Universitat d'Alacant
Universidad de Alicante

Estadística básica en Ciencias de la Salud

Andreu Nolasco Bonmatí

Joaquín Moncho Vasallo

Introducción

En este texto se presentan contenidos básicos de estadística para el entorno de las ciencias de la salud. Proviene de la experiencia docente de sus autores, profesores en asignaturas, cursos y seminarios impartidos en el entorno de las ciencias de la salud, dirigidos a profesionales que desarrollan su labor asistencial, de gestión, de docencia o de investigación en este ámbito.

Aunque la Estadística ha estado presente en la mayoría de titulaciones universitarias de Ciencias de la Salud desde hace décadas, es desde su inclusión como materia básica en los planes de estudio de las titulaciones universitarias de grado incluidas en la rama de conocimiento de Ciencias de la Salud cuando ha venido a consolidar su presencia en todas las titulaciones de esta rama, tal como había sido recomendado por la Organización Mundial de la Salud, al recoger una necesidad derivada de la creciente utilización de la estadística por parte de los profesionales de las Ciencias de la Salud en todas las parcelas de su campo de actuación. De esta forma, la Estadística habría pasado a desempeñar un papel básico en la formación de estos profesionales.

Si partimos de la finalidad de elaborar materiales docentes que puedan ser útiles a una buena parte de programas formativos en asignaturas de estadística en Ciencias de la Salud, habría que reflexionar, en primer lugar, sobre aquellos contenidos básicos que deberían incorporarse a los programas de las asignaturas a impartir. La Estadística es una ciencia que ha venido construyendo métodos y procedimientos para dar respuesta a diferentes problemas suscitados por otras Ciencias. En particular, las Ciencias de la Salud plantean un escenario y sujetos de trabajo particulares (la salud del ser humano). Así, las funciones asistencial y de gestión, tanto en su vertiente clínica como comunitaria, requieren herramientas básicas para manejar información generada en los puestos de trabajo (manejo e interpretación de la información de uno o varios sujetos, organización de datos, comprensión de información cuantitativa, redacción básica de informes,...), y para el acceso a información procedente de fuentes secundarias (lectura de artículos científicos, libros, informes, guías clínicas,...). Por su parte, la función de investigación requeriría de mayor profundización en aspectos esenciales del método científico en los que la estadística juega un papel importante cuando se utilizan métodos cuantitativos de investigación, debiendo profundizar en contenidos relacionados con la inferencia, útiles para la correcta valoración de las conclusiones obtenidas, y métodos específicos de estadística, con lo que responder preguntas de investigación más complejas. Así mismo, la labor docente del profesional de las ciencias de la salud requiere de unos conocimientos mínimos para poder trasmitir con rigor la información cuando ésta ha sido tratada con métodos estadísticos.
El siguiente paso trataría de delimitar el nivel de profundización en los contenidos elegidos, por ejemplo ¿qué nivel de profundización matemática alcanzar?. La respuesta está en alcanzar un equilibrio entre la cantidad y complejidad del lenguaje matemático utilizado y la exposición de ideas, problemas e interpretaciones en el lenguaje más próximo e intuitivo para el alumno.
Un último paso requeriría la exposición de los contenidos sobre ejemplos pertinentes y atractivos para el entorno profesional actual o futuro del alumno. La experiencia de los autores en investigación aplicada en el entorno de las Ciencias de la Salud ha guiado la elección de diversos problemas y situaciones para ejemplificar los contenidos de este texto.

Los materiales que se presentan en este documento parten de la reflexión sobre todos estos aspectos por parte de sus autores así como de la experiencia docente acumulada por los mismos durante varias décadas de contacto con alumnado en formación para ser futuros profesionales de las Ciencias de la Salud o profesionales en ejercicio con deseos de reciclar o aumentar sus conocimientos en estadística. Están organizados en cinco capítulos, respondiendo a la estructura habitual en los programas de esta materia.

El primer capítulo, 'Explorado los datos: Organización, descripción y presentación de la información' recoge elementos básicos de estadística descriptiva que ayuden a organizar, comprender e interpretar la información de un conjunto de datos . El segundo capítulo, 'El azar sobre nuestros datos: Midiendo la probabilidad', introduce conceptos, propiedades y teoremas básicos de la probabilidad de interés y uso frecuente en el ámbito de la salud. Los capítulos tres, 'Fundamentos para extraer conclusiones de los datos: La inferencia estadística', y cuatro, '¿Qué prueba estadística utilizamos? Selección de pruebas estadísticas

de inferencia: Aplicaciones básicas', se dirigen a presentar los procedimientos básicos generales de inferencia y una selección específica para dar respuesta a la mayoría de situaciones básicas de análisis inferencial de los datos. Por último, el capítulo cinco, 'Precauciones en el análisis de los datos', se revisan brevemente algunos tópicos y/o elementos a tener especialmente en cuenta en el análisis de los datos. Se ha incorporado un anexo que incluye las tablas de probabilidad de los modelos más utilizados.

Estos materiales han venido mostrando su utilidad como soporte docente en la mayor parte de las asignaturas impartidas en las titulaciones de grado o máster de Ciencias de la Salud de la Universidad de Alicante, como Enfermería, Nutrición, Investigación de Ciencias de la Salud, Salud Pública, Óptica,.., y en numerosos cursos de posgrado impartidos por los autores. Se trataría de un material básico en el sentido de que incluye los contenidos mínimos generalistas en estadística que un profesional de las Ciencias de la Salud debería recibir en su formación de grado para adquirir las competencias incluidas en su plan de estudios, especialmente las más relacionadas con esta materia. No obstante, cabe mencionar que el encaje de las diferentes disciplinas en los planes de estudio de las titulaciones o algunas características específicas de los mismos hace que no todos los contenidos aquí recogidos se correspondan con los contenidos presenciales impartidos en el aula, debiendo realizar en algunos casos una selección de los mismos.

Finalmente, hay que tener en cuenta que el carácter básico de estos materiales requerirá de su ampliación en programas formativos de mayor profundidad, generalmente vinculados a mayores necesidades para la investigación, como es el caso de algunos másters. En este caso, estos materiales deben ser complementados con otros generales o específicos que desarrollen métodos de mayor complejidad y sofisticación, generalmente métodos multivariantes probabilísticos (modelos lineales como regresión lineal múltiple, regresión logística, regresión de Poisson, de Cox, etc.), no probabilísticos (análisis factorial, análisis de cluster, etc.) u otros.

Alicante, otoño de 2016

Los autores

Andreu Nolasco Bonmatí y Joaquín Moncho Vasallo son profesores del Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia de la Universidad de Alicante en el que han venido desarrollando su labor como docentes e investigadores. Han impartido docencia en diversas titulaciones de Ciencias de la Salud (Medicina, Enfermería, Nutrición humana y dietética, Óptica, etc.) tanto en estudios de grado como en posgrado (máster y doctorado), en materias y/o asignaturas como Bioestadística, Estadística Avanzada, Demografía y Salud, Metodología de la Investigación, Desigualdades en Salud, Análisis de la mortalidad, etc.. Han venido desarrollando investigación en líneas como: Análisis de la Mortalidad, Geografía Sanitaria, Estadísticas Sanitarias, Encuestas de salud, Demografía y salud, Desigualdades en salud y otras. Su experiencia en la aplicación del método estadístico en el entorno de las Ciencias de la Salud proviene y se refleja en la dirección de numerosos proyectos de investigación, tesis doctorales, publicaciones científicas y el continuo contacto con el contexto sanitario a través del asesoramiento metodológico a diversas instituciones sanitarias (Administración sanitaria, Centros de Salud y Salud Pública, Hospitales y otras).

Índice

C A P I T U L O 1

EXPLORANDO LOS DATOS: ORGANIZACIÓN DESCRIPCIÓN Y PRESENTACIÓN DE LA INFORMACIÓN

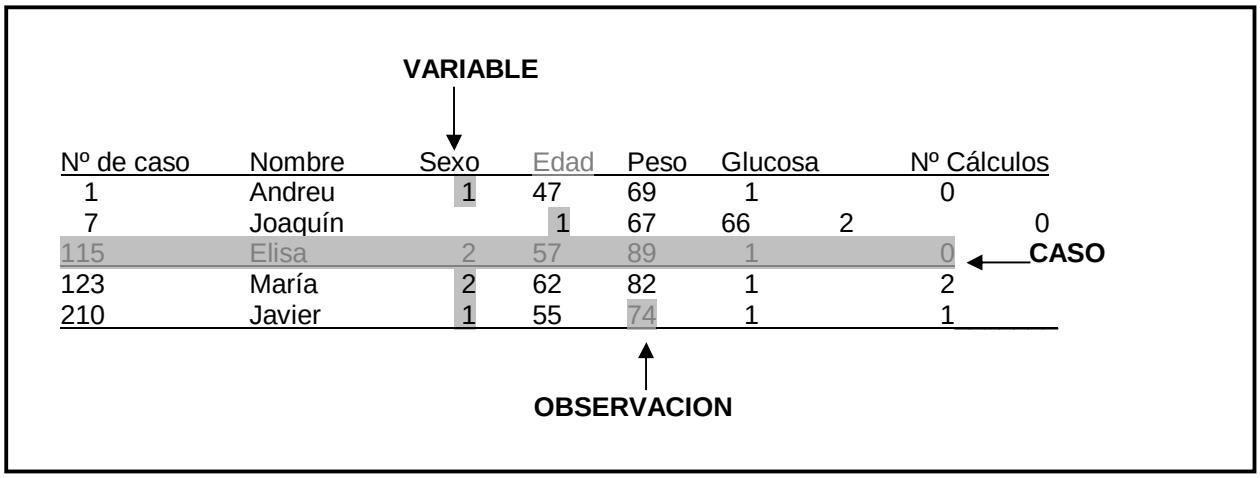
El análisis estadístico juega un papel esencial en el tratamiento de la información procedente de una investigación, abarcando diversos aspectos relativos a su descripción, o a la extracción de conclusiones y generalización de éstas que podamos realizar. A lo largo de cuatro capítulos se abordarán las técnicas y procesos básicos que permitirán analizar estadísticamente la información de un conjunto de datos. El problema a resolver en el presente capítulo se centra en el proceso de organización, descripción y presentación de la información necesaria para el estudio de las características de interés. Nos referiremos a esa información como el conjunto o *base de datos*. La calidad de las aplicaciones estadísticas realizadas dependerá en gran medida de un manejo correcto de tal información y una identificación adecuada de los diversos elementos estadísticos que la conforman, pudiendo extraer conclusiones erróneas como consecuencia de una identificación incorrecta.

1.1 ELEMENTOS BASICOS EN EL ANÁLISIS ESTADÍSTICO

Definiremos el concepto de base de datos como una colección de resultados de diversas características, estructurados de acuerdo con algún objetivo particular. Es muy frecuente que la gran cantidad de información generada en el entorno sanitario dé lugar a la configuración de diferentes bases de datos. Así, por ejemplo, la información de las historias clínicas de los pacientes atendidos en un servicio hospitalario pueden presentar diferentes estructuras según la finalidad, administrativa, de investigación, con vistas al tratamiento,...
Para definir los elementos estadísticos básicos nos basaremos en la información contenida en el cuadro 1.1, obtenida sobre cinco individuos en un estudio para determinar la presencia de cálculos biliares y otras características.

La información que se presenta en este cuadro está estructurada en forma de base de datos, en la que cada línea contiene los datos de cada uno de los individuos del estudio. Desde el punto de vista estadístico, una base de datos contiene diversos elementos estadísticos, a saber:

Individuo o elemento: La base de datos contiene información sobre las características de una colección de unidades que denominaremos individuos o elementos (pueden ser personas o cosas). En el cuadro 1.1, los elementos son cada una de las personas de las que se recoge información.



Cuadro 1.1.- Datos estructurados en forma de base de datos

Variable: Es una característica de interés sobre un elemento. Es esencial que sus resultados cambien entre diferentes elementos, pues de no hacerlo estaríamos ante una *constante*. En el cuadro 1.1, la edad es una variable. Según la forma en que las variables presentan sus resultados, éstas pueden ser *cualitativas* (presentan sus resultados en forma de estados o categorías, como en el caso del sexo, hombres-mujeres), o *cuantitativas* (presentan sus resultados en forma de valores numéricos, como en el caso de la edad o el número de cálculos). No obstante, en función de los posibles resultados, las variables cuantitativas pueden ser *continuas*, cuando éstos pueden ser cualquier valor numérico entre dos dados, pudiendo alcanzar un número infinito de valores distintos, o *discretas*, cuando sus resultados surgen, habitualmente, de un proceso de recuento, pudiendo tomar un número finito o infinito numerable de valores. En el ejemplo, la edad es una variable cuantitativa continua, puesto que cualquier edad entre dos dadas es válido, pudiendo medirla en las unidades que estimemos oportunas (años, meses, días,...). El número de cálculos es un ejemplo de variable cuantitativa discreta, presentando un número de resultados distintos que es numerable (0,1,2,3,...), y supuestamente finito. Un ejemplo de variable discreta, con un número de resultados infinito numerable puede ser el número de tomas de temperatura corporal realizadas con un termómetro clínico, 0 (se rompe en la primera), 1, 2, 3,...., ∞ (desconocemos el número, pudiendo ser tan grande como queramos).
Es frecuente convertir variables cuantitativas en cualitativas, agrupando los valores numéricos en diferentes categorías. Así, podríamos convertir el número de cálculos en una variable de dos categorías: 'No tiene' (se corresponde con el valor 0), 'Tiene' (se corresponde con los valores restantes, 0, 1, 2,...); de hecho, la variable glucosa aparece categorizada según ésta sea mayor a 140 mg/100ml o menor o igual a 140 mg/100ml.

Observación: Es la información de una variable sobre un individuo de la base de datos. También es denominada medida, valor o resultado. Así, en el ejemplo, 74 es la observación de la variable edad en el individuo llamado Javier.

