{36} & = & \frac{1}{6} & \times & \frac{1}{6} \end{array}$$

Otro ejemplo:

$$\begin{array}{ccccc} P^2(\texttt{dado[1]=par},\texttt{dado[2]=par}) & = & P(\texttt{dado[1]=par}) & \times & P(\texttt{dado[2]=par}) \\ \frac{9}{36} & = & \frac{3}{6} & \times & \frac{3}{6} \end{array}$$

En definitiva, al construir la probabilidad P^2 como producto de una misma probabilidad asumimos implícitamente que los sucesos relativos al resultado del primer dado son independientes de los suecos relativos al segundo. De esta manera se formaliza la condición (i).

Otro ejemplo más: consideremos 5 lanzamientos de una moneda simétrica. El espacio original es $\Omega = \{\texttt{C},\texttt{X}\}$, teniendo ambos sucesos elementales probabilidad 1/2. El espacio muestral es

$$\Omega^5 = \{\texttt{CCCCC},\texttt{CCCCX},\texttt{CCCXC},\texttt{CXXXC},\ldots,\texttt{XXXXX}\}$$

con un total de $2^5 = 32$ series o muestras aleatorias equiprobables, es decir, la probabilidad de cada uno de ellos es $(1/2)^5 = 1/32$

Repetimos que, a pesar de haber distinguido dos tipos diferentes de modelos probabilísticos no existe distinción formal entre ellos pues comparten los dos elementos esenciales: un conjunto de posibilidades y una función de probabilidad sobre el mismo. Denominamos modelo de probabilidad, a secas, a este marco teórico común. No debe pues preocuparnos si el modelo que tenemos entre manos es original o se deriva de otro como producto. De hecho y para simplificar la notación, hablaremos en todo caso de una probabilidad P, sea cual sea el tipo de espacio sobre el que se define.

3.2. Distribución de probabilidad

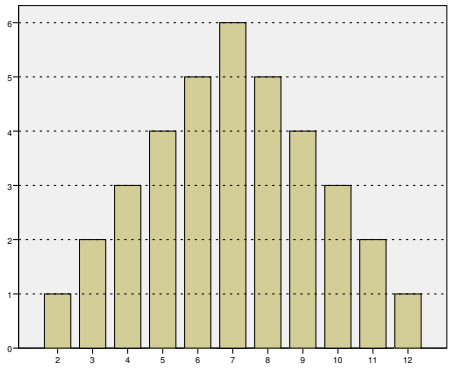
En la práctica, el estudio de un determinado fenómeno aleatorio no consiste en la descripción exhaustiva de los sucesos elementales derivados del mismo, sino en el análisis de uno o varios caracteres cuantitativos considerados. La medición X de cualquier carácter sobre cada suceso elemental constituye lo que denominamos variable aleatoria. Por lo tanto, si nuestro estudio se centra en un determinado carácter lo que nos importa realmente es determinar su distribución de probabilidad, lo cual significa conocer qué valores puede tomar la variable y con que probabilidad en cada caso. Se denomina también distribución teórica para distinguir la de la distribución de frecuencias estudiada en Estadística Descriptiva.

3.2.1. Función de probabilidad

Retomemos el ejemplo del lanzamiento de dos dados. Sabemos que en determinados juegos de azar no importa exactamente cuál ha sido el resultado de cada uno de los dados sino la suma X de ambas puntuaciones. Ése es un sencillo ejemplo de variable aleatoria, que puede tomar 11 valores diferentes, concretamente $\texttt{x}_1 = 2$, $\texttt{x}_2 = 3$, $\texttt{x}_3 = 4$,...,$\texttt{x}_{11} = 12$. Si suponemos una simetría perfecta,

podemos determinar su distribución de probabilidad contabilizando el número de casos favorables a cada resultado de la variable dividido por el número total de casos que presenta el espacio, es decir, 36:

$\mathtt{x}_i$	$P(X = \mathtt{x}_i)$
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

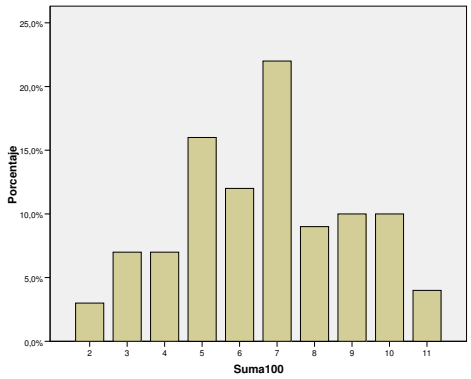


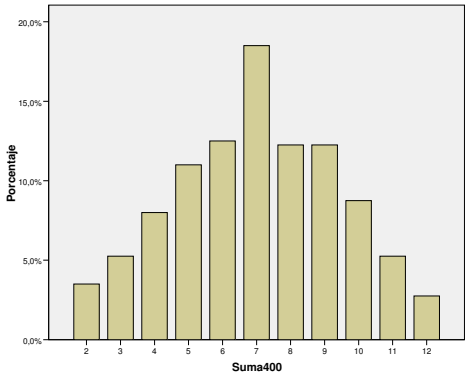
La función que asigna a cada posible valor su probabilidad se denomina función de probabilidad y caracteriza la distribución de la variable. Viene dada por la tabla anterior y guarda una gran similitud con la tabla de frecuencias relativas estudiada en el primer capítulo. Al igual que aquélla se ilustra mediante el denominado diagrama de barras para frecuencias relativas, ésta da lugar al diagrama que vemos a su derecha.

La diferencia entre este gráfico y el diagrama de barras estriba en que en este caso, en lugar de computarse la frecuencia relativa $\hat{p}_i$ para cada valor $\mathtt{x}_i$ de una muestra de tamaño n, se computa la probabilidad p_i de que la variable X tome el valor $\mathtt{x}_i$. Sin embargo, la Ley del Azar establece una clara relación entre ambas. Concretamente, si lanzamos ambos dados un total de n veces obtendremos una muestra aleatoria de tamaño n, lo cual supone n datos[1] comprendidos entre 2 y 12 correspondientes a sendas sumas. Podemos contabilizar las correspondientes frecuencias relativas $\hat{p}_i$. La Ley del Azar viene a decir que si n es grande, entonces

$$\text{(ii)} \qquad \boxed{\hat{p}_i \simeq p_i}$$

Efectivamente, simulamos mediante un programa estadístico 100 lanzamientos de dos dados; posteriormente, simulamos 400 lanzamientos de dos dados. Obtenemos los correspondientes diagramas de barras:





[1]En lo sucesivo cometeremos el abuso de identificar la muestra en sí con los datos que proporciona al medir la variable estudiada.

Se aprecia pues una confirmación empírica de la Ley del Azar, pues podemos observar cómo a medida que aumenta el numero de ejecuciones del experimento, las proporciones de resultados favorables a cada posible valor de la variable (frecuencias relativas) se aproximan a las probabilidades teóricas. Esto viene a confirmar la aptitud del modelo matemático asumido.

Además de la función de probabilidad, debemos mencionar otra función que también caracteriza la distribución de probabilidad de una variable aleatoria X. De hecho, se denomina `función de distribución`, y se define como aquélla que asigna a cada valor x la probabilidad

$$F(x) = P(X \leq x)$$

Recibe este nombre porque, a pesar de resultar menos intuitiva que la función de probabilidad y al contrario de ésta, puede definirse para cualquier tipo de variable, ya sea discreta y continua, por lo que es ideal para caracterizar cualquier distribución. La relación entre esta función y el diagrama de barras para frecuencias relativas acumuladas es idéntica a la que se da entre la función de probabilidad y el diagrama de frecuencias relativas.

3.2.2. Parámetros probabilísticos. Ley de Grandes Números

Así pues, hemos establecido ya la conexión clave entre la Estadística Descriptiva y la Probabilidad: la repetición n veces de un fenómeno aleatorio da lugar a una muestra aleatoria de n datos cuyas frecuencias relativas se van identificando progresivamente con las correspondientes probabilidades. Estamos entonces en condiciones de redefinir todos los valores típicos estudiados en Estadística Descriptiva en términos probabilísticos, en particular la media aritmética y varianza. Recordemos que definíamos en (1.1) la media aritmética de una muestra con valores $\mathtt{x}_1, \ldots, \mathtt{x}_k$ mediante

$$\overline{x} = \sum_i \mathtt{x}_i \hat{p}_i.$$

De esta forma, se define la `Esperanza` o media de una variable aleatoria X con valores posibles $\mathtt{x}_1, \ldots, \mathtt{x}_k$ mediante

$$\mathtt{E}[X] = \sum_i \mathtt{x}_i p_i$$

Se trata pues del centro de gravedad que se obtiene ponderando los datos por su probabilidades. Este parámetro suele denotarse por la letra griega μ. En el caso del lanzamiento de dos dados es claro que $\mu = 7$. También es claro que, a medida que el número de repeticiones del experimento aumenta, el valor de la media aritmética $\overline{x}$ se aproxima al de μ. Por ejemplo, en la muestra que se obtiene tras lanzar 100 veces el para de dados se obtiene como media aritmética de la suma de puntuaciones $\overline{x} = 6{,}670$. Sin embargo, tras 400 repeticiones se obtiene $\overline{x} = 7{,}008$. Ello es signo del cumplimiento de la Ley del Azar. En nuestro modelo matemático esta convergencia debe verificarse necesariamente en virtud de la denominada Ley de los Grandes Números. Lo expresamos así:

$$\text{(ii)} \qquad \boxed{\lim_{n\to\infty} \overline{x} = \mu}$$

De manera análoga podemos redefinir la varianza en términos de probabilidad. Recordemos que definíamos en (1.2) la varianza muestral mediante

$$s^2 = \sum_i (\mathtt{x}_i - \overline{x})^2 \cdot \hat{p}_i.$$

Así pues, se define la varianza probabilística $\mathtt{var}[X]$, que también se denota por la letra griega σ^2, mediante

$$\mathtt{var}[X] = \sum_i (\mathtt{x}_i - \mu)^2 \cdot p_i$$

Su raíz cuadrada es la desviación típica probabilística y se denota por σ. Se verifica pues que, en las condiciones mencionadas, $\lim_{n\to\infty} s = \sigma$. Lo mismo puede decirse en definitiva del resto de parámetros estudiados, cosa que omitimos.

3.2.3. Ejemplo: distribución binominal

Supongamos que el color de ojos (distinguimos únicamente entre claros y oscuros) depende únicamente de una gen cuyo alelo dominante `A` determina un color oscuro y cuyo alelo recesivo `a` determina un color claro. Consideremos una pareja donde ambos individuos son hererocigóticos `Aa`. El color de los ojos de un descendiente depende de qué alelos se combinen, con lo que el número de posibilidades es 4:

$$\Omega = \{\mathtt{AA},\mathtt{Aa},\mathtt{aA},\mathtt{aa}\}$$

Si asumimos la simetría en lo que respecta a este gen tanto en el proceso de meiosis como en el de la fecundación, podemos suponer que todas la posibilidades son equiprobables y que el color de ojos de un descendiente no condiciona el que pueda tener otro. La probabilidad de que un descendiente tenga ojos claros es pues 1/4. Supongamos que la pareja tiene 5 descendientes, lo cual nos conduce al complicado espacio muestral Ω^5. No obstante, sólo estamos interesados en este caso en conocer cuántos descendientes poseerán ojos claros.

Así pues, la variable aleatoria considerada X es el número de descendientes con ojos claros, que puede tomarlos valores 0,1,2,3,4 ó 5. Basta conocer la función de probabilidad para caracterizar la distribución de esta variable. Nos preguntamos, por ejemplo, cuál es la probabilidad de tener exactamente 2 descendientes con los ojos claros. Ese suceso puede verificarse de muchas formas, por ejemplo si tenemos la secuencia o muestra aleatoria `CCOOO`. La probabilidad de dicha combinación puede calcularse de dos formas: dividiendo el número de casos favorables en el espacio muestral Ω^5 por el número total de casos posibles, en este caso 27/1024; más fácil es calcularla multiplicando las probabilidades de ir obteniendo cada suceso de la secuencia, es decir,

$$\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{1024}$$

Pero hemos de tener en cuanta que no es ésta la única combinación que aporta dos descendientes con ojos claros, pues cualquier alteración del orden da lugar a otra igualmente válida e igualmente probable, por la conmutatividad del producto: `COCOO, OCOOC`, etc. La pregunta es: ¿cuántas combinaciones con `2C` entre 5 posibilidades pueden darse? La respuesta es clara si tenemos nociones básicas de combinatoria: $\binom{5}{2}$, es decir,

$$P(X = 2) = \binom{5}{2} \cdot \frac{27}{1024} = 0{,}26$$

Si este modelo matemático explica el fenómeno considerado, ¿qué debería suceder en la práctica? Pues que, dada una gran cantidad de parejas en esas condiciones y con 5 descendientes, aproximadamente el 26 % de las mismas debe tener dos descendientes con ojos claros. Generalizando los cálculos podemos decir que

$$P(X=j)=\binom{5}{j}\cdot\left(\frac{1}{4}\right)^{j}\cdot\left(\frac{3}{4}\right)^{5-j},\qquad j=0,1,2,3,4,5$$

Hemos construido pues la distribución de probabilidad. Podemos generalizar aún más los cálculos de la siguiente forma: si una variable X contabiliza el número de veces que se verifica cierto suceso, que ocurre con una probabilidad p, tras n repeticiones independientes del experimento, la probabilidad de que X tome un valor $j=0,1,\ldots,n$ es la siguiente:

$$P(X=j)=\binom{n}{j}\cdot p^{j}\cdot(1-p)^{n-j}$$

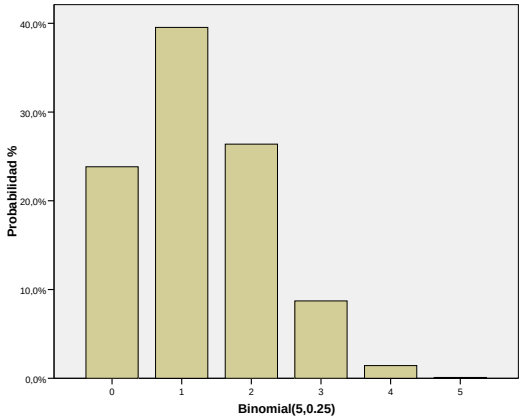
En ese caso, se dice que la variable X sigue un tipo o modelo de distribución Binomial de parámetros n y p, denotándose

$$X\sim B(n,p)$$

Si es así, tanto la media como la desviación típica pueden obtenerse directamente y sin demasiada dificultad (al menos en el caso de la media) conocidos n y p. Concretamente

$$\mu=np,\qquad \sigma=\sqrt{np(1-p)}$$

Así pues, en el ejemplo anterior puede decirse que el número de hijos de ojos claros sigue un modelo de distribución $B(5,1/4)$. Aplicando las fórmulas anteriores obtenemos $\mu=1{,}25$ y $\sigma=0{,}56$. Representamos a continuación su función de probabilidad



3.2.4. Distribuciones continuas. Distribución Normal

Lo dicho hasta ahora no es válido para cualquier tipo de variable aleatoria sino sólo para aquellas que pueden tomar una cantidad finita o al menos enumerable (ordenable) de valores, dando lugar a lo que denominamos distribuciones discretas. Por contra, las variables que pueden tomar cualquier valor en un intervalo (nótese que éstos no pueden enumerarse) darán lugar a las distribuciones continuas.

Como ejemplo podemos considerar un disco que presenta una marca en un punto de su perímetro y que gira en un viejo tocadiscos. Nos preguntamos en qué ángulo exacto de la circunferencia (medido en radianes) quedará la marca cuando el disco se detenga. La medida de dicho ángulo es

una variable aleatoria X con valores en el intervalo $[0, 2\pi)$. Podemos calcular diversas probabilidades por simetría: por ejemplo, la probabilidad de que la marca quede en el primer cuadrante es

$$\frac{\pi/2}{2\pi} = \frac{1}{4}$$

Sin embargo, podemos razonar fácilmente que la probabilidad de que X tome un valor exacto dentro del intervalo considerado es tan pequeña como se quiera, es decir, nula, lo cual podría resultar paradójico si se piensa que la marca debe detenerse en algún sitio concreto. Lo cierto es que nosotros no apreciamos un punto de parada sino un intervalo, que será más pequeño cuanto mayor sea nuestra precisión, y ese intervalo puede tener una probabilidad muy escasa pero nunca será nula. Esta paradoja es consecuencia de una pequeña pero insalvable discordancia entre la realidad percibida por los sentidos y el modelo matemático que la idealiza.

A la hora de formalizar este tipo de situaciones nos encontramos pues con los problemas inherentes a las mediciones sobre un continuo, por lo que se precisa una cierta familiaridad con las técnicas de integración y el lenguaje infinitesimal. Aquí no tiene sentido hablar de la función de probabilidad y los parámetros de la distribución no pueden definirse como vimos anteriormente. No obstante, sí que podemos establecer relaciones entre un incremento de la variable Δx y el correspondiente incremento de la probabilidad ΔP. En este caso concreto y por tratarse de una simetría pura, la relación entre ambos, $\Delta P/\Delta x$, es constante y vale $1/2\pi$. Sea o no constante, lo que nos interesa para medir probabilidades es el límite del cociente incremental en cada punto x del intervalo. La función definida en el lenguaje infinitesimal mediante

$$f(x) = \frac{dP}{dx}$$

se denomina función de densidad. En ese caso, tenemos que

$$dP = f(x)\, dx$$

Así pues, la probabilidad de que X pertenezca a un intervalo $[\mathtt{x}_1, \mathtt{x}_2]$ se obtiene integrando

$$P(\mathtt{x}_1 \leq X \leq \mathtt{x}_2) = \int_{\mathtt{x}_1}^{\mathtt{x}_2} dP = \int_{\mathtt{x}_1}^{\mathtt{x}_2} f(x)\, dx$$

La función de densidad caracteriza la distribución de la variable pues nos permite obtener la probabilidad de cualquier intervalo mediante el cálculo del área subyacente a la misma entre sus límites. Como dijimos en el caso discreto, se puede definir también la función de distribución mediante

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx$$

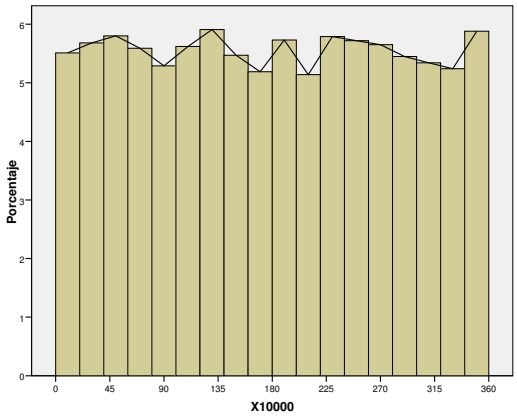
Caracteriza igualmente el modelo de probabilidad. Por otra parte, la densidad permite calcular los diferentes parámetros probabilísticos. Por ejemplo, si la media se definía mediante $\mu = \sum_i \mathtt{x}_i p_i$ en el caso discreto, en el caso continuo se define mediante

$$\mu = \int X\, dP = \int x f(x)\, dx$$

De manera análoga puede definirse la varianza.

En el ejemplo del disco, la función de densidad será la función constante $1/2\pi$ definida entre 0 y 2π y la media de la distribución es $\pi = 180$ grados. También es la mediana. Cuando la función de densidad (o la de probabilidad en el caso discreto) es constante se dice entonces que la distribución es uniforme.

Si la función de probabilidad de una distribución discreta guardaba una estrecha relación con el diagrama de barras de frecuencias relativas, la de densidad se vincula claramente al histograma. Efectivamente, simulemos mediante un programa estadístico el fenómeno anterior (midiendo el ángulo en grados) en 10.000 ocasiones y representemos mediante un histograma de frecuencias relativas los 10.000 valores entre 0 y 360 grados obtenidos.



Podemos observar que los diferentes intervalos considerados aportan rectángulos de áreas muy similares. Recordemos que las áreas de éstos son proporcionales a las frecuencias relativas de cada intervalo y éstas, por la Ley del Azar, deben ser parecidas a las que determina la distribución uniforme. Este efecto es más acusado cuantas más veces se repite el experimento. La media aritmética de las 10.000 mediciones es, por cierto, $\overline{x} = 179{,}88$ grados, muy cercana a $\mu = 180$.

El modelo de distribución continua más importante, por razones que veremos a continuación, es sin duda la distribución normal. Se dice que una variable aleatoria X sigue un modelo de distribución normal de parámetros μ y σ cuando su función de densidad es la denominada curva normal:

$$f(x) = (\sigma\sqrt{2\pi})^{-1} \cdot \mathrm{e}^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

Se denota $X \sim N(\mu, \sigma)$. En ese caso, puede demostrarse que μ y σ son, respectivamente, su media y desviación típica, de ahí la notación utilizada. Las probabilidades de los distintos intervalos se obtienen calculando las áreas subyacentes a la curva. De esta forma, puede comprobarse, por ejemplo, que la probabilidad de que la variable esté en el intervalo $(\mu - \sigma, \mu + \sigma)$ es 0.68, y en el intervalo $(\mu - 2\sigma, \mu + 2\sigma)$ es 0.95. ¿Nos suena esto de algo?

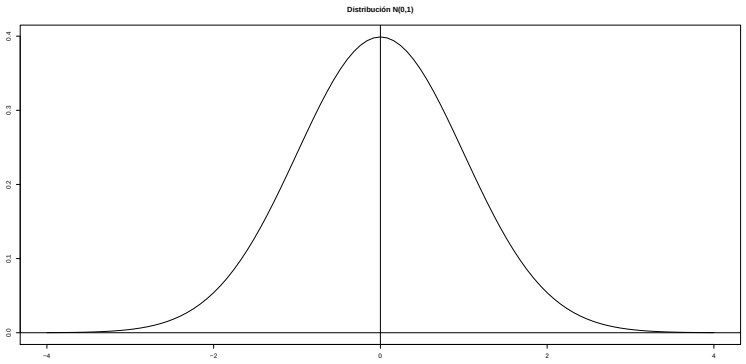
Desde el punto de vista gráfico, la media μ nos da el eje de simetría y la desviación típica indica, por supuesto, el grado de condensación. El área total subyacente a la curva es 1, como corresponde a una función de densidad. Se verifica en general, por las propiedades de la media y la desviación típica, que si X es una variable aleatoria de media μ y desviación típica σ, la variable

$$Z = \frac{X - \mu}{\sigma}$$

posee media 0 y desviación típica 1. Este proceso al que con frecuencia se someten las variables, sigan o no un modelo de distribución normal, se denomina tipificación o estandarización. Puede demostrarse que, si X sigue un modelo de distribución normal, también lo sigue cualquier trasformación afín de la misma y en particular su tipificación Z. Por lo tanto,

$$Z \sim N(0,1)$$

Este último modelo de distribución se denomina normal estándar.



La tipificación nos permite, entre otras cosas, calcular probabilidades correspondientes a cualquier normal a partir de la distribución normal estándar.

3.2.5. Distribuciones muestrales

En la primera sección del capítulo distinguimos entre el modelo de probabilidad asociado a una única ejecución del experimento aleatorio y el asociado a n ejecuciones del mismo. El segundo se denomina producto y está compuesto por el espacio de las muestras aleatorias de tamaño n y la probabilidad producto que rige el grado de verosimilitud de las mismas.

Se denomina variable aleatoria muestral a cualquier variable sobre el espacio producto de las muestras aleatorias. De la distribución de dicha variable decimos que es una distribución muestral. Se trata pues de un caso particular del concepto de distribución estudiado anteriormente. Concretamente, es una variable muestral la suma de las puntuaciones obtenidas por los dos dados, pues cada lanzamiento de dos dados puede considerarse una muestra aleatoria de tamaño $n = 2$ del fenómeno aleatorio consistente en el lanzamiento de uno. Su distribución, estudiada ya con detalle, es por lo tanto una distribución muestral. Si dividimos por 2 la suma de las puntuaciones estaremos hablando de la media aritmética de las mismas.

Por lo tanto y en general, la media aritmética valorada en el espacio de las muestras aleatorias de tamaño n es una variable muestral. Queremos decir con esto que no la entendemos como un simple número sino que puede variar de una muestra aleatoria a otra. Su distribución muestral determina entonces qué valores puede tomar y con qué probabilidades. Lo mismo puede decirse de la varianza y de todos los valores típicos estudiados en los capítulos 1 y 2. Así pues, desde este punto de visto más amplio, los parámetros descriptivos pueden entenderse como variables muestrales con sus correspondientes distribuciones, en cuyo caso se denotarán mediante las letras mayúsculas $\overline{X}$, S^2, $\tilde{X}$, etc. Una vez obtenida una muestra aleatoria concreta, la variable muestral

aportará el número correspondiente que se denota en minúscula: su media $\overline{x}$, su desviación típica s, su coeficiente de asimetría, su coeficiente de correlación si se trata de dos variables, etc.

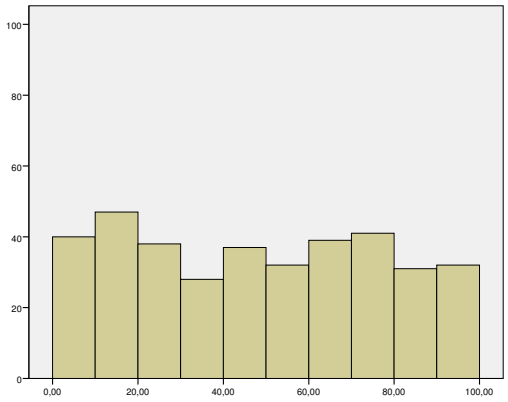
Dada una variable X de media μ y varianza σ^2, puede demostrarse que la esperanza y varianza de la media muestral definida sobre las muestral aleatorias de tamaño n son las siguientes:

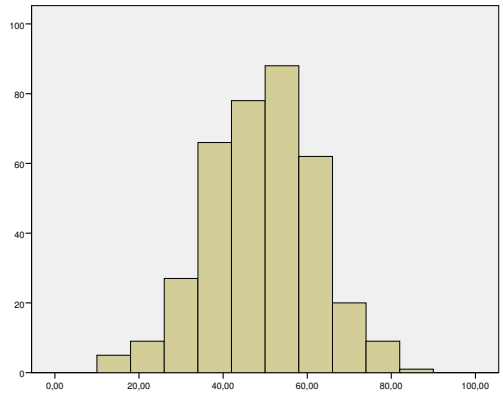
$$\mathtt{E}[\overline{X}] = \mu, \qquad \mathtt{var}[\overline{X}] = \frac{\sigma^2}{n}$$

Es decir, el valor medio esperado para $\overline{X}$ es la propia media probabilística de X pero su varianza es inversamente proporcional al tamaño muestral considerado. Por lo tanto, dado que la varianza expresa el grado de dispersión de los valores respecto a su media, se verifica que, si n es suficientemente grande, la probabilidad de que la media aritmética de una muestra aleatoria de tamaño n se aleje de la media probabilística será muy pequeña. Esto se parece desde luego a la condición

$$\text{(ii)} \qquad \boxed{\lim_{n\to\infty} \overline{x} = \mu}$$

Veamos otro ejemplo de distribución muestral: se estudia la media aritmética de cinco números entre 1 y 99 extraídos mediante un sorteo de lotería con reemplazamiento. Estamos pues hablando del modelo producto que corresponde a $n = 5$ repeticiones del fenómeno aleatorio consistente en extraer una bola entre 99 posibles. El valor de la bola sigue una distribución discreta uniforme con media $\mu = 50$ y desviación típica $\sigma = 28{,}6$. Así pues, la media muestral tendrá media 50 y desviación típica $\sigma/\sqrt{5} = 12{,}8$. Para ilustrar la diferencia entre ambas distribuciones vamos a imaginar que el sorteo se repite todos los día de un año, es decir, 365 veces. Vamos a anotar, por un lado, el resultado de la primera bola extraída cada día, que decimos sigue una distribución uniforme. Por otro lado, anotamos la media aritmética de las cinco bolas extraídas cada día. Los resultados simulados mediante un programa estadístico dan lugar a los siguientes histogramas:



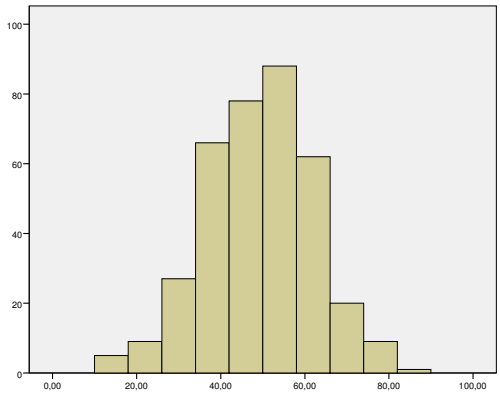


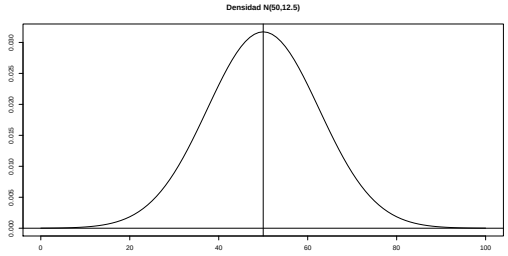
Comprobamos cómo efectivamente los datos correspondientes a la primera bola ofrecen una histograma relativamente plano, como corresponde a una distribución uniforme, cosa que realmente ocurre con las bolas restantes, pero no con la media aritmética de las cinco bolas, pues ésta se distribuye también de manera simétrica en torno a la media 50, pero más concentradamente. La explicación heurística puede ser la siguiente: en primer lugar, la distribución ha de ser simétrica respecto a 50, pues, por la simetría del fenómeno, nada nos hace pensar que los números mayores que 50 son más probables que los menores; en segundo lugar, el hecho de que se condensen más en torno a 50 se debe a que todas las posibles series o muestras aleatorias son equiprobables, pero

la mayoría aporta una media aritmética próxima a 50, pues para obtener una media aritmética extrema es necesario que todas las bolas lo sean. No queremos decir que la serie o muestra aleatoria $(1, 1, 1, 1, 1)$ sea menos probable que la serie $(49, 51, 47, 62, 36)$. Lo que sucede es que, por pura combinatoria, son más numerosas las series cuya media aritmética se aproxima a 50. Si se permite la expresión, los caminos al centro son más variados que los caminos a los extremos.

3.2.6. Teorema Central del Límite

En en el histograma de la derecha del ejemplo anterior se perfila una curva que a estas alturas debe ser ya familiar:





Se trata en efecto de la denominada curva normal, concretamente hablamos de la curva

$$N(50, 12.8)$$

que es la que corresponde según al media y varianza de $\overline{X}$. Es decir, la distribución de la media aritmética de las 5 bolas se aproxima a una distribución continua normal de media 50 y desviación típica 12,8. Esta aproximación a la distribución normal es más precisa cuanto mayor sea el tamaño de la muestras aleatorias. Realmente, puede demostrarse, y en eso consiste en Teorema Central del Límite, que esto sucede con carácter general. Es decir, que para muestras aleatorias suficientemente grandes, la media muestral de una variable X con media μ y varianza σ^2 sigue aproximadamente un modelo de distribución $N(\mu, \sigma/\sqrt{n})$. Tipificando obtenemos pues

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \overset{\text{aprox}}{\sim} N(0, 1)$$

Si el tamaño de muestra es grande, dado que $\lim_{n\to\infty} s = \sigma$, podemos sustituir en la anterior expresión la desviación típica probabilística σ por la muestral S, es decir,

$$\boxed{\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{\text{aprox}}{\sim} N(0, 1)} \tag{3.1}$$

Puede demostrarse también que si la distribución original de la variable es normal, la distribución muestral de la media aritmética será también exactamente normal. Por lo tanto, al tipificar obtendremos una $N(0, 1)$ exacta. Si sustituimos entonces σ por la desviación típica muestral S obtendremos una distribución muy similar que comentaremos a continuación: la distribución t-Student.

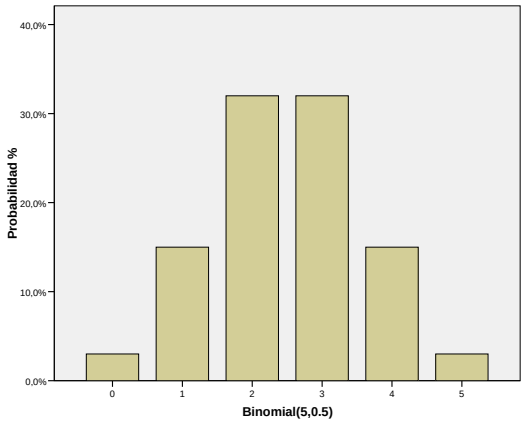
El resultado anterior, importantísimo, otorgará a la distribución normal un papel central en la Inferencia Estadística. Además, podría explicar por qué se observa con cierta frecuencia en la naturaleza. Cabe conjeturar, como apuntábamos en el primer capítulo, que cuando una variable aleatoria sigue una distribución aproximadamente normal se trata internamente del resultado de sumar una serie de variables o factores independientes. El caso es que esta distribución fue ya caracterizada por Gauss y Laplace al estudiar una variable que puede considerarse aleatoria: el error en la medición de parámetros astronómicos. De ahí que reciba comúnmente el nombre de campana de Gauss.