Caso o registro: Es el conjunto de observaciones correspondientes a un individuo de la base de datos. En el ejemplo, cada fila contiene toda la información de cada uno de los individuos de la base de datos, representando un caso o registro.

Tipos de datos: En la práctica, los datos que configuran la base de datos provienen de las observaciones de diferentes variables. Tales datos pueden ser de diferentes tipos. Así, los datos que surgen de un proceso de medición, representando cantidades, capacidades, o características similares pueden ser denominados *medidas*. En el cuadro 1.1, la edad o el peso son de este tipo. Los datos que surgen de un proceso de recuento o son frecuencias de ocurrencia de algún suceso se denominan *frecuenciales*. El número de cálculos biliares es de este tipo, pues surge del recuento de cálculos en cada individuo. Otro tipo de datos

surge de producir una ordenación en los individuos. Estos datos son denominados *ordinales*. El número de orden según la nota obtenida en un examen es de este tipo (1º, 2º, 3º,...). Por último, cuando cada individuo es asignado a una categoría de entre una colección de posibilidades, hablaremos de datos *categóricos*. Los datos del sexo son de este tipo, puesto que cada individuo es asignado a una de las dos posibles categorías de esta variable.

Población y muestra: Desde el punto de vista estadístico, resulta esencial caracterizar los datos incluidos en nuestra base en relación a los objetivos de nuestro estudio. Se trata de definir si la base de datos contiene las observaciones de *todos* los individuos sobre los que se pretende estudiar, caracterizar y extraer conclusiones acerca de las variables consideradas, en cuyo caso se dirá que se estudia a toda la población, o de *una parte* de ellos, en cuyo caso se dirá que se estudia una muestra. Así, una muestra será cualquier subconjunto de una población.

1.2 EXPLORANDO LOS DATOS: TABULACIONES Y GRÁFICOS

Una vez hemos recolectado los datos necesarios para nuestro estudio podemos proceder al análisis estadístico. Habitualmente, la primera fase de un estudio será la de inspeccionar la información contenida en los datos disponibles. La organización de los datos en una base, con las características que han sido descritas en el apartado anterior, es insuficiente. Para ello, y dada la dificultad para retener y manejar la información individual, los datos deben ser organizados y reducidos a estructuras o gráficos que nos permitan comprender su comportamiento, obteniendo adicionalmente medidas que, de una forma resumida y sintética, informen de los rasgos más destacables de las variables estudiadas. Este proceso suele ser denominado descripción de los datos. Así, se trata de contestar preguntas de interés acerca de los datos observados tales como:

- ¿Qué podemos decir acerca de la cantidad de hombres y mujeres que han formado parte del estudio? ¿Cuál es la categoría más frecuente?
- ¿Cuál es el porcentaje de individuos que consumen alcohol ocasionalmente? ¿Es diferente este porcentaje según el sexo de los individuos?
- ¿Cómo son las edades de los individuos del estudio? ¿Podemos construir una tabla donde se observe la distribución de las edades? ¿Es posible representar gráficamente esta distribución?
- Interesa caracterizar el comportamiento de la variable colesterol. Algunos investigadores apuntan la idea de que la mayor parte de los individuos tienden a alcanzar unos valores de colesterol *intermedios*, mientras que se hace más difícil encontrar individuos a medida que nos alejamos hacia valores *extremos*. ¿Es posible construir algún gráfico que permita visualizar su comportamiento y compararlo con este hipotético *patrón* o *modelo*?

Tabulaciones y distribuciones de frecuencias de una variable

Para comprender el comportamiento de una variable sobre un conjunto de datos se recurre a la organización de sus observaciones en una estructura que se denomina *tabla de distribución de frecuencias*. Ésta no es más que una tabla en la se disponen los diferentes valores o categorías de la variable a estudio acompañados de información sobre su frecuencia de aparición entre los individuos. Se trata de establecer a qué categoría pertenece o cuál es el valor alcanzado por cada uno de los individuos de la base de datos.

Para cualquier tipo de variable, la información básica que podemos incluir en la tabla está formada por las *frecuencias absolutas* (f_i) de los valores o categorías de las variables, obtenidas por recuento de los individuos que verifican cada uno de ellos, y las *frecuencias relativas* (fr_i) y *porcentajes* (p_i), medidas relativas de la frecuencia de aparición de cada valor o categoría entre los individuos de la base de datos. En el cuadro 1.2 observamos tales elementos sobre la variable sexo, obteniendo así la distribución de hombres y mujeres entre los individuos de un estudio. Las medidas relativas son independientes del número de

individuos de la base de datos, permitiendo la comparación con otras distribuciones de frecuencias obtenidas sobre otros conjuntos de datos.

SEXO	f_i	fr_i	p_i
Hombres	92	0,46	46
Mujeres	108	0,54	54
Total	200	1,00	100

$n = 200$

f_i = frecuencia absoluta categoría i

$$fr_i = \frac{f_i}{n} \quad ; \quad p_i = fr_i \times 100$$

Cuadro 1.2.- Tabla de distribución de frecuencias

La asignación de los individuos a las diferentes categorías o valores debe ser *exclusiva* y *exhaustiva*, es decir, cada uno de los individuos debe ser asignado a una y sólo una de las categorías o valores, mientras que todos los individuos deben ser asignados.

Cuando la variable es cuantitativa, suele ser estructurada de forma más sintética, pues la mayor variabilidad en sus observaciones, especialmente si la variable es continua, daría lugar a una tabla con información excesivamente detallada y con poca capacidad de resumen. La solución se obtiene tabulando la variable por intervalos. Se trata de definir una secuencia de intervalos de forma que un individuo será contabilizado en uno de ellos si el valor de la variable a estudio está incluido en él. Cada intervalo es definido a través de sus límites $[x_i , x_{i+1}[$, como el conjunto de valores de la variable, digamos X, tal que $x_i \leq X < x_{i+1}$. Con esta definición se dirá que los intervalos son abiertos por la derecha y cerrados por la izquierda.

Así, en el cuadro 1.3 encontramos la distribución de frecuencias de la variable edad agrupada en intervalos. El intervalo 20 - 25, p.ej., contiene todos aquellos individuos cuya edad es superior o igual a 20 e inferior a 25.

EDAD	f_i	p_i	P_i
20 - 25	13	6,5	6,5
25 - 30	26	13,0	19,5
30 - 35	14	7,0	26,5
35 - 40	16	8,0	34,5
40 - 45	13	6,5	41,0
45 - 50	20	10,0	51,0
50 - 55	17	8,5	59,5
55 - 60	25	12,5	72,0
60 - 65	21	10,5	82,5
65 - 70	24	12,0	94,5
70 - 75	11	5,5	100,0
Total	200	100	

Cuadro 1.3.- Distribución de frecuencias por intervalos

En algunos casos, el primer y último intervalo puede no verificar el procedimiento descrito para su construcción. Así, es frecuente ver expresiones ' < 25 ' o ' ≥ 70 ' en tales intervalos. Esta situación no es recomendable, salvo que sea estrictamente necesaria, puesto que cálculos posteriores pueden requerir del conocimiento de ambos límites. Otra situación frecuente es la de definir los intervalos para algunas variables, como es el caso de la edad, como 20 - 24, 25 - 29, 30 - 34,.....Esta definición es equivalente, en la práctica, a la establecida (hay que tener en cuenta que una persona tiene 24 años hasta que cumple los 25) pero puede resultar engañosa para cálculos posteriores, como por ejemplo para el cálculo del punto medio de un intervalo (el punto medio del intervalo 20-25 es 22,5, el mismo que para el intervalo 20-24 definido).

Por lo general, el último intervalo es considerado cerrado por la derecha. Como recomendación general, conviene que los intervalos tengan todos ellos la misma amplitud. En caso contrario, la interpretación de la

tabla puede resultar dificultosa y poco práctica.

Al igual que la amplitud, y relacionado con ella, el número de intervalos debe ser establecido en función del objetivo de la distribución de frecuencias. Como idea general debe prevalecer el hecho de que, a mayor número de intervalos, existirá una pérdida menor de información, pero a cambio la tabla resultante será menos práctica y manejable. Hay que tener en cuenta que los datos pueden ser clasificados en tantas categorías como valores distintos de la variable hasta un único intervalo en el que se encuentren todos los valores. Habitualmente, la mayoría de las situaciones son resueltas con un número de intervalos que puede oscilar entre 4 y 20.

Adicionalmente, si la variable es cuantitativa, la magnitud en que han ido apareciendo y agrupándose los individuos, a medida que nos referimos a edades mayores puede ser recogida a través de los *porcentajes acumulados* (P_i), obtenidos sumando acumulativamente los porcentajes. De igual forma pueden ser calculadas las frecuencias absolutas acumuladas, (F_i) y frecuencias relativas acumuladas (Fr_i) de los diferentes valores:

$$P_i = \sum_{j=1}^{i} p_j \quad ; \quad F_i = \sum_{j=1}^{i} f_j \quad ; \quad Fr_i = \sum_{j=1}^{i} fr_j$$

Gráficos sobre una variable

A través de la distribución de frecuencias es posible resumir el patrón de comportamiento de una variable. Sin embargo, la representación de este patrón en un gráfico puede ayudar a comprenderlo. Se entenderá por gráfico cualquier representación con símbolos, líneas, figuras geométricas o caracteres orientada a este fin. Desde este punto de vista, las representaciones gráficas posibles son muchas y muy diversas. Además, en la actualidad los paquetes de representación gráfica por ordenador ofrecen una gama muy amplia y variada de posibles representaciones (bidimensionales, tridimensionales, pictogramas,...). No obstante, podemos establecer algunos gráficos básicos para las diferentes variables.

Gráficos para variables cualitativas

La información disponible para este tipo de variables se concentra en sus categorías y las correspondientes frecuencias absolutas, relativas o porcentajes.

Las figuras 1.2 y 1.3 presentan las distribuciones de frecuencias de las variables sexo y consumo de alcohol obtenidas en un estudio sobre 200 individuos.

En la primera de ellas se ha representado los porcentajes de las diferentes categorías de la variable,

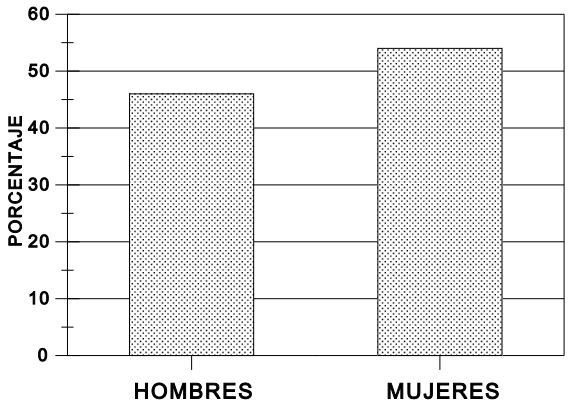


Figura 1.2.- Diagrama de barras

dibujando un paralelogramo hasta la altura correspondiente en cada una de ellas. Este tipo de representación gráfica recibe el nombre de *diagrama de barras*.

En la figura 1.3 se ha representado los porcentajes de las categorías de la variable consumo de alcohol, asignando un sector circular de forma proporcional al porcentaje alcanzado por cada una de ellas.
Este tipo de representación gráfica recibe el nombre de *diagrama de sectores*, aunque también es conocido como pastel o tarta.

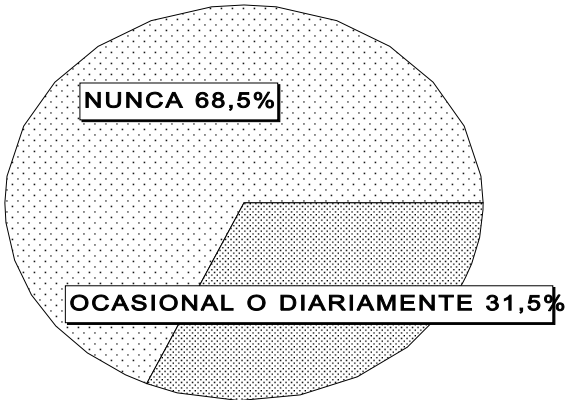


Figura 1.3.- Diagrama de sectores

Gráficos para variables cuantitativas

La información disponible para este tipo de variables es mayor que en el caso cualitativo. Por una parte dispondremos de información acerca de los valores de la variable, hecho que sugiere la utilización de alguna escala de medida para ellos. En segundo lugar, la información sobre la distribución de frecuencias de la variable puede ser
ampliada a través de los porcentajes y frecuencias absolutas o relativas acumuladas.

Las figuras 1.4 y 1.5 presentan la distribución de frecuencias de la variable edad, agrupando los valores en intervalos de 5 años de edad, y el nivel de colesterol, en intervalos de amplitud 40mg/100ml, respectivamente. Este gráfico recibe el nombre de *histograma.* En ambos casos, el eje de ordenadas representa los porcentajes de cada uno de los intervalos y en el eje de abcisas se ha ubicado los puntos medios de los intervalos, calculados como la semisuma de sus límites.

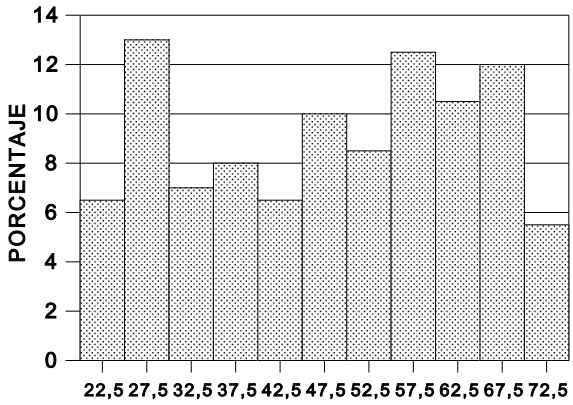


Figura 1.4.-Histograma para la edad

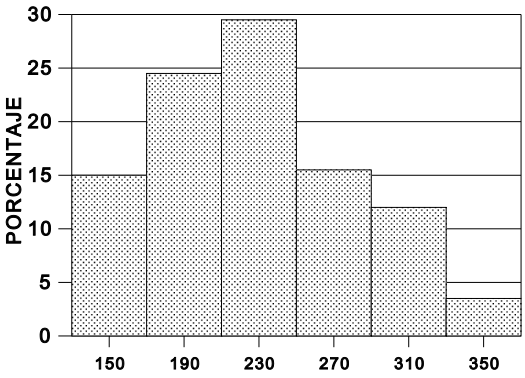


Figura 1.5.- Histograma para colesterol

Otro gráfico para una variable continua es el *polígono de frecuencias*. Su construcción requiere unir los puntos medios de los intervalos definidos, a la altura de su frecuencia absoluta, relativa o porcentaje. La figura 1.6 presenta el polígono de frecuencias de la variable colesterol.