El tamaño de muestra n requerido para que la distribución de la media muestral se aproxime satisfactoriamente al modelo normal depende de la distribución original de la variable y, especialmente, de su sesgo. De hecho, cuando la distribución es simétrica, como en el caso del ejemplo, se consigue la aproximación aún con muestras pequeñas. Sin embargo, cuanto mayor es la asimetría más costoso es conseguir que la media muestral se ajuste a un modelo normal. No existe pues una cota universal para el valor de n, aunque con frecuencia se conviene que con $n = 30$ no debemos tener problemas. Otros estadísticos más conservadores exigen muestras aleatorias de al menos 60 datos para tener garantías. Lo más razonable es observar previamente el histograma de la muestra y el coeficiente de asimetría g_1.

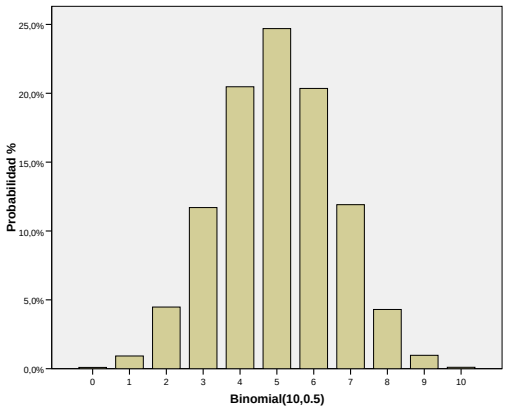
Sin ir más lejos, si una variable X sigue una distribución binomial $X \sim B(n, p)$, es decir, si recoge la suma de resultados favorables a un suceso con probabilidad p tras n ejecuciones del experimento aleatorio, la variable $\frac{1}{n}X$ recoge la media aritmética de una muestra aleatoria de tamaño n para la variable W que asigna un 1 si el resultado es favorable y un 0 si no lo es. En consecuencia, si n es suficientemente grande $\frac{1}{n}X$ seguirá aproximandamente una modelo de distribución normal y, por lo tanto, también será normal X. Dado que su media es np y su varianza $np(1-p)$, se verifica entonces

$$\boxed{B(n,p) \overset{\text{aprox}}{\approx} N(np, \sqrt{np(1-p)})}$$

El tamaño n requerido para que esta aproximación sea satisfactoria depende, según hemos dicho de la simetría de W y, en definitiva, de p. De hecho, para p próximo a $1/2$ se obtiene una distribución de W muy simétrica y, por lo tanto, una rápida convergencia. Tal es el caso de una distribución $B(5, 1/2)$, que se parece a la distribución $N(0.25, 1.11)$.

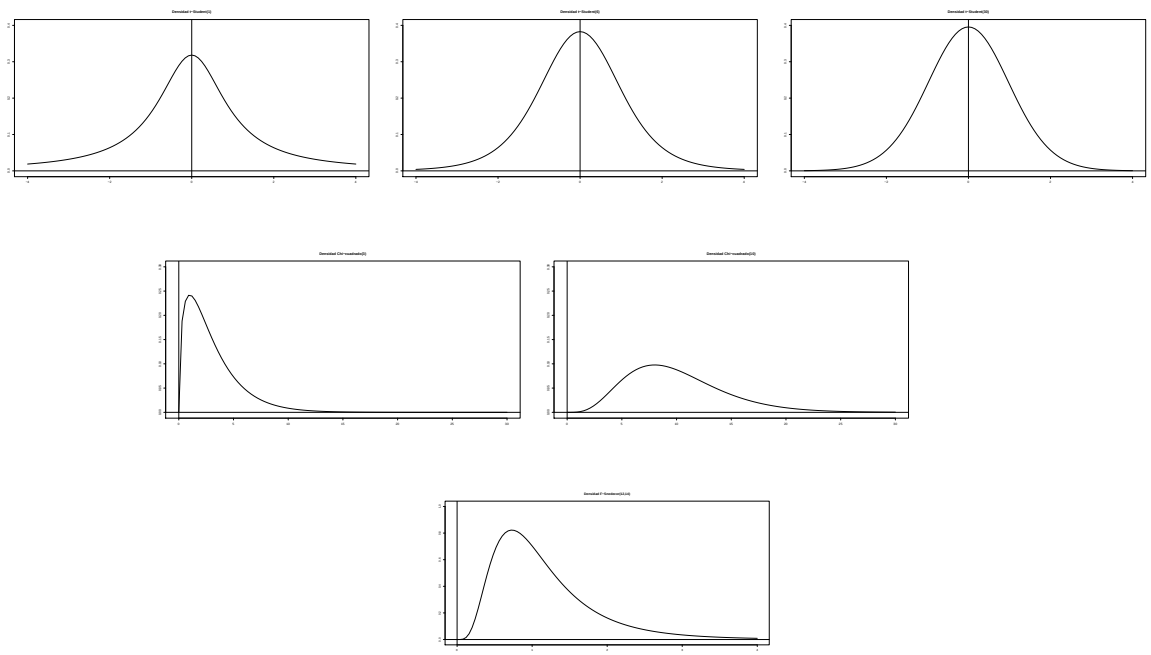


Veamos qué sucede con $B(10, 1/2)$, que debe parecerse a $N(5, 1.58)$.



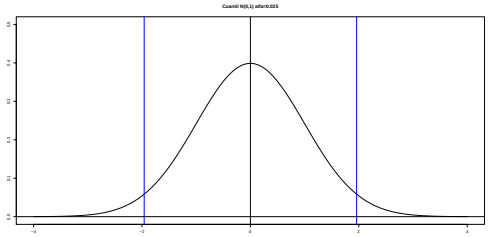
Si queremos aproximar probabilidades de la distribución discreta $B(10, 1/2)$ mediante la distribución continua $N(5, 1.58)$ parece razonable identificar cada valor entero de 0 a 10 con el intervalo de longitud 1 centrado en dicho valor. Es decir, la probabilidad que corresponde, por el ejemplo, al valor 3 según el modelo $B(10, 1/2)$ debe parecerse a la que corresponde al intervalo (2.5, 3.5) según el modelo $N(5, 1.58)$.

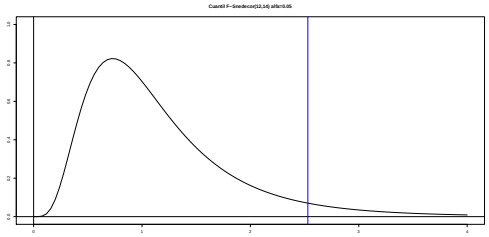
Podemos mencionar otras distribuciones continuas que se derivan de la distribución normal: primeramente, la distribución t-Student, que depende de un parámetro entero denominado grado de libertad y es parecida a la distribución $N(0, 1)$. De hecho, a partir de un grado de libertad 30 se consideran idénticas. Segundo, la distribución χ^2, que depende también de un grado de libertad. Por último, mencionamos la distribución F-Snedecor, que depende de dos grados de libertad. Para hacernos una idea de las distribuciones de las que hablamos mostramos a continuación las funciones de densidad de las t-Student con 1, 6 y 30 grados de libertad, de las χ^2 con 4 y 10 grados de libertad, y de la F-Snedecor con (12,14) grados de libertad:



Estas distribuciones aparecerán en la Inferencia Estadísticas como consecuencia de ciertas operaciones a las que se someterán los datos: así, la distribución t-Student, surge cuando se tipifica una variable normal pero sustituyendo su desviación típica probabilística por la muestral, obteniéndose por lo tanto una distribución similar a la $N(0,1)$; la distribución χ^2 cuadrado se obtiene a partir de la suma de cuadrados de normales. Recordemos que la suma de cuadrados es la forma habitual de medir errores en Estadística, de ahí que esta distribución esté precisamente asociada a la medición de diversos tipos de errores en sentido amplio. Concretamente, la varianza muestral sigue, salvo una constante, un modelo de distribución $\chi^2(n-1)$ cuando la distribución original de la variable es normal. Por último, la distribución F-Snedecor surge de la necesidad de dividir, es decir, comparar, errores o varianzas en sentido amplio, es decir, distribuciones χ^2. Sus grados de libertad se corresponden con los de las χ^2 del numerador y denominador.

Nos interesa especialmente conocer los cuantiles de estas distribuciones así como los de la $N(0,1)$. Nos referimos a un concepto estudiado ya en Estadística Descriptiva. El cuantil α de una distribución es el valor que deja una probabilidad α a su derecha y $1-\alpha$ a su izquierda. El caso más importante es el que corresponde a $\alpha = 0{,}005$. En el caso de las distribuciones simétricas, como $N(0,1)$ y t-Student, puede ser más interesante el caso 0,025 pues, entre dicho cuantil y su opuesto queda comprendida una probabilidad del 95 %. Mostramos a continuación los cuantiles $z_{0{,}025} = 1{,}96$ y $F_{0{,}005}(12,14) = 2{,}53$, correspondientes a las distribuciones $N(0{,}1)$ y F-Snedecor(12,14). Al término del manual se muestran unas tablas que podemos consultar para encontrar cuantiles de estos tipos d distribuciones.





3.3. Población, Inferencia y Probabilidad

Todo lo dicho hasta ahora en el presente capítulo resultará sin duda apasionante a un ludópata: dados, ruletas, loterías... No obstante, debemos preguntarnos qué utilidad práctica puede tener este estudio para el lector interesado en las Ciencias de la Salud, que es precisamente a quien va dirigido este breve manual. Así pues hemos llegado al punto clave del capítulo y posiblemente de la materia. Es el momento de entender cuál es la conexión entre el Cálculo de Probabilidades, dedicado al análisis de los fenómenos aleatorios y la Estadística entendida como tratamiento de la

Información relativa a poblaciones y variables. Describiremos brevemente cómo interviene en los dos problemas fundamentales de la Inferencia Estadística que abordaremos en el próximo capítulo: Estimación y Contraste de Hipótesis.

3.3.1. Probabilidad y Estimación

En primer lugar, en la introducción definimos Población en sentido amplio como el objeto de nuestro estudio. Aunque suponga una excesiva simplificación, hemos de reconocer que en el caso de las Ciencias de la Salud prevalece la acepción común del término como colectivo de individuos, ya sean personas en general, un colectivo de pacientes, animales o plantas de cierta especie, semillas o espermatozoides. El estudio consistirá concretamente en la descripción de una o varias variables. Por lo tanto, si tuviéramos acceso a las mediciones que aportan o aportaría la población Ω completa, es decir un censo, el estudio se restringiría a lo que hemos denominado Estadística Descriptiva y no se precisaría el concurso del Cálculo de Probabilidades.

Sin embargo y por desgracia, el conocimiento de los valores de toda la población Ω es poco menos que utópico, por lo que deben ser estimadas. En la práctica, aspiramos a estudiar los datos de una muestra de tamaño n extraída de dicha población, la cual se somete a las técnicas propias de la Estadística Descriptiva. La pregunta es ¿en qué medida podemos generalizar o inferir conclusiones relativas a la población Ω a partir de la descripción de una muestra de la misma? Pues resulta que, si la muestra es aleatoria, estamos en condiciones de hacerlo. ¿Qué quiere decir que la muestra sea aleatoria? Pues que los individuos que la componen hayan sido seleccionados mediante un fenómeno aleatorio equivalente a una lotería. Veremos cómo en esas condiciones la descripción de la muestra aporta conclusiones muy concretas respecto a la población total pero que vendrán expresadas lógicamente en términos probabilísticos.

Así, por ejemplo, si estamos estudiando la incidencia de una cualidad C, por ejemplo una enfermedad, que se da en cierta población Ω en una proporción p que queremos determinar, al escoger una muestra aleatoria de tamaño n, ¿cómo calcular la probabilidad de que cada individuo de la misma presente dicha cualidad? Teniendo en cuenta que todos los individuos son sucesos elementales equiprobables del sorteo, debe calcularse dividiendo el número de casos favorables por el número de casos posibles, es decir, el número de individuos de la población que presenta la cualidad C entre el número total e individuos de la población, y eso es precisamente la proporción p. Podríamos denotar $p = P(C)$. Es decir, identificamos proporción en la población con probabilidad en el sorteo de la muestra.

Siguiendo ese mismo razonamiento, si estudiamos una variable cuantitativa X, la media aritmética de la población, que se obtiene como suma de los valores que toma X en la misma ponderados por las frecuencias relativas o proporciones poblacionales (1.1), coincide con la media probabilística μ correspondiente a la medición de X sobre un individuo seleccionado por sorteo. Lo mismo podríamos decir de la varianza y de cualquier valor típico. Así pues, en este contexto identificamos los parámetros poblacionales con los probabilísticos.

El fenómeno aleatorio que realmente nos interesa no es el sorteo en sí sino la repetición n veces del mismo. De esta forma, la muestra aleatoria de tamaño n es un elemento del modelo aleatorio producto asociado. Los parámetros descriptivos de la muestra, como la media aritmética $\overline{X}$, la varianza S^2, las distintas, proporciones $\hat{P}_i$, etc., no son sino variables muestrales con sus correspondientes distribuciones. Recordemos que la muestra a estudiar es contingente, es decir, ha sido seleccionada de igual forma que podría haber sido seleccionada cualquier otra. De hecho

todas son equiprobables. De ahí que la media aritmética y demás parámetros descriptivos deban considerarse variables aleatorias con sus correspondientes distribuciones.

¿Y de qué sirve en definitiva que las muestras sean aleatorias? Al satisfacerse la Ley de Azar (ii), debe verificarse una aproximación de los parámetros muestrales a los correspondientes poblaciones o probabilísticos. Es decir,

$$\hat{p}_i \longrightarrow P_i$$

$$\overline{x} \longrightarrow \mu$$

$$s \longrightarrow \sigma$$

$$r^2 \longrightarrow \rho^2$$

Etcétera. Los parámetros muestrales serán pues estimaciones de los análogos poblacionales, y la aproximación a éstos será tanto mejor cuanto mayor sea el tamaño de la muestra. Pero no nos conformaremos con vagas expresiones al respecto. Por ejemplo, veremos en el próximo capítulo cómo el resultado (3.1) de la sección anterior puede servir para acotar de manera probable el error cometido en la aproximación de $\overline{x}$ a μ.

En definitiva, en el contexto de las Ciencias de la Salud debemos inclinarnos a interpretar el concepto de **probabilidad** no como una medida difusa y universal del grado de fe que tenemos en que algo ocurra, sino como una **proporción** respecto al total de la **población**.

3.3.2. Probabilidad y Contraste de Hipótesis

El problema de contraste de hipótesis consiste, como veremos con más detenimiento en el capítulo siguiente, en determinar la validez de un modelo teórico a la hora de explicar una cierta cantidad de observaciones experimentales. Haremos uso del lenguaje probabilístico para plantear y resolver el problema. Efectivamente, este tipo de problemas se afrontará identificando ese modelo teórico con un fenómeno aleatorio, es decir, con un modelo probabilístico ideal y explícito. Así, en el ejemplo 9 del capítulo siguiente, contrastaremos si en una determinada localidad actúan agentes ambientales capaces de influir en el sexo de la población. De no ser así, cabría pensar por simetría, dado que por cada espermatozoide portador del cromosoma `X` debe existir otro por tanto el cromosoma `Y`, que el sexo de una serie de n nuevos individuos puede considerarse el resultado de un fenómeno aleatorio equivalente al lanzamiento n veces de una moneda simétrica. Ésa será pues la hipótesis inicial a contrastar, de manera que debemos determinar si las observaciones obtenidas son o no probables según este modelo probabilístico concreto y, en función de eso, tomar una decisión razonable.

En otras ocasiones, como en el ejemplo 8 del capítulo siguiente, en el que se contrasta si la media poblacional de cierta variable presenta cierto valor concreto μ_0, no cabe pensar en otro fenómeno aleatorio que el propio muestreo (lotería), de manera que la hipótesis inicial se identifica con un valor medio μ_0 para el modelo probabilístico asociado. En definitiva, los fenómenos aleatorios que pueden interesarnos son la lotería, pues es teóricamente el procedimiento de selección de la muestra, y cualquier otro que pretenda identificar aproximadamente una hipótesis inicial a contrastar.

3.4. Cuestiones propuestas

1. Establecer un paralelismo entre todos los conceptos estudiados en Estadística Descriptiva y los estudiados en este capítulo.

2. Discutir si la medición de colesterolemia puede considerarse un fenómeno aleatorio.

3. Sospechamos que un dado está trucado. ¿Qué podemos hacer para contrastarlo?

4. Se cuenta con un bombo con 99.999 bolas iguales numeradas del 1 al 99.999. Se realiza el experimento consistente en hacer girar el bombo y extraer una bola cualquiera. Comenta las siguientes afirmaciones:

   - No sabemos nada acerca del resultado del experimento.
   - Si el resultado de una extracción es superior a 50.000, el resultado de la siguiente extracción (después del reemplazamiento) ha de ser inferior a 50.000.
   - A medida que vamos extrayendo bolas (con reemplazamiento), los valores de éstas se van aproximando a 50.000.

5. Calcula la probabilidad de obtener dos caras tras cinco lanzamientos de una moneda no trucada.

6. Si una variable aleatoria discreta X sigue una distribución de probabilidad del tipo $B(16, 0.5)$, calcular la probabilidad de que X sea mayor a igual que 14. Lo mismo si el modelo es $B(16, 0.8)$.

7. Se estudia una cierta variable bioquímica X (medida en gramos) sobre una determinada población. Se conoce que el valor de la media es $\mu = 1$ y el de la varianza es $\sigma^2 = 0$. ¿Cómo se interpretan estos datos?

8. Consideremos una variable $Z \sim N(0, 1)$. Calcula mediante las tablas:

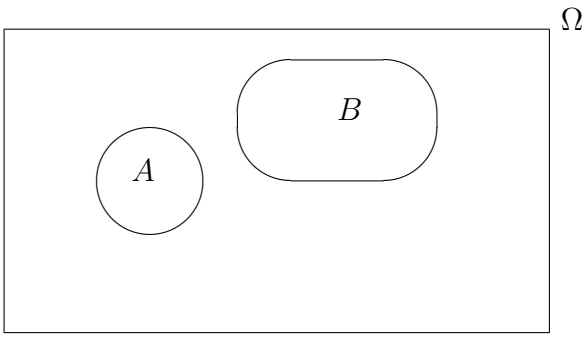
   - $P(Z < 0{,}5)$
   - $P(Z > 0{,}5)$
   - $P(Z < -0{,}5)$
   - $P(0{,}5 < Z < 1)$
   - $P(-1 < Z < 0{,}5)$

9. Se estudia determinado carácter cuantitativo sobre una población. La correspondiente variable X se distribuye aproximadamente según un modelo Normal, siendo su media 20 y desviación típica 5.

   - Calcula la proporción aproximada de individuos cuyo valor de la variable es inferior a 31.2.
   - Calcula la proporción aproximada de individuos cuyo valor de la variable está comprendido entre 30 y 20.

- Calcula la proporción aproximada de individuos cuyo valor de la variable es superior a 50.

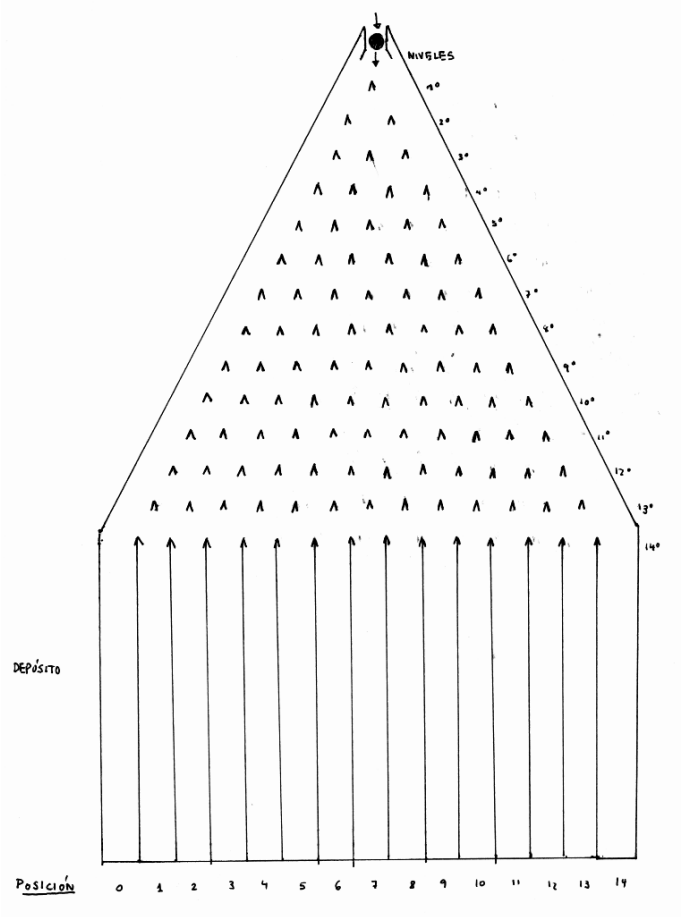
10. Se tiene dos variables $X \sim N(12, 4)$ e $Y \sim N(12, 2)$. Razonar (sin necesidad de cálculos) si son verdaderas o falsas cada una de las siguientes afirmaciones:
    - $P(X > 11) > P(Y > 11)$
    - $P(X \leq 12) = P(Y \geq 12)$
11. Si $Z \sim N(0, 1)$, calcula el número positivo z tal que $P(-z \leq X \leq z) = 0{,}95$. Entender que ese número es $z_{0{,}05/2}$.
12. Se tiene un procedimiento para medir glucosa en la sangre. Se comprueba que el método no es exacto, pues el contenido de glucosa medido en una determinada porción de sangre, según el procedimiento, difiere de su verdadero valor. Más aún, distintas mediciones de una misma porción de sangre aportaron distintos resultados, pero verificando la Ley del Azar. Se comprobó además que los distintos valores obtenidos se agrupan formando una Campana de Gauss, por lo que podemos considerar que la variable $X =$`error cometido tras medir glucosa en sangre` sigue un modelo de distribución Normal.
    - ¿Que hemos de hacer para averiguar lo posible acerca de la media y la desviación típica de dicha variable?
    - Supongamos conocidos los valores de μ y σ. Ordena por orden de conveniencia los siguientes casos:
        - $X \sim N(3, 1)$
        - $X \sim N(0, 4)$
        - $X \sim N(3, 4)$
        - $X \sim N(0, 1)$
13. Consideremos cierta variable bioquímica X que suponemos aleatoria. Se desea saber si se ajusta a un modelo de distribución Normal. Es conocido que el 50 % de los valores obtenidos experimentalmente es a 8, que el 20 % es superior a 10 y que el 2 % son inferiores a 6.
    - ¿Contradicen estos datos la normalidad de la variable?
    - ¿Puedes decir algo acerca del coeficiente de asimetría g_1?
14. Calcula la esperanza y la varianza de la variable aleatoria $X =$`resultado obtenido tras lanzar un dado no trucado`.
15. Calcula la probabilidad de obtener más de 6 caras tras 10 lanzamientos de una moneda no trucada. Calcula la probabilidad aproximada de obtener más de 60 caras tras 100 lanzamientos de una moneda no trucada.
16. Podemos considerar el espacio aleatorio Ω como una sección del plano, y cada suceso, por lo tanto, como un subconjunto de dicha sección. La probabilidad de cada suceso puede interpretarse entonces como la proporción de área que éste ocupa. De esta forma, el diagrama presenta dos sucesos disjuntos. La probabilidad (área) de la unión sera por tanto la suma de las probabilidades (áreas) de cada uno:



¿Cómo se expresarían gráficamente dos sucesos independientes? Recueda que A y B son independientes cuando $P(A \cap B) = P(A) \times P(B)$.

17. Describe la función de probabilidad de la distribución $B(6, 0.8)$.

18. Considera el ingenio que representa la figura:



¿Cuál es la probabilidad de que una bola introducida en la abertura superior termine en la posición 7 del depósito? Si se introducen un total de 200 bolas, que figura se formará en el depósito, una vez hayan caído todas?

19. En numerosas ocasiones hemos afirmado que, si una variable X sigue una distribución normal de media μ y desviación típica σ, la probabilidad de que una valor se encuentre entre $\mu - \sigma$ y $\mu + \sigma$ es, aprox., del 68 %. ¿Se trata de un hecho experimental o existe alguna forma de probarlo?

20. Cuando se habla de probabilidad, debemos estar necesariamente refiriéndonos a un fenómeno aleatorio. Por ejemplo; podemos hablar de la probabilidad de obtener cara tras lanzar una moneda, la probabilidad de que la suma de las puntuaciones de dos dados sea 7, etc. Sin embargo, con frecuencia se utilizan expresiones como la siguiente: `la probabilidad de que un individuo varón mayor de 18 años mida más de` $1{,}74m$ `es del 50 %`. ¿A qué fenómeno aleatorio nos estamos refiriendo en este caso?

Capítulo 4

Introducción a la Inferencia Estadística

La Estadística, como su propio nombre parece indicar, se concibe en principio para el tratamiento de la información relativa a grandes poblaciones, entendidas éstas como colectivos de individuos. Si bien el término de población puede considerarse hoy en día mucho más amplio, la acepción clásica del mismo es la que prevalece en las Ciencias de la Salud. En todo caso, sucede en la mayoría de las ocasiones que dicha población, entiéndase como se entienda, es demasiado grande o compleja, inabarcable, por lo que su descripción exhaustiva es infactible.

¿Cómo podemos paliar esta incapacidad? Pues, según hemos visto en el capítulo anterior, seleccionando aleatoriamente n individuos de la población, los cuales constituirán una muestra aleatoria de ésta. Nos permitimos el abuso de denominar igualmente por muestra a los datos que aportan esos individuos. Dichos datos serán sometidos a las técnicas descriptivas consideradas en los capítulos 1 y 2 para, posteriormente y en virtud de los métodos que estudiaremos a partir de ahora, inferir o generalizar conclusiones relativas a la población total. Esta nueva fase del estudio se denomina Inferencia Estadística y exige, como hemos dicho, que los componentes de la muestra hayan sido escogidos aleatoriamente. Sólo en esas condiciones estamos capacitados para extrapolar los resultados, pero siempre en términos probabilísticos.

El proceso de selección aleatoria de los integrantes de la muestra se denomina muestreo aleatorio. Existen realmente diferentes tipos de muestreos aleatorios, pero nosotros consideraremos únicamente el muestreo aleatorio simple. En el caso de una población en el sentido clásico del término, el muestreo aleatorio simple es equivalente a un sorteo de lotería en el que cada individuo de la población posee la misma probabilidad de ser seleccionado. De ahí que en lo sucesivo identifiquemos la probabilidad de que suceda algo en la población con la proporción de individuos de la misma que verifican ese algo.

El presente capítulo está dedicado a una muy breve explicación de los elementos fundamentales de la Inferencia Estadística, distinguiendo los dos problemas que pretende resolver: el de Estimación y el de Contraste de Hipótesis. En el capítulo siguiente expondremos una clasificación de las técnicas más populares de la Inferencia Estadística, siempre desde la perspectiva de las Ciencias de la Salud.

4.1. Problema de Estimación

Hemos distinguido dos tipos de parámetros o valores típicos: los muestrales o descriptivos, como $\overline{x}$, s, r, etc, y los probabilísticos, como μ o σ. En el caso de que el fenómeno aleatorio considerado sea el sorteo de una muestra aleatoria simple, sabemos que los parámetros probabilísticos coinciden con los parámetros descriptivos de la población, es decir, que μ es la media aritmética de toda la población, σ^2 es la varianza de toda la población, etc. De ahí que los denominemos a partir de ahora parámetros poblacionales.

Estos parámetros se suponen normalmente desconocidos pues la población suele ser inabarcable. Sin embargo, sabemos que los parámetros de la muestra aleatoria convergen a sus análogos poblacionales a medida que el tamaño de la misma tiende a infinito. Esto es lo que da sentido al muestreo aleatorio. El problema de Estimación tiene por objeto estimar o aproximar los parámetros probabilísticos a partir de otros calculados directamente a partir de la muestra. De esa forma, podemos decir por ejemplo que la media aritmética $\overline{X}$ de la muestra es un estimador de la media poblacional μ.

4.1.1. Criterios de Estimación

El problema de estimación es más complejo de lo que puede parecer a simple vista pues debemos establecer primeramente criterios para determinar si un estimador es aceptablemente bueno o si es peor que otro. Así, por ejemplo, puede llamar la atención el hecho de que la varianza muestral se haya definido de la forma

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

cuando lo natural hubiera sido dividir directamente por n. Y esto es así porque S^2 así definida (dividiendo por $n-1$) es un estimador insesgado de σ^2, lo cual quiere decir que es exacto por término medio.

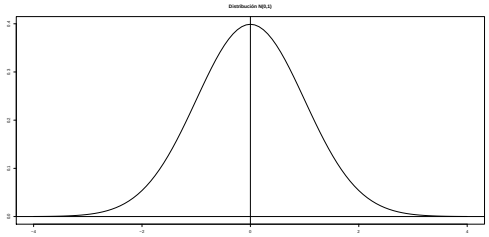
Los dos criterios más populares a la hora de justificar un estimador son el de Mínimos Cuadrados y el de Máxima Verosimilitud: el primero pretende minimizar el error cuadrático que se comete por término medio en la estimación, mientras que el segundo escoge el parámetro que hace la observación obtenida lo más verosímil posible. Precisamente, la varianza muestral dividiendo por n es el estimador de máxima verosimilitud de σ^2 cuando la variable considerada sigue un modelo de distribución normal. No obstante, en lo sucesivo no volveremos a hacer hincapié en estos aspectos.

4.1.2. Intervalos de confianza

Existen más parámetros por estimar, como veremos en el capítulo siguiente. Ahora nos centraremos en otro aspecto del problema de estimación. El valor concreto que aporta un estimador en la muestra obtenida se denomina estimación puntual. Así, dada una muestra aleatoria, $\overline{x}$ es una estimación puntual de μ. Por supuesto que dicha estimación está sometida a un error. No podemos esperar que coincida con el valor exacto del parámetro poblacional desconocido que estamos estimando. Sin embargo, nos gustaría precisar un probable margen máximo de error, de manera que podamos determinar un intervalo en el cual se encuentre seguramente el parámetro poblacional. ¿Cómo podemos construir ese intervalo? Veamos un ejemplo.

Se considera cierta variable cuantitativa X sobre una población Ω cuya media es μ. A través de una muestra aleatoria de tamaño n podemos estimar μ mediante su media aritmética $\overline{X}$ (recordamos que la diferencia entre $\overline{X}$ y $\overline{x}$ consiste en que la primera denota la variable muestral y la segunda el valor concreto de dicha variable para la muestra concreta que se estudia). Si el tamaño de muestra considerado es suficientemente grande (digamos $n > 30$), podemos aplicar el resultado (3.1) del capítulo anterior, de manera que

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim N(0,1)$$



En ese caso, se verifica entonces aproximadamente que

$$P\left(\frac{|\overline{X} - \mu|}{S/\sqrt{n}} \leq z_{\frac{0,05}{2}}\right) = 0{,}95$$

Es decir,

$$P\left(|\overline{X} - \mu| \leq z_{\frac{0,05}{2}} \frac{S}{\sqrt{n}}\right) = 0{,}95$$

Por lo tanto, la probabilidad de que, para una muestra aleatoria de tamaño n, la diferencia entre su media aritmética $\overline{x}$ y la media poblacional μ sea a lo sumo $z_{0,05/2}S/\sqrt{n}$ es del 95 %. Dicho de otra forma, en el 95 % de las posibles muestras de tamaño n que pueden extraerse de la población, la diferencia entre la media de la muestra y la de la población es a lo sumo $z_{0,05/2}S/\sqrt{n}$. Esa cantidad se denomina margen máximo de error al 95 % de confianza y se denota por $E_{\text{máx}}$.