La información sobre la distribución acumulada de una variable cuantitativa puede ser representada gráficamente a través del *polígono acumulativo de frecuencias*, representando los valores o intervalos y sus correspondientes frecuencias absolutas, relativas o porcentajes acumulados.

La figura 1.7 presenta el polígono acumulativo de frecuencias para la variable nivel de colesterol, uniendo los puntos medios de los intervalos, a la altura de sus porcentajes acumulados.

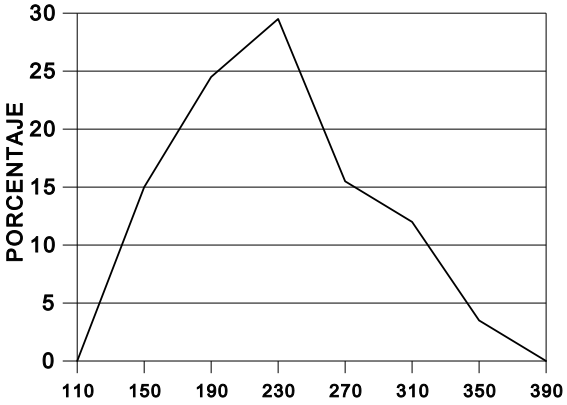


Figura 1.6.- Polígono de frecuencias

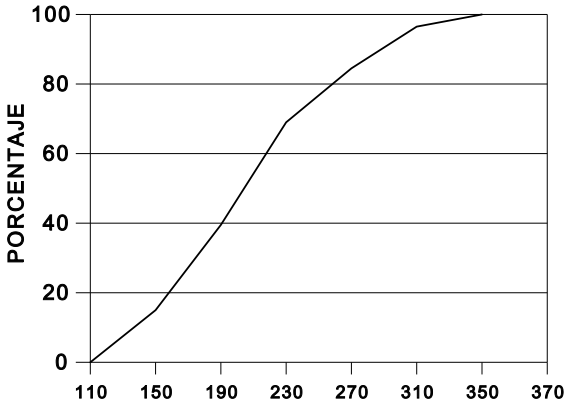


Figura 1.7.- Polígono acumulativo de frecuencias

1.3. TABULACIONES Y GRÁFICOS SOBRE DOS VARIABLES

Frecuentemente interesa inspeccionar la información conjunta de dos variables. Para ello podemos construir tablas de distribución de frecuencias conjuntas, o construir gráficos que informen sobre el comportamiento conjunto. Distinguiremos tres casos según el tipo de variables involucradas.

Variables cualitativas

Como ejemplo, sea un estudio en el que para las variables SEXO (hombre, mujer) y ALCOHOL (nunca, ocasional o diariamente), se observan las siguientes frecuencias en las diferentes parejas de categorías combinadas:

Hombre, Nunca toma alcohol	43	Hombre, Ocasional o diariamente	49
Mujer, Nunca toma alcohol	94	Mujer, Ocasional o diariamente	14

La forma habitual de estructurar esta información es a través de una tabla de *distribución de frecuencias conjuntas,* como se muestra en el cuadro 1.4, obteniendo adicionalmente las distribuciones de frecuencias de cada una de las variables por separado. Estas reciben el nombre de distribuciones de *frecuencias marginales* y recogen los totales por filas o columnas en la tabla. En general, el interés en estas situaciones se centra en el cálculo de los *porcentajes por filas y columnas*. Estos porcentajes (entre paréntesis en el cuadro 1.4, primer paréntesis por filas y segundo por columnas) recogen la información condicional de la distribución de cada una de las variables dada una categoría determinada de la otra. Así, entre los hombres, los porcentajes de consumo de alcohol son 46,7% y 53,3% para las categorías *nunca* y *ocasional o diariamente respectivamente,* mientras que en las mujeres, estos porcentajes son 87% y 13% respectivamente. La representación gráfica más habitual para la distribución conjunta de dos variables cualitativas es el *diagrama de barras combinado*.

	Hombre	Mujer	
Nunca	43 (31,4) (46,7)	94 (68,6) (87,0)	137 (68,5)
Ocas. o diaria.	49 (77,8) (53,3)	14 (22,2) (13,0)	63 (31,5)
	92 (46,0)	108 (54,0)	200 (100)

Cuadro 1.4.- Distribución de frecuencias conjuntas

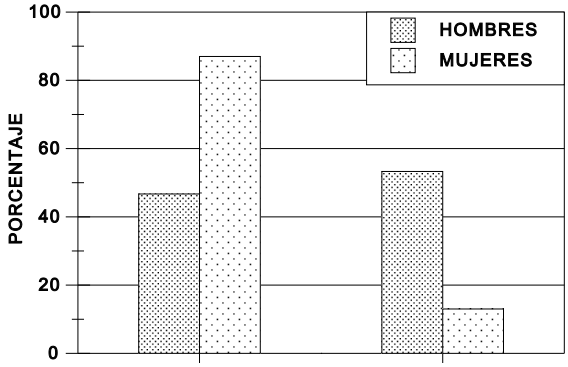


Figura 1.8.- Diagrama de barras combinado. Porcentajes por columnas

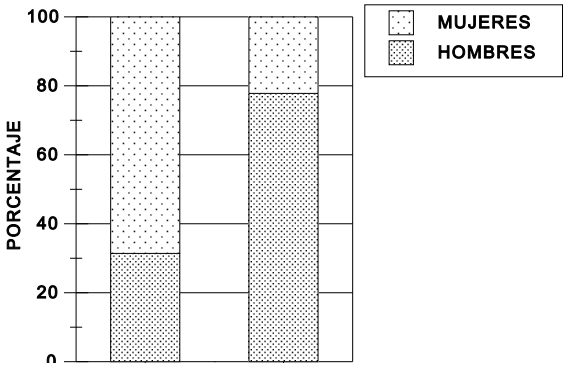


Figura 1.9.-Diagrama de barras combinado. Porcentajes por filas

Las figuras 1.8 y 1.9 presentan distintas construcciones de este tipo. En la 1.8 se ha representado los porcentajes por columnas y la 1.9 los porcentajes por filas. En ambos casos se visualiza rápidamente el desequilibrio existente entre hombres y mujeres en las diferentes categorías de consumo de alcohol.

Variables cuantitativas

La mayor variabilidad de este tipo de variables, junto con la información cuantitativa de sus valores sugiere mayor complejidad en el resumen de la información.

La información disponible será la de todas las parejas de valores (x_i , y_i), i=1,....,n, de la que se desprenderá la distribución conjunta de las variables. La figura 1.10 presenta el *diagrama de dispersión* de las variables nivel de colesterol y edad. Para su construcción se ha representado en el eje de abcisas la edad y en el de ordenadas el nivel de colesterol. Cada uno de los puntos corresponde a la representación de la pareja de coordenadas (x_i , y_i) (x_i = edad del individuo i-ésimo, y_i = colesterol del individuo i-ésimo). Este tipo de diagrama permite visualizar el comportamiento conjunto de las variables, intentando resumir la forma en que se relacionan o vinculan. En el ejemplo se observa un comportamiento de conjunto que sugiere una tendencia a alcanzar mayores valores de colesterol según se incrementa la edad.

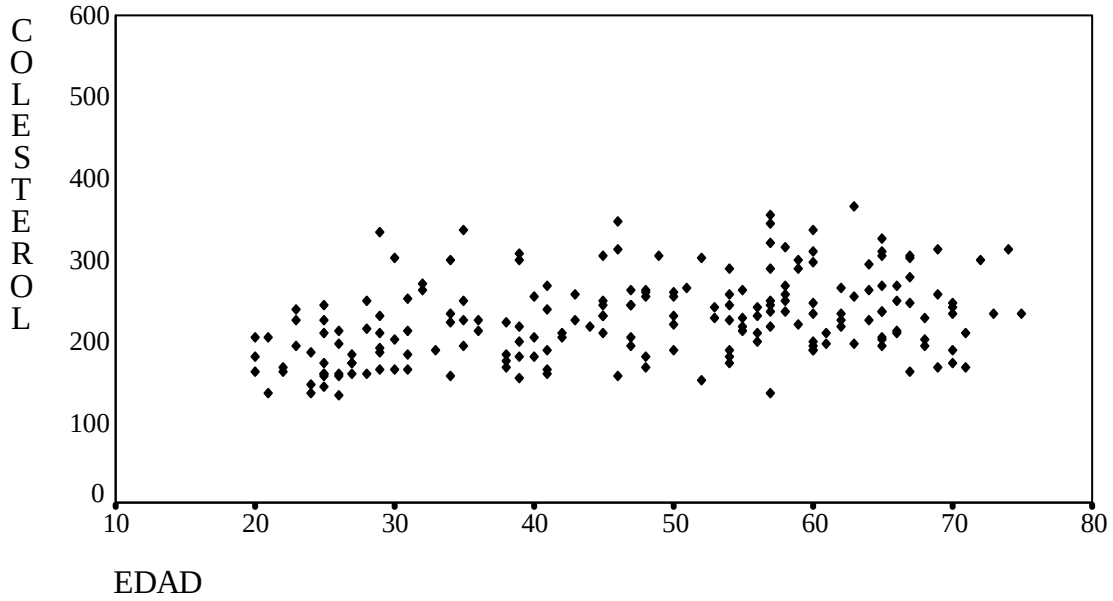


Figura 1.10.- Diagrama de dispersión

Otra forma de abordar la inspección del comportamiento conjunto puede ser categorizando cada una de las variables a través de la definición de los intervalos correspondientes y generando la tabla de distribución de frecuencias conjunta tal como se ha descrito para variables cualitativas.

Una variable cualitativa y otra cuantitativa

Cuando los datos se refieren a una variable cualitativa, por ejemplo el consumo de alcohol (nunca, ocasional o diariamente) y el nivel de colesterol, es posible utilizar tablas o representaciones gráficas para resumir el comportamiento conjunto. En el cuadro 1.5 y figura 1.11 se presenta las distribuciones de frecuencias del colesterol según consumo de alcohol y la representación gráfica de éstas a través de los polígonos de frecuencias correspondientes. Puede observarse la tendencia a alcanzar con mayor frecuencia valores más elevados de colesterol en el grupo definido por el consumo ocasional o diario de alcohol.

ALCOHOL

	Nunca		Ocas. o diariam.		Total	
Colest	f_i	p_i	f_i	p_i	f_i	p_i
130-170	27	19,7	3	4,8	30	15,0
170-210	30	21,9	19	30,2	49	24,5
210-250	42	30,7	17	27,0	59	29,5
250-290	20	14,6	11	17,5	31	15,5
290-330	15	10,9	9	14,3	24	12,0
330-370	3	2,2	4	6,3	7	3,5
Total	137	100,0	63	100,0	200	100,0

Cuadro 1.5.- Distribución de frecuencias de una variable cuantitativa según categorías de una cualitativa

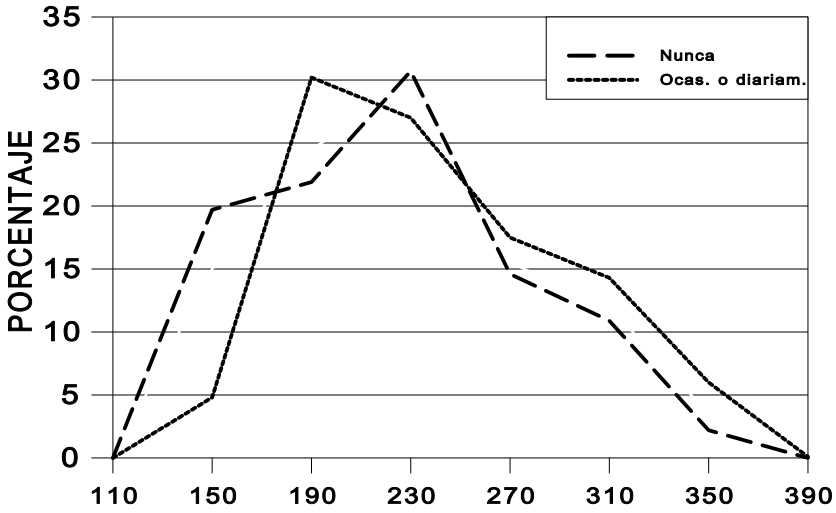


Figura 1.11.- Polígonos de frecuencias según categorías de una variable cualitativa

1.4 EXPLORANDO LOS DATOS: RESUMEN DE LA INFORMACION

La tabulación y gráficos sobre los datos es una primera aproximación a la comprensión de su comportamiento. Sin embargo, las variables que conforman nuestra base de datos pueden presentar, individual o conjuntamente, rasgos específicos que nos pueda interesar conocer. Ello sugiere la conveniencia de disponer de ciertas medidas que nos permitan captar de forma resumida tales rasgos. Cuando los datos proceden de la observación de variables cuantitativas, es posible construir de una forma natural, dado el carácter cuantitativo, funciones matemáticas que se utilizarán para resumir la información que contienen. Si éstos proceden de la observación de variables cualitativas, las únicas medidas posibles son las proporciones o porcentajes de las diferentes categorías o combinaciones de categorías de la variable, o las razones entre proporciones o porcentajes. Por consiguiente, las medidas expuestas en este apartado van dirigidas al resumen de información de variables cuantitativas. Así, se intentará contestar preguntas tales como:

263	239	230	241	180	235	167	263
163	218	197	224	191	233	166	137
304	222	212	188	206	248	188	244
195	185	176	266	214	225	250	303
211	180	184	227	263	366	195	305
268	311	302	232	160	210	181	154
206	299	262	156	298	236	307	206
199	256	228	261	226	188	158	187
235	251	209	231	237	200	211	173
237	260	167	202	337	314	188	159
305	216	185	225	259	251	271	237
235	250	188	306	313	244	164	227
159	217	209	355	219	164	211	244
344	268	226	310	223	233	229	301
206	228	180	232	254	137	214	200
242	167	251	169	316	258	302	133
163	218	326	334	203	206	172	252
161	182	195	173	162	256	145	193
269	160	156	301	214	232	169	268
322	290	241	203	289	259	244	247
266	314	294	193	250	193	137	160
156	299	198	188	290	257	219	220
278	245	209	263	172	245	226	190
254	336	239	196	247	204	146	164
347	212	234	187	209	151	174	226

Cuadro 1.6.- Datos de colesterol

- Una de las variables recogidas en un estudio es el nivel de colesterol (en mg/100ml) (ver cuadro 1.6). Además del interés de esta variable por sí misma, su importancia como factor de riesgo de enfermedad cardiovascular sugiere a los

investigadores profundizar en su comportamiento. Se desea obtener medidas que resuman la información contenida en el conjunto de observaciones de la base de datos, y que, a ser posible, puedan ser comparadas por otros investigadores. Entre las posibles medidas a calcular, se desea obtener alguna que *represente* al conjunto de los datos en un sentido de promedio o centro. Así mismo, se desea una medida que nos informe del nivel de *homogeneidad o parecido* en el colesterol entre los individuos que han dado lugar a la base de datos.

- Un problema adicional para esta variable deriva de la definición de *normocolesterolemia* (nivel de colesterol por debajo de cierta cifra, denotando normalidad) e *hipercolesterolemia* (nivel de colesterol por encima de cierta cifra, denotando un colesterol excesivamente elevado). Estas calificaciones dependen de que el nivel de colesterol de un individuo se encuentre por debajo o sea superior a una cifra predeterminada. Diversos estudios clínico-epidemiológicos han ido modificando a lo largo del tiempo su valor (240 mg/100ml, 220 mg/100ml o 200 mg/100ml como cifra más reciente). Los investigadores se plantean algunas cuestiones de interés: ¿cúal es el porcentaje de individuos que serían calificados como hipercolesterolémicos para cada uno de los posibles puntos de corte? ¿atendiendo a un criterio puramente empírico, cuál debería ser el punto de corte tal que fueran calificados como hipercolesterolémicos el 5% de los individuos? ¿Idem para el 10%?

- Previamente, los investigadores habían inspeccionado la distribución de frecuencias de la variable colesterol, inspeccionando la forma de esta distribución para intervalos de amplitud 40 mg/100ml de colesterol, observándose una cierta tendencia a que los valores se desplacen con mayor frecuencia hacia la derecha (valores más elevados) que hacia la izquierda. ¿Es posible calcular alguna medida que capte esta situación?.

Parámetros y estadísticos

Las medidas calculadas sobre los datos disponibles recibirán una calificación diferente según los datos sean los de todos los individuos que se quiere caracterizar o sobre los que se pretende extraer conclusiones, en cuyo caso diremos que hemos observado a toda la población, o únicamente los de una parte de estos individuos, en cuyo caso diremos que hemos observado una muestra. Desde el punto de vista estadístico es esencial diferenciar estas situaciones, pues condicionan decisivamente el nivel de aplicación de diferentes procedimientos. Así, mientras que si se ha observado a toda la población las técnicas de descripción estadística (distribuciones de frecuencias, gráficos, medidas resumen de información) aportarán toda la información deseada, en el caso de que la observación sea parcial, es decir una muestra, será necesario aplicar procedimientos estadísticos más complejos para extraer conclusiones sobre la población.
Cuando las medidas para resumir la información de una o más variables sean calculadas sobre datos de una población recibirán el nombre de *parámetros*, mientras que si lo son sobre datos de una muestra recibirán el nombre de *estadísticos.* En la mayor parte de las ocasiones se dispone sólo de una muestra de observaciones, por lo que los parámetros poblacionales pueden resultar de interés pero serán desconocidos, mientras que sí serán calculables los estadísticos muestrales, siendo deseable la extracción de conclusiones sobre toda la población. Este proceso se conoce con el nombre de *inferencia estadística*, y aspectos tales como la estricta y clara definición de la población o el proceso de selección de los individuos que conforman la muestra, conocido como *proceso de muestreo,* son esenciales para su correcta aplicación. En capítulos sucesivos se profundizará sobre estos elementos.

Medidas de tendencia central

Las medidas de tendencia central tienen como objetivo describir, a través de un valor numérico, la localización de las observaciones. Son valores que representan, según diferentes criterios, la posición donde se concentran los datos observados. La comparación de estas medidas, calculadas para una misma variable, entre diferentes conjuntos de datos puede indicarnos las diferencias en la posición de los valores entre los conjuntos considerados. Presentamos a continuación las medidas más usuales.

Media

La media, o media aritmética para diferenciarla de otras posibles medias, es la medida de tendencia central más utilizada.. Diremos que dada una variable genérica X, y un conjunto de n observaciones de esta variable $\{x_1, x_2, x_3,...,x_n\}$, la media es la suma de todas las observaciones dividida por su número:

$$\textbf{Media} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Si calculamos la media de la variable colesterol, sobre los datos que se presentan en el cuadro 1.6, obtendremos:

$$\bar{x} = \frac{263 + 163 + 304 + 195 + \ldots + 226}{200} = 227{,}38 \text{ mg/100ml}$$

Para un determinado conjunto de datos, la media es única, interviniendo en su cálculo todos y cada uno de los valores de la variable. Este hecho confiere a la media el ser más informativa que otras medidas de tendencia central. Sin embargo, la existencia de datos atípicos puede incrementar o disminuir su valor de forma notable, convirtiéndola en poco representativa del centro de la estructura de los datos. Esto sucederá en mayor o menor medida cuando los datos presenten estructuras no simétricas. Para resolver esta situación pueden calcularse *medias ajustadas de orden p* (0<p<100), utilizando el p% central de las observaciones (una vez ordenadas de menor a mayor).

Mediana

La mediana de un conjunto de n observaciones de una variable es aquel valor tal que la cantidad de datos inferiores a él es igual a la cantidad de datos superiores. Es una medida que busca el centro de la estructura de los datos bajo la idea de distribuir las observaciones en dos conjuntos de igual número. Para entender el concepto y proceder a su cálculo es necesario partir de que el conjunto de observaciones de la variable es ordenado (generalmente de menor a mayor), de esta forma la mediana puede ser definida como:

Mediana = Md = Valor de la observación que ocupa la posición o rango

$$r_{Md} = \frac{n+1}{2}$$

En caso de que r_{Md} no sea entero, Md se calcula como la semisuma de los valores anterior y posterior. Para ejemplificar su cálculo, considérese la secuencia de observaciones de la variable tiempo de estancia en un centro hospitalario, para 12 individuos, ordenados de menor a mayor (ver cuadro 1.7).

1, 1, 2, 4, 6, 10, 10, 14, 15, 15, 18, 121

Cuadro 1.7.- Días de estancia de 12 individuos

La mediana será el valor que ocupe la posición

$$r_{Md} = \frac{n+1}{2} = \frac{12+1}{2} = 6{,}5$$

posición que, al no ser exacta, nos lleva a calcularla como la semisuma de los valores que ocupan las posiciones 6 (10) y 7 (10), por lo que

Md = 10 días

Al igual que en el caso de la media, existe una única mediana para un conjunto determinado de datos. Sin embargo, la mediana no utiliza para su cálculo todos los valores de las observaciones de la variable, lo que le confiere menor capacidad informativa. A cambio la mediana no se verá afectada por observaciones extremas. Este último resultado la hace especialmente apropiada para captar la localización de un grupo de observaciones de una variable con estructura asimétrica.

Moda

La moda de un conjunto de observaciones de una variable es aquel valor que se presenta con mayor frecuencia. En el caso de que la variable sea cuantitativa continua, su mayor variabilidad puede hacer que la inspección de los valores individuales nos lleve a que éstos se repiten con escasa frecuencia. En este caso, resulta conveniente agrupar la variable en una distribución por intervalos, hablando entonces de *clase o intervalo modal* (nos referimos a intervalos de igual amplitud)

Moda = Mo = Valor de la variable con mayor frecuencia

Intervalo modal = Intervalo de la distribución con mayor frecuencia

La moda puede utilizarse con datos cualitativos, representando entonces la categoría más frecuente de la variable. En la mayoría de las ocasiones, especialmente en el caso de variables cuantitativas continuas, interesa detectar si los valores de la variable se concentran en torno a un cierto valor de la variable o éstos tienden a concentrarse alrededor de más de un valor. Estas situaciones son reconocidas como *unimodales* o *multimodales*. La inspección del polígono de frecuencias permite detectar estas situaciones. Las figuras 1.12 y 1.13 presentan los polígonos de frecuencias de una situación unimodal y otra bimodal.

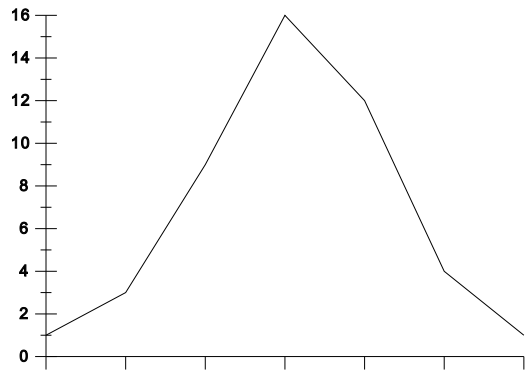


Figura 1.12.- Distribución unimodal

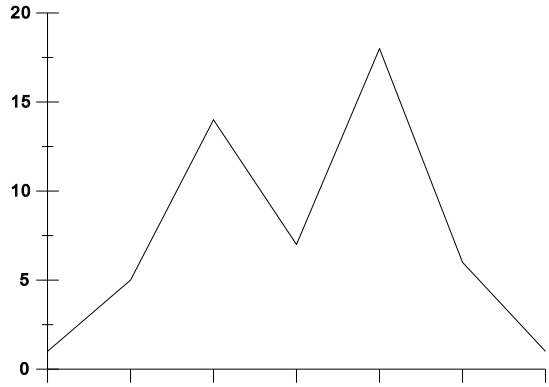


Figura 1.13.- Distribución bimodal

Medidas de dispersión

Las medidas de tendencia central informan acerca de la localización de los valores de las observaciones de una variable. Sin embargo, esta información es insuficiente para comprender el comportamiento de la variable. Situaciones claramente diferenciadas pueden dar lugar a medidas de tendencia central iguales, por lo que éstas sólo pueden ser utilizadas parcialmente como resumen de la información.
Se denominan medidas de dispersión aquellas que pretenden captar y resumir la mayor o menor variabilidad, la mayor o menor concentración, homogeneidad o parecido entre las observaciones de la variable. Se presenta a continuación las medidas de dispersión más frecuentes. Como en las medidas de tendencia central, se partirá de una variable genérica, X, y de un conjunto de n observaciones $\{x_1, x_2, x_3, x_4,\ldots,x_n\}$.

Rango o recorrido

Se define como la diferencia entre el mayor y el menor valor de la variable:

$$\textbf{Rango} = R = x_{max} - x_{min}$$

Los valores máximo y mínimo de la variable colesterol, referida en el cuadro 1.6, son 133 y 366 mg/100ml respectivamente. Con ello, el rango para esta variable sobre las 200 observaciones es

$$R = 366 - 133 = 233 \text{ mg/100ml}$$

es decir, todas las observaciones se encuentran en este recorrido. El rango es una medida de cálculo sencillo y rápido, puesto que depende sólo del mayor y menor valor de la variable. Pero debido a ello es escasamente informativa de lo que sucede con el resto de observaciones, afectándose por la existencia de observaciones extremas.

Varianza y desviación típica o estándar

Es la medida de variabilidad más utilizada. La idea para su construcción surge de cuantificar las distancias, y por consiguiente la variabilidad, entre los valores de la variable a través de su diferencia respecto de una medida central como es la media:

$$\textbf{Varianza} = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Debido a que la varianza no está expresada en las mismas unidades que la variable, sino en unidades al cuadrado, se define la desviación típica o estándar como la raíz cuadrada positiva de la varianza:

$$\textbf{Desviación típica o estándar} = s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Sobre los datos del cuadro 1.7, para los que la media en días de estancia es de 18,1 días, la varianza y desviación típica serían:

$$s^2 = \frac{(1-18,1)^2 + (1-18,1)^2 + (2-18,1)^2 + ... + (121-18,1)^2}{12} = 995,4 \text{ días}^2$$

$$s = \sqrt{s^2} = \sqrt{995,4} = 31,6 \text{ días}$$

Tanto la varianza como la desviación típica deben ser mayores o iguales a 0. El valor 0 sólo se alcanzará en aquellos casos en que los datos alcancen el mismo valor. Representan una cuantificación absoluta de la variabilidad o dispersión de los datos, es decir, dependiente de su localización (media) y sus unidades de medida. Esto hace que sus valores para diferentes variables o conjuntos de datos no sean comparables.
Las expresiones expuestas para la varianza y la desviación típica se refieren al conjunto de datos interpretado como una población. Representan las medidas descriptivas resumen de la variabilidad de ese conjunto de datos. Cuando éste es contemplado como una muestra de una determinada población, y el objetivo a través de la varianza y la desviación típica es calcular un valor que resuma la variabilidad, pretendiendo aproximarse al verdadero valor poblacional, deben ser utilizadas las expresiones para la varianza y desviación típica:

$$\text{Varianza} = s_c^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \quad ; \quad \text{Desviación típica} = s_c = \sqrt{s_c^2}$$

que resultan ser los *estimadores* de los respectivos parámetros poblacionales. Cuando n es grande, la diferencia es mínima entre las expresiones expuestas, pero si n es pequeño, puede haber una diferencia notable. Hay que tener en cuenta que algunos paquetes estadísticos para ordenador o calculadoras científicas realizan los cálculos con ambas expresiones o a veces con sólo una de ellas.