De esta forma, el verdadero aunque desconocido valor de μ quedará dentro del intervalo $\overline{X} \pm z_{0,05/2}S/\sqrt{n}$ en el 95 % de las posibles muestras de tamaño n. Dada una muestra concreta de tamaño n, se dice entonces que

$$\overline{x} \pm z_{\frac{0,05}{2}} \frac{s}{\sqrt{n}}$$

es un intervalo de confianza al 95 % para la media μ. El valor de $z_{0,05/2}$ es, por cierto, 1.96. Cuando construimos un intervalo de confianza al 95 % estamos asumiendo una probabilidad de error o riesgo del 5 %. ¿Por qué el 5 %? Pues por nada en especial, pero existe un convenio tácito en la Estadística de considerar algo como raro o poco probable cuando la probabilidad de que ocurra sea inferior al 0.05, seguramente por ser una cantidad pequeña y redonda. De ahí que lo más habitual sea construir intervalos al 95 % de confianza. No obstante, podemos admitir otras opciones con niveles de riesgo diferentes. En general, si se denota por α la probabilidad de error (en el caso anterior tendríamos $\alpha = 0{,}05$) el intervalo de confianza a nivel $(1-\alpha) \times 100\,\%$ para la media será

$$\overline{x} \pm z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

Los valores alternativos más frecuentes para el nivel de riesgo son $\alpha = 0{,}01, 0{,}001$. También se asumen en ocasiones riesgos mayores, como $\alpha = 0{,}10$.

Podemos construir intervalos de confianza para otros parámetros poblacionales como la varianza, el coeficiente de determinación, la pendiente de regresión, etc. No obstante, en la mayoría de las ocasiones será necesario suponer ciertas condiciones en el modelos de distribución considerado.

Ejemplo 8: [Intervalo de confianza para una media]

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

Así pues ya tenemos una estimación puntual de la media μ: la media aritmética $\overline{x} = 162{,}3$. El margen máximo de error al 5 % 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 $162{,}3 \pm 1{,}6$. En definitiva, podemos afirmar con una confianza del 95 % que la media de altura de la población se encuentra entre 160.7cm y 163.9cm.

Observemos que, en general, no sólo el intervalo sino el propio margen máximo de error depende de la muestra obtenida. En el caso de que la varianza poblacional σ^2 fuese conocida, el margen máximo de error podría calcularse mediante

$$E_{\text{máx}} = z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

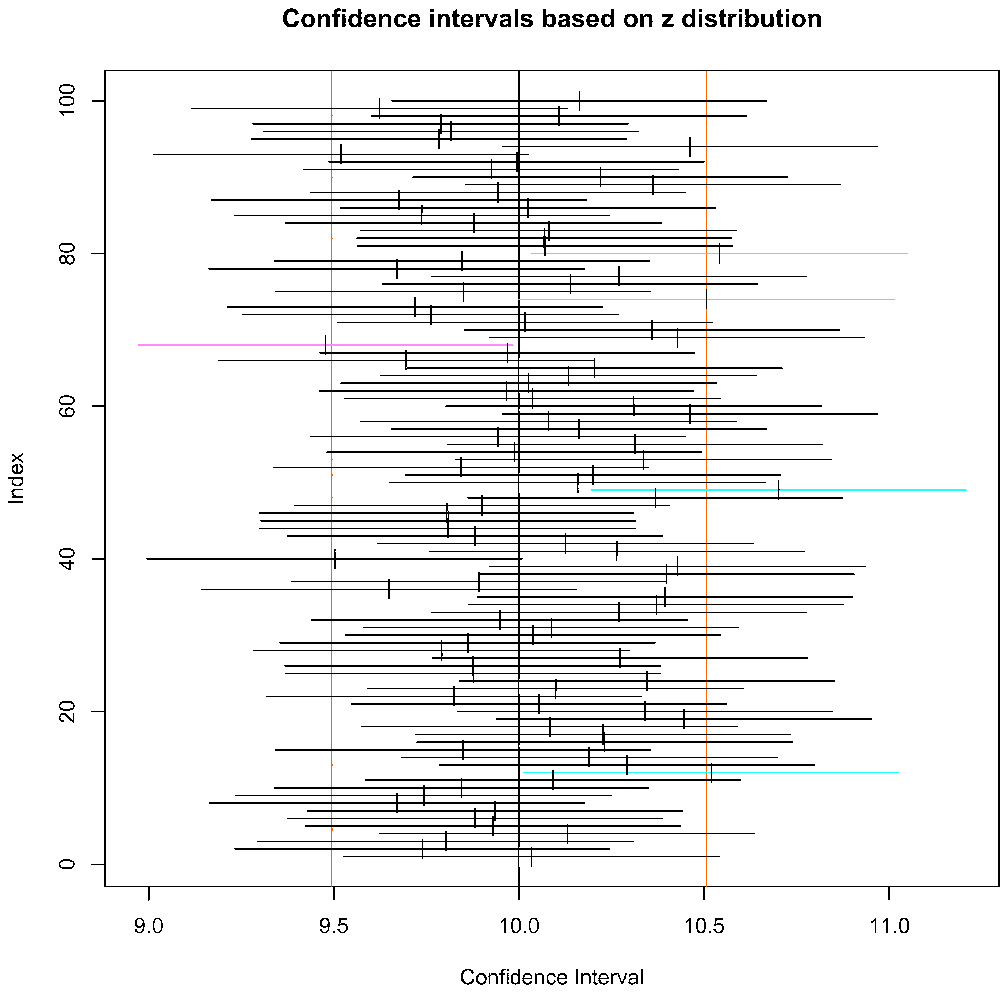
con lo que dependería de la muestra únicamente a través de su tamaño n. Este hecho será de utilidad en el siguiente capítulo para determinar el tamaño de muestra requerido en función del probable margen máximo de error que estemos dispuesto a asumir en la estimación del parámetro poblacional. También puede ser de utilidad para ilustrar qué entendemos exactamente por intervalo al 95 % de confianza.

Efectivamente, supongamos que la media y desviación típica de la población (parámetros que en la práctica son desconocidos) fueran $\mu = 10$ y $\sigma = 2$. Según la expresión anterior, el margen máximo de error en la estimación de μ con una confianza del 95 % es

$$E_{\text{máx}} = 1{,}96 \frac{2}{\sqrt{60}} = 0{,}51$$

Así pues, un intervalo de confianza al 95 % para μ a partir de una muestra de tamaño $n = 60$ será de la forma $\overline{x} \pm 0{,}51$. A continuación simulamos mediante un programa estadístico la extracción de 100 muestras aleatorias diferentes, cada una de las cuales aporta una media aritmética distinta, pues es una variable muestral. Según lo dicho, debemos esperar que aproximadamente 95 de ellas disten de la verdadera media $\mu = 10$ a lo sumo 0.51 unidades, de manera que sus correspondientes intervalos de confianza contendrán efectivamente a la media de la población. En este caso observamos que en cuatro ocasiones (las líneas horizontales coloreadas) las muestras seleccionadas han diferido de

$\mu = 10$ más de 0.51 unidades, de manera que los intervalos de confianza al 95 % construidos a partir de estas cuatro muestras inducirían a error. El 5 % residual debe pues entenderse como la proporción de muestras cuyos intervalos asociados no contendrán realmente a μ por el hecho de que son extremas. Que sean extremas quiere decir que están compuestas por valores demasiado altos o demasiado bajos, lo cual es poco probable. Pero si tenemos la mala suerte de que ocurra nos haremos una idea equivocada de la media μ de a población.



Si queremos que esta probabilidad de error sea menor, por ejemplo del 1 %, basta tomar $\alpha = 0{,}01$. Ello se traducirá en un mayor margen de error para la estimación y, en consecuencia, en intervalos de confianza más amplios. Es decir, que lo que ganamos en seguridad lo perdemos en precisión. Si queremos mayor seguridad sin disminuir la precisión o viceversa, nos vemos obligados a trabajar con muestras más amplias. De hecho, basta echar un vistazo a la ecuación (4.1) para entender que, a medida que el tamaño de muestra tiende a infinito, el margen máximo de error tiende a 0. Es una forma de decir que

$$\boxed{\lim_{n\to\infty} \overline{x} = \mu}$$

4.2. Problema de contraste de hipótesis

En esta ocasión no nos ocupa la estimación de un parámetro poblacional, sino evaluar la validez de un determinado modelo teórico para explicar el comportamiento de nuestros datos, denominado hipótesis inicial. La decisión ha de tomarse pues de manera razonable, a partir de la información que presta una muestra aleatoria de tamaño n. Denominamos test de hipótesis al algoritmo o proceso matemático preestablecido al que se someten los n datos de la muestra y que desemboca en la toma de decisión. No existe, evidentemente, un único test para cada problema de decisión. Desde un punto de vista teórico, el problema de decisión consiste en establecer criterios razonables de comparación de tests, y determinar qué test es el mejor, según el criterio establecido.

4.2.1. Planteamiento del problema.

Para entender las dificultades a las que nos enfrentamos y los elementos de un problema de contraste de hipótesis, consideraremos un problema sencillo.

Ejemplo 9:[Contraste bilateral para una probabilidad o proporción]

Se estudia si en una pequeña localidad existen factores de tipo ambiental capaces de modificar la distribución natural del sexo en los recién nacidos. Para ello se tienen en cuenta los 10 nacimientos acaecidos a lo largo del último año. Los resultados son los siguientes: `HHVHVHHHHH`

Partiremos de la **hipótesis inicial** de que en el pueblo no ocurre nada que pueda influir en el sexo del bebé. Esta hipótesis debe ser juzgada o contrastada mediante la muestra estudiada, es decir, los 10 nacimientos registrados. Analizaremos si estos datos suponen una prueba significativa contra la hipótesis inicial, que se denota por H_0 . Como alternativa a esta hipótesis puede ocurrir que ciertos agentes ambientales favorezcan el nacimiento de hembras o bien el nacimiento de varones. La primera regla del contraste de hipótesis puede formularse así:

(1) `La decisión que tomemos respecto a` H_0 `dependerá exclusivamente de un valor numérico calculado a partir de la muestra`

En consecuencia, toda la información que la muestra, en nuestro caso la secuencia `HHVHVHHHHH`, puede aportar en lo referente a este problema de decisión concreto, debe quedar resumido en un único número. En general, dado que la muestra es contingente y puede llegar a ser en principio de 2^{10} tipos diferentes (en este caso), deberíamos hablar más bien de una variable numérica $\mathcal{H}$ que pudiera tomar diferentes valores según la muestra observada pero que nos permitiera distinguir si la hipótesis inicial se está verificando o no. Esta variable se denominará estadístico de contraste, y el valor concreto que tome para la muestra observada se denominará **valor experimental**, denotándose por $\mathcal{H}_{exp}$. De la capacidad de ese valor experimental para resumir la información dependerá la **potencia** del test que diseñemos a partir del mismo, es decir, la capacidad para detectar la falsedad de la hipótesis inicial en el caso de que dé. Quiere decir esto que si la muestra es pequeña o no somos capaces de resumirla satisfactoriamente con un solo número será difícil probar la posible falsedad de H_0.

¿Cuál es el número más adecuado para resumir la información de nuestra muestra `HHVHVHHHHH`? Por ejemplo, podríamos proponer la media arítmética de las posiciones de los varones. En este caso, los nacimientos varones están registrados en las posiciones tercera y quinta, por lo que dicha media

será 4. ¿Sirve este número para discriminar entre el cumplimiento y la violación de la hipótesis inicial? No lo parece, más bien al contrario. Podemos intuir que el orden en que se producen los nacimientos no guarda relación con el problema planteado, es decir, que cualquier permutación de nuestra secuencia debería conducir a una misma decisión. De hecho, no parece razonable decidir una cosa para la secuencia `HHVHVHHHHH` y la contraria para `VHHHHHHHVH`. Si pretendemos que la decisión final se base en un único número, debemos estar dispuestos a desechar aquella información que no sea relevante para el problema de decisión planteado. En este caso, el orden en que aparecen los diferentes varones y hembras no parece serlo. Luego, si el orden no importa, parece claro que lo que realmente interesa de la muestra a la hora de decidir si hay alteraciones en el sexo de los recién nacidos es el número total de hembras (o, equivalentemente, el número de varones; o también la proporción de varones o hembras). Esta variable si parece apropiada para la toma de decisión porque nos permite evaluar si la hipótesis inicial se cumple o no. Si se cumpliera cabría esperar un valor experimental próximo a 5. Así pues, si definimos $\mathcal{H}$ como el número de varones entre los 10 nacimientos registrados tendremos para nuestra muestra un valor experimental $\mathcal{H}_{exp} = 8$.

Esta es en definitiva la información que aporta nuestra muestra. Suponiendo que el estadístico de contraste sea adecuado para discriminar entre el cumplimiento y la violación de la hipótesis inial (cosa que es cierta en este caso), parece pues claro que, dado que un valor céntrico (próximo a 5) es acorde con dicha hipótesis, lo contrario, es decir, un valor **extremo** (próximo a 0 ó 10), vendría a contradecirla. La siguiente cuestión es qué entendemos exactamente por extremo. De ello dependerá la decisión final que adoptemos a partir de la muestra. Teniendo en cuenta que el test es por definición un algoritmo automático, la frontera a partir de la cual el valor experimental o, equivalentemente, la muestra en sí, es considerada extrema debe quedar explicitado previamente a la observación de la misma. Sólo así podremos decidir si contradice la hipótesis inicial el hecho de obtener 8 hembras. La segunda regla del contraste de hipótesis la formularemos así:

(2) `A través del estadístico de contraste` $\mathcal{H}$`, la hipótesis inicial` H_0 `debe traducirse en una distribución de probabilidad concreta.`

Veamos de qué modelo probabilístico estaríamos hablando en este caso. Sabemos que el sexo del bebé depende de si el espermatozoide que fecunda el óvulo porta el cromosoma `X` o el `Y`. En principio, cabe pensar por simetría que la proporción de espermatozoides `X` es idéntica a la de espermatozoides `Y`. Supongamos además que ambos tipos de espermatozoides se comportaran igualmente en el proceso de fecundación[1]. Entonces, si ningún otro factor externo influye en la fecundación o el posterior desarrollo sexual del embrión y el feto, cabría equiparar la secuencia de $n = 10$ nacimientos con una serie de 10 lanzamientos independientes de una moneda simétrica, por lo que podríamos hablar de una probabilidad $p = 0{,}50$ de que el bebé sea hembra. Así podemos expresar la hipótesis inicial

$$H_0 : p = 0{,}50$$

En ese caso, según estudiamos en el capítulo anterior, que la hipótesis inicial sea cierta equivale a que $\mathcal{H}$ se distribuya según un modelo de probabilidad $B(10, 0.50)$, es decir

$$H_0 : \ \mathcal{H} \sim B(10, 0.50)$$

Nos preguntamos anteriormente si el hecho de obtener un valor experimental $\mathcal{H}_{exp} = 8$ podía considerarse extremo desde el punto de vista de la hipótesis inicial, lo cual podría traducirse

[1]Lo cual es mucho suponer. Ver la cuestión 14.

en una contradicción de la misma. Dado el modelo de probabilidad asociado a dicha hipótesis, podríamos formular la pregunta de manera equivalente así: ¿podríamos obtener 8 caras tras 10 lanzamientos con una moneda simétrica? Evidentemente, la respuesta es sí. Incluso, teóricamente existe la posibilidad de obtener 100 caras consecutivas con una moneda simétrica. Otra cosa es que ello es muy poco probable por lo que, si ocurriera, una persona razonable se inclinaría a pensar que la moneda tiene tendencia a resultar cara, porque ese modelo teórico explica mejor las observaciones. Llegamos así a la tercera regla, la fundamental, en el contraste de hipótesis y, posiblemente, el axioma principal de la Estadística:

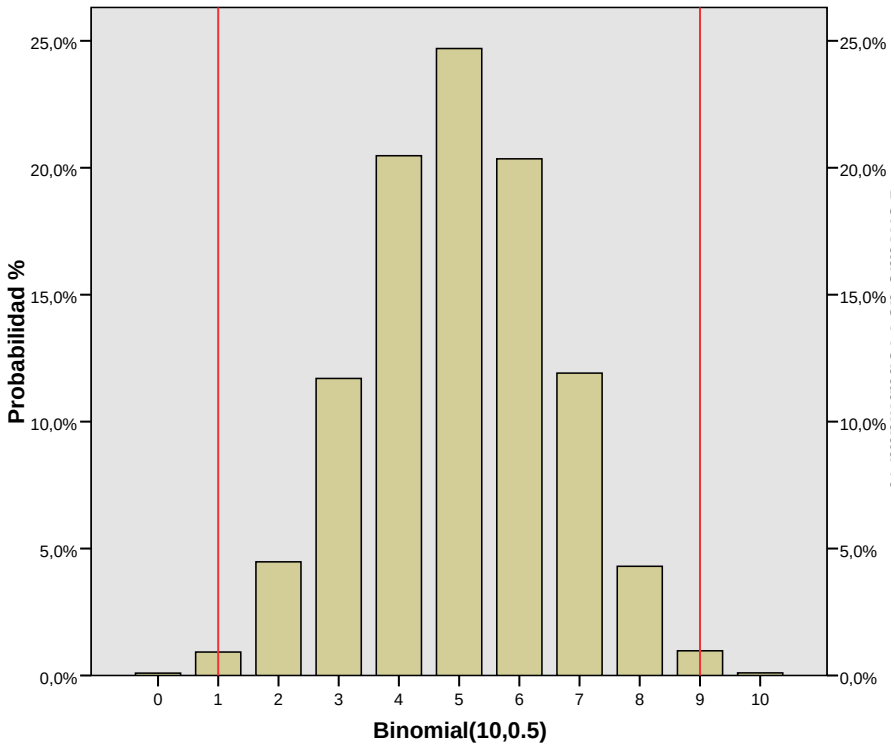
(3) Principio de Máxima Verosimilitud:

`Debemos optar por un modelo probabilístico que haga verosímil nuestra observación. Es decir, si la observación es rara según un modelo teórico deberíamos pensar en rechazarlo en favor de otro que la explique mejor.`

Así pues, siguiendo este axioma, diseñar un test de hipótesis significa determinar *a priori* cuáles de los posibles valores experimentales se considerarán `raros` según el modelo inicial (si el estadístico de contraste es adecuado estos valores quedarán bien explicados por un modelo teórico alternativo a dicha hipótesis), de manera que, si el valor experimental resulta ser finalmente uno de ellos se rechazará H_0. Este conjunto de valores se denomina **región crítica**, mientras que el resto es la región de aceptación de H_0.

Pues bien, ya hemos comentado con anterioridad que en Estadística se conviene en considerar un suceso raro cuando su probabilidad es inferior a 0,05. Esta afirmación deja mucho que desear. Basta pensar en una distribución continua donde cualquier valor concreto se verifica con probabilidad 0. Se trata más bien de determinar una región cuyos elementos sumen una probabilidad de a los sumo 0.05. Teniendo en cuenta lo dicho anteriormente, esa región debe ser extrema, es decir, alejada del valor central 5 y, además, simétrica respecto a 5 porque no tenemos ninguna razón para privilegiar alguno de los lados. Efectivamente, no parecería razonable, por ejemplo, que obtener 10 hembras condujera a rechazar la hipótesis inicial $p = 0{,}50$ pero que 10 varones (0 hembras) no lo hiciera, o que 7 hembras condujera a aceptarla pero 7 varones no. Por lo tanto, la región crítica será un conjunto extremo, raro (con probabilidad igual o inferior a 0.05) y, al menos en este caso, simétrico.

Para construirlo nos valdremos del conocimiento de la función de probabilidad de la distribución $B(10, 0.50)$ cuyo diagrama de barras es el siguiente



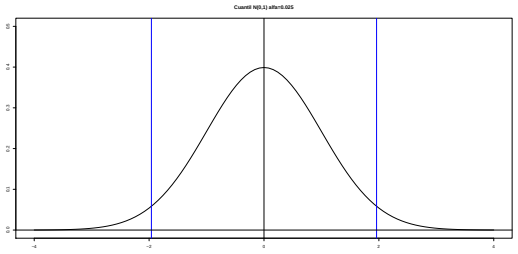
Según este modelo de probabilidad, los valores 0 y 10 pueden obtenerse con probabilidad 0.001. La suma de ambos valores es 0.002 (si 10 pertenece a la región crítica también debe pertenecer 0), es decir, la probabilidad de obtener un resultado tan extremo como 10 es de 0.002, lo cual sería entonces raro y nos llevaría a rechazar el modelo inicial. Por lo tanto, 0 y 10 deben formar parte de la región crítica. No obstante, dado que esta probabilidad es inferior a 0.05 puede que tengamos margen para ampliar la región crítica con otros valores no tan extremos. Si añadimos el par 1 y 9 (el 1 debe conducir a la misma decisión que el 9), dado que cada uno presenta probabilidad 0.01, obtendríamos una región que sumaría una probabilidad de 0.022. Ésa es la probabilidad de de obtener un resultado al menos tan extremo como 9. Es por lo tanto poco verosímil según la hipótesis inicial por lo que, si ocurriera realmente, nos llevaría a rechazar dicha hipótesis. Por lo tanto, 1 y 9 deben estar también incluidos en la región crítica.

¿Podemos ampliarla aún incluyendo los valores 8 y 2? Dado que la probabilidad de obtener 8 es de 0.044 (la de 2 también), obtendríamos una suma acumulada de 0.110. Por lo tanto, obtener un resultado al menos tan extremo como 8 presenta una probabilidad aceptable según el modelo o hipótesis inicial, por lo que la región constituida por 0,1,2,8,9 y 10 ya no podría considerarse rara para este modelo. Por lo tanto, 8 y 2 no pueden estar incluidos en la región crítica, que queda configurada finalmente por los valores 0,1,9 y 10. Es decir, la muestra contradirá significativamente la hipótesis inicial si el número de hembras es 0,1,9 o 10 (de ahí las línea rojas que se muestran en el gráfico) o, equivalentemente, si nacen 9 o 10 hembras o bien 9 o 10 varones. En tal caso diremos que el `resultado del test es significativo`, lo cual querrá decir que la observación supone una prueba clara contra H_0.

En el caso concreto del ejemplo, donde el número de hembras es $\mathcal{H}_{exp} = 8$, aplicando este test obtenemos un resultado no significativo, es decir, la observación no llega a ser los suficientemente extraña desde el punto de vista de la hipótesis inicial por lo que no logra contradecirla con claridad

y, por lo tanto, no permite rechazarla. Por lo tanto, la muestra estudiada no permite concluir que en el pueblo ocurra algo especial que altere la distribución simétrica del sexo en los nacimientos. Ciertamente, se han obtenido bastantes hembras (8 de 10), pero no son suficientes porque pueden achacarse al azar, es decir, aunque no existieran factores que alterasen la simetría sería verosímil que de 10 nacimientos 8 sean hembras, por pura suerte. De haber logrado una hembra más ya no podríamos decir lo mismo y habría supuesto un resultado significativo, con lo que concluiríamos que en el pueblo concurren circunstancias que alteran la simetría original.

En definitiva, un test de hipótesis viene a delimitar los límites del azar a la hora de explicar una observación según un modelo concreto: la hipótesis inicial. En el caso de estadísticos de contrastes que sigan distribuciones continuas el problema es más sencillo. Por ejemplo, si tenemos una distribución $N(0,1)$ la región crítica quedará delimitada por el cuantil $z_{0,05/2}$ y su opuesto $-z_{0,05/2}$, pues la probabilidad de obtener una valor más extremo que éstos es exactamente 0.05, como se indica en el gráfico.



En los casos como éste en que la región crítica queda delimitada por un el cuantil, dado que el mismo se calcula a partir de un modelo de probabilidad teórico, éste se denomina **valor teórico**, en contraposición con el valor experimental. De hecho, el test de hipótesis consiste en comparar un valor experimental que proporciona la muestra con un valor teórico que proporciona la hipótesis inicial.

En definitiva, los elementos de un test de hipótesis son, en general, los siguientes:

1. **Valor experimental:** un número que resumirá en lo posible la información relevante que aporta la muestra.

2. **Hipótesis inicial:** se traduce en un modelo probabilístico teórico cuya validez se juzga mediante una muestra. Si la muestra (su valor experimental) resulta extrema y poco verosímil según este modelo supondrá una prueba significativa contra el mismo y tendremos que rechazarlo.

3. **Potencia:** capacidad del test de detectar la falsedad de la hipótesis inicial. Cuanto más información relativa al problema sea capaz de recoger el valor experimental mayor será la potencia del test.

4. **Región crítica:** es un conjunto tal que la pertenencia del valor experimental al mismo es poco verosímil (probabiliad inferior a 0.05) según la hipótesis inicial. Lo deseable es que esa circunstancia se explique adecuadamente mediante algún modelo alternativo, lo cual sucede con conjuntos extremos en cierto sentido siempre y cuando el valor experimental posea capacidad de discriminar entre el cumplimiento y la obligación de la hipótesis inicial.

La pertenencia a la región crítica debe implicar pues el rechazo de la hipótesis inicial en favor de algún modelos alternativo. En el caso continuo la región crítica se construye a partir de un cuantil denominado valor teórico.

4.2.2. P-valor

Este concepto es fundamental porque viene a expresar el resultado final de un problema de contraste de hipótesis, lo cual puede convertirlo en el parámetro definitivo en un estudio más envergadura.

Tal y como hemos construido el test, aceptaremos o rechazaremos la hipótesis inicial dependiendo exclusivamente de si el valor experimental está o no dentro de una región extrema y poco probable, estableciendo un tope del 5 % para lo que llegamos a entender como raro o poco probable. Debemos tener en cuenta primeramente que esta cota es puramente convencional. En segundo lugar, no es exactamente lo mismo que el valor experimental este muy cerca de la región crítica, como ha sido el caso de nuestro ejemplo, que lejos de ésta, aunque en ambos casos la decisión final sea la misma: aceptar H_0.

Retomando el ejemplo anterior, no sería lo mismo, aunque ambos casos habría conducido a aceptar la hipótesis inicial (en el pueblo no ocurre nada especial), que nacieran 5 u 8 hembras. Tampoco sería lo mismo, aunque ambos casos habría conducido a rechazar la hipótesis inicial (en el pueblo pasa algo), que nacieran 9 o 10 hembras. Al margen del tope que delimita el 5 % de casos extremos, constituido por el par 1-9, nos gustaría dar una medida del grado de verosimilitud de nuestra muestra según la hipótesis inicial. Esto nos lo proporciona el denominado P-valor o probabilidad de significación, que se define como la probabilidad de obtener una muestra al menos tan extrema como la nuestra según la hipótesis inicial.

Concretamente, con la secuencia de nacimientos obtenida tenemos 8 hembras y 2 varones. La probabilidad de obtener un caso igual de extremo o más que el nuestro, es decir, la probabilidad de obtener 8,2,9,1,10 o 0 hembras según el modelo $B(10, 0.5)$ es $P = 0{,}110$. Ese es el P-valor que corresponde a nuestra muestra. Al ser mayor que 0.05 se dice que es un resultado no significativo. Si hubiéramos obtenido, por ejemplo, 9 hembras, habría que sumar las probabilidades de obtener 9,1,10 y 0, obteniendo $P = 0{,}022$. Este sí sería un resultado significativo porque $P < 0{,}05$, lo cual equivale a que 9 se encuentre en la región crítica. Que este valor de P sea inferior al anterior nos habla de una muestra menos verosímil según la hipótesis inicial. De hecho, es tan poco verosímil que nos invita a rechazar la hipótesis. Sin embargo, si el número de hembras hubiera sido 10 (el caso más extremo) el P-valor sería 0.002, lo cual nos habla de una situación casi inverosímil según la hipótesis inicial, lo cual invitaría a rechazarla pero con mayor seguridad que en el caso $P = 0{,}022$.

En definitiva, no sólo distinguiremos entre resultados significativos ($P < 0{,}05$) y no significativos ($P > 0{,}05$) sino que hablaremos de distintos grados de significación. Concretamente, lo más habitual es clasificar el resultado según la tabla siguiente:

$0{,}05 <$	P		Resultado no significativo
$0{,}01 <$	P	$\leq 0{,}05$	Resultado significativo
$0{,}001 <$	P	$\leq 0{,}01$	Resultado muy significativo
	P	$\leq 0{,}001$	Resultado altamente significativo

Ya hemos dicho que establecer el tope de lo que consideramos como raro en el 5 % es puramente convencional. De hecho, en ocasiones estamos dispuestos a ser más críticos con la hipótesis inicial elevando ese tope al 10 %, lo cual supondría ampliar la región crítica. Eso se hace en ocasiones para compensar la ventaja que para la hipótesis inicial supone una muestra pequeña. En otras circunstancias, queremos adoptar una actitud más conservadora disminuyendo el tope al 1 % o incluso al 0.1 %, lo cual implicaría reducir la región crítica. Esto ocurre cuando rechazar la hipótesis inicial puede implicar en la práctica serias modificaciones por lo que precisamos estar especialmente convencidos de ello. El tope que establezcamos finalmente, que suele ser del 5 %, se denomina nivel de significación del test, y se definiría formalmente como la probabilidad de la región crítica según el modelo H_0.

Pero en definitiva, debemos tener presente que una misma muestra puede conducir a distintas conclusiones según el grado de riesgo que estemos dispuestos a asumir respecto a la hipótesis inicial. Sin embargo, ello no debe suponer ninguna complicación si conocemos el P-valor correspondiente a nuestra muestra, porque sabemos en qué medida es verosímil para H_0 y ése es el resultado final del test de hipótesis. Que esa verosimilitud sea considerada grande o pequeña es completamente subjetivo, aunque se convenga en establecer 0.05 como cota habitual. El P-valor es el resultado final que aporta un programa estadístico cuando aplica un contraste de hipótesis.

$$\texttt{Muestra} \leftrightarrow \texttt{Valor experimental} \leftrightarrow P\texttt{-valor}$$

Si el estadístico de contraste sigue una distribución continua, obtener el P-valor es más sencillo aún. Se trata de calcular el área de las colas que delimitan el valor experimental por un lado y su opuesto por otro en la densidad de la distribución. Retomemos los datos del ejemplo 8. Supongamos que, por estudios previo, se considera que la estatura media de las mujeres de esa franja de edades es de $\mu_0 = 164{,}5cm$. Vamos a utilizar los $n = 40$ datos de nuestra muestra para contrastar la veracidad de dicha afirmación.

$$H_0 : \mu = 164{,}5$$

En primer lugar, debemos calcular a partir de la muestra un número (valor experimental) que resuma lo mejor posible la información relevante para este problema. Aunque no deje de ser intuitivo, es necesario un conocimiento más profundo de la Estadística para poder determinar el valor experimental. No obstante, no parece mala idea que la media aritmética intervenga en el cálculo, aunque también sabemos que por sí sola difícilmente podrá resumir suficientemente la información, por lo que la desviación típica se antoja también necesaria. Tal vez un único número calculado a partir de ambas podría recoger suficiente información, al menos en lo relativo a este problema, máxime teniendo en cuenta (3.1). En definitiva, el estadístico de contraste que proponemos en este caso es el siguiente

$$\mathcal{T} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

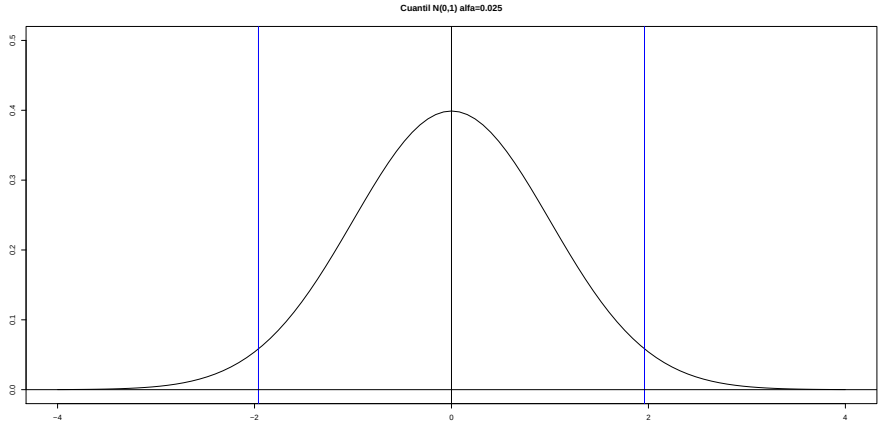
Según (3.1), si la media de la población μ coincidiera efectivamente con $\mu_0 = 164{,}5$, el estadístico de contraste seguiría (aproximadamante) un modelo de distrubución $N(0, 1)$. Por lo tanto, la hipótesis inicial puede expresarse de la forma

$$H_0 : \mathcal{T} \sim N(0, 1)$$

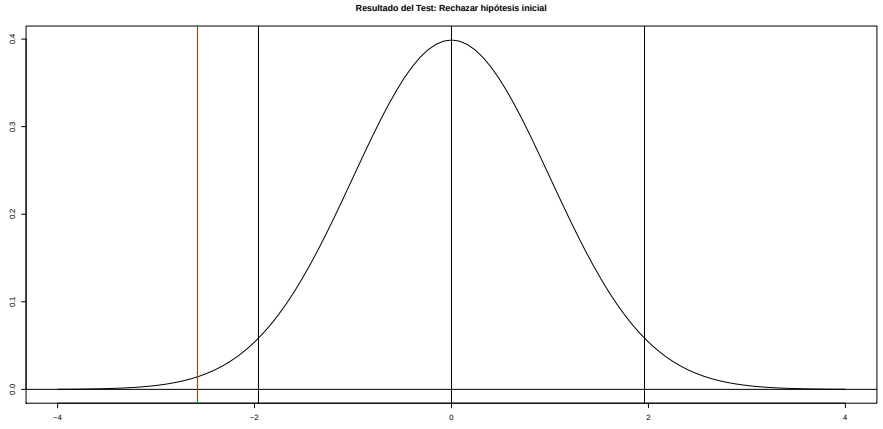
El valor experimental, que se obtiene sustituyendo en la expresión anterior los datos correspondientes a nuestra muestra es

$$\mathcal{T}_{exp} = \frac{162{,}3 - 164{,}5}{5{,}2/\sqrt{40}} = -2{,}67$$

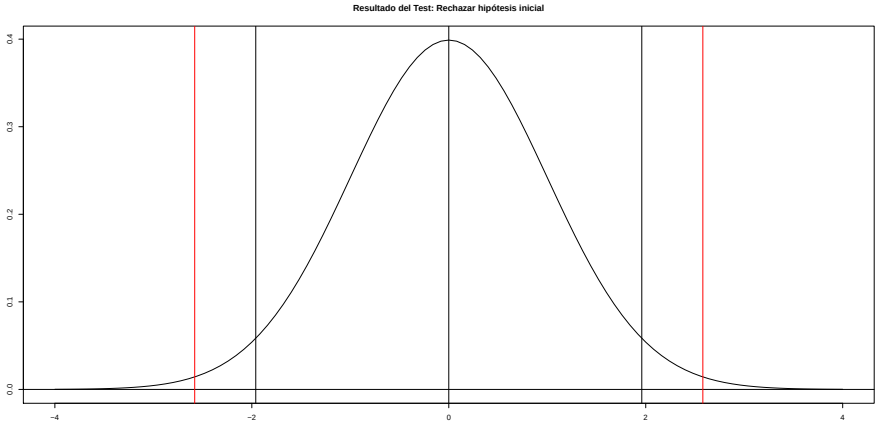
Si la hipótesis inicial fuera cierta, es decir, si $\mu = 164{,}5$, cabría esperar que la media aritmética de la muestra fuera próxima a 164.5 o, equivalentemente, un valor experimental próximo a 0. Si la verdadera media μ dista mucho de μ_0, la distribución de $\mathcal{T}$ será normal pero no de media 0 sino $\mu - \mu_0$. Si esta diferencia es positiva, los valores alejados de 0 por la derecha, que eran poco verosímiles según el modelo $N(0,1)$, quedarán mejor explicados según ese modelo alternativo. Si la diferencia $\mu - \mu_0$ es negativa, quedarán mejor explicados los valores a la izquierda. Así pues, queda claro que un valor experimental lejano a 0, es decir, extremo, debe conducir a rechazar la hipótesis inicial. También parece razonable, por simetría, que la decisión que corresponde un determinado valor experimental debe ser la misma que corresponde a su opuesto. Así pues, la región crítica del test estará constituida por dos colas simétricas respecto a 0 de la curva $N(0,1)$ que sumen un área o probabilidad de 0.05. La cola de la derecha debe ser pues la que queda delimitada por el cuantil $z_{0{,}05/2}$ y la de la izquierda, por el valor $-z_{0{,}05/2}$, como se indica en la figura



Si el valor experimental queda en las colas, debemos rechazar la hipótesis inicial. Si queda entre ellas, la aceptaremos. En nuestro caso ocurre lo primero pues $\mathcal{T}_{exp} = -2{,}67$



Es decir, rechazamos la hipótesis inicial cuando $|\mathcal{T}_{exp}| > z_{0,05/2}$. En este caso, ocurre efectivamente que 2,67 > 1,96. No obstante, conviene en todo caso calcular el P-valor, que hemos definido como la probabilidad, según H_0 (es decir, según el modelo $N(0,1)$), de obtener un resultado al menos tan extremo como $\mathcal{T}_{exp}$. Viendo el anterior gráfico, se trataría del área que queda a la izquierda de la línea roja multiplicada por 2, puesto que habría que sumarle la cola simétrica.



El valor exacto de P se obtendría pues resolviendo la ecuación

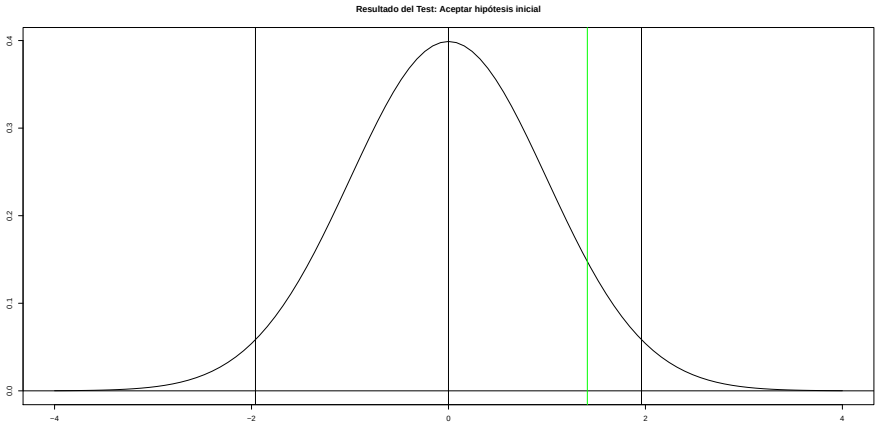
$$z_{\frac{P}{2}} = |\mathcal{T}_{exp}|$$

aunque cualquier programa estadístico lo calcula automáticamente. En este caso se obtiene $P =$ 0,0076. Como cabía espera, el P-valor es inferior a 0.05. Es más, incluso queda por debajo de 0.01, lo que supone un resultado muy significativo. Para valorarlo en su justa medida debemos tener muy presente el significado del P-valor:

> El P-valor viene a ser una medida del grado de verosimilitud de la muestra desde el punto de vista del modelo inicial, es decir, informa de lo rara y extrema que es nuestra muestra según dicha hipótesis.

Así pues, cuanto más pequeño sea el P-valor, mayor será la contradicción entre nuestra muestra y la hipótesis inicial y, en consecuencia, más significativas serán las pruebas en su contra. En este caso, tenemos pruebas muy significativas contra la hipótesis inicial de que la media sea 164.5, lo que nos induce a pensar que esta hipótesis es falsa.

Imaginemos que los datos de la muestra hubieran aportado una media de 165.7cm con una desviación típica de 3.8cm. En ese caso, habríamos obtenido un valor experimental $\mathcal{T}_{exp} = 1{,}41$, que representamos a continuación con una línea verde



Como podemos apreciar, queda dentro de la región de aceptación de la hipótesis inicial pues $|\mathcal{T}_{exp}| \leq 1{,}96$. Por lo tanto, el correspondiente P-valor debe ser superior a 0.05. Concretamente se tiene que $P = 0{,}1586$. En definitiva, esta muestra no aportaría una prueba significativa contra la hipótesis inicial. Es decir, no estamos en condiciones de rechazarla. La diferencia existente entre la media supuesta, 164.5cm, y la que presenta la muestra, 165.7cm, puede explicarse por el azar inherente al muestreo.

4.2.3. Relación entre test de hipótesis e intervalo de confianza

Los propios datos originales del ejemplo 8 pueden servir para evidenciar una clara vinculación entre el intervalo al 95 % de confianza y el test de hipótesis considerado, dado que ambos han sido construidos partiendo del resultado (3.1) del capítulo anterior

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \overset{\text{aprox}}{\sim} N(0,1)$$

Recordemos que, con estos datos, el intervalo de confianza al 95 % para la media de nuestra población es $(160.7, 163.9)$. Es decir, con una confianza del 95 % y, por lo tanto, asumiendo ua probabilidad de error de 5 %, afirmamos que la media μ se encuentra entre esos límites. De no ser así, significaría que nuestra muestra estaría entre el 5 % de muestras más extremas, cosa que por principio nos negamos a pensar. Dado que μ_0=164.5 queda fuera del intervalo, debemos entender entonces que nuestra media no puede ser μ_0, que es precisamente lo que se concluye tras aplicar el test de hipótesis. Si queremos analizar el porqué de esta coincidencia basta tener en cuenta que μ_0 pertenece al intervalo de confianza al 95 % cuando

$$|\overline{X} - \mu_0| \leq E_{\text{máx}}$$

es decir, cuando

$$\frac{|\overline{X} - \mu_0|}{S/\sqrt{n}} \leq z_{\frac{0{,}05}{2}}$$

o, equivalentemente, cuando

$$|\mathcal{T}_{exp}| \leq z_{\frac{0{,}05}{2}},$$

que es lo que tiene que ocurrir para que el resultado del test de hipótesis sea no significativo. En nuestro caso el resultado es $P = 0{,}0078$, que es muy significativo. Puede probarse igualmente que eso equivale a que μ_0 quede fuera del intervalo al 99 % de confianza para la media, mayor que el anterior. Como no se trata de un resultado altamente significativo, podemos comprobar que μ_0 sí queda dentro del intervalo al 99.9 % de confianza. La regla es sencilla: el nivel de riesgo α que se asume al construir el intervalo debe concordar con la probabilidad asociada a la región crítica del test, es decir, con su nivel de significación.

Este vínculo entre intervalos de confianza y tests de hipótesis puede extrapolarse a muchos de los métodos que estudiaremos en el siguiente capítulo, y será de especial interés a la hora de contrastar la hipótesis inicial igualdad entre las medias μ_1 y μ_2 de sendas poblaciones. Efectivamente, en ese caso, la hipótesis inicial equivale a que la diferencia entre ambas medias sea nula. El programa estadístico proporcionará, además del resultado del test correspondiente en forma de P-valor, un intervalo al 95 % de confianza para la diferencia de medias $\mu_1 - \mu_2$. De esta forma, el resultado será significativo cuando el valor 0 quede fuera de dicho intervalo. Pero el intervalo tiene la virtud añadida de expresar el tamaño de la diferencia entre ambas medias.

4.2.4. Hipótesis alternativa: contrastes bilaterales y unilaterales

Hasta ahora nos hemos centrado en la hipótesis inicial H_0 y hemos hablado en términos muy vagos de su alternativa. Hemos entendido como hipótesis alternativa cualquier modelo teórico diferente del inicial H_0. En el caso del ejemplo 9, si imponemos una serie de supuestos formales, esa familia de modelos se expresaría mediante $p \neq 0{,}5$. Esa hipótesis (o familia de hipótesis) se denota por H_1. Así pues el contraste se plantea de la forma

$$\begin{cases} H_0: & p = 0{,}50 \\ H_1: & p \neq 0{,}50 \end{cases}$$

En el caso de los datos del ejemplo 8, las hipótesis a contrastar son

$$\begin{cases} H_0: & \mu = 164{,}5 \\ H_1: & \mu \neq 164{,}5 \end{cases}$$

También podemos contrastar si la media de dos poblaciones, μ_1 y μ_2, son o no diferentes. En tal caso, la hipótesis inicial es $H_0 : \mu_1 = \mu_2$, mientras que la alternativa es la negación de H_0, es decir $H_1 : \mu_1 \neq \mu_2$.

Sin embargo, en ocasiones tenemos una idea más clara y por lo tanto restrictiva de la hipótesis alternativa. Volvamos al ejemplo 9: existe la teoría de que ciertos contaminantes ambientales no sólo están afectando a la capacidad de reproducción masculina sino que incluso está impidiendo que los embriones y fetos masculinos prosperen. En definitiva, de ser eso cierto, existiría una mayor tendencia a los nacimientos de niñas en las zonas con mayor exposición ($p > 0{,}50$). Supongamos que nuestro pueblo es una de esas zonas y que lo hemos seleccionado como muestra para contrastar dicha teoría. En ese caso, la hipótesis inicial es, como siempre[2], $H_0 : p = 0{,}50$. Sin embargo, la hipótesis alternativa no es la negación de la anterior sino $H_1 : p > 0{,}50$. Así pues, nos planteamos

[2]Porque esta hipótesis debe identificarse con un fenómeno aleatorio concreto a partir del cual podamos calcular probabilidades.

el contraste

$$\begin{cases} H_0: & p = 0{,}50 \\ H_1: & p > 0{,}50 \end{cases}$$

Contrastes de este tipo se denominan **unilaterales** en contraposición de con los anteriores, denominados **bilaterales**. ¿En qué afecta este matiz al diseño del test de hipótesis? Pues viene a romper la simetría en la región crítica. En este caso, un número elevado de hembras en la secuencia de nacimientos puede resultar raro según la hipótesis inicial pero verosímil según la alternativa considerada, lo que debe conducirnos a optar por esta última. Sin embargo, un escaso número de hembras (muchos varones) puede resultar raro para la hipótesis inicial pero lo será mucho más para la alternativa, por lo que el Principio de Máxima Verosimilitud nos conduce a aceptar H_0. Por lo tanto, la región crítica para este contraste debe estar constituida exclusivamente por los valores extremos y raros a la derecha de 5. Como no hay que sumar la probabilidad de las dos colas estamos en principio condiciones de ampliar la región crítica por este lado, es decir, vamos a ser más críticos que en el planteamiento bilateral si la muestra presenta más hembras que varones. Por contra, si presenta más varones que hembras la decisión será automáticamente H_0.

¿Cómo afecta este nuevo diseño al P-valor? Pues en el caso de que haya más varones que hembras no se define siquiera. Si el número de hembras es mayor, el P-valor será la probabilidad de obtener una valor tan grande al menos como ése. Como no hay que considerar la región simétrica a la izquierda de 5, esta probabilidad será exactamente la mitad del P-valor correspondiente al contraste bilateral. Por lo tanto, con los datos de nuestro ejemplo, tendríamos

$$P = \frac{0{,}110}{2} = 0{,}055$$

Vemos que el P-valor ha disminuido, lo que supone un resultado más crítico hacia H_0, aunque sigue sin ser significativo.

Ni que decir tiene que pueden considerarse hipótesis alternativas del tipo $H_1 : p < 0{,}50$. En ese caso, la región crítica quedaría a la izquierda y la regla para obtener el P-valor sería la misma pero al contrario. También podemos considerar hipótesis del tipo $H_1 : \mu_1 < \mu_2$, $H_1 : p_1 > p_2$, etcétera.

4.3. Cuestiones propuestas

1. Se estudia cierta variable X. Una muestra de tamaño n aportó un determinado intervalo de confianza para la media μ de la variable al 95 % de confianza. Razona si el intervalo de confianza al 99 % ha de ser más o menos amplio que el anterior.

2. En una muestra de 100 pacientes con infarto se ha medido el GOT a las 12 horas, obteniéndose una media de 80 y una desviación típica de 120. Construir un intervalo de confianza al 95 % para la media de todos los infartados. Según estudios anteriores el valor medio del GOT es de 85. Contrasta esta hipótesis calculando el correspondiente P-valor. Relacionar el resultado obtenido con el intervalo de confianza anterior.

3. Supongamos que el tiempo utilizado en la atención a un paciente es una variable aleatoria. Se pretende determinar de la manera más precisa posible el tiempo medio esperado de atención a partir de una muestra supuestamente aleatoria de tamaño 50 que aportó una media aritmética de 34 minutos con una desviación típica de 2.3 minutos. ¿Qué podemos

hacer? Según los organismos públicos el tiempo medio de atención no excede de los 30 minutos. Contrastar dicha hipótesis inicial calculando el P-valor.

4. En el contraste de hipótesis del ejemplo 9 se ha obtenido un P-valor de 0.110, lo cual supone un resultado no significativo. ¿Significa eso que se ha demostrado que no existen en el pueblo factores ambientales que alteren la simetría en el sexo de los bebes? Si no se está de acuerdo con dicha afirmación, qué deberíamos hacer?

5. Diseñar un test de hipótesis para contrastar la hipótesis inicial anterior pero partiendo en esta ocasión de una muestra de 100 nacimientos. Indica que P-valor se obtendrá si la proporción de hembras en la muestra es del 80 %.

6. Cuando hemos construido el test para el contraste bilateral de una media hemos afirmado que el estadístico de contraste
$$\mathcal{T} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$
calculado a partir de la media aritmética y la desviación típica, podría recoger suficiente información de la muestra, al menos en lo relativo a este problema. Decimos "podría"porque ello ocurre bajo ciertas condiciones. ¿Puedes intuir en que condiciones el estadístico $\mathcal{T}$ es idóneo?

7. ¿En qué sentido puede influir el tamaño de la muestra en un test de hipótesis?

8. Al contrastar la hipótesis inicial $\mu = 164{,}5$ con los datos del ejemplo 8 se ha obtenido un resultado muy significativo. ¿Estamos entonces seguros de que la media de la población difiere de 164.5?

9. En un problema de contraste de hipótesis se obtiene como resultado final $P > 0{,}05$. ¿Significa eso que se ha demostrado la autenticidad de H_0? ¿Cómo debe interpretarse un valor $P < 0{,}05$? ¿Cuál ha de ser el valor de P para tener certeza absoluta de la falsedad de H_0?

10. Partiendo de una muestra aleatoria de tamaño $n = 250$ de una variable, se obtuvo como resultado $\overline{x} = 13{,}1$ y $s = 2{,}2$. El intervalo al 95 % de confianza para la media es el siguiente:

    A: (11.73 , 14.44)

    B: (12.09 , 14.10)

    C: (12.82 , 13.37)

    D: (2.74 , 3.75)

    E: (2.72 , 3.77)

11. En relación con la cuestión anterior, se plantea el problema de decidir si la media de la variable estudiada es igual a 13.3. Indica si son verdaderas o falsas cada una de las siguientes afirmaciones:

    A: El resultado del test a partir de los datos de la muestra no es significativo.

    B: Tenemos una confianza del 95 % de que la hipótesis inicial se da con absoluta seguridad.

C: El resultado del test a partir de los datos de la muestra es muy significativo.

12. ¿Por qué podemos afirmar que el resultado del test para contrastar la hipótesis inicial $H_0 : \mu = \mu_0$ es muy significativo cuando μ_0 queda fuera del intervalo al 99 % de confianza para la media μ?

13. Con los datos del problema 8, describe la región crítica del test para contrastar la hipótesis inicial $H_0 : \mu = 164{,}5$ contra la alternativa unilateral $H_1 : \mu < 164{,}5$. Calcula de manera directa el P-valor correspondiente a este contraste.

14. Aunque en el ejemplo 8 hemos supuesto que, si no concurren agentes ambientales externos, la proporción de nacimientos de varones ha de ser idéntica a la de nacimientos de hembras, lo cierto es que se sabe desde hace tiempo que esto no es así y que, de hecho, depende de la composición étnica de la población. Tradicionalmente, se ha venido registrando en Europa Occidental año tras año proporciones de nacimientos varones en torno al 51 %. Por lo tanto, si se aplicara un test para contrastar la hipótesis inicial $H_0 : p = 0{,}50$, el resultado del mismo sería significativo, ¿o no? Comenta de manera crítica esta última afirmación.

Capítulo 5

Métodos de Inferencia Estadística

Después de dos capítulos eminentemente teóricos estamos en disposición de abordar un somero estudio de las técnicas más utilizadas en Inferencia Estadística. En buena parte, se trata de una continuación de los capítulos 1 y 2 dedicados a la Estadística Descriptiva, con la salvedad de que, en esta ocasión, no nos conformaremos con la descripción de la muestra pues nuestro propósito es extraer, a partir de ésta, conclusiones relativas a la población de la que procede, suponiendo que haya sido seleccionada de la manera más aleatoria posible.

Por el capítulo anterior debemos saber a grandes rasgos a qué tipo de problemas nos enfrentamos y cómo se pretenden solucionar. En éste intentaremos concretar distinguiendo entre el estudio aislado de una variable, tanto cuantitativa como cualitativa, el de relación entre dos variables, bien sean ambas cuantitativas, cualitativas o mezcla de ambos tipos, y el de relación entre más de dos variables. No se trata de un análisis exhaustivo de cada uno de ellos sino más bien de una clasificación donde se indicará, en cada caso, las motivaciones y el tipo de tratamiento que han de seguir los datos, dando por supuesto que los aspectos relativos al cálculo deben ser solucionados mediante un programa estadístico. 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 correcta de un programa estadístico. Los pormenores de los diferentes métodos pueden encontrarse en la bibliografía recomendada.

Métodos paramétricos versus métodos no paramétricos

A lo largo de este capítulo tendremos la ocasión de comprobar que la mayor parte de nuestras inferencias están relacionadas con los parámetros poblacionales media μ y varianza σ^2, u otro como el coeficiente de correlación poblacional ρ que se obtiene a partir de la covarianza y de las varianzas. Este interés esta claramente vinculado con la distribución normal. Efectivamente, sabemos de la importancia que en general posee el parámetro media, y que éste debe complementarse con alguna medida de dispersión para poder caracterizar la distribución de los datos. La varianza desempeña ese papel, al menos en en el caso e 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 métodos de inferencia que proponemos son válidos aunque no se dé la normalidad. 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 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 la variables y no se busca por lo tanto el conocimiento de los parámetros que las caracterizan, de ahí que se denomine Estadística no Paramétrica. Nosotros nos centraremos 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 éste sea inviable, bien por las condiciones de la distribución, bien por el escaso número de datos. El esquema a seguir en la mayoría de nuestros problemas es el siguiente:

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

Para decidir si la distribución original de los datos es o no normal contamos con los denominados tests de normalidad que introducuiremos en la siguiente sección. Respecto al tamaño de muestra requerido para que ésta sea considerada suficientemente grande, sabemos que se suele manejar la cota $n = 30$.

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 muestra. Son por lo tanto métodos robustos ante la presencia de valores extremos (como suede con el cálculo de la mediana). No obstante, para un estudio más detallado remitimos al lector a la bibliografía recomendada. Por último, antes de empezar con la exposición de las diferentes estudios a considerar, mostramos un breve esquema de los mismos agrupados en las siete secciones a considerar. Algunos de ellos apenas se tratarán pues quedan fuera de los contenidos de las asignaturas que este manual pretende cubrir.

¿Qué se mide en la población?	**Problemas estadísticos relacionados**
Una variable cuantitativa	Contraste para una media o varianza Intervalo de confianza para una media Límites de normalidad (diagnóstico)
Una variable cualitativa	Contraste para una proporción Intervalo de confianza para una proporción
Dos variables cuantitativas	Comparación dos medias muestras apareadas Análisis de regresión-correlación
Dos variables cualitativas	Tablas de contingencia. Test χ^2 Comparación de dos proporciones Factores de riesgo de una enfermedad Validez métodos de diagnóstico enfermedad
Explicativa cualitativa y respuesta cuantitativa	Comparación dos medias muestras indeptes. Compración de más de dos medias (anova)
Explicativa cuantitativa y respuesta cualitativa	Regresión logística
Más de dos variables	Regresión múltiple, Análisis de la Covarianza Anova multifactorial Manova, Multidimensional Scaling

5.1. Estudio de una variable cuantitativa

Esta sección, junto con la siguiente, son continuación del capítulo 1. Además, muchos de sus contenidos ya han sido tratados a título ilustrativo en el capítulo anterior. Dada un variable numérica medida en una población nos puede interesar en principio cualquiera de los valores típicos estudiados en el capítulo, aunque nos centraremos en la media μ y la varianza σ^2, fundamentalmente en el primero.

5.1.1. Inferencias para la media

Ya sabemos que la media μ se estima mediante la media aritmética de la muestra. También sabemos, por lo visto en el capítulo anterior, construir intervalos de confianza y tests para contrastar hipótesis del tipo

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu > \mu_0 \end{cases}$$

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \end{cases}$$

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

para algún valor concreto de μ_0. Vimos incluso un ejemplo numérico (ejemplo 8). Por otra parte, en las propias salidas de estadística descriptiva podemos encontrar intervalos de confianza para la media al 95 %. Por ejemplo, en el estudio de la concentración de Ozono cerca de Seattle a partir de $n = 35$ datos se obtuvo lo siguiente:

Descriptivos

		Estadístico
Media		171,66
Intervalo de confianza para la media al 95%	Límite inferior	167,62
	Límite superior	175,69
Media recortada al 5%		170,44
Mediana		168,00
Varianza		137,997
Desv. típ.		11,747
Mínimo		160
Máximo		212
Rango		52
Amplitud intercuartil		15
Asimetría		1,646
Curtosis		2,999

Podemos apreciar que el intervalo de confianza al 95 % para la media μ de la concentración e O_3 es (167.62 , 175.69). Supongamos además que es conocido que la media de concentración de ozono en el área de Los Ángeles es de 220 (partes por millón). Deseamos contrastar si, por término medio, existen diferencias entre Seattle y Los Ángeles. Podemos proponer pues el contraste

$$\begin{cases} H_0: & \mu = 220 \\ H_1: & \mu \neq 220 \end{cases}$$

Lo resolvemos calculando el valor experimental

$$t_{exp} = \frac{\overline{x} - 220}{s/\sqrt{n}} = \frac{171{,}66 - 220}{11{,}74/\sqrt{35}} = -24{,}36$$

que se comparará con un valor teórico $z_{0{,}05/2}$ de la tabla $N(0,1)$ (si tenemos en cuenta el intervalo de confianza obtenido, deducimos e antemano que la decisión debe ser H_1). No obstante, podemos precisar que en este caso obtenemos $P < 0{,}001$, por lo que la muestra escogida supone una prueba altamente significativa de que los niveles medios de contaminación difieren. Seguramente, un experto en el tema no contempla en ningún caso que el nivel de ozono sea mayor en Seattle que en Los Ángeles y pretenda probar directamente que es significativamente menor. En ese caso, las hipótesis a contrastar habrían sido las siguientes:

$$\begin{cases} H_0: & \mu = 220 \\ H_1: & \mu < 220 \end{cases}$$

La resolución de este nuevo problema consiste en dividir entre 2 el P-valor anterior, con lo que la conclusión sería la misma aunque más contundente si cabe. En general, no debe obsesionarnos demasiado si el contraste debe ser unilateral o bilateral. El investigador que plantea el problema no tendrá ninguna duda al respecto y la resolución del caso unilateral es automática a partir de la del caso bilateral.

Todo esto lo vimos en el capítulo anterior. También dijimos que tanto el test como el intervalo de confianza son válidos sea cual sea la distribución de la variable en la población estudiada, siempre y cuando la muestra escogida sea lo suficientemente grande, pues nos basamos siempre en la aproximación de la distribución muestral de $\overline{X}$ al modelo normal, según se indica en (3.1). Solemos exigir $n \geq 30$.

No obstante, para garantizar que el test, denominado de Student, es óptimo según los criterios estadísticos principales, precisamos que la distribución de la variable original se ajuste satisfactoriamente a un modelo normal, cosa que no sucede en nuestro caso (basta echar un vistazo al coeficiente de asimetría). Insistimos en que una muestra grande puede paliar en parte la violación de este supuesto. Si no estamos en condiciones de asumir normalidad y la muestra es pequeña, ni el test ni el intervalo de confianza son válidos en el sentido de que las probabilidades de error que se les supone no son correctas. Por contra, si la distribución de la variable se ajusta satisfactoriamente a un modelo normal no encontraremos métodos más adecuado que el de Student, sea cual sea el tamaño de la muestra. Recordamos por último que, en tal caso y si el tamaño e muestra es pequeño $n < 30$ se utilizar el cuantil de la tabla t-Student en lugar del de la tabla $N(0,1)$, pues pueden diferir bastante.

5.1.2. 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 las Ciencias de la Salud, lo cual no quiere decir que deba dar por descontado.

Precisamente, existen diversos métodos, como el de Kolmogorov-Smirnov, el de Shapiro-Wilk, el χ^2 o el 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 está vinculados a aspectos gráficos. Por ejemplo, el test χ^2 analiza cuantitativamente si el histograma de frecuencias relativas se asemeja al que correspondería a una distribución normal; Kolmogorov-Smirnov analiza si el histograma de frecuencias relativas acumuladas se parece a la función de distribución de un modelo Normal, etc. También existe un método basado 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 observado la contradigan claramente. Nótese que una muestra pequeña y por lo tanto con escasa información difícilmente podrá conducir a rechazar la hipótesis inicial de normalidad. Este hecho puede contrarrestarse considerando significativos los resultados $P < 0{,}10$ o incluso $P < 0{,}20$. Por contra, si la muestra es muy grande, los tests propuestos serán muy potentes y detectaran la menor violación del supuesto de Normalidad. Dado que entendemos que ese supuesto es ideal y que nuestros métodos son razonablemente válidos para aproximaciones aceptables de la distribución al modelo Normal, deberíamos estar dispuestos a reducir en esos casos el nivel de significación.

5.1.3. Tamaño de muestra requerido en la estimación

En ocasiones estamos interesados en determinar de antemano el tamaño de muestra que se requiere aproximadamente para poder estimar la media con cierto grado de precisión y de confianza establecidos. Por ejemplo, en el caso del ozono obtuvimos como intervalo de confianza al 95 % para la media (167.62 , 175.69). Eso quiere decir que el margen máximo de error que otorgamos a nuestra estimación es $E_{\text{máx}}$ =4.03, con una confianza del 95 %. Imaginemos que nuestra intención fuera estimar la media con un margen máximo de error de una unidad con una confianza del 95 %, cosa que no se ha conseguido en este caso. ¿Qué deberíamos hacer? Tener en cuenta que

$$E_{\text{máx}} = z_{\alpha/2}\frac{s}{n}$$

Por lo tanto, fijado $\alpha = 0{,}05$, el margen de error de la estimación depende únicamente del tamaño de la muestra y de su desviación típica. Esta última es impredecible pues se trata de una variable muestral. No obstante, podemos seleccionar una primera muestra denominada muestra piloto cuya desviación típica puede servirnos de referencia, de manera que, asumiendo cierta inexactitud, el margen de error $E_{\text{máx}}$ dependerá únicamente del tamaño de muestra. Concretamente, será inversamente proporcional a su raíz cuadrada. En nuestro ejemplo, la muestra estudiada podría servir

como muestra piloto, de manera que el tamaño de muestra aproximado que se requiere para que $E_{\text{máx}}$ sea a lo sumo 1 se obtiene resolviendo la inecuación

$$1 \leq 1{,}96\frac{11,74}{\sqrt{n}}$$

La solución es $n \geq 529$. Esta solución es sólo aproximada, entre otras cosas porque la desviación típica de una muestra de ese tamaño no será igual a 11.74, aunque esperamos que se parezca. El caso es que ya sabemos de que orden debe ser el tamaño de la muestra requerida si tenemos esas exigencias.

5.1.4. Inferencias para la varianza

El interés de la varianza de una distribución estriba en que viene a complementar el papel que desempeña la media a la hora de caracterizarla. No obstante, hemos de reconocer que las inferencias respecto a la varianza desempeñan un papel secundario por dos razones. Primero porque la media es el parámetro central y es más importante. De hecho, la varianza se denomina en algunas contextos parámetro ruido o fantasma, lo que viene a significar molesto. Segundo, porque está excesivamente vinculado al supuesto de normalidad tanto en lo que respecta a su interpretación como en lo referente a la validez de los métodos propuestos para su estimación y contraste. Efectivamente, sabemos ya que los métodos de inferencia para la media son válidos para muestras grandes aunque la distribución original no sea normal, pero no podemos decir lo mismo de la varianza.

Existen no obstante circunstancias excepcionales en las que la varianza se convierte en el parámetro estrella. Nos referimos a los problemas relacionados con la precisión de un método de medición. Efectivamente, es muy común asumir que la distribución del error en la medición de una variable (peso, colesterolemia, etc) sigue una distribución normal donde su media indica un error que se comete sistemáticamente y, por lo tanto, fácilmente corregible, mientras que su varianza expresa el grado de precisión del método de medida. Así, podemos plantearnos, hipótesis del tipo

$$\left\{\begin{array}{ll} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 < \sigma_0^2 \end{array}\right. \qquad \left\{\begin{array}{ll} H_0: & \sigma^2 = \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{array}\right.$$

Estos contrastes se resuelven haciendo uso de la tabla teórica del modelo χ^2 y, repetimos, son sensibles a la violación del supuesto de normalidad.