Coeficiente de variación

La varianza y desviación típica representan medidas absolutas de la dispersión de un conjunto de observaciones de una variable. Su interpretación depende de la unidad de medida de la variable así como de su localización. Esto hace que las desviaciones típicas o varianzas de variables distintas sobre un mismo conjunto de datos o de la misma variable sobre conjuntos de datos diferentes no sean comparables, no pudiendo afirmar en qué caso hay mayor o menor variabilidad. Para resolver este problema se puede recurrir al siguiente coeficiente:

$$\text{Coeficiente de Variación} = CV = \frac{s}{\bar{x}} (x100)$$

que, al dividir desviación típica por la media elimina las unidades de medida y el efecto de la localización de la variable, resultando así una medida relativa de la variabilidad de los datos. Suele ser expresado en porcentaje, pudiendo alcanzar valores entre 0 e ∞. Sobre los datos del cuadro 1.7, para los que la media y desviación típica resultaban ser de 18,1 y 31,6 días, respectivamente, el coeficiente de variación será:

$$CV = \frac{31,6}{18,1} = 2,02 \ (\times 100) = 202 \ \%$$

El resultado refleja la alta variabilidad de los datos, producida, fundamentalmente por la observación con valor 121 días.

Percentiles

Un percentil de orden q (0<q<100) es un valor de la variable tal que el porcentaje de observaciones cuyo valor es inferior o igual a ése es precisamente q. Esta definición es de fácil comprensión si observamos la figura 1.14

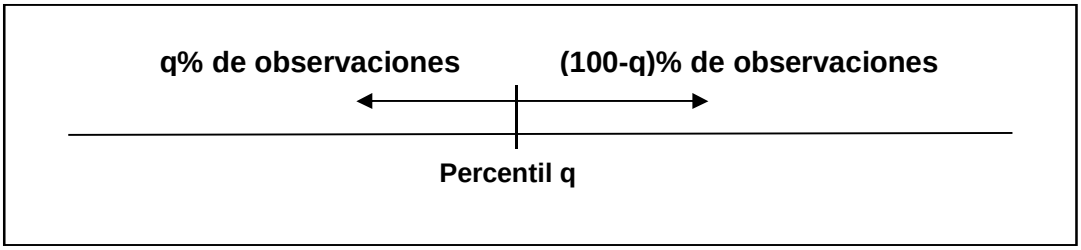


Figura 1.14.- Distribución de observaciones según el percentil de orden q

El cálculo de un determinado percentil requiere la ordenación (en general de menor a mayor) de los valores de la variable, y la determinación del valor de la variable que verifica la definición establecida. Sin embargo, existen tres situaciones en las que la aplicación de esta definición puede ser problemática, (1) con números pequeños de observaciones, (2) con valores repetidos, y (3) cuando el percentil no es único. Para comprender estas situaciones consideremos cuatro observaciones de una variable X, cuyos valores son 12, 12, 36, 39. No hay ningún valor de la variable que verifique la definición de percentil para el orden q=25. Para el orden q=75, existe una infinitud de valores, p.ej. 36,5, 37, y 38,126543, que satisfacen todos ellos la definición establecida. Si el número de observaciones es grande, estos problemas tienden a desaparecer. Una definición consistente con la definición enunciada de percentil de orden q es la de que éste sería el valor de la variable con rango o posición :

$$r_q = \frac{q}{100}(n+1)$$

una vez ordenadas de menor a mayor las observaciones de la variable, aproximando cuando éste no sea entero a través del promedio ponderado entre los valores que ocupen los rangos anterior y posterior:

$$p_q = (1-f)\, x_i + f\, x_{i+1}$$

donde f es la parte fraccionaria de r_q , x_i, y x_{i+1} son los valores que ocupan rangos anterior y posterior. Este método de cálculo es el utilizado en buena parte de los paquetes estadísticos para ordenador personal.
Sobre los datos del cuadro 1.7 los percentiles 5, 10, 25, 75, 90 y 95 serán los valores que ocupen los rangos:

Percentil	Rango	Valor
p_5	(5/100) 13 =0,65	No existe
p_{10}	(10/100) 13 = 1,30	0,70 . 1 + 0,30 . 1 = 1
p_{25}	(25/100) 13 = 3,25	0,75 . 2 + 0,25 . 4 = 2,5
p_{75}	(75/100) 13 = 9,75	0,25 . 15 + 0,75 . 15 = 15
p_{90}	(90/100) 13 = 11,7	0,30 . 18 + 0,70 . 121 = 90,1
p_{95}	(95/100) 13 =12,35	No existe

Algunos autores ubican los percentiles como medidas de tendencia central. Su utilización puede ser adecuada como resumen de la localización. La mediana coincide con el percentil de orden 50. Sin embargo,

en el contexto sanitario, la utilización más frecuente de estas medidas es la de variabilidad o dispersión. Especialmente en el caso de la determinación de valores de *normalidad* (no probabilística sino biomédica o sanitaria), se recurre a ciertos percentiles extremos (generalmente de orden 2,5, 5, 95, 97,5) para establecer puntos de corte que sitúan porcentajes de población que se encuentran en las zonas de valores más elevados o bajos de la variable (p. ej. peso, talla, ácido úrico, colesterol, etc.). Algunos percentiles reciben nombres específicos, como es el caso de los *deciles* (percentiles 10, 20, 30,...,90), *cuartiles* (percentiles 25, 50 y 75, dividen la distribución de las observaciones en cuatro regiones con igual porcentaje de casos), y, en general, *quintiles* (dividen en cinco regiones), *sextiles*, etc.

Medidas de forma

Las medidas de forma pretenden resumir una característica distinta de la localización y la dispersión de las observaciones de la variable.

Se trata de resumir si los datos presentan una distribución más o menos simétrica o con un menor o mayor apuntamiento. Para cuantificar el grado de asimetría puede calcularse el *coeficiente de asimetría*:

$$As = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\frac{n}{s^3}}$$

donde s es la desviación típica de la variable.

La interpretación del coeficiente es como sigue:

- As = 0 → Simetría (Figura 1.15)
- As > 0 → Asimetría positiva (Figura 1.16)
- As < 0 → Asimetría negativa (Figura 1.17)

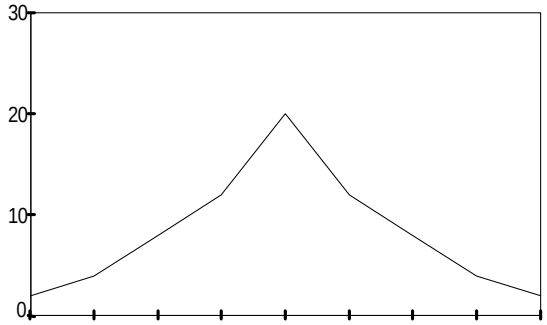


Figura 1.15.- Distribución simétrica

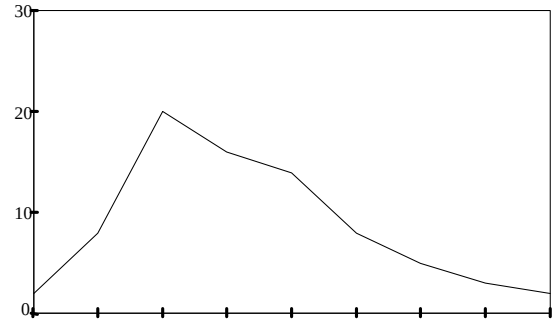


Figura 1.16.- Distribución asimétrica positiva

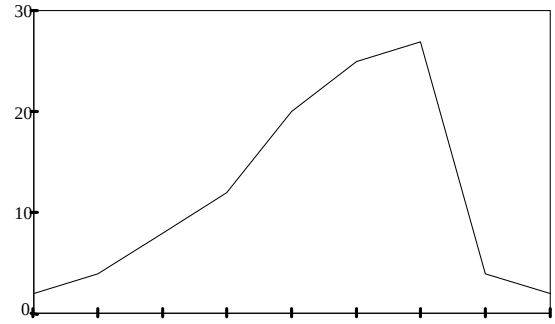


Figura 1.17.- Distribución asimétrica negativa

Otra característica de la forma de la distribución de la variable es su mayor o menor apuntamiento (ver figura 1.18). La cuantificación del grado de apuntamiento puede realizarse a través del *coeficiente de curtosis*:

$$Cu = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\frac{n}{s^4}}$$

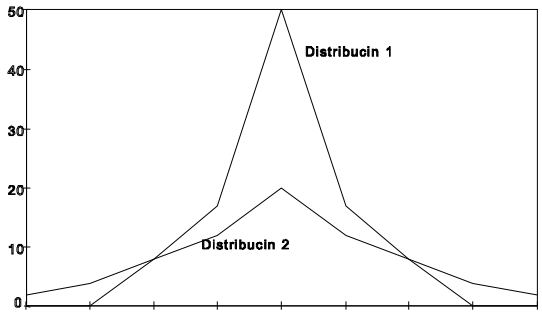


Figura 1.18.- Dos casos de apuntamiento

donde s es la desviación típica de la variable.

La interpretación del coeficiente de curtosis es ligeramente diferente a la del coeficiente de asimetría, puesto que no existe una situación equivalente a la simétrica, pudiendo hablar de mayor o menor curtosis únicamente. No obstante, es frecuente comparar el valor de Cu con la curtosis de la curva de probabilidad normal (se definirá más adelante), cuyo valor es 3, hablando entonces de:

Cu = 3 Distribución mesocúrtica (apuntamiento semejante al del modelo normal)

Cu > 3 Distribución leptocúrtica (más apuntada que la curva normal)

Cu < 3 Distribución platicúrtica (menos apuntada que la curva normal)

Las medidas de forma son utilizadas con frecuencia para tener un resumen descriptivo de la mayor o menor normalidad (como modelo de probabilidad) de la variable, puesto que los valores para una variable que siga este modelo son $\mathbf{As = 0, Cu = 3}$. Algunos programas de análisis trabajan restándole 3 al coeficiente de curtosis anterior, interpretándose entonces su resultado sobre el valor 0 y no sobre 3.

$$Cu = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\frac{n}{s^4}} - 3$$

Cálculo de medidas resumen para datos tabulados

Las definiciones y cálculos para las diferentes medidas resumen han sido expuestos sobre las observaciones individuales $\{x_1, x_2, x_3,...,x_n\}$ de una variable. Estas son las definiciones y cálculos exactos. Cuando los datos disponibles están tabulados en forma de tabla de distribución de frecuencias, las expresiones de cálculo de algunas de las medidas deben ser adaptadas a esta situación. En el cuadro 1.8 se presenta las fórmulas adecuadas.

El cálculo será solo aproximado si la tabulación es en forma de intervalos. En ese caso xm_i es el punto medio del intervalo i. En otro caso, las fórmulas son adecuadas, y xm_i representará el valor i-ésimo de la variable. En el cálculo de la mediana o percentiles de orden q, x_i representa el límite inferior del intervalo al que debe pertenecer la mediana o el percentil de que se trate, una vez inspeccionadas las frecuencias acumuladas de la variable y considerado que la mediana y el percentil q dejarán n/2 y qn/100 observaciones a su izquierda, respectivamente. De la misma manera, a_i representa la amplitud de ese intervalo.

Medidas de tendencia central	Medidas de dispersión	Medidas de forma
$\bar{x} = \frac{\sum_{i=1}^{k} xm_i f_i}{n}$	$R = xm_{max} - xm_{min}$	$As = \frac{\sum_{i=1}^{k} (xm_i - \bar{x})^3 f_i}{\frac{n}{s^3}}$
$Md = x_i + \left(\frac{n/2 - F_{i-1}}{F_i - F_{i-1}} \right) a_i$	$s^2 = \frac{\sum_{i=1}^{k} (xm_i - \bar{x})^2 f_i}{n}$	
	$s = \sqrt{s^2}$ $\quad CV = \frac{s}{\bar{x}} (x100)$	
Mo = Intervalo modal = Intervalo para el que f_i es máxima	$P_q = x_i + \left(\frac{\frac{qn}{100} - F_{i-1}}{F_i - F_{i-1}} \right) a_i$	$Cu = \frac{\sum_{i=1}^{k} (xm_i - \bar{x})^4 f_i}{\frac{n}{s^4}}$

Cuadro 1.8.- Cálculo de estadísticos descriptivos para datos agrupados

Relación entre variables cuantitativas

La idea de asociación o relación entre dos variables cuantitativas es más intuitiva que entre cualitativas. Ideas del tipo de 'a más talla más peso', o 'el colesterol depende de la edad', son intuitivamente comprensibles. Se trata de obtener alguna medida con capacidad para detectar si los valores de una de las variables suelen ir acompañados de valores de otra de las variables, en el sentido de a mayor valor de una variable mayor (o menor de la otra) como norma general. Sin embargo, aunque aparentemente sea más sencillo comprender la idea de lo que queremos medir, en la práctica la dificultad es superior, debido a que, en definitiva, se trata de buscar una relación matemática que conecte ambas variables y evaluar su pertinencia. Este proceso es conocido como análisis de la regresión entre dos variables. Así, dadas, dos variables X e Y, trata de encontrar la ecuación $\mathbf{y = f(x)}$ que relacione a ambas variables. Cuando la ecuación propuesta es la de una línea recta, es decir $\mathbf{y = \alpha + \beta x}$, es posible calcular un coeficiente que nos permita medir el grado de relación lineal (se supone una ecuación lineal entre ambas variables). Este coeficiente es el de *correlación lineal de Pearson*:

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

Es un coeficiente estandarizado, cuyo rango de valores es $-1 \leq r \leq 1$. Valores próximos a 0 indicarían ausencia de relación lineal, pudiendo existir relación bajo otro modelo de regresión (no lineal). Los valores -1 y 1 representan las relaciones lineales perfectas, en sentido inverso (a mayor valor de una variable menor de la otra) o directo (a mayor valor de una variable mayor de la otra). En las figuras 1.19, 1.20 y 1.21 se visualizan las ideas de relación lineal directa, inversa y ausencia de relación lineal.