5.1.5. Diagnóstico clínico I: límites de normalidad

En las Ciencias de la salud es muy frecuente asumir que la distribución de una variable cuantitativa X de interés en la población sana es normal, porque de hecho así sucede de manera aproximada en muchas ocasiones. No obstante, consideraciones de este tipo deberían fundamentarse el el resultado de algún test de normalidad. El hecho es que, si asumimos dicha hipótesis, la aplicación del Principio de Máxima Verosimilitud se traduce en una sencilla herramienta para diagnosticar patologías. Podemos considerar que un individuo no está sano cuando la puntuación que obtiene en la variable estudiada es extrema según el modelo de distribución que corresponde a la población sana.

En algunas ocasiones pueden resultar patológicos los valores excesivamente altos, en otras los excesivamente bajos, o también ambos casos. Por ejemplo, en la tercera situación, para diagnosticar la patología deberíamos marcas dos límites, uno inferior y otro superior, de manera que el sobrepasarlos sea muy poco probable para un individuo de la población sana. Estas cotas se denominan límites de normalidad o tolerancia. Primeramente debemos acalarar qué entendemos exactamente por extremo, es decir, hay que determinar un valor α =0.05, 0.01, 0,001 y construir un intervalo de la forma $\mu \pm \mathtt{I}$ tal que $P(\mu - \mathtt{I} \leq X \leq \mu + \mathtt{I}) = 1 - \alpha$. En ese caso, si la puntuación que corresponde al individuo queda fuera del intervalo se considerará un dato extremo, lo cual supondrá un positivo en el diagnóstico. Lo ventajoso de suponer la normalidad es que, dado α, el valor de $\mathtt{I}$ será proporcional a la desviación típica σ, es decir, el intervalo será de la forma

$$(\mu - \mathtt{k}_\alpha \sigma, \mu + \mathtt{k}_\alpha \sigma)$$

Concretamente, se trata de buscar el valor de $\mathtt{k}_\alpha$ tal que

$$\alpha = P\left(\frac{|X - \mu|}{\sigma} > \mathtt{k}_\alpha\right) = P(|Z| > \mathtt{k}_\alpha)$$

donde $Z \sim N(0,1)$. Por lo tanto, $\mathtt{k}_\alpha = z_{\alpha/2}$. En el caso de que se considere patológico sólo un valor excesivamente alto habría que considerar el límite superior de normalidad $\mu + z_\alpha \sigma$; si lo patológico es un valor excesivamente bajo tomaríamos el límite inferior de normalidad $\mu - z_\alpha \sigma$.

En la práctica nos enfrentamos al problema de que se desconocen lo valores reales de μ y σ. Tan sólo tenemos sus estimaciones $\overline{x}$ y s a partir de una muestra de tamaño n. Por ello los límites de tolerancia debe corregirse sutilmente atendiendo al tamaño de la muestra y al nivel de confianza $1 - \alpha'$ que consideremos conveniente, de manera que, a la postre, los límites son de la forma

$$\overline{x} \pm \mathtt{k}_{[\alpha, n, \alpha']} \cdot s$$

También existen técnicas para construir límites de tolerancia en distribuciones no normales.

Por último, citamos ciertos límites de normalidad facilitados por los Servicios de Bioquímica y Hematología de un hospital universitario español.

Glucosa (mg/dl)	[70,110]
Urea (mg/dl)	[10,40]
Colesterol (mg/dl)	[150,200]
Hematocrito (%)	[36,46]
Eosinófilos (%)	<4

5.2. Estudio de una variable cualitativa

El mejor ejemplo para entender cómo se trata estadísticamente una variable cualitativa podría ser el estudio de una serie de lanzamientos de una moneda. Efectivamente, el resultado del lanzamiento es una variable cualitativa con dos categorías: cara y cruz. No obstante, dado que el lector debe estar bastante harto a estas alturas de los lanzamientos de monedas y demás objetos consideraremos mejor el estudio de una cualidad que puede presentarse o no en los individuos de una población, como puede ser la presencia de una enfermedad E.

Sabemos que en ese caso, la proporción p de individuos enfermos equivale a la probabilidad de que un individuo seleccionado aleatoriamente (mediante sorteo) la padezca. La manera de estimarla será pues extrayendo una muestra aleatoria de tamaño n y calculando su frecuencia relativa $\hat{p}$, es decir, la proporción de individuos de la muestra que padecen la enfermedad E. Si además estamos interesados en construir intervalos de confianza para p o en resolver, como en el ejemplo 9, contrastes de hipótesis del tipo

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p > p_0 \end{cases} \qquad \begin{cases} H_0: & p = p_0 \\ H_1: & p < p_0 \end{cases} \qquad \begin{cases} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{cases}$$

podemos hacer uso de las técnicas de inferencias para la media estudiadas anteriormente. Efectivamente, basta percatarse de que el estudio de una cualidad, como estar o no enfermo, se corresponde con una variable cuantitativa X sobre la población que asigna valor 1 cuando el individuo está enfermo y 0 cuando no lo está. De esta forma, la media μ de dicha variable coincide con la proporción de enfermos en la población. Igualmente, la media aritmética $\overline{x}$ de la muestra coincide con la frecuencia relativa de la misma $\hat{p}$. Puede demostrarse que, en ese caso, la desviación típica muestral es igual a $s = \sqrt{\hat{p}(1-\hat{p})}$, aunque eso no importe a la hora de realizar las inferencias mediante un programa estadístico. Tan sólo debemos construir la variable 0-1 y proceder como si de una media se tratase, teniendo en cuenta lo aprendido en la sección 3.2.6. En todo caso, tener en cuenta que se precisa un tamaño de muestra grande, en especial cuando la cualidad estudiada sea demasiado rara o demasiado frecuente, pues ello comportaría un fuerte sesgo en la variable X. En ese sentido solemos exigir que tanto $n\hat{p}$ como $n(1-\hat{p})$ sean superiores a 5.

Para hacernos una idea, podemos decir que, si deseamos contrastar si una moneda es simétrica a partir de una serie de 100 lanzamientos de la misma, una cantidad de caras menor que 40 o mayor que 60 supondrá un resultado significativo según el test que proponemos. Es decir, que 61 caras tras 100 lanzamientos de la moneda es una prueba significativa contra la hipótesis de simetría.

También podemos plantearnos el problema de determinación del tamaño de muestra necesario para alcanzar cierto grado de precisión en la estimación. Se procede de igual forma, resolviendo la inecuación

$$E_{\text{máx}} \leq \mathsf{z}_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

El valor de $\hat{p}$ puede ser aproximado previamente mediante una muestra piloto. También podemos tener en cuenta el hecho fácilmente demostrable de que, $\hat{p}(1-\hat{p}) \leq 1/4$ y resolver la inecuación

$$E_{\text{máx}} \leq \mathsf{z}_{\alpha/2}\sqrt{\frac{1}{4n}}$$

aunque el valor de n obtenido en ese caso peque por exceso.

Ejemplo 10:[Proporción de alérgicos al polen de gramíneas]

Se desea saber qué tamaño mínimo debe tener una muestra seleccionada aleatoriamente en la población española para poder estimar con un margen máximo de error del 1 % proporción de alérgicos, para una confianza del 95 %.

En principio, ya que no disponemos de un muestra piloto para una primera estimación de la proporción p de alérgicos, consideraremos la última inecuación:

$$0{,}01 \leq 1{,}96\sqrt{\frac{1}{4n}}$$

cuya solución es $n \geq 9604$. Es decir, necesitamos realizar una prueba de reacción cutánea a al menos 9.604 individuos escogidos de manera arbitraria. No obstante, este procedimiento es muy conservador, en especial cuando la cualidad estudida es poco frecuente (que no es nuestro caso). Lo mejor sería tomar una muestra aleatoria piloto, por ejemplo de 100 individuos. Supongamos que, de estos 100, 18 dan positivo en la prueba cutánea. Entonces, podremos disponer de la primera inecuación

$$0{,}01 \leq 1{,}96\sqrt{\frac{0{,}015 \cdot 0{,}85}{n}}$$

cuya solución es $n \geq 4337$. El tamaño sigue siendo bastante elevado. Si no estamos en condiciones de afrontar un diseño de este tipo, tendremos que rebajar nuestras pretensiones en lo referente a la precisión.

5.3. Estudio de relación de dos variables cuantitativas

Esta sección es, en buena parte, continuación de la sección 2.1. Supongamos que sobre una determinada población se miden dos variables cuantitativas X e Y, de medias μ_1 y μ_2 y varianzas σ_1^2 y σ_2^2. Por analogía a lo visto para una muestra en Estadística Descriptiva, podemos definir la covarianza y coeficiente de correlación lineal probabilísticos, es decir, de la población, los cuales se denotan por las correspondientes letras griegas σ_{XY} y ρ_{XY}, que se interpretan de manera análoga. Lo mismo podemos decir del coeficiente de determinación poblacional ρ_{XY}^2. En este contexto, podemos plantearnos dos tipos de problemas.

5.3.1. Comparación de medias con muestras apareadas

Supongamos que queremos comparar las medias μ_1 y μ_2 de las dos variables. Nos referimos a contrastes del tipo

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases} \qquad \begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 < \mu_2 \end{cases} \qquad \begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases}$$

Esto puede ser interesante cuando las dos variables expresen una evolución temporal o el resultado de una medición en circunstancias diferentes, siempre efectuadas sobre los mismos individuos o, en su defecto, en individuos prácticamente idénticos (de ahí el nombre de muestras apareadas).

Ejemplo 11:[Comparación de medias a partir de muestras apareadas]

Se pretende determinar si el hecho de dejar la bebida comporta una reducción en la presión sistólica media de las personas alcohólicas. Para ello se midieron las presiones sistólicas $X =$`antes` de dejar la bebida y $Y =$`después` de dos meses de abstención, con los siguientes resultados

Antes	140	165	160	162	175	190	170	175	155	160
Después	145	150	150	160	170	175	160	165	145	170

Se trata pues de resolver el contraste

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

Para resolver el problema basta considerar la variable diferencia $D = X - Y$, de media $\mu_D = \mu_1 - \mu_2$ y contrastar, en este caso, las hipótesis

$$\begin{cases} H_0: & \mu_D = 0 \\ H_1: & \mu_D > 0 \end{cases}$$

según sabemos hacerlo, pues estamos en condiciones idénticas a las de la sección primera. El programa estadístico lo resuelve automáticamente. El test se basa por lo tanto en la media aritmética y desviación típica de las diferencias. Para que el test sea válido se precisa que la muestra sea grande o bien que a distribución de la diferencia sea aproximadamente Normal, cosa que supondremos en esta ocasión. A tal suposición se debería llegar por conocimientos previos de las variables o bien mediante un test normalidad, aunque teniendo en cuenta que, con tan sólo 10 datos el test de normalidad tendrá escasa capacidad para detectar una posible anormalidad. Por eso debemos tener presente en todo momento la alternativa no paramétrica que en esta caso se denomina Test de los rangos con signos de Wilcoxon.

También puede construirse un intervalo de confianza para la diferencia de las medias $\mu_1 - \mu_2$, de manera que la aceptación de la hipótesis inicial $\mu_1 = \mu_2$ mediante el test con el nivel de significación habitual del 5 % equivale al hecho de que el intervalo a nivel de confianza del 95 % contenga al 0. Veamos el resultado según el programa estadístico. En primer lugar se obtienen las sigientes medias aritméticas para cada variable:

$$\overline{x} = 165{,}2, \quad \overline{y} = 150{,}0$$

De esta forma, la media aritmética de la variable $D = X - Y$ es

$$\overline{d} = 6{,}2$$

Tras aplicar el test estudiado en el apartado 5.1.1 con $\mu_0 = 0$ se obtiene el valor experimental $t_{exp} = 2{,}36$ que, al compararlo con los diferentes cuantiles de la distribución t-Student con 9 grados de libertad, proporciona el P-valor

$$P = 0{,}021$$

Elresultado es pues significativo, es decir, se observa un descenso significativo en el nivel medio de presión sistólica tras dos meses sin ingerir alcohol. También podemos construir un intervalo de confianza al 95 % para la media de D, es decir, para la diferencia de medias $\mu_1 - \mu_2$.

$$IC95\,\% = (0.26, 12.15)$$

Nótese que el descenso medio observado en la muestra es de 6.2 puntos, aunque al 95 % de confianza se le asigna un intervalo que va de 0.26 a 12.15. En todo caso, el 0 no es un valor posible, es decir, que no puede asumirse que las medias permanezcan constantes. El resultado del test de Wilcoxon es, por cierto, $P = 0.024$. La conclusión sería pues la misma.

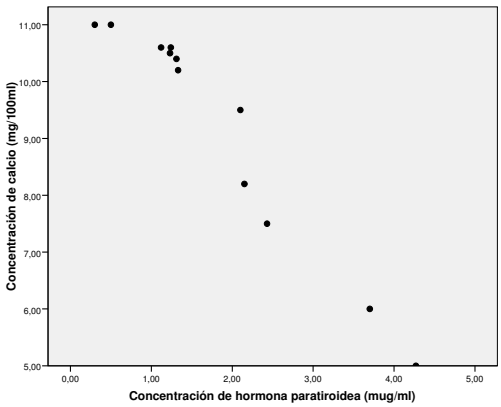
5.3.2. Problema de regresión-correlación

A continuación abordamos desde un punto de vista poblacional el estudio de relación entre dos variables cuantitativas, es decir, pretendemos determinar en qué medida una de ellas puede explicar la variabilidad de la otra y de qué forma. Desde el punto de vista técnico, los métodos de inferencia que utilizaremos exigen diversos supuestos estadísticos. Para simplificarlo al máximo nos centraremos en el más trascendental, aquél cuya violación puede conducir a conclusiones drásticamente erróneas: que la relación entre las variables, si es que se da, sea de tipo lineal. Esto puede observarse directamente a través de un diagrama de dispersión para una muestra. De no ser así habrá que buscar trasformaciones de las variables que sí cumplan ese requisito, como vimos en el capítulo 2; en otras ocasiones la violación de la linealidad se debe a la necesidad de incluir en el modelo nuevas variables cuantitativas o cualitativas, pero eso corresponde a la última sección de este capítulo. También contamos con un método no paramétrico alternativo basado en el denominado coeficiente de correlación de Spearman.

Test de correlación

Ejemplo 12:[Test de correlación]

Se estudia la relación entre $X =$`concentració de hormona paratiroidea` (μg/ml) e $Y =$`concentración de calcio en sangre` (mg/100ml), a partir de una muestra de tamaño $n = 12$ que aportó el siguiente diagrama de dispersión:



Parece claro que la relación en este caso es, de existir, de tipo lineal. Supongamos que dicha muestra es aleatoria y que queremos determinar si, efectivamente, esta relación aparentemente lineal que se aprecia en la muestra puede extrapolarse al resto de la población. Que exista relación lineal a nivel poblacional equivale a que la variable X tenga capacidad para explicar linealmente alguna parte de la variabilidad de Y. Por analogía a lo que vimos en el capítulo 2 desde el punto de vista muestral, estaríamos afirmando que el coeficiente de determinación poblacional ρ^2 es mayor que 0. El caso contrario equivaldría en estas condiciones a la independencia entre ambas variables, pues el valor de X no tendría capacidad alguna para explicar Y (estamos suponiendo, insistimos, que sólo caben explicaciones de tipo lineal). Así pues, contrastar si existe correlación lienal a nivel poblacional equivale a contrastar las hipótesis siguientes:

$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

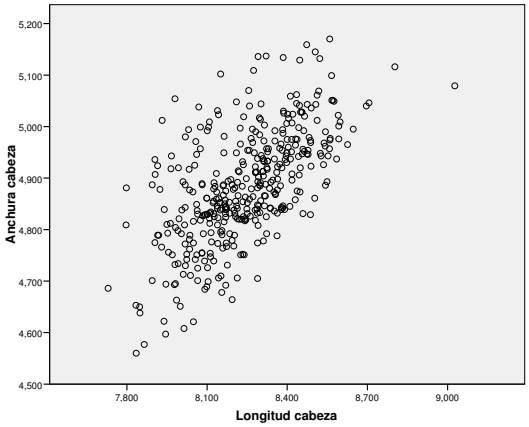
El test que resuleve este contraste, denominado test de correlación lineal, se basa en el valor del coeficiente de correlación muestral r, que podemos considerar como una estimación de ρ. También tiene en cuenta el tamaño de la muestra. Concretamente, consiste en comparar el valores experimental siguiente:

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

que se valorará tomando como referencia la distribución t-Student con $n-2$ grados de libertad, es decir, se rechazará la hipótesis inicial de independencia cuando $t_{exp} > t_{0,05/2}$. El valor teórico puede buscarse en la tabla $N(0,1)$ si $n > 30$. En definitiva, decidimos que hay correlación lineal a nivel poblacional cuando en la muestra se observa una correlación lineal clara en relación con su tamaño n. Hay que tener en cuenta que el valor de r no será nunca 0 en la práctica, es decir, que incluso en el hipotético caso de que las variables fueran independientes, siempre se observaría en la muestra algún grado de correlación, pero sería atribuible al azar inherente al muestreo. El test viene a cuantificar el margen que estamos dispuestos a atribuir al azar. Pero ese margen depende, como vemos, del tamaño de la muestra n. Efectivamente, si la muestra es pequeña el margen debe ser amplio, mientras que para muestras grandes el margen es estrecho pues r debe estar próximo a ρ, de ahí que incluso un valor de r relativamente bajo puede ser significativo para muestras grandes.

En nuestro ejemplo contamos con una muestra muy pequeña ($n = 12$). Sin embargo, se observa una correlación tan fuerte en la misma ($r = -0{,}973$) que el resultado del test no ofrece lugar a dudas ($P < 0{,}001$). Es decir, tenemos una prueba altamente significativa de que la concentración de hormona paratiroidea se relaciona inversamente con la de calcio y la afirmación hace referencia a la población en general.

Otro ejemplo puede ser el estudio de relación entre la longitud y la anchura en las cabezas de espermatozoides de una población animal a partir de una muestra de 391 datos, cuyo diagrama de dispersión aparece a continuación. En este caso el valor del coeficiente de correlación muestral es $r = 0{,}625$. Tras aplicar el test de correlación vuelve a obtenerse un resultado significativo, incluso más significativo que la vez anterior aunque el grado de correlación observado en la muestra sea más débil. Pero es que estamos si cabe más seguros que antes de que la tendencia lineal que se aprecia en el diagrama no se puede explicar exclusivamente por el azar. En definitiva, queda claro que las cabezas de los espermatozoides guardan ciertas proporciones largo-ancho, aunque no universales, pues eso implicaría una correlación lineal perfecta. :



Recta de regresión poblacional

Si existe relación lineal a nivel poblacional entre las dos variables puede ser interesante calcular o, mejor dicho, estimar los parámetros de la recta $y = \alpha + \beta x$ que mejor explica la variabilidad de Y a partir de la de X. Nótese además que la independencia bajo nuestros supuestos equivale a $\beta = 0$, pues una recta constante significa una nula capacidad de predicción port parte de X. De hecho, el test que contrasta dicha hipótesis es el propio test de correlación.

El caso es que los parámetros muestrales a y b definidos en la sección 2.1 son estimaciones puntuales de α y β, respectivamente. En el caso de la hormona paratiroidea la recta de regresión quedaría estimada mediante $y = 12{,}2 - 1{,}6x$.

Predicciones

Dado un valor concreto x_0 para la variable X, el valor $\hat{y} = a + bx_0$ sirve para estimar tanto el valor que le correspondería a x_0 según la recta poblacional como el valor real de Y que correspondería a un individuo con $X = x_0$. Se trata pues de una predicción. Podemos incluso construir un intervalo de confianza a nivel $1 - \alpha$ para dicha predicción:

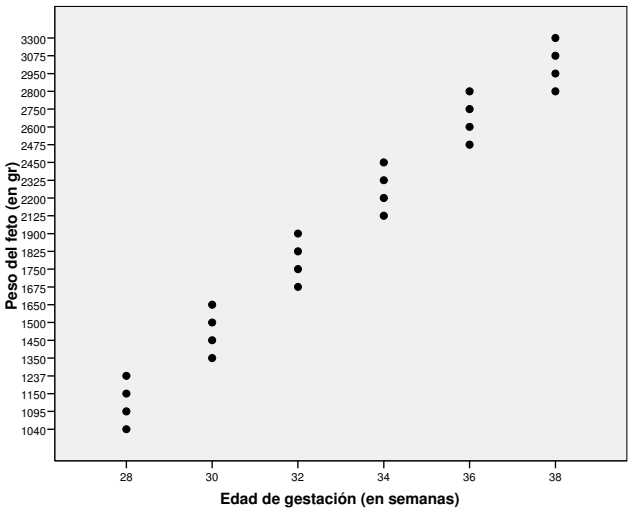
$$\hat{y} \pm t_{\alpha/2(n-2)}\sqrt{s_y^2(1-r^2)\left[1 + \frac{1}{n} + \frac{1}{n}\frac{(\overline{x} - x_0)^2}{s_x^2}\right]}$$

Lo importante es percatarse de que la magnitud del error máximo que podemos atribuir a nuestra predicción y, por lo tanto, la fiabilidad del mismo, depende , además del nivel de significación escogido, de tres factores:

- El término $s_y^2(1 - r^2)$ que expresa la parte de variabilidad de Y explicada linealmente por X, es decir, lo "buena"que es la regresión. Cuanto mejor es la regresión, más fiable es la predicción.
- El tamaño de muestra n. Cuanto mayor es la muestra, mejor es la predicción.
- El término $(\overline{x} - x_0)^2/s_x^2$, que expresa la distancia relativa del punto donde se efectúa a predicción del centro aritmético de los datos de X. A medida que nos alejamos del centro perdemos fiabilidad, cosa lógica pues la regresión se resuelve localmente.

¿Regresión o Correlación?

Para terminar con esta sección, distinguiremos dos tipos de diseños que pueden utilizarse en problemas de este tipo: el que hemos estudiado considera dos variables aleatorias medidas sobre una muestra de datos. Este diseño se denomina de correlación. El otro, denominado de regresión pura, consiste en controlar de antemano los valores que correspondientes a la variable X, que actúa únicamente como explicativa. Para cada valor fijo de X considerado se toma una muestra de individuos que lo verifiquen y se les mide el valor de Y. Es lo que sucede en el estudio del peso de fetos con edad de gestación entre 28 y 38 semanas:



Observamos claramente cómo la variable edad de gestación, que puede conocerse perfectamente, está controlada, no así el peso del feto, que es el que pretendemos explicar. En diseños de este tipo no tiene sentido hablar del coeficientes de correlación poblacional ρ dado que sólo tenemos una variable aleatoria. En particular, no tiene sentido hablar de independencia. Por contra, sí tiene sentido contrastar la hipótesis inicial $\beta = 0$, que viene a significar lo mismo. Salvo estos matices teóricos, los métodos de inferencia que se utilizan en ambos diseños resultan ser idénticos, por lo que hablaremos simplemente de problema de regresión-correlación, que abarca y trata ambos casos indistintamente. La única diferencia en términos prácticos radica en que en un problema de regresión pura los papeles de las variables X e Y no pueden permutarse. Es el precio de controlar de la variable explicativa X, cosa que también puede reportar interesantes beneficios.

5.4. Estudio de relación entre dos variables cualitativas

Esta sección puede considerarse como una continuación, desde un punto de vista poblacional o probabilístico, de la sección 2.2. Podemos distinguir diversos apartados, todos ellos de gran interés en las Ciencias de la Salud.

5.4.1. Test χ^2

Recordamos que en la sección 2.2 se estudiaba la relación a nivel muestral entre dos variables cualitativas o categóricas. Los datos relativos a la muestra se organizan mediante lo que denominamos tabla de contingencia. Mostramos a continuación la que corresponde a los datos del ejemplo 6:

`Nivel cloroplastos`

`Nivel` 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

Conocemos un parámetro muestral, el coeficiente de Contingencia C de Pearson, que pretende cuantificar el grado de correlación que se aprecia en la muestra. En nuestro caso debe estar comprendido entre 0 y 0.816, pues se trata de una tabla 3×3. Concretamente, $C = 0.444$, lo cual indica un nivel medio de correlación en la muestra. Otra medida del grado de correlación es la denominada distancia χ^2_{exp}. Ambas medidas se relacionan mediante

$$C = \sqrt{\frac{\chi^2_{exp}}{\chi^2_{exp} + n}} \qquad \chi^2_{exp} = n\frac{C^2}{1 - C^2}$$

En este nuevo contexto y suponiendo que la muestra fuera aleatoria, nos interesa saber si esa asociación o correlación que se aprecia en la misma puede extrapolarse al total de la población, es decir, ¿podemos afirmar que la contaminación influye en la salud de los árboles? Estamos pues contrastando, al igual que en problema de correlación-regresión, la hipótesis inicial de independencia contra la hipótesis alternativa de asociación.

$$\begin{cases} H_0 : \texttt{Independencia} \\ H_1 : \texttt{Asociación} \end{cases}$$

El test que resuelve el contraste, denominado Test χ^2, tiene como valor experimental la propia distancia

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

que se comparará con el cuantil $\chi^2_{0,05}$ de la distribución χ^2 con $(r-1)(s-1)$ grados de libertad. En definitiva, el resultado depende del grado de asociación observado en la muestra, que se cuantifica por C y del tamaño de la misma. Un valor pequeño de C propicia un valor pequeño de χ^2_{exp}, pero dependiendo también de n. Al igual que sucediera en el problema de correlación-regresión, aunque C sea distinto de 0 hemos de dar a la hipótesis inicial de independencia un margen achacable al azar. Este margen se cuantifica mediante la tabla de la distribución χ^2. Nótese pues que el coeficiente de contingencia de Pearson C desempeña en la decisión un papel muy similar al del coeficiente de correlación r.

El test que hemos visto no es sino una aplicación particular de un test más general, el test χ^2 propiamente dicho, que contrasta si las frecuencias observadas en distintas categorías se diferencian

claramente de las que cabría esperar según un determinado modelo probabilístico. En nuestro caso estamos comparando las observaciones O_{ij} con los valores que cabría observar en caso de independencia, E_{ij}.

En el ejemplo anterior se obtiene $\chi^2_{exp} = 14.7$ que, al confrontarlo con la distribución χ^2 con 4 grados de libertad, da lugar a un resultado muy significativo $P = 0{,}005$. Es decir, que la muestra estudiada constituye una prueba muy significativa de que la contaminación en SO_2 y la salud en las hojas de los árboles se relacionan.

En el caso de una tabla 2×2 podemos obtener χ^2_{exp} a partir del coeficiente ϕ mediante

$$\chi^2_{exp} = n\phi^2$$

En este caso, debe confrontarse con la tabla de la distribución χ^2 con un grado de libertad. De esta forma, en el ejemplo 7, cuya tabla de contingencia mostramos a continuación, se obtiene un resultado altamente significativo, es decir, que queda claro que existe relación entre la vacunación y la incidencia de la hepatitis.

		`Vacunación`		
	(2 × 2)	No	Sí	Total
`Hepatitis`	Sí	70	11	81
	No	464	538	1002
	Total	534	549	1083

Para poder aplicar el test χ^2 se precisa una cierta cantidad de datos por casilla. En el caso de las tablas 2×2, si alguna de las celdas presenta pocos datos conviene aplicar como alternativa el **test exacto de Fisher**, disponible en cualquier programa estadístico.

5.4.2. Comparación de dos proporciones

Estudiamos a continuación un método alternativo al test χ^2 para determinar si existe relación entre dos variables cualitativas dicotómicas. Se trata pues de otra forma de tratar una tabla 2×2 como la anterior.

La variable dicotómica `vacunación` divide la población objeto del estudio en dos subpoblaciones: la de los individuos vacunados y la de los no vacunados, lo cual nos permite considerar las proporciones

$$p_1 = P\big(\texttt{hepatitis} \mid \texttt{no vacunados}\big) \qquad p_2 = P\big(\texttt{hepatitis} \mid \texttt{vacunados}\big)$$

Es decir, la proporción de individuos vacunados que ha contraído la hepatitis y la proporción de individuos no vacunados que no la han contraído. Al contrario que en la sección 2.2 donde se utilizó la notación $\hat{P}$ para hacer referencia a las proporciones, simples o condicionales, aquí las mismas se denotan mediante P para dejar patente que estamos hablando de la proporción (o probabilidad) calculada a partir de la población completa, no de una muestra de tamaño n. No obstante, las proporciones muestrales

$$\hat{p}_1 = \hat{P}\big(\texttt{hepatitis} \mid \texttt{no vacunados}\big) \qquad \hat{p}_2 = \hat{P}\big(\texttt{hepatitis} \mid \texttt{vacunados}\big)$$

pueden considerarse estimaciones de las anteriores, supuesto que los individuos hayan sido seleccionados aleatoriamente, al menos en lo que respecta a la variable `incidencia de la hepatitis`. De esta forma tenemos

$$\hat{p}_1 = \frac{70}{534}, \qquad \hat{p}_2 = \frac{11}{549}$$

Es decir, que en la muestra de no vacunados llegan a contraer la hepatitis un 13.1 %, mientras que en la de vacunados la contrae un 2.0 %. A nivel muestral se observa pues una relación entre el hecho de estar vacunado y la incidencia de la hepatitis. Nos preguntamos si estamos en condiciones de generalizar esta conclusión al global de la población, es decir, si la vacunación disminuye la probabilidad de contraer hepatitis. Se tata pues de contrastar las hipótesis

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 > p_2 \end{cases}$$

Es perfectamente discutible si el contraste debe ser bilateral o unilateral y, en en el último caso, cuál debería ser la hipótesis inicial. Lo que nos ocupa en este momento es el test que lo resuelve, que es un caso particular del test de Student o del de Welch, que se estudia en la sección 5.4 aplicado a una variable 0-1. En este caso el resultado es, al igual que sucediera con el test χ^2, significativo, es decir, que la muestra estudiada aporta una clara evidencia de que la vacunación influye en la incidencia de la hepatitis. Que los resultados de ambos tests sean similares es lo esperable. También puede aplicarse el método de Student para construir un intervalo de confianza para $p_1 - p_2$

Los dos apartados que siguen se encuadran también el estudio de las tablas 2×2.

5.4.3. Factores de riesgo

Nos centramos en esta ocasión en una situación particular, 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 o la propia hepatitis, mientras que la otra es la ausencia o presencia de un posible factor de riesgo `FR`, como el hecho de fumar o el de no estar vacunado contra la hepatitis. En ese caso pueden estudiarse diferentes parámetros de interés. En primer lugar, podemos definir la **prevalencia** como la proporción de individuos enfermos $P(\texttt{E})$ en un instante dado en la población. En segundo lugar, podemos definir la **incidencia** de la enfermedad como la proporción de individuos que enferman a lo largo de un periodo de tiempo. Se pueden distinguir distintas incidencias, por ejemplo, la incidencia entre los individuos con factor de riesgo o la incidencia entre los que no lo presentan. La posibilidad de estimar estos parámetros depende del diseño escogido a la hora de seleccionar los individuos de la muestra. De esta forma, distinguiremos tres tipos de diseños:

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

(b) **Estudios de seguimiento o de cohortes:** se selecciona una muestra de individuos expuesta al factor de riesgo y otra no expuesta para estudiar su evolución a lo largo de un

periodo de tiempo que suele ser largo, anotándose cuáles llegan a contraer la enfermedad en cada caso. Este diseño permite estimar las incidencias de la enfermedad para ambos grupos, $P(\texttt{E}|\texttt{FR})$ y $P(\texttt{E}|\overline{\texttt{FR}})$, para compararlas de diversas formas.

(c) **Estudios retrospectivos o de caso-control:** en un determinado momento se escoge una muestra de enfermos (caso) y otra de sanos (control), para a continuación averiguar qué individuos han estado expuestos al factor de riesgo. Suelen ser los menos costosos: los de prevalencia requieren muestras más grandes para que puedan registrarse suficientes enfermos; los de cohortes requieren de un seguimiento a lo lo largo del tiempo. En contrapartida, los estudios caso-control no permitirán estimar prevalencias, incidencias ni medidas relacionadas. Por contra, si podemos estimar las proporciones del tipo $P(\texttt{FR}|\texttt{E})$, $P(\overline{\texttt{FR}}|\texttt{E})$, etc, que será de utilidad para estimar el `Odds ratio`.

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

(2×2)	Sí factor	No factor	Total
Sí enfermo	a	b	a+b
No enfermo	c	d	c+d
Total	a+c	b+d	n

En el ejemplo 7, la enfermedad estudiada es la hepatitis y el posible factor de riesgo el hecho de no estar vacunado. Se supone que estamos ante un estudio de cohortes.

Como hemos dicho anteriormente, en un estudio de cohortes pueden estimarse las incidencias de la enfermedad por grupos a través de la tabla. Concretamente:

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

y se entenderán respectivamente como el riesgo de contraer la enfermedad si se está expuesto al factor y en caso contrario. En un estudio caso-control podemos estimar, por ejemplo,

$$\hat{P}(\texttt{FR}|\texttt{E}) = \frac{a}{a+b}$$

que se interpreta como la probabilidad de que un enfermo presente el factor de riesgo. A partir de este tipo de parámetros calculamos otros de mayor interés práctico:

Riesgo atribuible

Se define el riesgo atribuible al factor, también denominado diferencia de incidencia, como la diferencia entre las incidencias o proporciones poblacionales de enfermos, es decir,

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

Obviamente, no aspiramos a calcular el verdadero RA sino a estimarlo mediante una muestra lo mayor posible, es decir, mediante nuestra tabla. Este parámetro puede estimarse en estudios de cohortes. Según vimos en el apartado anterior estamos en condiciones de construir un intervalo

de confianza para el verdadero RA o contrastar si es positivo, es decir, podemos determinar si el posible factor de riesgo lo es realmente.

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

$$\hat{RA} = 13{,}1\,\% - 2{,}0\,\% = 11{,}1\,\%$$

y sabemos ya que es significativo, es decir, que el hecho de no estar vacunado constituye realmente un factor de riesgo. El número se interpreta de la siguiente forma: el porcentaje de enfermos entre los no vacunados es 11.1 puntos superior al de lo vacunados.

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 al factor propiamente, entendiendo que una parte de los que están expuesto 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

Es seguramente la más intuitiva de todas las medidas de riesgo. Se trata de determinar en qué medida incrementa el factor de riesgo la incidencia de la enfermedad, es decir. Para ello se define el riesgo relativo mediante

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

que, suponiendo que el estudio sea de cohortes, se estima directamente a partir de la tabla mediante

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

Para los datos de la hepatitis tendríamos la siguiente estimación

$$\hat{RR} = \frac{13{,}1}{2{,}0} = 6{,}55$$

Es decir, en esta muestra se observa que el hecho de no estar vacunado aumenta 6.55 veces la proporción de enfermos. Este número se considera una estimación del riesgo relativo poblacional RR. Es obvio que el hecho de que el factor considerado no guarde relación con la enfermedad equivale a $RR = 1$. Estamos en condiciones de contrastar la hipótesis inicial $H_0 : RR = 1$ frente la aternativa $H_1 : RR \neq 1$ comparando con $\chi^2_{0,05}(1)$ el valor experimental

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

donde

$$s^2_{log\hat{RR}} = \frac{c}{a(a+c)} + \frac{d}{b(b+d)}$$

En nuestro caso,

$$s^2_{log\hat{RR}} = 0{,}101, \quad \chi^2_{exp} = 34{,}97, \quad P < 0{,}001$$

En definitiva, es claro que el factor supone en general un riesgo de cara a padecer la enfermedad.

Odds Ratio

Necesitamos una medida de riesgo que pueda estimarse a partir de un estudio caso-control. La que propondremos a continuación es válida tanto para los estudios de cohortes como para los de caso-control. Se define primeramente el `Odd` asociado al factor riesgo mediante

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

Igualmente, podemos definir el `Odd` asociado a la ausencia del factor mediante

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

Entonces, definimos su cociente, denominado `Odds ratio` mediante

$$OR = \frac{O_{FR}}{O_{\overline{FR}}}$$

Obviamente, la no influencia del factor se correspondería $OR = 1$. Cuanto mayor sea OR más claro será el riesgo que comporta el factor. En principio, esta medida puede ser estimada a partir de la tabla únicamente en un estudio de cohortes de la forma

$$\hat{OR} = \frac{a}{c} : \frac{b}{d}$$

Nótese que esa expresión equivale al siguiente cociente de productos cruzados:

$$\hat{OR} = \frac{ad}{bc} \tag{5.1}$$

el cual puede expresarse también mediante

$$\hat{OR} = \frac{a}{b} : \frac{c}{d}$$

Este último término puede considerarse una estimación perfectamente válida en un estudio de caso-control del parámetro

$$\frac{P(\mathtt{FR}|\mathtt{E})}{P(\mathtt{FR}|\overline{\mathtt{E}})} : \frac{P(\overline{\mathtt{FR}}|\mathtt{E})}{P(\overline{\mathtt{FR}}|\overline{\mathtt{E}})}$$

que, por el mismo razonamiento[1], equivale a OR. En definitiva, OR es una medida de riesgo que puede estimarse tanto en los estudios de cohortes como en los de casos-control mediante la expresión (5.1). Por ejemplo, en el caso de la hepatitis, que se un estudio de cohortes, obtendríamos

$$\hat{OR} = \frac{70 \cdot 518}{11 \cdot 464} = 7{,}10$$

Si el estudio se diseña como caso-control, si la muestra no es muy numerosa, es frecuente obtener un $\hat{OR}$ mayor que el $\hat{RR}$ que se habría obtenido de haber optado por un diseño de cohortes. Por otra parte y al igual que ocurre con RR, que el factor no guarde relación con la enfermedad implica $OR = 1$. Esta hipótesis inicial puede contrastarse comparando con $\chi^2_{0,05}(1)$ el valor experimental

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

donde

$$s^2_{log\hat{OR}} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

En nuestro caso,

$$s^2_{log\hat{RR}} = 0{,}109, \quad \chi^2_{exp} = 35{,}24, \quad P < 0{,}001$$

En definitiva, es claro que el factor supone en general un riesgo de cara a padecer la enfermedad.

5.4.4. Diagnóstico Clínico II: sensibilidad y especificidad

Otra cuestión de gran interés en epidemiología y relacionada 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 la sección 5.1 vimos cómo construir un test de diagnóstico partiendo de una variable cuantitativa X, de manera que si el resultado para un individuo excede del límite de normalidad que se determina para la mayoría de los individuos sanos se diagnostica como enfermo. El propio planteamiento revela inmediatamente la presencia de probabilidades de error, tanto para individuos sanos como enfermos.

Efectivamente, cae dentro de los posible, aunque sea poco probable, que un individuo sano presente un valor extremo para dicha variable, por lo que sería diagnosticado erróneamente como enfermo. Sería un `falso positivo`. También es posible que un individuo enfermo presente un valor normal para la variable, por lo que sería diagnosticado erróneamente como sano. Sería un falso negativo. Para calcular las probabilidades de cometer ambos tipos de errores deberíamos conocer las distribuciones exactas de la variable en ambas poblaciones, si cupiera hablar de ellas. Otro método más realista, que es el que consideraremos nosotros, pasa por la estimación a partir de una muestra de gran tamaño. En todo caso, sea cual sea el procedimiento utilizado para el diagnóstico, nuestra primera intención es estimar la sensibilidad del test, es decir, las probabilidad (proporción) de que un enfermo $\mathtt{E}$ resulte positivo, y la especificidad o probabilidad de que un sano $\overline{\mathtt{E}}$ dé negativo.

$$\mathtt{Sensibilidad} = \hat{P}(+|\mathtt{E}) \qquad \mathtt{Especificidad} = \hat{P}(-|\overline{\mathtt{E}})$$

[1] Regla de Bayes

Ejemplo 12: [Sensibilidad y especificidad de un test]

Se aplica un test diagnóstico a 1000 individuos, 200 de los cuales sabemos que están enfermos mientras que de los 900 restantes sabemos que están sanos. Los resultados son los siguientes:

`Resultado del test`

`Enfermedad`	(2 × 2)	+	-	Total
	Sí	120	80	200
	No	90	710	800
	Total	210	790	1000

A partir de los datos de la muestra obtenemos las siguientes estimaciones de la sensibilidad y especificidad:

$$\texttt{Sensibilidad} = \frac{120}{200} = 0{,}600 \qquad \texttt{Especificidad} = \frac{710}{800} = 0{,}887$$

Es decir, la proporción de falsos negativos en la muestra es del 40.0 % y la de falsos positivos del 11.3 %. Los parámetros más interesantes del test son los valores predictivos positivo $VP+$ y negativo $VP-$. EL primero es la probabilidad de que un positivo esté realmente enfermo y el segundo, la probabilidad de que un negativo esté realmente sano. Si se trata de un estudio de prevalencia, es decir, si los 1000 individuos hubieran sido escogidos arbitrariamente sin tener en cuenta la presencia o ausencia de la enfermedad, podríamos estimar estas probabilidades o proporciones poblacionales a partir de las proporciones muestrales de forma obvia. El problema, al igual que sucede en el cálculo de riesgos, radica en que la variable `enfermedad` suele estar controlada para que el estudio sea viable, es decir, que 200/1000 sería una estimación completamente errónea de la prevalencia por lo que la estimación del valor predictivo positivo resultaría muy elevada. Es la patología propia de los estudios de caso-control.

No obstante, la prevalencia de una enfermedad es una tasa epidemiológica que puede haber sido estimada previamente mediante un estudio de tipo transversal. A partir de ese dato y de la sensibilidad y especificidad obtenidas mediante la tabla anterior, podemos hacer de la denominada Regla de Bayes para conseguir las proporciones buscadas. La Regla de Bayes es un método trivial, consecuencia directa de las igualdades (2.1), para reconstruir unas proporciones condicionales a partir de las proporciones condicionales inversas. En general se verifica:

$$\begin{aligned}
\hat{P}(A|B) &= \frac{\hat{P}(A \cap B)}{\hat{P}(B)} = \frac{\hat{P}(B \cap A)}{\hat{P}(B)} \\
&= \frac{\hat{P}(B \cap A)}{\hat{P}(B \cap A) + \hat{P}(B \cap \overline{A})} \\
&= \frac{\hat{P}(B|A)\hat{P}(A)}{\hat{P}(B|A)\hat{P}(A) + \hat{P}(B|\overline{A})\hat{P}(\overline{A})}
\end{aligned}$$

En nuestro caso, al aplicar la Regla se obtiene

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

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

Así, si la enfermedad considerada en el ejemplo presenta una prevalencia del 2 %, tendremos:

$$VP+ = \frac{0{,}60 \times 0{,}02}{0{,}60 \times 0{,}02 + 0{,}113 \times 0{,}98} = 0{,}097$$

$$VP- = \frac{0{,}887 \times 0{,}98}{0{,}40 \times 0{,}02 + 0{,}887 \times 0{,}98} = 0{,}990$$

El test del ejemplo parece ser 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 consultar la bibliografía recomendada, en es especial Cobo, Muñoz, González (2007).

5.5. Relación entre una variable cualitativa y otra cuantitativa

En esta importante sección estudiaremos diversos problemas en los que están involucradas una variable cualitativa y otra cuantitativa. En algunos casos será la variable cualitativa la que pretende explicar a la cuantitativa; en otros sucede al contrario.

5.5.1. El test de Student y otros métodos relacionados

Aunque son varios los tests estadísticos bautizados con este nombre nos referimos en esta ocasión al más popular, que se utiliza a la hora de comparar las medias de dos subpoblaciones. Tenemos por lo tanto una variable explicativa que es cualitativa y dicotómica, pues divide la población en dos partes, y una variable respuesta X cuantitativa. Queremos determinar si ambas están relacionadas.

En nuestro caso solemos utilizar este método para contrastar si una terapia o medicamento funciona. Podemos pensar en un principio que el tratamiento busca una mejora en un carácter cuantificable de la población, como la tensión arterial, el nivel de colesterol, etc. Tenemos pues una variable cualitativa o factor F que distingue los individuos tratados (grupo caso) de los no tratados (grupo control), y una variable cuantitativa X que, posiblemente, guarde relación con el factor. En ese sentido podemos distinguir dos circunstancias extremas: una, que el factor no tenga capacidad alguna de explicar la variabilidad de X, lo cual querrá decir que el tratamiento es absolutamente inútil; en el polo opuesto, que el factor explique completamente la variabilidad de X, lo cual querrá decir que los individuos sin tratamiento tienen todos la misma puntuación y los del tratamiento otra puntuación constante diferente de la anterior. Esa situación no parece muy realista. Nos conformaremos con dilucidar si el factor modifica la media de la distribución de X en cada una de las categorías que determina, es decir, nos planteamos contrastes del tipo

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases} \qquad \begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} \qquad \begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 < \mu_2 \end{cases}$$

donde μ_1 y μ_2 son las medias de la variable X en las dos categorías o subpoblaciones del factor: caso y control. Evidentemente, en este apartado no sólo se enmarca los estudios caso-control sino también cualquier comparación entre dos tratamientos o entre la media de una variable medida en dos poblaciones. En definitiva, en las ocasiones en las que se pretende explicar una variable cuantitativa mediante un factor dicotómico. Los detalles de los tests que proponemos a continuación pueden encontrarse en la bibliografía recomendada.

Test de Student

Partiremos de la información que aporten sendas muestras de tamaños n_1 y n_2. El procedimiento más habitual para resolver el problema es el denominado test de Student para muestras independientes, que consiste en contrastar con la tabla t-Student con n_1+n_2-2 grados e libertad, o con la $N(0,1)$ si $n_1+n_2-2>30$, el valor experimental

$$t_{exp}=\frac{\overline{x}_1-\overline{x}_2}{\sqrt{s_c^2\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}, \tag{5.2}$$

siendo

$$s_c^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

La decisión dependerá pues de la diferencia entre las medias aritméticas de las muestras junto con la magnitud de sus varianzas y los tamaños de las mismas. Podemos construir un intervalo de confianza para la diferencia de medias $\mu_1-\mu_2$ consistente con el test anterior en el sentido de que el test decide H_0 en el contraste bilateral a nivel de significación α si, y sólo si, el valor 0 está comprendido en el intervalo de confianza a nivel $1-\alpha$ para $\mu_1-\mu_2$.

El test de Student puede considerarse en cierto sentido óptimo si la variable X es normal en ambas categorías y con idénticas varianzas. El primer supuesto puede contrastarse mediante sendas pruebas de normalidad aunque debemos tener en cuenta que el test sigue siendo válido aunque no se verifique la normalidad si ambas muestras son suficientemente grandes.

Tests de Snedecor y Levene

El supuesto de igualdad de varianzas puede contrastarse, es decir, podemos contrastar las hipótesis siguientes

$$\begin{cases} H_0: & \sigma_1^2=\sigma_2^2 \\ H_1: & \sigma_1^2\neq\sigma_2^2 \end{cases}$$

mediante el test de Snedecor, que compara el valor experimental $F_{exp}=s_1^2/s_2^2$ con la tabla de la F-Snedecor. Pero este test requiere también del supuesto de normalidad y, al contrario que el de Student, es bastante sensible ante su violación. El test e Levene es una variante del mismo que se muestra más robusto.

Caso de varianzas distintas

Si no puede asumirse la igualdad de las varianzas tenemos varias opciones. Primeramente, una variante del test de Student consistente en estimar las varianzas de cada grupo por separado

$$t_{exp} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n_2}}}$$

Otra opción es el test de Welch, que considera el mismo valor experimental (5.2) que en el test de Student pero corrige el grado de libertad de la distribución t-Student teórica; por último, si las muestras son grandes y de similar tamaño, el test de Student sigue siendo válido aunque no puedan asumirse los supuestos de normalidad ni igualad de varianzas.

Test de Mann-Whitney-Wilcoxon

Aún así, puede suceder que no puedan asumirse esos supuestos y las muestras no sean lo suficientemente grandes como para constrarrestar dicha carencia. En tal caso, tenemos la opción de aplicar el método no paramétrico de la suma de rangos de Wilcoxon, también conocido como de Mann-Whitney. La idea es simplísima: si el tratamiento no influye en la distribución de la variable cuantitativa, al mezclar los datos de los dos grupos los rangos o posiciones de los datos deben repartirse de forma aleatoria, de manera que los rangos medios de ambos grupos sean similares. De no ser así cabrá pensar en que el tratamiento altera la distribución de la variable. La gran ventaja de este método radica en que el nivel de significación es válido independientemente de la distribución concreta de la variable cuantitativa, siempre que sea continua. Sin embargo, tiende a dar significaciones demasiado bajas (menor potencia) en diversas circunstancias.

Comparación de dos proporciones

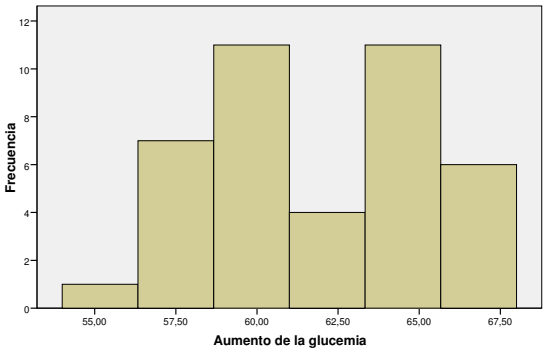
Ya dijimos en la sección anterior que el test de Student puede servir para contrastar la diferencia entre dos proporciones si el carácter cualitativo dicotómico estudiado se asocia a una variable cuantitativa X que toma un 1 si se da la cualidad y un 0 si no se da. De esta forma, la igualdad de proporciones equivale a la igualdad de las medias de esta variable. Así, en el ejemplo de la vacuna contra la hepatitis se obtiene una diferencia significativa entre las dos proporciones de enfermos: la de la muestra vacunada y la de la muestra no vacunada.

Veamos un ejemplo práctico de comparación entre dos grupos mediante los diversos test considerados:

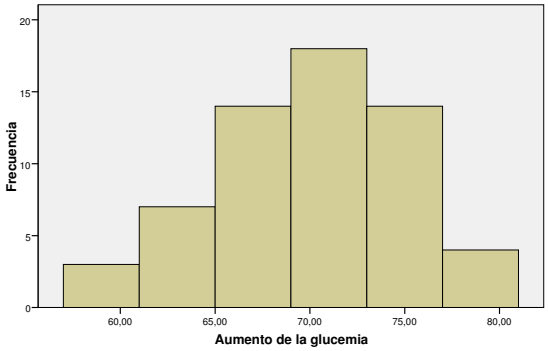
Ejemplo 13:[Comparación de dos grupos]

Se pretende determinar si la edad es un factor a tener en cuenta a la hora de explicar los mecanismos de regulación de la glucemia. Para ello se clasifica a los individuos en jóvenes o mayores en función de determinada edad de corte. Se toman entonces una muestra de 40 jóvenes y otra de 60 mayores, a los cuales se les mide la glucemia en ayunas. A continuación ingieren una bebida muy azucarada y permanecen en reposo una hora, transcurrida la cual se les vuelve a medir la glucemia que, lógicamente, aumenta (este hecho puede contrastarse perfectamente mediante un test de comparación de medias para muestras apareadas). Los resultados de la variable `Aumento de la glucemia` en ambas muestras son los siguientes:

para edad= Jóvenes

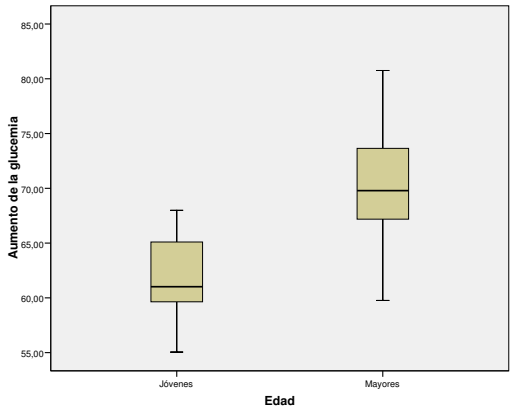


para edad= Mayores



Aumento de la glucemia

Edad			Estadístico
Jóvenes	Media		61,9511
	Intervalo de confianza para la media al 95%	Límite inferior	60,7916
		Límite superior	63,1106
	Media recortada al 5%		61,9614
	Mediana		61,0193
	Varianza		13,144
	Desv. típ.		3,62549
	Mínimo		55,05
	Máximo		67,99
	Rango		12,94
	Amplitud intercuartil		5,59
	Asimetría		,054
	Curtosis		-1,077
Mayores	Media		69,9547
	Intervalo de confianza para la media al 95%	Límite inferior	68,6684
		Límite superior	71,2410
	Media recortada al 5%		69,9810
	Mediana		69,7932
	Varianza		24,794
	Desv. típ.		4,97941
	Mínimo		59,76
	Máximo		80,76
	Rango		21,00
	Amplitud intercuartil		6,55
	Asimetría		-,157
	Curtosis		-,330



En el aspecto puramente inferencial, mostramos los resultados de los test de Shapiro-Wilk (prueba de normalidad), Levene (prueba de igualdad de varianzas), Student (comparación de las medias asumiendo igualdad de varianzas), Welch (comparación de las medias sin asumir igualdad de varianzas) y Mann-Whitney (alternativa no paramétrica):

Test	P-valor
Shapiro-Wilk	P=0.108(J) P=0.593(M)
Levene	P=0.126
Student	P<0.001
Welch	P<0.001
Mann-Whitney	P<0.001

Vamos a extraer conclusiones de los resultados obtenidos. Llaman poderosamente la atención las diferencias entre los diagramas de cajas. No obstante, esta apreciación debe confirmarse mediante un test de hipótesis adecuado. Las condiciones ideales para aplicar el test de Student son las siguientes:

1. Normalidad de la distribución para ambos grupos. Esto se ha contrastado mediante el test de Shapiro-Wilk, obteniéndose en ambos casos resultados no significativos que vienen a apoyar la hipótesis inicial de normalidad, más aún teniendo en cuenta que el test posee potencia suficiente para detectar la hipótesis alternativa dado el tamaño de las muestras.

2. Igualdad de las varianzas. Se ha contrastado mediante el test de Levene con resultado no significativo, lo cual apoya la hipótesis inicial de igualdad de varianzas. El test aplicado es totalmente válido pues las distribuciones son aproximadamente normales.

En ese caso, procede aplicar el test de Student que aporta un resultado altamente significativo, por lo que rechazamos la hipótesis inicial de igualdad de medias. Es decir, que las muestras consideradas constituyen una prueba muy clara de que la media de aumento de la glucemia varía con la edad. Se confirma pues la impresión que nos produce la descriptiva anterior. Podemos obtener también un intervalo de confianza para la diferencia de medias. Concretamente, al 95 % de confianza tenemos el intervalo siguiente

$$\mu_2 - \mu_2 \in (6.2, 9.8)$$

es decir, que los mayores tienen un aumento medio entre 6.2 y 9.8 puntos mayor que el de los jóvenes con una confianza del 95 %.

Si la prueba de igualdad de varianzas hubiera aportado un resultado significativo, habría procedido aplicar el test de Welch, cuyo resultado es el mismo. Por otro lado, aunque alguna de las pruebas de normalidad hubiera resultado significativa, las pruebas paramétricas (Student-Welch) habrían seguido siendo válidas por la amplitud de los tamaños de muestra y los aspectos de los histogramas, que no presentan claras anormalidades. Realmente, con este diseño podemos aplicar en todo caso el test de Student. Si los tamaño de muestra hubieran sido menores y se hubieran apreciado fuertes sesgos y valores extremos, lo detectara o no la prueba de normalidad, habría convenido aplicar la prueba no paramétrica de Mann-Whitney. En todo caso nunca esta de más conocer el resultado de este test. En este caso es el mismo que el de los anteriores porque la situación es clarísima: estamos completamente seguros de que en los mayores aumenta más la glucemia.

Con frecuencia los métodos no paramétricos aportan resultados menos significativos que los análogos paramétricos pues, al contar con menos suposiciones pueden poseen menor potencia o capacidad para detectar la hipótesis alternativa. De ahí que cuando sea ése nuestro propósito intentemos hacer lo posible para aplicar el test de Student, más potente.

5.5.2. Anova de una vía

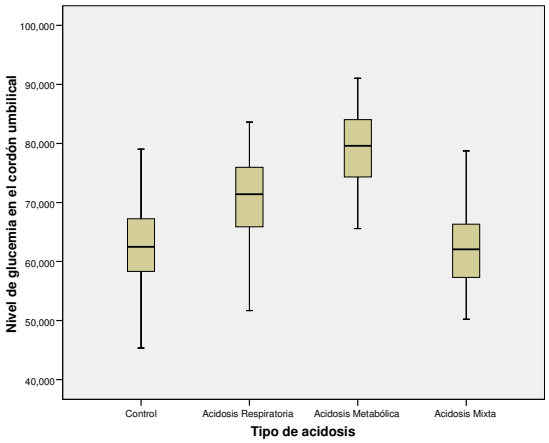
Afrontamos ahora un estudio algo más general que el del apartado anterior. Se trata de explicar una variable cuantitativa X mediante un factor cualitativo con la diferencia de que, en esta ocasión, posee más de dos categorías. Tal puede ser el caso de un estudio para contrastar la eficacia de un tratamiento donde las diferentes categorías pueden corresponderse con distintas dosis o variedades de la terapia a las que se suma un caso control o placebo.

Por lo demás, el problema y los supuestos son idénticos: se trata de seleccionar una muestra por cada categoría y contrastar la igualdad de las medias mediante un test denominado anova de una vía que compara la variabilidad detectada entre las categorías con la variabilidad detectada dentro de las categorías. Si la primera es suficientemente grande en relación con la segunda se rechaza la hipótesis inicial de igualdad de medias. El valor experimental de confrontará con la tabla de cuantiles de la distribución F-Snedecor con ciertos grados de libertad. Los supuestos que se requieren en principio son la normalidad en todas las categorías y la igualdad de las varianzas. De no verificarse estas propiedades puede recurrirse al test de Brown-Forsythe o al método no paramétrico de Kruskal-Wallis. No obstante, al igual que sucede con el test de Student, el anova tiene un excelente comportamiento si las muestras son grandes. Conviene también que los tamaños sean similares o incluso iguales.

Rechazar la hipótesis inicial de igualdad de medias equivale a afirmar que el factor influye en la variable cuantitativa estudiada. Convendría aclarar en qué sentido. Para ello se comparan las categorías por parejas mediante las denominadas comparaciones múltiples. Existen diferentes familias de comparaciones múltiples, las más clásicas son las de Scheffé, Bonferroni y Tuckey. La dos primeras pecan de conservadoras, es decir, aportan resultados menos significativos de lo debido; la tercera requiere que los tamaños de muestra sean iguales.

Ejemplo 15:[Comparación de cuatro grupos]

Se pretende determinar si la acidosis influye en la glucemia medida en el cordon umbilical y en qué sentido, distinguiendo entre acidosis metabólica, respiratoria y mixta. Tenemos por lo tanto una variable cuantitativa, glucemia, y un factor con tres categorías a las que se les une la de los recién nacidos sanos. Se toman en este caso sendas muestras de 50 datos cada una. El diseño es pues inmejorable. Mostramos algunos de los resultados que pueden obtenerse mediante un programa estadístico.



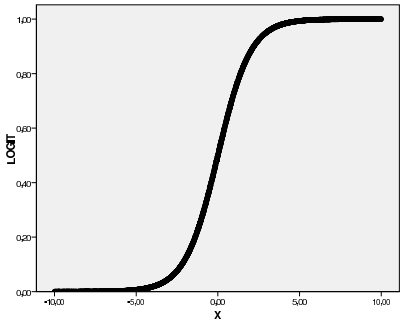
Anova→ $P < 0001$
Tuckey: Control-Mixta→ $P > 0{,}05$. Todas las demás son significativas.
Kruskal-Wallis→ $P < 0001$.

Las conclusiones son claras: tanto el anova como el test no paramétrico coinciden en señalar que el factor influye en la glucemia. Tras aplicar las comparaciones múltiples mediante el método de Tuckey obtenemos que existen diferencias significativas entre las medias de todos los grupos excepto entre el control y la acidosis mixta, que no llegan a diferenciarse claramente. Así pues, la cuestión propuesta se resuelve así: la acidosis mixta no influye claramente en la glucemia, la respiratoria sí y la metabólica más aún. ¿En qué sentido? Basta echar un vistazo al diagrama de cajas.

5.5.3. Regresión logística simple

Vamos a considerar ahora el problema recíproco: una variable cuantitativa X que podría explicar un factor cualitativo, es decir, pretendemos clasificar a un individuo respecto a varias categorías a partir de la observación de un variable cuantitativa. Se trata pues de un problema de regresión, aunque no lineal. Este problema se enmarca dentro de otro más general denominado Análisis Discriminante en el que podemos contar con tantas variables cuantitativas como deseemos. En esta ocasión restringiremos a 2 el número de categorías del factor, es decir, se trata e una variable dicotómica del tipo 0-1. Por ejemplo, podemos plantearnos de qué forma cierta variable bioquímica pude explicar la incidencia (1) o ausencia (0) de determinada enfermedad.

Esto puede resolverse mediante una cierta función que asigne a cada valor de X un número entre 0 y 1. Si el valor resultante es cercano al 1 nos decantaremos por esa categoría. Una función de este tipo no puede ser lineal. Una de las opciones más utilizada para desempeñar ese papel es la denominada función logística, que dependerá de ciertos parámetros que habrá que obtener a partir e los datos. El aspecto de una función logística es el siguiente:



5.6. Relaciones entre más de dos variables

Hemos estudiado hasta ahora los métodos más utilizados en la Estadística Básica. Para facilitar una mejor visión global de la materia terminaremos el capítulo con una breve reseña de las técnicas que se sitúan en en siguiente nivel de complejidad. tanto estas técnicas como la regresión logística que hemos introducido anteriormente no figuran en los contenidos de las asignaturas que cubre este manual, de ahí que se expongan a título meramente informativo por dar coherencia al conjunto. Para más información, remitimos al lector a la bibliografía propuesta.

5.6.1. Regresión múltiple

Ya hemos estudiado el problema de regresión simple, consistente en explicar una variable cuantitativa a través de otra a su vez cuantitativa. Cuando tenemos varias variables explicativas cuantitativas se dice que es un problema de regresión múltiple. En ese caso, pretendemos establecer ecuaciones del tipo

$$y \simeq \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q$$

La magnitud del coeficiente β_j guarda relación con la trascendencia de la variable X_j a la hora de explicar la variabilidad de Y. Nótese que, cuantas más variables explicativas introduzcamos mayores posibilidades tendremos de explicar satisfactoriamente la variable respuesta aunque, por contra, generaremos un diseño más complejo y aparatoso.

Desde el punto de vista teórico hay pocas diferencias entre la regresión simple y la múltiple. Se define un nuevo parámetro denomina coeficiente de correlación múltiple que expresa la proporción de variabilidad explicada conjuntamente, así como unos coeficientes denominados de correlación parcial que expresan la proporción de variabilidad que cada variable explica de manera aislada; de estos últimos dependerán los resultados de los contrastes parciales en los que se dilucida si las diferentes variables introducidas en el modelo son útiles para mejorar la explicación de la variable respuesta.

Si, además, consideramos simultáneamente varias variables respuesta estaremos hablando de un problema de regresión multivariante. Por ejemplo, podemos estudiar la relación existente entre las variables morfólogicas de los espermatozoides y las referentes a la motilidad, calculando por un lado medidas de correlación y, por otro, parámetros que ayuden a explicar en lo posible unas variables a partir de las otras.

5.6.2. Diseños multifactoriales

En la última sección se ha considerado el problema de explicar una variable cuantitativa mediante un factor cualitativo, dando lugar a lo que conocemos como análisis de la varianza o anova de una vía. Al igual que sucede en el apartado anterior, podemos diseñar el experimento introduciendo más factores con la intención de explicar mejor la variable respuesta. En el caso de que sean dos los factores hablaremos de anovas de dos vías. Aquí, los contrastes de interés será los relacionados con la capacidad de los factores para explicar la variable respuesta.