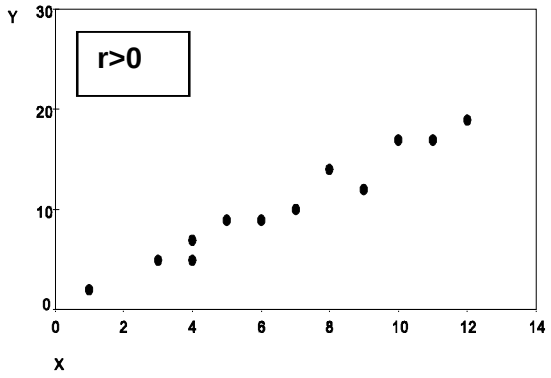


Figura 1.19.- Relación lineal directa

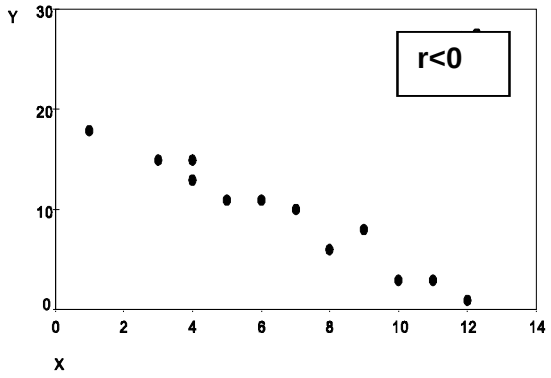


Figura 1.20.- Relación lineal inversa

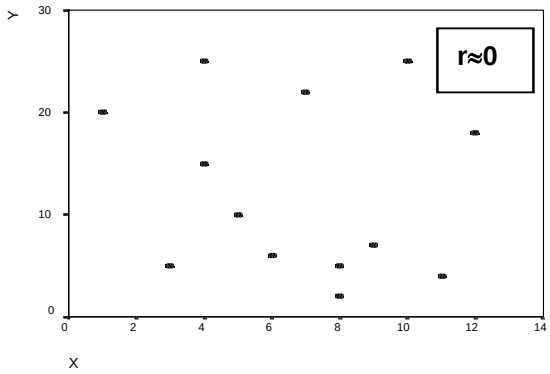


Figura 1.21.- Ausencia de relación lineal

C A P I T U L O 2

EL AZAR SOBRE NUESTROS DATOS: MIDIENDO LA PROBABILIDAD

Los elementos introducidos en el capítulo precedente, relativos a la recolección, organización, resumen y exploración descriptiva de los datos recogidos son una base inicial de gran importancia en la utilización de los datos para el análisis estadístico. En este capítulo se introducirán conceptos básicos de la teoría de la probabilidad. Esencialmente, se trata de poner en evidencia el origen no determinista o aleatorio de la información disponible, así como de establecer ciertos resultados útiles para la construcción de modelos encaminados a medir la posibilidad de que nuestras variables presenten unos u otros resultados con mayor o menor frecuencia. Se tratará en definitiva de establecer las bases para la medición de la *probabilidad.* Adicionalmente, los conceptos básicos desarrollados en este capítulo sustentarán, en buena parte, los argumentos necesarios para la aplicación rigurosa de la *inferencia estadística*.

A modo de ejemplo, suponga que sobre 200 individuos estudiados nos preguntamos lo siguiente:

- Si elegimos un individuo cualquiera de los 200 ¿Son su peso, talla, o la presencia o no de cálculos resultados predecibles sin error?
- ¿Tiene sentido preguntar acerca de la posibilidad de diferentes combinaciones de resultados, como por ejemplo 'Ser consumidor de alcohol y pesar más de 70 Kg', o 'tener más de 30 años o un nivel de colesterol superior a 200 mg/100ml'.
- ¿Es posible cuantificar la posibilidad de que al elegir un individuo al azar de entre los 200, éste verifique los diferentes resultados enunciados en el punto anterior u otros?.
- El mencionado valor cuantitativo, ¿dependerá según sea referido a hombres o a mujeres? ¿El hecho de tener la glucosa por encima de 140 mg/100ml favorece la presencia de cálculos?

Estas y otras situaciones de interés justifican la presentación de los conceptos básicos de probabilidad que a continuación se desarrollan. En definitiva, se trata de reconocer las situaciones de incertidumbre y proponer axiomas, leyes y propiedades que faciliten su cuantificación.

2.1 CARACTERISTICAS ALEATORIAS Y ESPACIOS MUESTRALES

Si nos fijamos en las preguntas realizadas en el apartado anterior se puede aceptar sin dificultad que los resultados observados en las diferentes características estudiadas para cada uno de los individuos no serían predecibles sin error antes de ser observadas. En efecto, el sexo, edad, consumo de alcohol, talla, peso, etc., de cualquiera de los individuos resultaban imposible de conocer antes de ser observadas o medidas. Se dirá entonces que las características estudiadas son *características aleatorias*. Esta idea de aleatoriedad no debe ser entendida como la distribución del azar por igual para cada posible resultado, simplemente refleja la intervención del azar en el sentido en que somos incapaces de predecir con exactitud los resultados que se producirán, no siendo necesario que cualquier resultado tenga la misma posibilidad de producirse. Así, si elegimos un individuo cualquiera al azar de entre nuestros 200, podemos pensar sin dificultad que su peso tiene mayor posibilidad de encontrarse entre 50 y 80 Kg que de ser superior a 150 Kg. La pregunta clave, aceptando esta idea de aleatoriedad es ¿se puede cuantificar la posibilidad de ocurrencia de los diferentes resultados o combinaciones de resultados de una o más características aleatorias? Previo a establecer alguna medida orientada a este fin, es necesario revisar algunos conceptos básicos que nos ayuden a reconocer las situaciones de incertidumbre y los elementos que las componen.

Espacio muestral

Se entenderá por espacio muestral al conjunto de todos los posibles *resultados simples* de una característica aleatoria. Los espacios muestrales podrán ser *finitos* o *infinitos*, según la característica aleatoria posea un número finito o infinito de posibles resultados simples y el conjunto de resultados simples que forma el espacio muestral debe ser *exhaustivo y exclusivo*, es decir debe contener todos los posibles resultados y sólo uno de ellos se producirá. Además, los espacios muestrales pueden ser *univariantes*, cuando los resultados simples proceden de una única característica aleatoria, o *multivariantes*, cuando se refieren a dos o más características consideradas simultáneamente.

- Referido a los 200 individuos de la base de datos del anexo 1, si consideramos las características aleatorias 'consumo de alcohol', 'presencia o no de cálculos biliares', o 'talla', los correspondientes espacios muestrales para cada una de estas características serán:

Característica aleatoria	Espacio muestral
Consumo de alcohol	{ nunca, ocasionalmente }
Presencia de cálculos	{ si, no }
Talla (cm)	{ 173, 168, 154, 165,...}

- Si consideramos las características aleatorias 'consumo de alcohol' y 'nivel de glucosa', el espacio muestral bivariante asociado será:

Características aleatorias	Espacio muestral
Consumo de alcohol y nivel de glucosa	{ nunca y glucosa $\leq$ 140, ocasionalmente y glucosa $\leq$ 140, nunca y glucosa > 140, ocasionalmente y glucosa > 140 }

Sucesos

A menudo el interés se centra en ciertos resultados simples del espacio muestral. Como ejemplo, si consideramos la característica aleatoria 'nivel de colesterol' como medida continua en mg/100ml, el espacio muestral es el conjunto de posibles medidas que se pueden obtener sobre los sujetos. Sin embargo para ciertos estudios puede interesar cuantificar la incertidumbre sobre la situación 'tener el colesterol por encima de 200 mg/100ml', situación reconocida como 'hipercolesterolemia'. Se dirá que el interés se centra en un determinado *suceso* cuando nos refiramos a un subconjunto de resultados simples del espacio muestral.

Así, el suceso 'hipercolesterolemia' considerado se configura como el subconjunto de todos los resultados simples del nivel de colesterol que superan el valor 200 mg/100ml:

hipercolesterolemia = { niveles de colesterol > 200 mg/100ml }

y se producirá cuando el nivel de colesterol presente cualquiera de los resultados simples con valor superior a 200 mg/100ml (213, 267, 314, etc.).
En muchas ocasiones los sucesos de interés coinciden con resultados simples del espacio muestral.

Suceso imposible y suceso seguro

Dado el espacio muestral vinculado a una característica aleatoria, se puede definir dos sucesos que representan los extremos de la incertidumbre acerca de tal característica. Así, se define:

Ω = *Suceso seguro* = Suceso que se producirá con seguridad sea cual sea el resultado simple de la característica aleatoria

Φ = *Suceso imposible* = Suceso que no se producirá nunca sea cual sea el resultado simple de la característica aleatoria

☛ Si elegimos un individuo al azar de entre los 200 considerados anteriormente, el suceso

Ω = Medir menos de 500 cm es seguro, puesto que sea cual sea la talla éste se verificará.

El suceso **Φ** = Medir más de 500 cm es imposible, puesto que no se verificará nunca

Operaciones entre sucesos

Frecuentemente el interés sobre las situaciones de incertidumbre se centra en dos o más sucesos, vinculados a través de diferentes operaciones. Para definir las principales operaciones entre sucesos, se considerará, de forma genérica, dos sucesos derivados de un espacio muestral, digamos A y B: Las principales operaciones entre sucesos son:

Unión de sucesos

Se representará como $\mathbf{A} \cup \mathbf{B}$. Se trata de un suceso nuevo que se produce cuando se presenta A, ó B ó ambos. Es un suceso con mayor posibilidad de presentarse que cualquiera de los que lo forman.

Considérese la tabla adjunta. En ella se describe la distribución de frecuencias de las características aleatorias 'consumo de alcohol' y 'nivel de glucosa' (esta última según sea inferior o igual o superior a 140 mg/100ml) de los 200 individuos considerados. Sean los sucesos

O = Consumir alcohol ocasionalmente, y
M =Tener el nivel de glucosa superior a 140 mg/100ml.

Alcohol \ Glucosa	≤ 140	> 140
Nunca	122	14
Ocasionalmente	58	5

La unión de ambos sucesos se producirá cuando un individuo consuma alcohol ocasionalmente ó tenga el

nivel de glucosa por encima de 140 ó verifique ambos sucesos:

$\mathbf{O} \cup \mathbf{M}$ = Suceso que se produce si ocurre O ó M ó ambos.

Para comprobar que es un suceso más fácil de ocurrir que cualquiera de los que lo forman, baste percibir que el número de individuos que verifican $\mathbf{O} \cup \mathbf{M}$ es 77, es decir los que verifican O, más los que verifican M, menos los que verifican ambos sucesos, puesto que son contados dos veces, número superior a cualquiera de los que verifican O (63 casos), ó M (19 casos).

Intersección de sucesos

Se representará como $\mathbf{A} \cap \mathbf{B}$. Se trata de un suceso nuevo que se produce cuando A y B ocurren. Es un suceso con menor posibilidad de producirse que cualquiera de los que lo forman.

A partir de la tabla descrita en la unión de sucesos, considérese el suceso $\mathbf{O} \cap \mathbf{M}$. Este suceso se producirá cuando un individuo consuma alcohol ocasionalmente y su nivel de glucosa sea superior a 140 mg/100ml:

$\mathbf{O} \cap \mathbf{M}$ = Suceso que se produce cuando se verifica O y se verifica M

El número de individuos sobre el que se produce este suceso es 5, inferior al número de cualquiera de los que lo forman.

Suceso complementario

Dado un suceso A, se define su complementario, $\bar{\mathbf{A}}$, como el suceso que se produce cuando no se produce A, es decir cuando el resultado simple del espacio muestral no está contenido en A.

Para la característica aleatoria 'peso en Kg', dado el suceso A = pesar más de 60 Kg, el suceso complementario será:

$\bar{\mathbf{A}}$ = pesar menos o igual de 60 Kg

que se producirá cuando el peso de un individuo presente cualquier resultado simple por debajo o igual a 60 Kg.

Propiedades y definiciones derivadas de las operaciones básicas entre sucesos

Algunas propiedades o definiciones elementales son las siguientes:

- $\mathbf{A} \cup \bar{\mathbf{A}} = \Omega$

Es evidente que dado un suceso A se verificará él o su complementario

- $\mathbf{A} \cap \bar{\mathbf{A}} = \Phi$

La ocurrencia de un suceso y su complementario es imposible

- **Sucesos mutuamente excluyentes**

Una colección de dos o más sucesos son mutuamente excluyentes si la intersección entre cualesquiera dos de ellos es el suceso imposible. Dada una colección de k sucesos, digamos $\{ E_1 , E_2, E_3, ..., E_k \}$, serán mutuamente excluyentes si:

$$\mathbf{E_i \cap E_j = \Phi} \quad \text{para cualquier i y j}$$

- **Partición de un espacio muestral**

Una colección de sucesos mutuamente excluyentes es una partición del espacio muestral de interés si la unión de todos ellos es el suceso seguro. Dada una colección de k sucesos, digamos $\{ E_1 , E_2, E_3, ..., E_k \}$, mutuamente excluyentes, serán una partición si:

$$\mathbf{E_1 \cup E_2 \cdots \cup E_k = \bigcup_{i=1}^{k} E_i = \Omega}$$

es decir, si alguno de ellos se produce con seguridad

2.2 MEDIDA DE PROBABILIDAD

Puesto que no se puede establecer con seguridad cual será el resultado de una característica aleatoria previamente a su realización, parece razonable intentar cuantificar la posibilidad o verosimilitud de que se presenten unos u otros sucesos. Para un suceso cualquiera del espacio muestral asociado a la característica aleatoria, el valor cuantitativo que mide su verosimilitud es denominado *probabilidad*. Esta idea básica y fácilmente comprensible no está exenta de complejidad. Así, aspectos tales como la interpretación del concepto de probabilidad y las condiciones bajo las que debe ser evaluada, o el propio cálculo de las probabilidades de determinados sucesos ha sido objeto de discusión a lo largo de la historia de la evolución de la teoría de la probabilidad.
Se abordará a continuación algunos de estos aspectos, así como los axiomas y propiedades básicas de la medida de la probabilidad.