Por ejemplo, podemos intentar averiguar si las diferencias en el crecimiento de una semilla obedecen sólo al tipo de fertilizante o depende también del factor terreno. Para ello se puede diseñar un experimento en el que se mide el crecimiento experimentado por las semillas en diferentes combinaciones fertilizante-terreno. Suponemos entonces que el crecimiento de una semilla k con fertilizante i y terreno j es

$$Y_{ijk} = \theta + \alpha_i + \beta_j + \varepsilon_{ijk}$$

donde θ expresa una componente común a todas las categorías de fertilizante y terreno, α_i denota la influencia particular del fertilizante i en el crecimiento y β_j la influencia del terreno j. Por último, ε_{ijk} denota la componente puramente aleatoria del experimento, pues es de esperar que diferentes semillas creciendo en las mismas condiciones aporten distintas medidas y esas diferencias las consideraremos aleatorias. Este tipo de diseño se denomina aditivo pues suponemos que los efectos de los factores se suman entre sí. Es una suposición análoga a la de linealidad en un problema de regresión. En definitiva, en el modelo aditivo propuesto, que el tipo de fertilizante sea realmente una fuente de variabilidad en el crecimiento equivale a la falsedad de la hipótesis

$$H_0^1 : \alpha_1 = \ldots = \alpha_a = 0$$

que puede contrastarse mediante un test similar al anova de una vía que toma como referencia la distribución F-Snedecor. Algo similar ocurre con el contraste de la hipótesis

$$H_0^2 : \beta_1 = \ldots = \beta_b = 0$$

cuya falsedad equivale a la influencia del terreno en el crecimiento de las semillas. En general, cabe suponer que exista una interacción entre el terreno y el fertilizante que rompa la aditividad, de manera que la ecuación anterior se expresaría mediante

$$Y_{ijk} = \theta + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

En el modelo resultante, denominado bifactorial completo podemos contrastar si realmente se da la aditividad, que se correspondería con la veracidad de la hipótesis inicial

$$H_0^3 : (\alpha\beta)_{11} = \ldots = (\alpha\beta)_{ab} = 0$$

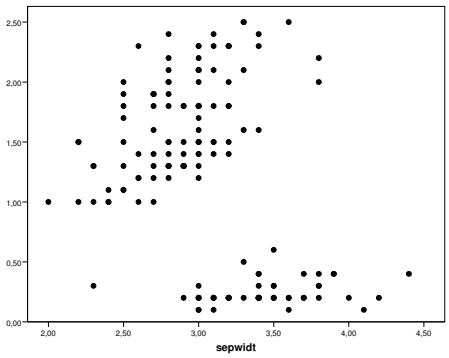
El número de factores que pueden introducirse en el diseño no tiene límite en principio. Hemos de tener en cuenta que la introducción de muchos factores propicia una reducción de la parte de variabilidad de Y que atribuimos al azar, lo cual es positivo, pero debe ir aparejada a la selección de grandes muestras, pues es necesario cruzar las diferentes categorías de los factores considerados, salvo que se apliquen argucias como el diseño de cuadrados latinos.

5.6.3. Análisis de la covarianza

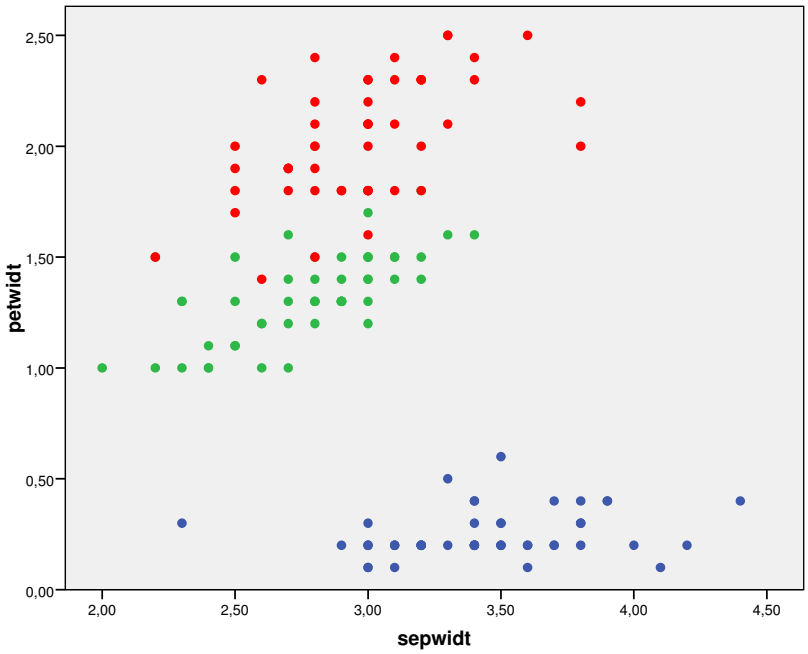
Se trata de una fusión entre la regresión y el análisis de la varianza pues se pretende explicar una variable respuesta cuantitativa mediante uno o más factores y una o más variables cuantitativas (covariables). Realmente, los tres problemas considerados, es decir, regresión, análisis de la varianza (anova) y análisis de la covarianza (ancova) se formalizan mediante un mismo modelo matemático denominado Modelo Lineal.

Ejemplo 15:[Ancova]

Presentamos a continuación el gráfico de dispersión para las variables `X=anchura de sépalo` y `Y=anchura de pétalo` medidas en 150 flores diferentes.



Se intuye la presencia de un factor cualitativo que influye en la proporción que guardan sépalos y pétalos. Efectivamente, si controlamos el factor especie, distinguiendo entre `virginica`, `vesicolor` y `setosa` tenemos el gráfico en color que a parece a continuación



Parece claro que la relación entre pétalos y sépalos ha de estudiarse por separado en cada especie, o por lo menos en setosa (azul). Con ese diseño conseguimos explicar el 72.7 % de la varibilidad de la anchura del pétalo, mientras que, si no distinguimos entre especies, sólo explicamos el 13.4 %. Igualmente, si pretendemos explicar el variabilidad del pétalo únicamente a partir de la especie, sin tener en cuenta la medida del sépalo, lo conseguimos en un 62 %.

5.6.4. Análisis de la varianza multivariante

Se trata en esta ocasión de explicar no una sino un grupo de variables cuantitativas a partir de uno o varios factores cualitativos. Suele denominarse manova y en esencia es parecido al anova aunque adolece de una mayor complejidad, además de la inherentes a los cálculos que no deben preocuparnos pues damos por sentado que contamos con un programa estadístico para llevar a cabo el estudio. De lo contrario es materialmente imposible.

5.6.5. Análisis discriminante

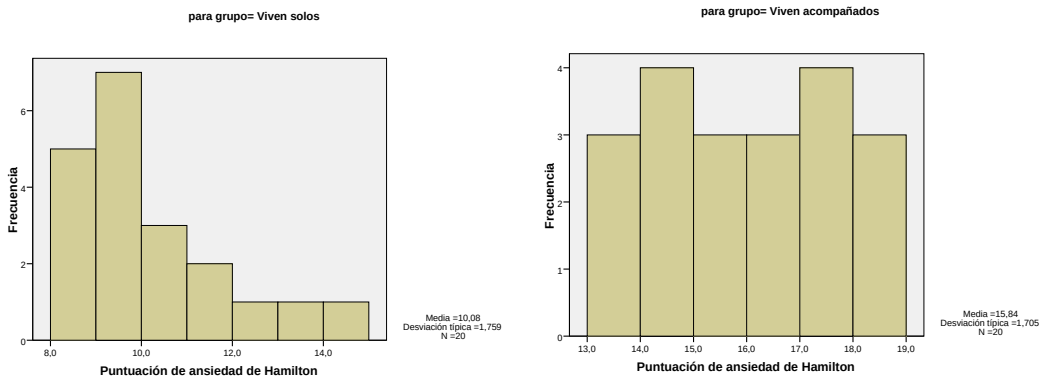
Es el reverso del manova con un factor, pues pretende explicar una variable cualitativa a partir de varias variables explicativas cuantitativas. Es decir, se tata de clasificar a un individuo respecto a varias categorías posibles a partir de la observación de varias variables numéricas. Podemos pensar, por ejemplo, en un diagnóstico a partir de una analítica o hemograma, en predecir si un individuo padecerá determinado tipo de tumor o en un problema taxonómico. También podemos clasificar un espermatozoide como viable o no viable en virtud de sus medidas morfológicas o de motilidad.

Para acabar proponemos una clasificación para los métodos considerados y algunos más, en función de la naturaleza y número de las variables explicativas y respuesta.

Explicativa	Respuesta	Problema
cuantitativa	cuantitativa	regresión simple
cualitativa	cualitativa	tabla contingencia
cualitativa biaria	cuantitativa	dos tratamientos
cuantitativa	cualitativa binaria	logística simple
cuantitativas	cuantitativa	regresión múltiple
cualitativas	cuantitativa	anova
cuantitativas+cualitativas	cuantitativa	ancova
cuantitativas	cuantitativas	regresión multivariante
cualitatitivas	cuantitativas	manova
cuantitativas+cualitativas	cuantitativas	mancova
cuantitativas	cualitativa	análisis discriminante

5.7. Cuestiones propuestas

1. Consideremos nuevamente el estudio de la puntuación de ansiedad de Hamilton en un grupo de 20 personas que viven solas y otras tantas que viven acompañadas. Los respectivos diagramas de caja se muestran en el capítulo 1. Vemos que aspecto tienen los histogramas:



Se indican a continuación los resultados del test de normalidad ed Shappiro-Wilk para ambas muestras, del test de Levene de igualdad de varianzas, y de los test de Student, Welch y Mann-Whitney de comparación de medias (o valores centrales).

Test	P-valor
Shapiro-Wilk	P=0.015(solos) P=0.272(acompañados)
Levene	P=0.746
Student	P<0.001
Welch	P<0.001
Mann-Whitney	P=0.004

A partir de los mismos y suponiendo que ambas muestras son aleatorias, contestar la siguiente pregunta: ¿existe relación entre el tipo de vida (en soledad o en compañía) y el nivel de ansiedad? Indicar claramente en qué se basa la conclusión obtenida.

2. Considera los dos estudios siguientes:

   (a) Comparando los niveles medios de pepsina en jugo gástrico de dos formas diferenciadas de una enfermedad mediante el test de Student, se obtuvo una diferencia altamente significativa ($P = 0{,}00001$) partiendo de muestras aleatorias independientes de tamaños 120 y 110.

   (b) Comparando los niveles medios de O_3 en dos regiones del planeta mediante el test de Student, se obtuvo una diferencia altamente significativa ($P = 0{,}00001$) partiendo de muestras de tamaño 20 y 25.

   Responde a las siguientes preguntas:

   - ¿En cuál de los dos estudios se está más seguro de que existen diferencias entre los niveles medios?
   - Si suponemos que las dispersiones (varianzas) de todas las variables en juego son similares, ¿en cuál de los dos estudios será mayor la diferencia entre las medias aritméticas de los grupos estudiados

3. Se desea contrastar la hipótesis de que una determinada cualidad se verifica en el 50 % de la población. Para ello se toma una muestra aleatoria de n individuos y se calcula qué proporción presenta dicha cualidad en la muestra, obteniéndose un 51 %. Tras aplicar el test adecuado, se obtuvo $P = 0{,}0032$. ¿Qué decisiones corresponderán a unos niveles de significación del 5 %, 50 % y 0.0001 %? En términos muy prácticos, ¿qué conclusión se extrae de este resultado? ¿Qué podemos decir del tamaño de la muestra? Contesta con la mayor concreción posible.

4. Se pretende comparar las proporciones de alérgicos de dos poblaciones A y B. Para ello se toma una muestra de cierto tamaño n en la población A, resultando ser alérgicos el 12 % de éstos. Se toma, de manera independiente, otra muestra del mismo tamaño en la población B, resultando alérgicos el 22 %. Tras aplicar el test de Student se obtuvo $P > 0{,}05$. Comenta brevemente la implicación práctica del resultado.

   Volvemos a repetir el experimento tomando sendas muestras de un mismo tamaño m (en principio distinto de n). Se obtiene entonces unas proporciones de alérgicos del 12 % y 16 % respectivamente. Tras aplicar el mismo test se obtiene $P < 0{,}01$. Comenta brevemente la implicación práctica del resultado.

   Existe una aparente contradicción entre ambos resultados. ¿Cuál es? ¿Puede justificarse este hecho de alguna forma? ¿Qué podemos decir concretamente de n y m?

5. Un investigador desea probar que el valor medio de una determinada variable bioquímica en una región A es mayor que en una región B. ¿Con cuál de los siguientes resultados se sentiría más satisfecho?

   A : $P > 0{,}05$ con muestras pequeñas.

   B : (0,03 , 0,91), intervalo al 95 % para la diferencia de medias $\mu_A - \mu_B$.

   C : (1,83 , 1,91) intervalo al 99 % para la diferencia de medias $\mu_A - \mu_B$.

$D:$ $P > 0{,}01$ para el test de comparación de medias.

$E:$ $r = -0{,}99$ y $b = -1$.

6. Se pretende contrastar si las medias de dos variables normales con varianza común son iguales o no. Tras seleccionar sendas muestras independientes, de tamaños $n_1 = 20$ y $n_2 = 27$, y aplicar el test de Student, se obtuvo $t_{exp} = 2{,}001$. Elige la respuesta correcta:

   A: $P > 0{,}05$

   B: $0{,}01 < P < 0{,}05$

   C: $0{,}001 < P < 0{,}01$

   D: $P < 0{,}001$

   E: Ninguna de las anteriores respuestas es correcta.

7. Se estudia, de forma cualitativa, la posible relación entre el grado de adicción al tabaco (se distingue entre alto, medio o bajo) y el grado de capacidad pulmonar (se distingue también entre alto, medio o bajo) en un determinado grupo humano. Para ello, se seleccionaron un total de 200 individuos de dicho grupo, que se distribuyeron de la siguiente forma:

Capacidad pulmonar

Grado adicción	(3 × 3)	Alto	Medio	Bajo
	Alto	5	10	45
	Medio	10	50	20
	Bajo	20	22	18

- ¿Qué porcentaje de fumadores de alta adicción tienen una capacidad pulmonar baja?

  A: 30 %

  B: 75 %

  C: 54.21 %

  D: 22.5 %

  E: 45 %

- El valor correspodiente al coeficiente de Pearson es, en este caso, $C = 0{,}46096$. Indica cuál es la única afirmación verdadera:

  A: El valor de C en una tabla 3×3 ha de estar entre -1 y $+1$. El caso $C = +1$ se traduce en un máximo grado de relación entre los caracteres. El caso $C = -1$ expresa independencia de los caracteres.

  B: El valor de C en una tabla 3×3 ha de estar entre 0 y 0,8165. El caso $C = 0{,}8165$ se traduce en un máximo grado de relación. El caso $C = 0$ expresa un grado nulo de relación.

  C: El valor de C en una tabla 3×3 ha de estar entre 0 y 0,7071. El caso $C = 0{,}7071$ se traduce en un máximo grado de relación. El caso $C = 0$ expresa un grado nulo de relación.

D: El valor de C en una tabla 3×3 ha de estar entre -1 y $+1$. El caso $C = -1$ y $C = +1$ se traduce en un máximo grado de relación. El caso $C = 0$ expresa un grado nulo de relación.

E: El valor de C en una tabla 3×3 ha de estar entre 0 y 1. El caso $C = 0$ se traduce en un máximo grado de relación entre los caracteres. El caso $C = 1$ expresa un grado nulo de relación entre los caracteres.

- Supongamos que los 200 individuos han sido seleccionados de manera aleatoria. ¿Cuál será el resutado tras aplicar el test χ^2 para contrastar la hipótesis inicial de independencia de los caracteres?

  A: $P < 0{,}001$

  B: $0{,}001 < P < 0{,}01$

  C: $0{,}01 < P < 0{,}05$

  D: $0{,}05 < P$

  E: $0{,}05 < P < 0{,}01$

- ¿Cómo debemos interpretar el resultado anterior?

  A: No podemos optar por ninguna de las hipótesis

  B: Para un nivel de significación $\alpha = 0{,}05$, no nos atrevemos a rechazar la hipótesis inicial de independencia.

  C: Para un nivel de significación $\alpha = 0{,}001$, sí optamos por la hipótesis alternativa, es decir, parece claro que las medias de las variables son distintas.

  D: Para un nivel de significación $\alpha = 0{,}001$, no optamos por la hipótesis alternativa, es decir, las medias de las variables son iguales.

  E: Para un nivel de significación $\alpha = 0{,}001$, sí optamos por la hipótesis alternativa, es decir, parece claro que existe relación entre los caracteres.

Seguimos con el estudio de la relación entre el hábito de fumar y la capacidad pulmonar, sólo que esta vez lo realizaremos cuantitativamente: sobre cada uno de los 200 individuos (los mismos de antes) se miden ahora las variables X=`Consumo de nicotina al día` (mg) e Y=`Tiempo de resistencia en una inmersión bajo el agua` (seg).

- Según los resultados obtenidos en el caso cualitativo, ¿cuál de los siguientes valores del coeficiente de correlación lineal r te resulta más verosímil? En caso de duda, observa con detenimiento la anterior tabla 3×3.

  A: $r = 2{,}31$

  B: $r = 0{,}66$

  C: $r = -0{,}57$

  D: $r = -2{,}14$

  E: Ambos estudios no guardan ninguna relación.

- ¿Cuál de las rectas de regresión siguientes te parece más verosímil?

  A: $y = 20{,}31 - 5{,}67x$

  B: $y = 20{,}31 + 5{,}67x$

C: $y = -234{,}21 - 450{,}2x$

D: $y = 234{,}21 + 45{,}2x$

E: Ambos estudios no guardan ninguna relación.

8. Considera nuevamente los datos del problema 2.12. Aplica el test χ^2 e interpreta el resultado en términos muy prácticos.

9. Considera los datos del problema 2.10. Aplica el test χ^2 e interpreta el resultado en términos muy prácticos. Dado que se está considerando la exposición al agente radioactivo como un posible factor de riesgo, estimar el riesgo atribuible, el riesgo relativo el odd ratio.

10. En un estudio sobre el efecto de ciertos tratamientos químicos sobre semillas de cereales durante su almacenamiento, se probaron tres tratamientos químicos diferentes sobre una muestra de semillas. Los resultados quedan recogidos en la siguiente tabla de contingencia:

	Germinación correcta	Germinación defectuosa	Sin germinar
Tratamiento A	23	9	6
Tratamiento B	21	4	3
Tratamiento C	34	24	17

¿Entre qué valores estará comprendido el coeficiente C? El valor del mismo resulta ser $C = 0{,}2296$. Valóralo. El resultado del test χ^2 es $P > 0{,}005$. ¿Qué hipótesis se contrasta? Extrae las conclusiones oportunas en términos muy prácticos.

11. A partir de los datos del problema 2.7 decidir si existe correlación entre la edad y la presión sistólica. Es necesario calcular un P-valor e interpretarlo. Predecir la presión sistólica que cabría esperar para un individuo de 50 años y para otro de 5. Valorar qué estimación es más fiable.

12. Se pretende establecer la relación entre la temperatura t de determinadas aguas residuales (medida en grados C) y la concentración s de cierta sustancia tóxica (medida en mg/l). En la siguiente tabla se dan los resultados de las mediciones de la concentración s para ocho temperaturas distintas:

$17C$	$22C$	$27C$	$32C$	$37C$	$42C$	$47C$	$52C$
9,6	12,0	15,1	18,1	25,0	28,9	36,5	48,6
10,1	12,7	15,8	19,0	23,4	28,2	34,8	47,5
9,2		14,7	18,5	24,0		35,7	
		15,5				38,1	

Indicar de qué tipo de problema se trata y resolverlo mediante un programa estadístico, extrayendo las conclusiones oportunas. Estima puntualmente y mediante un intervalo de confianza al 95 %, la concentración que corresponde a una temperatura de $20C$.

13. ¿A qué nos referimos exactamente cuando afirmamos que el nivel de urea debe encontrarse en el intervalo [10,40], en las unidades que corresponda? ¿Cómo podemos llegar a una conclusión de ese tipo?

14. Utilizar los datos del problema 1.12 para determinar límites de normalidad en la concentración del cadmio. Considerar un 95 % tanto para la proporción de población de referencia como para la confianza deseada.

15. Se pretende probar la eficicacia de cierto fertilizante denominado auxina. Para ello se emplearon dos grupos de plantas escogidos al azar e independientemente. Un grupo fue tratado con auxina mientras que el otro, cultivado bajo condiciones idénticas, fue dejado sin tratamiento como control. Los resultados, medida la altura de las plantas, fueron los siguientes:

$$\text{Auxina} \begin{cases} n_1 = 10 \\ \overline{x}_1 = 15{,}2cm \\ s_1 = 8{,}7cm \end{cases} \qquad \text{Control} \begin{cases} n_2 = 20 \\ \overline{x}_2 = 13{,}4cm \\ s_2 = 6{,}9cm \end{cases}$$

Formular las hipótesis a contratar y aventurar una primer pronóstico a tenor de lo observado en las muestras. Discutir qué test puede ser el más adecuado para tomar la decisión. En este caso, tanto el resultado del test de Student como el del test de de Mann-Whitney es $P > 0{,}05$ y el intervalo de confianza al 95 % para la diferencia de medias es $(-0.9\ ,\ 5.5)$. Extraer las conclusiones oportunas en términos muy prácticos. Discutir si es precisa alguna modificación en el diseño del experimento.

16. Se pretende probar la influencia del estrés sobre la glándula suprarrenal. Para ello, un grupo de ratones fue sometido a una serie de situaciones de tensión, que produjeron una respuesta de temor. Después de cierto periodo de tiempo bajo estas condiciones, los ratones fueron comparados con los del grupo control que no había sido sometido a tensión. Los datos obtenidos fueron los siguientes:

Control	3.35	3.60	3.75	4.15	3.60			
Tensión	5.60	6.25	7.45	5.05	4.56	4.5	3.9	4.3

Indicar de qué tipo de problema se trata y plantear las hipótesis a contrastar. Discutir cuál es el test adecuado para constratarlas y ejecutarlo con la ayuda de un programa estadístico. Interpretar la solución obtenida.

17. Se pretende averiguar en qué medida influye en el crecimiento del crisantemo la cantidad del fertilizante $MgNH_4PO_4$ utilizada. Para ello, se suministró a 22 ejemplares de crisantemo dicho fertilizante en 4 diferentes concentraciones, anotándose en cada caso el crecimiento en cm. Los resultados fueron los siguientes:

50 gr/bu	17.2	13.0	14.0	14.2	21.6	
100 gr/bu	13.0	14.0	23.6	14.0	17.0	22.2
200 gr/bu	15.8	17.0	27.0	19.6	18.0	20.2
400 gr/bu	15.8	18.8	26.0	21.1	22.0	

Resolver el problema mediante un programa estadístico e interpretar los resultados en términos muy prácticos.

18. Se realizó un estudio sobre el efecto del ejercicio físico sobre la concentración de colesterol en sangre, a partir de una muestra aleatoria de 11 pacientes. Se obtuvieron las siguientes lecturas (en miligramos de triglicérido por 100 mililitros de sangre) previas y posteriores al ejercicio:

Sujeto	**Previo**	**Posterior**
1	68	95
2	77	90
3	94	86
4	73	58
5	37	47
6	131	121
7	77	136
8	24	65
9	99	131
10	629	630
11	116	104

Indicar de qué tipo de problema se trata y resolverlo mediante un programa estadístico.

19. Se pretende determinar si existe relación y de qué tipo entre el pulso y la temperatura de un enfermos. Tomamos un grupo de observación de 15 enfermos en los que medimos las pulsaciones por minuto (p/m), X, y la temperatura en grados centígrados (C), Y, obteniéndose los siguientes resultados:

X	70	65	80	60	75	85	70	65	80	85	65	65	75	70	70
Y	36,5	36,5	37	36	37	37,5	37	36	37,5	37	36	36,5	36,5	36,5	37

¿De qué tipo de problema se trata? Resolverlo mediante un programa estadístico. Obtener también la estadística descriptiva de ambos grupos de datos por separado.

20. Se sospecha que la presencia de cierto gen predispone a desarrollar un determinado tipo de tumor que presenta una prevalencia del 0.5 %. Se seleccionaron 1000 enfermos y otros tantos individuos sanos, obteniéndose la siguiente tabla de contingencia:

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

Calcular el coeficiente ϕ e interpretarlo. Aplicar el test χ^2 obteniendo P e interpretarlo en términos muy prácticos. ¿Cómo se interpretaría un resultado no significativo? Estimar el riesgo atribuible, el riesgo relativo y el odd ratio.

21. 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:

	+	-	Total
Enfermos	730	20	750
Sanos	50	200	250
Total	780	220	1000

Estimar la sensibilidad y especificidad de la prueba diagnóstico, así como las proporciones de falsos positivos y falsos negativos. Estimar los valores predictivos positivos y negativos. Valorar los resultados en términos muy prácticos. Dado que se trata de una tabla 2×2, podemos aplicar el test χ^2. ¿Serías capaz de aventurar si el resultado del mismo será significativo?

22. Disponemos de otro procedimiento diagnóstico B para la misma enfermedad. Sus resultados tars aplicarlo a los mismo individuos son los siguientes:

	+	-	Total
Enfermos	610	140	750
Sanos	3	247	250
Total	613	387	1000

Estimar nuevamente la sensibilidad, especificidad y los valores predictivos positivos y negativo. Valorar los resultados y compararlos con los del procedimiento A.

23. Se pretende probar que el tiempo de reacción ante un estímulo auditivo bajo dos condiciones radicalmente distintas F y Q es diferente. Para ello se ha selecionado aleatoriamente una muestra de 9 personas, las cuales han sido estimuladas en primer lugar bajo la condición F y, transcurido un tiempo de reposo, bajo la condición Q. Los tiempos de reacción en centésimas de segundo fueron los siguientes:

Individuo	1	2	3	4	5	6	7	8	9
Situación F	14	12	9	13	15	17	13	12	13
Situación Q	17	14	13	15	16	16	16	15	13

¿De qué tipo de problema se trata? ¿Cuáles son las hipótesis a contrastar? Discutir qué test es el más adecuado para contrastarlas. Tras aplicar dicho test se obtiene como resultado $0.001 < P < 0.01$. Interpretarlo en términos muy prácticos.

24. Para estudiar una determinada variable bioquímica en la sangre, se escogió un muestra aleatoria de 20 individuos a los que se les midió la variable con los resultados siguientes:

162	222	245	195	204
240	157	164	183	192
179	191	192	171	146
147	131	248	176	207

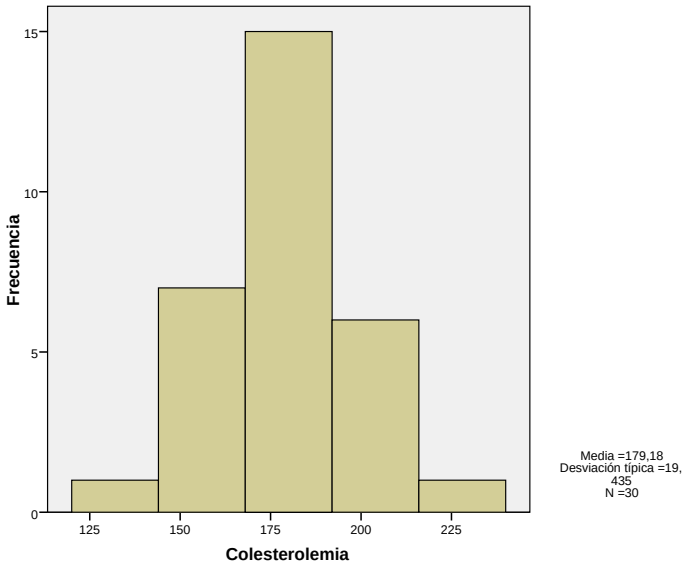
Mediante un programa estadístico obtener y valorar un estudio descriptivo completo. Indicar un intervalo al 95 % de confianza para la media de dicha variable e indicar qué tamaño aproximado de muestra sería necesario para estimar la media con un error máximo de dos unidades al 95 % de confianza.

25. Que la mortalidad infantil disminuye en aquellos lugares donde llega la medicina moderna está claramente contrastado. Indica factores concretos que puedan ser la causa del descenso de la mortalidad y plantear un diseño estadístico que tenga como objetivo valorar aisladamente la contribución real de esos factores a la mejora de la salud.

26. Idear un estudio estadístico con el objeto de evaluar la influencia de una central nuclear en la salud de las poblaciones vecinas.

27. Se presentan a continuación las mediciones de colesterol en sangre (mg/dl) en 30 personas sanas, así como el correspondiente histograma:

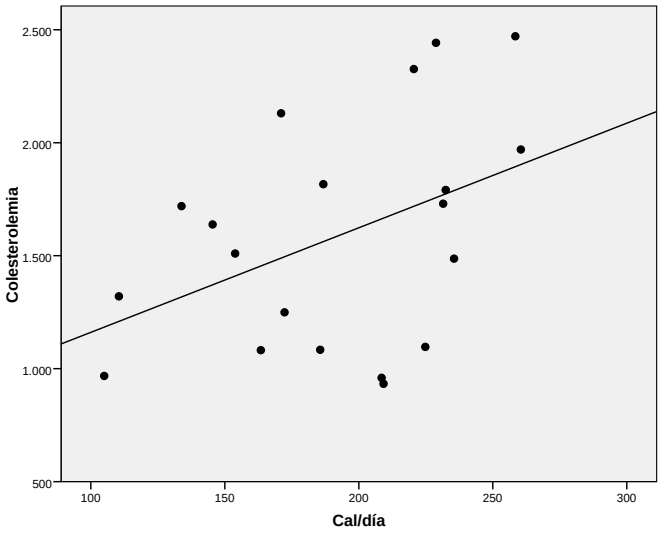
    180, 201, 152, 174, 164, 174, 157, 203, 207, 156, 179, 188, 181, 188, 183, 174, 197, 181, 181, 186, 156, 132, 196, 202, 188, 161, 171, 227, 167, 170.



Responder las siguientes preguntas:

- ¿Qué puedes decir acerca de los coeficientes de simetría y aplastamiento?
- Si aplicamos a estos datos el test de normalidad de Shappiro-Wilk, ¿qué tipo de resultado cabe esperar: significativo o no significativo? ¿Por qué?
- ¿Crees que la mediana será sensiblemente superior o inferior a la media?
- Calcula un par de valores típicos que resuman lo mejor posible la información que contienen estos datos.
- Suponiendo que los 30 individuos estudiados constiyuyan una muestra aleatoria de la población sana, construye un intervalo que contenga con una confianza del 95 % el colesterol medio de dicha población.
- Construye un intervalo que contenga aproximadamente al 95 % del total de la población sana. ¿Qué podemos decir entonces de un individuo cualquiera que presente un nivel de colesterol de 227? Matiza la afirmación si lo consideras necesario.

28. Seguimos con el colesterol. Un estudio pretende dilucidar si el nivel de colesterol (mg/dl) está relacionado con la ingesta media diaria de calorías. Para ello se sometió a estudio durante un mes una muestra de 20 individuos. Transcurrido ese tiempo se anotó en cada caso el consumo medio de calorías por día (X) y el nivel de colesterol final (Y), obteniéndose el siguiente diagrama de dispersión.



Responder las siguientes preguntas:

- Uno de los siguientes números es el valor del coeficiente de correlación lineal r. Escoge el que te resulte más verosímil: 0.42, -0.42, 0.98, -0.98, 1.37 ó -1.37.
- Indica entonces qué proporción de la variabilidad del colesterol es explicada en la muestra por la ingesta de calorías.
- Se desea decidir si podemos hablar de una correlación entre ambas variables a nivel poblacional. El test de correlación aportó el resultado $P = 0{,}060$. Indica qué afirmación te parece más correcta y matízala si lo consideras oportuno:
  (A) La muestra estudiada constituye una prueba significativa de que el consumo de calorías aumenta el nivel de colesterol.
  (B) La muestra estudiada no nos permite afirmar que el consumo de calorías influya en el nivel de colesterol.
  (C) La muestra estudiada constituye una prueba significativa de que el consumo de calorías disminuye el nivel de colesterol.
  (D) Hemos demostrado que existe relación entre el consumo de calorías y el nivel de colesterol.
  (E) Hemos demostrado que no existe relación alguna entre el consumo de calorías y el nivel de colesterol.
- La recta de regresión lineal muestral tiene por ecuación $y = 697 + 4{,}3x$. Estimar el nivel de colesterol de un individuo que consuma diariamente 200 cal. Valorar la fiabilidad de dicha predicción.

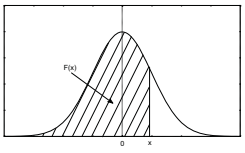
Tablas

Distribución Binomial hasta $n = 10$

		p									
n	k	**0.050**	**0.100**	**0.150**	**0.200**	**0.250**	**0.300**	**0.350**	**0.400**	**0.450**	**0.500**
2	0	0.903	0.810	0.723	0.640	0.563	0.490	0.423	0.360	0.303	0.250
2	1	0.095	0.180	0.255	0.320	0.375	0.420	0.455	0.480	0.495	0.500
2	2	0.003	0.010	0.023	0.040	0.063	0.090	0.123	0.160	0.203	0.250
3	0	0.857	0.729	0.614	0.512	0.422	0.343	0.275	0.216	0.166	0.125
3	1	0.135	0.243	0.325	0.384	0.422	0.441	0.444	0.432	0.408	0.375
3	2	0.007	0.027	0.057	0.096	0.141	0.189	0.239	0.288	0.334	0.375
3	3	0.000	0.001	0.003	0.008	0.016	0.027	0.043	0.064	0.091	0.125
4	0	0.815	0.656	0.522	0.410	0.316	0.240	0.179	0.130	0.092	0.062
4	1	0.171	0.292	0.368	0.410	0.422	0.412	0.384	0.346	0.299	0.250
4	2	0.014	0.049	0.098	0.154	0.211	0.265	0.311	0.346	0.368	0.375
4	3	0.000	0.004	0.011	0.026	0.047	0.076	0.111	0.154	0.200	0.250
4	4	0.000	0.000	0.001	0.002	0.004	0.008	0.015	0.026	0.041	0.063
5	0	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031
5	1	0.204	0.328	0.392	0.410	0.396	0.360	0.312	0.259	0.206	0.156
5	2	0.021	0.073	0.138	0.205	0.264	0.309	0.336	0.346	0.337	0.313
5	3	0.001	0.008	0.024	0.051	0.088	0.132	0.181	0.230	0.276	0.313
5	4	0.000	0.000	0.002	0.006	0.015	0.028	0.049	0.077	0.113	0.156
5	5	0.000	0.000	0.000	0.000	0.001	0.002	0.005	0.010	0.018	0.031
6	0	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016
6	1	0.232	0.354	0.399	0.393	0.356	0.303	0.244	0.187	0.136	0.094
6	2	0.031	0.098	0.176	0.246	0.297	0.324	0.328	0.311	0.278	0.234
6	3	0.002	0.015	0.041	0.082	0.132	0.185	0.235	0.276	0.303	0.313
6	4	0.000	0.001	0.005	0.015	0.033	0.060	0.095	0.138	0.186	0.234
6	5	0.000	0.000	0.000	0.002	0.004	0.010	0.020	0.037	0.061	0.094
6	6	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.008	0.016
7	0	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008
7	1	0.257	0.372	0.396	0.367	0.311	0.247	0.185	0.131	0.087	0.055
7	2	0.041	0.124	0.210	0.275	0.311	0.318	0.298	0.261	0.214	0.164
7	3	0.004	0.023	0.062	0.115	0.173	0.227	0.268	0.290	0.292	0.273
7	4	0.000	0.003	0.011	0.029	0.058	0.097	0.144	0.194	0.239	0.273
7	5	0.000	0.000	0.001	0.004	0.012	0.025	0.047	0.077	0.117	0.164
7	6	0.000	0.000	0.000	0.000	0.001	0.004	0.008	0.017	0.032	0.055
7	7	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.008
8	0	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004
8	1	0.279	0.383	0.385	0.336	0.267	0.198	0.137	0.090	0.055	0.031
8	2	0.051	0.149	0.238	0.294	0.311	0.296	0.259	0.209	0.157	0.109
8	3	0.005	0.033	0.084	0.147	0.208	0.254	0.279	0.279	0.257	0.219
8	4	0.000	0.005	0.018	0.046	0.087	0.136	0.188	0.232	0.263	0.273
8	5	0.000	0.000	0.003	0.009	0.023	0.047	0.081	0.124	0.172	0.219
8	6	0.000	0.000	0.000	0.001	0.004	0.010	0.022	0.041	0.070	0.109
8	7	0.000	0.000	0.000	0.000	0.000	0.001	0.003	0.008	0.016	0.031
8	8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004
9	0	0.630	0.387	0.232	0.134	0.075	0.040	0.021	0.010	0.005	0.002
9	1	0.299	0.387	0.368	0.302	0.225	0.156	0.100	0.060	0.034	0.018
9	2	0.063	0.172	0.260	0.302	0.300	0.267	0.216	0.161	0.111	0.070
9	3	0.008	0.045	0.107	0.176	0.234	0.267	0.272	0.251	0.212	0.164
9	4	0.001	0.007	0.028	0.066	0.117	0.172	0.219	0.251	0.260	0.246
9	5	0.000	0.001	0.005	0.017	0.039	0.074	0.118	0.167	0.213	0.246
9	6	0.000	0.000	0.001	0.003	0.009	0.021	0.042	0.074	0.116	0.164
9	7	0.000	0.000	0.000	0.000	0.001	0.004	0.010	0.021	0.041	0.070
9	8	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.008	0.018
9	9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002
10	0	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001
10	1	0.315	0.387	0.347	0.268	0.188	0.121	0.072	0.040	0.021	0.010
10	2	0.075	0.194	0.276	0.302	0.282	0.233	0.176	0.121	0.076	0.044
10	3	0.010	0.057	0.130	0.201	0.250	0.267	0.252	0.215	0.166	0.117
10	4	0.001	0.011	0.040	0.088	0.146	0.200	0.238	0.251	0.238	0.205
10	5	0.000	0.001	0.008	0.026	0.058	0.103	0.154	0.201	0.234	0.246
10	6	0.000	0.000	0.001	0.006	0.016	0.037	0.069	0.111	0.160	0.205
10	7	0.000	0.000	0.000	0.001	0.003	0.009	0.021	0.042	0.075	0.117
10	8	0.000	0.000	0.000	0.000	0.000	0.001	0.004	0.011	0.023	0.044
10	9	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.002	0.004	0.010
10	10	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001

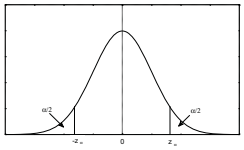
Distribución $N(0,1)$

Tabla III.2: Distribución normal $N(0,1)$. Función de distribución



x	$F(x)$	x	$F(x)$	x	$F(x)$	x	$F(x)$	x	$F(x)$
0.01	0.5040	0.61	0.7291	1.21	0.8869	1.81	0.9649	2.41	0.9920
0.02	0.5080	0.62	0.7324	1.22	0.8888	1.82	0.9656	2.42	0.9922
0.03	0.5120	0.63	0.7357	1.23	0.8907	1.83	0.9664	2.43	0.9925
0.04	0.5160	0.64	0.7389	1.24	0.8925	1.84	0.9671	2.44	0.9927
0.05	0.5199	0.65	0.7422	1.25	0.8944	1.85	0.9678	2.45	0.9929
0.06	0.5239	0.66	0.7454	1.26	0.8962	1.86	0.9686	2.46	0.9931
0.07	0.5279	0.67	0.7486	1.27	0.8980	1.87	0.9693	2.47	0.9932
0.08	0.5319	0.68	0.7517	1.28	0.8997	1.88	0.9699	2.48	0.9934
0.09	0.5359	0.69	0.7549	1.29	0.9015	1.89	0.9706	2.49	0.9936
0.10	0.5398	0.70	0.7580	1.30	0.9032	1.90	0.9713	2.50	0.9938
0.11	0.5438	0.71	0.7611	1.31	0.9049	1.91	0.9719	2.51	0.9940
0.12	0.5478	0.72	0.7642	1.32	0.9066	1.92	0.9726	2.52	0.9941
0.13	0.5517	0.73	0.7673	1.33	0.9082	1.93	0.9732	2.53	0.9943
0.14	0.5557	0.74	0.7704	1.34	0.9099	1.94	0.9738	2.54	0.9945
0.15	0.5596	0.75	0.7734	1.35	0.9115	1.95	0.9744	2.55	0.9946
0.16	0.5636	0.76	0.7764	1.36	0.9131	1.96	0.9750	2.56	0.9948
0.17	0.5675	0.77	0.7794	1.37	0.9147	1.97	0.9756	2.57	0.9949
0.18	0.5714	0.78	0.7823	1.38	0.9162	1.98	0.9761	2.58	0.9951
0.19	0.5753	0.79	0.7852	1.39	0.9177	1.99	0.9767	2.59	0.9952
0.20	0.5793	0.80	0.7881	1.40	0.9192	2.00	0.9772	2.60	0.9953
0.21	0.5832	0.81	0.7910	1.41	0.9207	2.01	0.9778	2.61	0.9955
0.22	0.5871	0.82	0.7939	1.42	0.9222	2.02	0.9783	2.62	0.9956
0.23	0.5910	0.83	0.7967	1.43	0.9236	2.03	0.9788	2.63	0.9957
0.24	0.5948	0.84	0.7995	1.44	0.9251	2.04	0.9793	2.64	0.9959
0.25	0.5987	0.85	0.8023	1.45	0.9265	2.05	0.9798	2.65	0.9960
0.26	0.6026	0.86	0.8051	1.46	0.9279	2.06	0.9803	2.66	0.9961
0.27	0.6064	0.87	0.8078	1.47	0.9292	2.07	0.9808	2.67	0.9962
0.28	0.6103	0.88	0.8106	1.48	0.9306	2.08	0.9812	2.68	0.9963
0.29	0.6141	0.89	0.8133	1.49	0.9319	2.09	0.9817	2.69	0.9964
0.30	0.6179	0.90	0.8159	1.50	0.9332	2.10	0.9821	2.70	0.9965
0.31	0.6217	0.91	0.8186	1.51	0.9345	2.11	0.9826	2.71	0.9966
0.32	0.6255	0.92	0.8212	1.52	0.9357	2.12	0.9830	2.72	0.9967
0.33	0.6293	0.93	0.8238	1.53	0.9370	2.13	0.9834	2.73	0.9968
0.34	0.6331	0.94	0.8264	1.54	0.9382	2.14	0.9838	2.74	0.9969
0.35	0.6368	0.95	0.8289	1.55	0.9394	2.15	0.9842	2.75	0.9970
0.36	0.6406	0.96	0.8315	1.56	0.9406	2.16	0.9846	2.76	0.9971
0.37	0.6443	0.97	0.8340	1.57	0.9418	2.17	0.9850	2.77	0.9972
0.38	0.6480	0.98	0.8365	1.58	0.9429	2.18	0.9854	2.78	0.9973
0.39	0.6517	0.99	0.8389	1.59	0.9441	2.19	0.9857	2.79	0.9974
0.40	0.6554	1.00	0.8413	1.60	0.9452	2.20	0.9861	2.80	0.9974
0.41	0.6591	1.01	0.8438	1.61	0.9463	2.21	0.9864	2.81	0.9975
0.42	0.6628	1.02	0.8461	1.62	0.9474	2.22	0.9868	2.82	0.9976
0.43	0.6664	1.03	0.8485	1.63	0.9484	2.23	0.9871	2.83	0.9977
0.44	0.6700	1.04	0.8508	1.64	0.9495	2.24	0.9875	2.84	0.9977
0.45	0.6736	1.05	0.8531	1.65	0.9505	2.25	0.9878	2.85	0.9978
0.46	0.6772	1.06	0.8554	1.66	0.9515	2.26	0.9881	2.86	0.9979
0.47	0.6808	1.07	0.8577	1.67	0.9525	2.27	0.9884	2.87	0.9979
0.48	0.6844	1.08	0.8599	1.68	0.9535	2.28	0.9887	2.88	0.9980
0.49	0.6879	1.09	0.8621	1.69	0.9545	2.29	0.9890	2.89	0.9981
0.50	0.6915	1.10	0.8643	1.70	0.9554	2.30	0.9893	2.90	0.9981
0.51	0.6950	1.11	0.8665	1.71	0.9564	2.31	0.9896	2.91	0.9982
0.52	0.6985	1.12	0.8686	1.72	0.9573	2.32	0.9898	2.92	0.9982
0.53	0.7019	1.13	0.8708	1.73	0.9582	2.33	0.9901	2.93	0.9983
0.54	0.7054	1.14	0.8729	1.74	0.9591	2.34	0.9904	2.94	0.9984
0.55	0.7088	1.15	0.8749	1.75	0.9599	2.35	0.9906	2.95	0.9984
0.56	0.7123	1.16	0.8770	1.76	0.9608	2.36	0.9909	2.96	0.9985
0.57	0.7157	1.17	0.8790	1.77	0.9616	2.37	0.9911	2.97	0.9985
0.58	0.7190	1.18	0.8810	1.78	0.9625	2.38	0.9913	2.98	0.9986
0.59	0.7224	1.19	0.8830	1.79	0.9633	2.39	0.9916	2.99	0.9986
0.60	0.7257	1.20	0.8849	1.80	0.9641	2.40	0.9918	3.00	0.9987

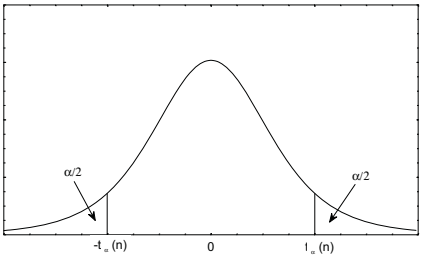
Tabla III.1: Distribución normal $N(0,1)$. Cuantiles.



α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
0.01	2.576	0.21	1.254	0.41	0.824	0.61	0.510	0.81	0.240
0.02	2.326	0.22	1.227	0.42	0.806	0.62	0.496	0.82	0.228
0.03	2.170	0.23	1.200	0.43	0.789	0.63	0.482	0.83	0.215
0.04	2.054	0.24	1.175	0.44	0.772	0.64	0.468	0.84	0.202
0.05	1.960	0.25	1.150	0.45	0.755	0.65	0.454	0.85	0.189
0.06	1.881	0.26	1.126	0.46	0.739	0.66	0.440	0.86	0.176
0.07	1.812	0.27	1.103	0.47	0.722	0.67	0.426	0.87	0.164
0.08	1.751	0.28	1.080	0.48	0.706	0.68	0.412	0.88	0.151
0.09	1.695	0.29	1.058	0.49	0.690	0.69	0.399	0.89	0.138
0.10	1.645	0.30	1.036	0.50	0.674	0.70	0.385	0.90	0.126
0.11	1.598	0.31	1.015	0.51	0.659	0.71	0.372	0.91	0.113
0.12	1.555	0.32	0.994	0.52	0.643	0.72	0.358	0.92	0.100
0.13	1.514	0.33	0.974	0.53	0.628	0.73	0.345	0.93	0.088
0.14	1.476	0.34	0.954	0.54	0.613	0.74	0.332	0.94	0.075
0.15	1.440	0.35	0.935	0.55	0.598	0.75	0.319	0.95	0.063
0.16	1.405	0.36	0.915	0.56	0.583	0.76	0.305	0.96	0.050
0.17	1.372	0.37	0.896	0.57	0.568	0.77	0.292	0.97	0.038
0.18	1.341	0.38	0.878	0.58	0.553	0.78	0.279	0.98	0.025
0.19	1.311	0.39	0.860	0.59	0.539	0.79	0.266	0.99	0.013
0.20	1.282	0.40	0.842	0.60	0.524	0.80	0.253	1.00	0.000

Distribución t-Student

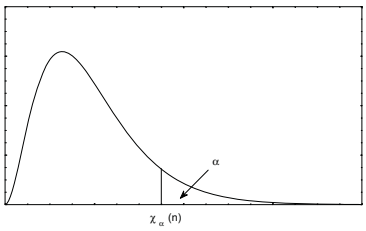
Tabla IV: Distribución t de Student



$t_\alpha(n)$	α							
n	0.5	0.2	0.1	0.05	0.025	0.02	0.01	0.001
1	1.000	3.078	6.314	12.706	25.452	31.821	63.656	636.611
2	0.816	1.886	2.920	4.303	6.205	6.965	9.925	31.602
3	0.765	1.638	2.353	3.182	4.177	4.541	5.841	12.923
4	0.741	1.533	2.132	2.776	3.495	3.747	4.604	8.610
5	0.727	1.476	2.015	2.571	3.163	3.365	4.032	6.869
6	0.718	1.440	1.943	2.447	2.969	3.143	3.707	5.959
7	0.711	1.415	1.895	2.365	2.841	2.998	3.499	5.408
8	0.706	1.397	1.860	2.306	2.752	2.896	3.355	5.041
9	0.703	1.383	1.833	2.262	2.685	2.821	3.250	4.781
10	0.700	1.372	1.812	2.228	2.634	2.764	3.169	4.587
11	0.697	1.363	1.796	2.201	2.593	2.718	3.106	4.437
12	0.695	1.356	1.782	2.179	2.560	2.681	3.055	4.318
13	0.694	1.350	1.771	2.160	2.533	2.650	3.012	4.221
14	0.692	1.345	1.761	2.145	2.510	2.624	2.977	4.140
15	0.691	1.341	1.753	2.131	2.490	2.602	2.947	4.073
16	0.690	1.337	1.746	2.120	2.473	2.583	2.921	4.015
17	0.689	1.333	1.740	2.110	2.458	2.567	2.898	3.965
18	0.688	1.330	1.734	2.101	2.445	2.552	2.878	3.922
19	0.688	1.328	1.729	2.093	2.433	2.539	2.861	3.883
20	0.687	1.325	1.725	2.086	2.423	2.528	2.845	3.850
21	0.686	1.323	1.721	2.080	2.414	2.518	2.831	3.819
22	0.686	1.321	1.717	2.074	2.405	2.508	2.819	3.792
23	0.685	1.319	1.714	2.069	2.398	2.500	2.807	3.768
24	0.685	1.318	1.711	2.064	2.391	2.492	2.797	3.745
25	0.684	1.316	1.708	2.060	2.385	2.485	2.787	3.725
26	0.684	1.315	1.706	2.056	2.379	2.479	2.779	3.707
27	0.684	1.314	1.703	2.052	2.373	2.473	2.771	3.690
28	0.683	1.313	1.701	2.048	2.368	2.467	2.763	3.674
29	0.683	1.311	1.699	2.045	2.364	2.462	2.756	3.659
30	0.683	1.310	1.697	2.042	2.360	2.457	2.750	3.646
35	0.682	1.306	1.690	2.030	2.342	2.438	2.724	3.591
40	0.681	1.303	1.684	2.021	2.329	2.423	2.704	3.551
45	0.680	1.301	1.679	2.014	2.319	2.412	2.690	3.520
50	0.679	1.299	1.676	2.009	2.311	2.403	2.678	3.496
55	0.679	1.297	1.673	2.004	2.304	2.396	2.668	3.476
60	0.679	1.296	1.671	2.000	2.299	2.390	2.660	3.460
65	0.678	1.295	1.669	1.997	2.295	2.385	2.654	3.447
70	0.678	1.294	1.667	1.994	2.291	2.381	2.648	3.435
75	0.678	1.293	1.665	1.992	2.287	2.377	2.643	3.425
80	0.678	1.292	1.664	1.990	2.284	2.374	2.639	3.416
85	0.677	1.292	1.663	1.988	2.282	2.371	2.635	3.409
90	0.677	1.291	1.662	1.987	2.280	2.368	2.632	3.402
95	0.677	1.291	1.661	1.985	2.277	2.366	2.629	3.396
100	0.677	1.290	1.660	1.984	2.276	2.364	2.626	3.390
∞	0.674	1.282	1.645	1.960	2.241	2.326	2.576	3.291

Distribución χ^2

Tabla V: Distribución chi-cuadrado $\chi^2(n)$



$\chi^2_\alpha(n)$	α												
n	0.001	0.005	0.01	0.025	0.05	0.1	0.2	0.8	0.9	0.95	0.975	0.99	0.995
1	10.83	7.88	6.64	5.02	3.84	2.71	1.64	0.06	0.02	0.00	0.00	0.00	0.00
2	13.82	10.60	9.21	7.38	5.99	4.61	3.22	0.45	0.21	0.10	0.05	0.02	0.01
3	16.27	12.84	11.35	9.35	7.82	6.25	4.64	1.01	0.58	0.35	0.22	0.11	0.07
4	18.47	14.86	13.28	11.14	9.49	7.78	5.99	1.65	1.06	0.71	0.48	0.30	0.21
5	20.52	16.75	15.09	12.83	11.07	9.24	7.29	2.34	1.61	1.15	0.83	0.55	0.41
6	22.46	18.55	16.81	14.45	12.59	10.65	8.56	3.07	2.20	1.64	1.24	0.87	0.68
7	24.33	20.28	18.48	16.01	14.07	12.02	9.80	3.82	2.83	2.17	1.69	1.24	0.99
8	26.13	21.96	20.09	17.54	15.51	13.36	11.03	4.59	3.49	2.73	2.18	1.65	1.34
9	27.88	23.59	21.67	19.03	16.92	14.68	12.24	5.38	4.17	3.32	2.70	2.09	1.73
10	29.59	25.19	23.21	20.49	18.31	15.99	13.44	6.18	4.87	3.94	3.25	2.56	2.16
11	31.27	26.76	24.73	21.92	19.68	17.28	14.63	6.99	5.58	4.57	3.82	3.05	2.60
12	32.92	28.30	26.22	23.34	21.03	18.55	15.81	7.81	6.30	5.23	4.40	3.57	3.07
13	34.54	29.83	27.69	24.74	22.36	19.81	16.99	8.63	7.04	5.89	5.01	4.11	3.56
14	36.13	31.33	29.15	26.12	23.69	21.07	18.15	9.47	7.79	6.57	5.63	4.66	4.07
15	37.71	32.81	30.58	27.49	25.00	22.31	19.31	10.31	8.55	7.26	6.26	5.23	4.60
16	39.26	34.27	32.01	28.85	26.30	23.54	20.47	11.15	9.31	7.96	6.91	5.81	5.14
17	40.80	35.73	33.41	30.20	27.59	24.77	21.62	12.00	10.08	8.67	7.56	6.41	5.70
18	42.32	37.16	34.81	31.53	28.87	25.99	22.76	12.86	10.86	9.39	8.23	7.01	6.26
19	43.83	38.59	36.20	32.86	30.15	27.21	23.90	13.72	11.65	10.12	8.91	7.63	6.84
20	45.33	40.01	37.57	34.18	31.41	28.41	25.04	14.58	12.44	10.85	9.59	8.26	7.43
21	46.81	41.41	38.94	35.48	32.68	29.62	26.17	15.44	13.24	11.59	10.28	8.90	8.03
22	48.28	42.81	40.30	36.79	33.93	30.82	27.30	16.31	14.04	12.34	10.98	9.54	8.64
23	49.74	44.19	41.65	38.08	35.18	32.01	28.43	17.19	14.85	13.09	11.69	10.19	9.26
24	51.19	45.57	42.99	39.37	36.42	33.20	29.56	18.06	15.66	13.85	12.40	10.86	9.89
25	52.64	46.94	44.32	40.65	37.66	34.39	30.68	18.94	16.47	14.61	13.12	11.52	10.52
26	54.07	48.30	45.65	41.93	38.89	35.57	31.80	19.82	17.29	15.38	13.84	12.20	11.16
27	55.49	49.66	46.97	43.20	40.12	36.75	32.91	20.70	18.11	16.15	14.57	12.88	11.81
28	56.91	51.01	48.29	44.47	41.34	37.92	34.03	21.59	18.94	16.93	15.31	13.56	12.46
29	58.32	52.35	49.60	45.73	42.56	39.09	35.14	22.47	19.77	17.71	16.05	14.25	13.12
30	59.72	53.69	50.90	46.99	43.78	40.26	36.25	23.36	20.60	18.49	16.79	14.95	13.78
40	73.43	66.79	63.71	59.35	55.77	51.81	47.27	32.34	29.05	26.51	24.43	22.16	20.70
50	86.69	79.52	76.18	71.44	67.52	63.18	58.17	41.45	37.68	34.76	32.35	29.70	27.99
60	99.65	91.98	88.41	83.32	79.10	74.41	68.98	50.64	46.45	43.18	40.47	37.48	35.53
70	112.37	104.25	100.46	95.05	90.55	85.54	79.72	59.89	55.32	51.73	48.75	45.43	43.26
80	124.90	116.37	112.37	106.66	101.90	96.60	90.42	69.20	64.27	60.38	57.14	53.53	51.16
90	137.28	128.35	124.16	118.17	113.17	107.59	101.07	78.55	73.28	69.11	65.63	61.74	59.18
100	149.53	140.23	135.86	129.60	124.37	118.52	111.68	87.94	82.35	77.91	74.20	70.05	67.31

Nota: La tabla proporciona el valor $\chi^2_\alpha(n)$ que deja a su derecha área α bajo la densidad de la distribución χ^2 con n grados de libertad.

Distribución F-Snedecor

Tabla VI: Tablas de la distribución $F(m,n)$

Tabla VI.2: Distribución $F(m,n)$. $\alpha = 0.05$. 1ª fila= g.l. denominador. 1ª columna= g.l. numerador.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	161.45	18.51	10.13	7.71	6.61	5.99	5.59	5.32	5.12	4.96	4.84	4.75	4.67	4.60	4.54	4.49	4.45	4.41	4.38	4.35
2	199.50	19.00	9.55	6.94	5.79	5.14	4.74	4.46	4.26	4.10	3.98	3.89	3.81	3.74	3.68	3.63	3.59	3.55	3.52	3.49
3	215.71	19.16	9.28	6.59	5.41	4.76	4.35	4.07	3.86	3.71	3.59	3.49	3.41	3.34	3.29	3.24	3.20	3.16	3.13	3.10
4	224.58	19.25	9.12	6.39	5.19	4.53	4.12	3.84	3.63	3.48	3.36	3.26	3.18	3.11	3.06	3.01	2.96	2.93	2.90	2.87
5	230.16	19.30	9.01	6.26	5.05	4.39	3.97	3.69	3.48	3.33	3.20	3.11	3.03	2.96	2.90	2.85	2.81	2.77	2.74	2.71
6	233.99	19.33	8.94	6.16	4.95	4.28	3.87	3.58	3.37	3.22	3.09	3.00	2.92	2.85	2.79	2.74	2.70	2.66	2.63	2.60
7	236.77	19.35	8.89	6.09	4.88	4.21	3.79	3.50	3.29	3.14	3.01	2.91	2.83	2.76	2.71	2.66	2.61	2.58	2.54	2.51
8	238.88	19.37	8.85	6.04	4.82	4.15	3.73	3.44	3.23	3.07	2.95	2.85	2.77	2.70	2.64	2.59	2.55	2.51	2.48	2.45
9	240.54	19.38	8.81	6.00	4.77	4.10	3.68	3.39	3.18	3.02	2.90	2.80	2.71	2.65	2.59	2.54	2.49	2.46	2.42	2.39
10	241.88	19.40	8.79	5.96	4.74	4.06	3.64	3.35	3.14	2.98	2.85	2.75	2.67	2.60	2.54	2.49	2.45	2.41	2.38	2.35
11	242.98	19.40	8.76	5.94	4.70	4.03	3.60	3.31	3.10	2.94	2.82	2.72	2.63	2.57	2.51	2.46	2.41	2.37	2.34	2.31
12	243.91	19.41	8.74	5.91	4.68	4.00	3.57	3.28	3.07	2.91	2.79	2.69	2.60	2.53	2.48	2.42	2.38	2.34	2.31	2.28
13	244.69	19.42	8.73	5.89	4.66	3.98	3.55	3.26	3.05	2.89	2.76	2.66	2.58	2.51	2.45	2.40	2.35	2.31	2.28	2.25
14	245.36	19.42	8.71	5.87	4.64	3.96	3.53	3.24	3.03	2.86	2.74	2.64	2.55	2.48	2.42	2.37	2.33	2.29	2.26	2.22
15	245.95	19.43	8.70	5.86	4.62	3.94	3.51	3.22	3.01	2.85	2.72	2.62	2.53	2.46	2.40	2.35	2.31	2.27	2.23	2.20
16	246.46	19.43	8.69	5.84	4.60	3.92	3.49	3.20	2.99	2.83	2.70	2.60	2.51	2.44	2.38	2.33	2.29	2.25	2.21	2.18
17	246.92	19.44	8.68	5.83	4.59	3.91	3.48	3.19	2.97	2.81	2.69	2.58	2.50	2.43	2.37	2.32	2.27	2.23	2.20	2.17
18	247.32	19.44	8.67	5.82	4.58	3.90	3.47	3.17	2.96	2.80	2.67	2.57	2.48	2.41	2.35	2.30	2.26	2.22	2.18	2.15
19	247.69	19.44	8.67	5.81	4.57	3.88	3.46	3.16	2.95	2.79	2.66	2.56	2.47	2.40	2.34	2.29	2.24	2.20	2.17	2.14
20	248.01	19.45	8.66	5.80	4.56	3.87	3.44	3.15	2.94	2.77	2.65	2.54	2.46	2.39	2.33	2.28	2.23	2.19	2.16	2.12
21	248.31	19.45	8.65	5.79	4.55	3.86	3.43	3.14	2.93	2.76	2.64	2.53	2.45	2.38	2.32	2.26	2.22	2.18	2.14	2.11
22	248.58	19.45	8.65	5.79	4.54	3.86	3.43	3.13	2.92	2.75	2.63	2.52	2.44	2.37	2.31	2.25	2.21	2.17	2.13	2.10
23	248.83	19.45	8.64	5.78	4.53	3.85	3.42	3.12	2.91	2.75	2.62	2.51	2.43	2.36	2.30	2.24	2.20	2.16	2.12	2.09
24	249.05	19.45	8.64	5.77	4.53	3.84	3.41	3.12	2.90	2.74	2.61	2.51	2.42	2.35	2.29	2.24	2.19	2.15	2.11	2.08
25	249.26	19.46	8.63	5.77	4.52	3.83	3.40	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.28	2.23	2.18	2.14	2.11	2.07
26	249.45	19.46	8.63	5.76	4.52	3.83	3.40	3.10	2.89	2.72	2.59	2.49	2.41	2.33	2.27	2.22	2.17	2.13	2.10	2.07
27	249.63	19.46	8.63	5.76	4.51	3.82	3.39	3.10	2.88	2.72	2.59	2.48	2.40	2.33	2.27	2.21	2.17	2.13	2.09	2.06
28	249.80	19.46	8.62	5.75	4.50	3.82	3.39	3.09	2.87	2.71	2.58	2.48	2.39	2.32	2.26	2.21	2.16	2.12	2.08	2.05
29	249.95	19.46	8.62	5.75	4.50	3.81	3.38	3.08	2.87	2.70	2.58	2.47	2.39	2.31	2.25	2.20	2.15	2.11	2.08	2.05
30	250.09	19.46	8.62	5.75	4.50	3.81	3.38	3.08	2.86	2.70	2.57	2.47	2.38	2.31	2.25	2.19	2.15	2.11	2.07	2.04
40	251.14	19.47	8.59	5.72	4.46	3.77	3.34	3.04	2.83	2.66	2.53	2.43	2.34	2.27	2.20	2.15	2.10	2.06	2.03	1.99
50	251.77	19.48	8.58	5.70	4.44	3.75	3.32	3.02	2.80	2.64	2.51	2.40	2.31	2.24	2.18	2.12	2.08	2.04	2.00	1.97
60	252.20	19.48	8.57	5.69	4.43	3.74	3.30	3.01	2.79	2.62	2.49	2.38	2.30	2.22	2.16	2.11	2.06	2.02	1.98	1.95
70	252.50	19.48	8.57	5.68	4.42	3.73	3.29	2.99	2.78	2.61	2.48	2.37	2.28	2.21	2.15	2.09	2.05	2.00	1.97	1.93
80	252.72	19.48	8.56	5.67	4.41	3.72	3.29	2.99	2.77	2.60	2.47	2.36	2.27	2.20	2.14	2.08	2.03	1.99	1.96	1.92
90	252.90	19.48	8.56	5.67	4.41	3.72	3.28	2.98	2.76	2.59	2.46	2.36	2.27	2.19	2.13	2.07	2.03	1.98	1.95	1.91
100	253.04	19.49	8.55	5.66	4.41	3.71	3.27	2.97	2.76	2.59	2.46	2.35	2.26	2.19	2.12	2.07	2.02	1.98	1.94	1.91
120	253.25	19.49	8.55	5.66	4.40	3.70	3.27	2.97	2.75	2.58	2.45	2.34	2.25	2.18	2.11	2.06	2.01	1.97	1.93	1.90
∞	254.30	19.50	8.53	5.63	4.37	3.67	3.23	2.93	2.71	2.54	2.41	2.30	2.21	2.13	2.07	2.01	1.96	1.92	1.88	1.84

Bibliografía recomendada

- **M. Andrés y Juan de Luna.** (2007) Bioéstadística para las ciencias de la Salud. Ed. Norma.
- **M. Andrés y Juan de Luna.** (1995) 50 ± 10 horas de Bioestadística. Ed. Norma.
- **F. Carmona.**(2005) Modelos Lineales. Ed. e-UMAB.
- **E. Cobo, P. Muñoz y J.A. González.**(2007) Bioestadística para no estadísticos. Ed. Elsewier/Masson.
- **Macía Antón, Lubin y Rubio de Lemus.** (1997) Psicología Matemática. UNED.
- **J. S. Milton.** Estadística para Biología y Ciencias de la Salud. Ed. Interamericana. McGraw-Hill.
- **A.G. Nogales.** (2004) Bioestadística Básica. Ed. abecedario.
- **Norman y Steiner** (1996) Bioestadística Ed. Mosby/Doyma Libros.
- **B. Visauta.** (1998) Análisis estadístico con SPSS para Windows. Ed. McGraw Hill.
- http://www.hrc.es/bioest/M_docente.html#tema3. Hospital Ramón y Cajal