Definición axiomática

Desde un punto de vista estrictamente matemático, puede definirse la probabilidad como el criterio o regla que permite asignar a cada suceso aleatorio, digamos A, de un espacio muestral, un valor numérico, verificando los siguientes *axiomas*[1]:

- Para cualquier suceso A,

$$\mathbf{p(A) \geq 0}$$

Se establece que la probabilidad de un suceso sea un número no negativo

- Dados dos sucesos A y B, tales que $\mathbf{A \cap B = \Phi}$,

$$\mathbf{p(A \cup B) = p(A) + p(B)}$$

Si dos sucesos son mutuamente excluyentes, la probabilidad de que se presente uno o el otro se debe obtener como la suma de las probabilidades respectivas.

- Dado Ω, suceso seguro,

1 Un axioma es una propuesta que no deriva de ningún otro enunciado. A partir de su aceptación, si se considera razonable, se deriva propiedades, teoremas u otros resultados

$$p(\Omega) = 1$$

Se atribuye la probabilidad máxima en la escala establecida al suceso seguro.

Los tres axiomas enunciados, conocidos como axiomas de Kolmogorov, establecen un marco inicial razonable para derivar matemáticamente propiedades y teoremas que faciliten el cálculo de probabilidades.

Propiedades básicas de la probabilidad

A partir de los axiomas expuestos se deriva múltiples propiedades, entre las que cabe destacar, por su efecto práctico inmediato en el cálculo de probabilidades, las siguientes:

- Dado un suceso A,

$$0 \leq p(A) \leq 1$$

El intervalo [0,1] representa la escala de medida de la probabilidad de cualquier suceso.

- Dados los sucesos A y su complementario,

$$p(\bar{A}) = 1 - p(A)$$

La probabilidad del complementario de un suceso puede ser obtenida a partir del conocimiento de la del original, restando de 1. Una consecuencia inmediata es que la probabilidad del suceso imposible es 0, $p(\Phi) = 0$.

- Dados dos sucesos A y B,

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

Esta propiedad es conocida como ley aditiva y establece una forma general para calcular la probabilidad de la unión de sucesos sean o no mutuamente excluyentes.

- Dados dos sucesos A y B,

$$p(\overline{A \cup B}) = p(\bar{A} \cap \bar{B}) \qquad p(\overline{A \cap B}) = p(\bar{A} \cup \bar{B})$$

Estas dos propiedades, derivadas de la teoría de conjuntos, permiten cambiar situaciones enunciadas sobre uniones o intersecciones de sucesos complementarios a expresiones en las que desaparecen los complementarios.

Para ejemplificar el uso de los axiomas o las propiedades básicas enunciadas, se considerará la siguiente situación. Supóngase que en el servicio de neurología de un hospital público se ha estimado la probabilidad de que un paciente diagnosticado de demencia presente dos tipos de alteraciones, *Leucaraiosis* (L) y *Atrofia cortical* (A). Las correspondientes probabilidades son p(L) = 0,50 y p(A) = 0,40. Además, se estima que la probabilidad de que presente ambas alteraciones es 0,10. Sobre un paciente diagnosticado de demencia y a partir de esta información, el equipo del servicio se plantea averiguar las probabilidades de los siguientes sucesos:

- 'Que no presente leucaraiosis'. Se trata del suceso complementario de presentar leucaraiosis, y, por lo tanto:

$$p(\bar{L}) = 1 - p(L) = 1 - 0{,}50 = 0{,}50$$

- 'Que no presente atrofia cortical'. Idem como el caso anterior:

$$p(\bar{A}) = 1 - p(A) = 1 - 0{,}40 = 0{,}60$$

- 'Que presente alguna de las alteraciones'. Se trata del suceso unión de ambas alteraciones, es decir que presente leucaraiosis o atrofia cortical. Si tenemos en cuenta que $p(A \cap L) = 0{,}10$, tendremos:

$$p(A \cup L) = p(A) + p(L) - p(A \cap L) = 0{,}40 + 0{,}50 - 0{,}10 = 0{,}80$$

- 'Que no presente ninguna de las alteraciones'. Se trata de la intersección de los sucesos complementarios, es decir, que no presente leucaraiosis ni atrofia cortical:

$$p\left(\bar{A} \cap \bar{L}\right) = p(\overline{A \cup L}) = 1 - p(A \cup L) = 1 - 0{,}80 = 0{,}20$$

- 'Que esté exento de alguna de las alteraciones'. Se trata de la unión de los complementarios de los sucesos, es decir, o no presenta leucaraiosis o no presenta atrofia cortical:

$$p(\bar{A} \cup \bar{L}) = p(\overline{A \cap L}) = 1 - p(A \cap L) = 1 - 0{,}10 = 0{,}90$$

Interpretación y asignación de la probabilidad

Ya se ha establecido que la probabilidad de cualquier suceso será un número real perteneciente al intervalo [0,1]. Supóngase que el suceso A denota la ocurrencia de que 'un paciente responde favorablemente a cierto tratamiento' y que p(A) = 0,80. Las preguntas son ¿cómo se interpreta este valor, y cómo se determina en la práctica? Se presenta a continuación dos interpretaciones de la probabilidad que proporcionan puntos de vista útiles, abordando a continuación algunos aspectos sobre cómo establecer los valores de las probabilidades de sucesos.

Interpretación objetiva

Desde esta interpretación, la probabilidad de un suceso es la frecuencia relativa de ocurrencia de ese suceso en infinitas realizaciones de la característica aleatoria que lo produce, bajo las mismas condiciones. De acuerdo con esta interpretación objetiva, la probabilidad del suceso mencionado, p(A) = 0,80, debería ser entendida como que en un número indefinidamente grande de pacientes , con condiciones clínicas, de edad, de sexo, etc., semejantes a las consideradas en el momento actual, aproximadamente en el 80% de ellos se habría producido una respuesta favorable al tratamiento. Esta interpretación es reconocida como interpretación *frecuentista.* La limitación obvia de la interpretación objetiva de la probabilidad es que es aplicable únicamente a sucesos que se pueden repetir bajo los mismos condicionantes. Esto quiere decir que si el suceso considerado, en lugar de referirse a un paciente genérico, se refiriera a Matías López, de 27 años de edad, diabético, residente habitualmente en Tenerife, etc., éste es un suceso difícilmente repetible, y la probabilidad enunciada podría no ser adecuada para él.

Interpretación subjetiva

Puesto que existen muchos sucesos de interés que no son repetibles bajo las mismas condiciones causales, existirán muchas ocasiones en las que la interpretación objetiva no podrá ser aplicada. La interpretación subjetiva (también denominada *bayesiana o personal*) interpreta la probabilidad de un suceso como el grado de creencia personal en su posible ocurrencia, condicional a las condiciones causales particulares del momento en que ésta está siendo evaluada. Desde esta aproximación, la probabilidad de

que la respuesta al tratamiento sea favorable en Matías López, de 27 años de edad, diabético, residente habitualmente en Tenerife, etc., podría ser asignada en función de las condiciones particulares interpretadas por el evaluador (por ejemplo el personal médico o de enfermería), que podría establecer un valor personal p(A) = 0,97. Aunque desde el punto de vista del método científico esta interpretación podría ser contemplada como carente de capacidad de generalización de resultados, ello no es así, al tratarse de una interpretación más amplia, que incluye como posible evaluación de probabilidades a la derivada de la interpretación objetiva.

Asignación de probabilidades

En la práctica, tanto si se parte de una interpretación objetiva como si ésta es subjetiva, el problema es común, ¿cómo evaluar las probabilidades de los sucesos de interés?. En la mayoría de las ocasiones el cálculo de los valores de probabilidad se basa en la frecuencia relativa de los sucesos, evaluada a través de la evidencia empírica previa o de la obtenida en experiencias diseñadas *ad hoc*.
Esta cuantificación corresponde a la interpretación objetiva y es aparentemente la más sencilla, puesto que libera al investigador de la necesidad de construir elementos (diversos tipos de preguntas, cuestionarios, etc.) diseñados para captar la probabilidad personal. No obstante, debe tenerse en cuenta las limitaciones de esta forma de medir las probabilidades, fundamentalmente por lo que se refiere a la necesidad de que las observaciones de las características aleatoria se realicen bajo condiciones causales semejantes, lo cual obliga a reflexionar sobre las condiciones en las que están siendo observadas las realizaciones de los sucesos. Una ventaja de esta forma de medir las probabilidades estriba en que información en forma de porcentajes (información relativa de ocurrencia de un suceso) puede ser convertida en probabilidades sin dificultad.
En definitiva, el método de cálculo derivado de la interpretación objetiva suele ser el más utilizado, pero, especialmente en el ámbito sanitario, en el que la gran variabilidad biológica del ser humano sugiere gran dificultad para aceptar condiciones constantes, conviene tener presente las limitaciones de esta interpretación.

Cuantificación de probabilidades en términos de *odds*

Una forma utilizada con cierta frecuencia para expresar la probabilidad de un suceso se basa en establecer la *razón de su probabilidad a la del suceso complementario*. Esta idea, que deriva del entorno de los juegos de azar (apuestas), es conocida en el ámbito anglosajón como *odds*. Dado un suceso A, se tiene:

$$\mathbf{odds(A) = \frac{p(A)}{1 - p(A)}}$$

La medida así obtenida representa la razón (o exceso si es mayor que 1) entre la probabilidad de que se presente A frente a que no se presente. Esta forma puede ser utilizada para cuantificar la probabilidad de un suceso, puesto que, conocida su *odds*, se tendrá:

$$\mathbf{p(A) = \frac{odds(A)}{1 + odds(A)}}$$

Así, por ejemplo, supóngase que se sabe que es 5 veces más probable que un individuo esté afectado por una determinada patología (A) que no lo esté. ¿Cuál es la probabilidad de que esté afectado por esa patología?

La información disponible se expresa en términos de *odds*(A) = 5. Por lo tanto:

$$\mathbf{p(A) = \frac{odds(A)}{1 + odds(A)} = \frac{5}{6} = 0{,}833}$$

2.3 PROBABILIDAD CONDICIONADA E INDEPENDENCIA

A partir de los conceptos expuestos se ha formalizado las ideas iniciales sobre probabilidades de resultados y sucesos del espacio muestral de una característica aleatoria. Sin embargo, conviene tener presente un fenómeno frecuente en el entorno sanitario y otros entornos: existen sucesos que modifican la probabilidad de otros sucesos. Así, es aceptado que es más probable encontrar un individuo hiperuricémico entre los hombres que entre las mujeres, o que es más probable desarrollar ciertas enfermedades si se es fumador que si no se fuma. De la misma forma, existen sucesos que no alteran la probabilidad de otros, como sucede al pensar que la probabilidad de que se produzca una infección postoperatoria no se verá alterada por el número de hermanos del paciente. Los argumentos expuestos conducen a las definiciones que a continuación se expone.

Probabilidad condicionada

Dados dos sucesos aleatorios, A y B, se define la probabilidad de A condicionada por B como la probabilidad de que se presente el suceso A cuando está presente el suceso B:

p(A / B) = probabilidad de A cuando ocurre B

Hay que tener en cuenta que al evaluar probabilidades condicionadas, el suceso A juega el papel de suceso condicionado y el B de condicionante, no siendo éste último un suceso aleatorio.
En muchas ocasiones el condicionante establecido conduce a contemplar la probabilidad de A entre un subconjunto de individuos, aquellos que verifican B.

Así, considérese la distribución de frecuencias expuesta en la tabla adjunta.
Supóngase que extraemos al azar un individuo cualquiera de entre los 200 considerados y sean los sucesos O=Consumir alcohol ocasionalmente, y M=Tener el nivel de glucosa superior a 140 mg/100ml. Se desea averiguar la probabilidad de que el individuo seleccionado tenga su nivel de glucosa por encima de 140 mg/100ml, condicionado por el hecho de que consume alcohol ocasionalmente.

Alcohol \ Glucosa	≤ 140	> 140
Nunca	122	14
Ocasionalmente	58	5

Se deberá determinar:

p(M / O) = probabilidad de M cuando ocurre O

actuando el suceso O como condicionante y el M como condicionado. Desde una aproximación frecuentista puede ser estimada como:

$$p(M/O) = \frac{\text{nº individuos que verifican M}}{\text{nº individuos que verifican O}} = \frac{5}{63} = 0{,}079$$

Nótese que si no consideramos la condición O, la probabilidad del suceso M será:

$$p(M) = \frac{\text{nº individuos que verifican M}}{\text{nº individuos en total}} = \frac{19}{200} = 0{,}095$$

..

Ley multiplicativa

La siguiente ecuación establece una forma de evaluar la probabilidad de sucesos condicionados por otros en función de probabilidades de intersecciones y de sucesos no condicionados:

Dados dos sucesos A y B,

$$p(A / B) = \frac{p(A \cap B)}{p(B)} \quad \text{si } p(B) \neq 0$$

$$p(B / A) = \frac{p(A \cap B)}{p(A)} \quad \text{si } p(A) \neq 0$$

Las ecuaciones establecidas pueden ser utilizadas, despejando, para determinar la probabilidad de la intersección de sucesos cuando se conoce las probabilidades condicionales y las de los propios sucesos no condicionados.

Considérese la situación ya descrita, referente a ciertas alteraciones neurológicas en pacientes diagnosticados de demencia. Se tenía probabilidades de Leucaraiosis, p(L) = 0,50, de Atrofia cortical, p(A) = 0,40, y de ambas alteraciones 0,10. Para un determinado paciente al azar, del que se sabe que presenta Leucaraiosis, se quiere determinar la probabilidad de que posea Atrofia cortical.

Se trata de obtener la probabilidad del suceso L, bajo la condición de verificar A. El suceso A no es aleatorio, pues conocemos ya el resultado, existiendo presencia de Atrofia cortical. Se tendrá, por la ley multiplicativa:

$$p(A / L) = \frac{p(A \cap L)}{p(L)} = \frac{0{,}10}{0{,}50} = 0{,}20$$

probabilidad inferior a la que tendría si no dispusiéramos de la información sobre la presencia de Leucaraiosis.

Sucesos independientes

En las definiciones anteriores se ha introducido la necesidad de considerar la 'influencia' de unos sucesos sobre otros al evaluar la probabilidad de que éstos se presenten. Como se ha visto las probabilidades de diversos sucesos pueden ser distintas si son evaluadas bajo la presencia de un condicionante a si lo son sin considerar este condicionante. Sin embargo existen situaciones en las que esto no tiene porque suceder.

Dados dos sucesos A y B, se dirá que son independientes si:

$$p(A / B) = p(A) \qquad p(B / A) = p(B)$$

es decir, si ninguno de ellos condiciona o modifica la probabilidad del otro.

El estudio de la independencia entre sucesos, o de forma más general entre características o variables aleatorias es de máximo interés en el entorno sanitario y biomédico. De hecho, para muchas de las técnicas inferenciales que se describirán más adelante el objetivo subyacente o explícito es ese.

Propiedades sobre probabilidades condicionales

Dados dos sucesos A y B, se verifican los siguientes resultados

- A y B serán independientes si y sólo si $\mathbf{p(A \cap B) = p(A) \cdot p(B)}$

Esta ecuación surge como aplicación inmediata de la ley multiplicativa al considerar la definición de independencia. Es una ecuación que puede ser utilizada como definición de independencia, pues si se verifica también se verificará la definición de independencia.

- Si A y B son independientes también lo serán las parejas $\overline{A}, B; \overline{A}, \overline{B}$; y $A, \overline{B}$.

- Dada la probabilidad de uno de los sucesos condicionado por el otro, p(A / B), se tiene que:

$$\mathbf{p(\overline{A} / B) = 1 - p(A / B)}$$

es decir, la evaluación de la probabilidad del complementario de un suceso condicionado por otro requiere que se mantenga el mismo condicionante.

Para ejemplificar las ideas sobre condicionalidad e independencia, considérese la siguiente situación:
En un estudio sobre cumplimentación terapéutica de pacientes hipertensos se estimó en un 80% el porcentaje de enfermos que seguían correctamente el tratamiento. Para mejorar la adherencia al tratamiento (cumplimentación) se diseñó una estrategia consistente en recuerdos telefónicos periódicos y apoyo familiar. Dicha estrategia se aplicó sobre el 50% de los enfermos. A posteriori se determinó que el 40% de los casos seguían bien el tratamiento y se les había aplicado la estrategia. A partir de esta información, se discutirá si los sucesos aplicar la estrategia (E) y seguir correctamente el tratamiento (T) son independientes.

Por la información disponible, y para un paciente hipertenso cualquiera, se tiene que:

$$\mathbf{p(T) = 0{,}80} \qquad \mathbf{p(E) = 0{,}50} \qquad \mathbf{p(T \cap E) = 0{,}40}$$

Para que los sucesos T y E sean independientes es suficiente que $\mathbf{p(T \cap E) = p(T) \cdot p(E)}$

$$\mathbf{p(T \cap E) = 0{,}40}$$

$$\mathbf{p(T) \cdot p(E) = 0{,}80 \cdot 0{,}50 = 0{,}40}$$

resultando que sí son independientes, es decir la probabilidad de que un paciente siga correctamente el tratamiento no se ve condicionada por el hecho de habérsele aplicado la estrategia de mejora.

2.4 TEOREMAS BASICOS: TEOREMA DE LA PROBABILIDAD TOTAL Y TEOREMA DE BAYES

Como consecuencia de la definición de probabilidad condicionada, se enuncian dos *teoremas*[2] cuya utilidad se desprende de la necesidad de obtener probabilidades de ciertos sucesos en función de las probabilidades de otros. Una situación particular en la que los teoremas se muestran especialmente útiles

2 Un teorema se enuncia en base a unas condiciones tales que si se verifican siempre se desprende un resultado.

es para el diagnóstico y/o detección de enfermedad, casos en los que la incertidumbre sobre el estado de salud o enfermedad puede verse modificada a través de alguna intervención encaminada a tal fin. Este aspecto será desarrollado en el siguiente apartado.
Los teoremas básicos son el teorema de la probabilidad total y el de Bayes. Ambos teoremas parten de las mismas condiciones, y su enunciado es el siguiente:

Dada una partición del espacio muestral, digamos $\{A_i\}_{i=1}^{k}$, y otro suceso aleatorio cualquiera, digamos B, se tiene que:

- $p(B) = \sum_{i=1}^{k} p(B \cap A_i) = \sum_{i=1}^{k} p(B / A_i) \cdot p(A_i)$ ***(Teorema de la probabilidad total)***

- $p(A_i / B) = \frac{p(B / A_i) \cdot p(A_i)}{p(B)} = \frac{p(B / A_i) \cdot p(A_i)}{\sum_{i=1}^{k} p(B / A_i) \cdot p(A_i)}$ **para cualquier A_i** ***(Teorema de Bayes)***

El teorema de la probabilidad total permite obtener la probabilidad de un suceso cuando lo que se conoce es su probabilidad condicionada por una familia de sucesos que es una partición del espacio muestral. El teorema de Bayes permite invertir el sentido de la condicionalidad al evaluar probabilidades, resultado especialmente útil en el entorno sanitario, como se verá más adelante.

Para ejemplificar estos teoremas, supóngase que en cierta zona industrial de una comunidad se ha calificado las diferentes industrias de cierto ramo de producción según de tipo I, II y III en función del mayor o menor nivel de medidas de seguridad e higiene de que dispongan. Cierto tipo de accidente (A) se da, sobre los trabajadores de este ramo, en porcentajes del 1%, 5% y 10% según la industria sea calificada como de tipo I, II o III. A su vez, el 60% de las industrias son de tipo I, el 35% de tipo II y el 5% de tipo III. Se desea calcular la probabilidad de que un trabajador de este ramo sufra un accidente de este tipo.

Si identificamos por I, II, y III los sucesos que representan que una industria cualquiera sea del tipo correspondiente, se tendrá que la probabilidad de que una industria cualquiera sea de los tipos establecidos será:

$$p(I) = 0{,}60 \qquad p(II) = 0{,}35 \qquad p(III) = 0{,}05$$

siendo los sucesos {I, II, III} una partición del espacio muestral, pues cada industria puede ser calificada sólo de un tipo, y todas las industrias reciben alguna calificación. Por otra parte, la probabilidad de que se produzca un accidente será,

$$p(A / I) = 0{,}01 \quad p(A / II) = 0{,}05 \quad p(A / III) = 0{,}10$$

siendo éstas probabilidades condicionales. La probabilidad global de que se produzca un accidente A puede ser obtenida por el teorema de la probabilidad total:

$$p(A) = p(A / I) \cdot p(I) + p(A / II) \cdot p(II) + p(A / III) \cdot p(III) =$$
$$= 0{,}01 \cdot 0{,}60 + 0{,}05 \cdot 0{,}35 + 0{,}10 \cdot 0{,}05 = 0{,}0285$$

Supongamos ahora que llega un individuo que trabaja en esa zona industrial y que presenta el accidente A. Se desea averiguar la probabilidad de que trabaje en un industria de tipo III.

Se trata de calcular p(III / A), que, por el teorema de Bayes será:

$$p(III / A) = \frac{p(A / III) \cdot p(III)}{p(A)} = \frac{0,10 \cdot 0,05}{0,0285} = 0,1754$$

Nótese que la probabilidad deseada, p(III / A), es la inversa de p(A / III), dato disponible inicialmente.

2.5 APLICACION DE LOS TEOREMAS BASICOS AL DIAGNOSTICO Y/O DETECCION DE ENFERMEDAD

Se puede decir que una característica fundamental en la determinación del estado de enfermedad (o salud) a lo largo de la vida de una persona es la incertidumbre, entendida ésta como la ausencia de seguridad absoluta sobre tal estado. En estas condiciones, una herramienta habitual del entorno sanitario son las *pruebas diagnósticas*. Se entenderá por prueba diagnóstica para una determinada enfermedad como un conjunto de intervenciones sobre un individuo encaminadas a 'determinar' la presencia o no de enfermedad o su grado. El objetivo de una prueba diagnóstica debe ser la reducción del nivel de incertidumbre sobre el estado de enfermedad de la persona. Idealmente, una prueba diagnóstica sería exacta si de la información que proporciona (resultados de la prueba) se desprende con exactitud el estado de enfermedad, haciendo desaparecer la incertidumbre inicial. Hay que decir que estas ideas y conceptos tienen aplicación en otros ámbitos menos sanitarios. Por ejemplo, podemos hablar de pruebas para detectar la presencia de un agente contaminante en un alimento, o la presencia de cierto grado de azúcar en un melón, u otros. Las expresiones, fórmulas, y relaciones que a continuación se exponen pueden ser aplicadas a cualquier situación de detección de la presencia o ausencia de un resultado.

Espacio muestral asociado a enfermedad y prueba diagnóstica

Se supondrá para el desarrollo posterior que la situación aleatoria procede de la consideración de los resultados de la característica aleatoria bivariante *enfermedad-prueba diagnóstica*, para la que se simplificará los posibles resultados a:

Resultados en enfermedad	Resultados en la prueba diagnóstica
Enfermo (**E**)	*Positivo* (**+**)
No enfermo ($\bar{\mathbf{E}}$)	*Negativo* (**-**)

Los resultados en enfermedad pueden ser los de estar o no realmente afectado por la enfermedad considerada, mientras que los resultados de la prueba diagnóstica se reducirán a que del conjunto de intervenciones se desprende una posible calificación de enfermedad (resultado positivo) o se desprende la posible ausencia de enfermedad (negativo).

El espacio muestral bivariante, construido con los resultados simples posibles será:

$$\textbf{Espacio muestral} = \left\{ \mathbf{E} \cap +, \mathbf{E} \cap -, \bar{\mathbf{E}} \cap +, \bar{\mathbf{E}} \cap - \right\}$$

que constituye todos los resultados simples generados a partir de la consideración bivariante de los resultados en enfermedad y prueba diagnóstica.
Suponga que una nueva prueba para el diagnóstico de cierta patología ha sido ensayada sobre 200 enfermos por esa causa y sobre 800 personas libres de la enfermedad. El resultado de la prueba ha sido positivo (la prueba detecta enfermedad) en 190 de los enfermos, y ha sido negativo (la prueba no detecta enfermedad) en 680 de los 800 no enfermos por esa causa. A partir de estos datos se desea estructurar el espacio muestral bivariante asociado a enfermedad y prueba diagnóstica, identificando los sujetos que verifican sus resultados simples.
La estructuración de la información relativa al espacio muestral suele realizarse en forma de tabla de doble entrada:

ESPACIO MUESTRAL ENFERMEDAD-PRUEBA		**Resultados en enfermedad**		Total
		E	_	
Resultados	+	190	120	310
de la prueba	-	10	680	690
diagnóstica	Total	200	800	1000

La frecuencia conjunta expresada en cada celda recoge el número de individuos que verifican cada uno de los resultados simples del espacio muestral, generado a partir de las intersecciones entre los resultados en enfermedad y en prueba diagnóstica.

Probabilidades de interés sobre sucesos del espacio muestral asociado a enfermedad y prueba diagnóstica

En la práctica, los resultados de la prueba diagnóstica dependerán del estado en enfermedad, mientras que, una vez realizada la prueba diagnóstica, las probabilidad de los resultados en enfermedad debe verse modificada según el resultado obtenido en la prueba diagnóstica. Así, en el caso del ejemplo del apartado anterior, surgen diversas preguntas de interés como por ejemplo ¿puede decirse que esa prueba es ´valida' para aplicarla sobre personas cuyo estado respecto a la enfermedad estudiada es desconocido e incierto?, ¿qué sucedería si utilizamos esa prueba en una campaña de detección precoz de la enfermedad en población general, donde se estima que la enfermedad está presente en el 0,5% de los individuos?, ó, si la enfermedad estudiada se presentase en cierto servicio hospitalario con frecuencia del 30% de los casos, ¿influiría este hecho sobre la posible aplicación de la prueba de forma rutinaria?. La respuesta a estas preguntas pasa por conocer diversas probabilidades condicionales.

Las probabilidades de los sucesos de mayor interés son:

Probabilidad de resultados de la prueba diagnóstica

- $p(+ / E)$ *Sensibilidad de la prueba*
- $p(- / \bar{E})$ *Especificidad de la prueba*
- $p(+ / \bar{E})$ *Falso positivo de la prueba*
- $p(- / E)$ *Falso negativo de la prueba*
- $p(+)$ *Probabilidad de resultado positivo de la prueba*
- $p(-)$ *Probabilidad de resultado negativo de la prueba*

Probabilidad de resultados en enfermedad

- $p(E / +)$ *Valor predictivo positivo*
- $p(\bar{E} / -)$ *Valor predictivo negativo*
- $p(\bar{E} / +)$ *Valor predictivo falso positivo*
- $p(E / -)$ *Valor predictivo falso negativo*
- $p(E)$ *Probabilidad de enfermedad (prevalencia o probabilidad 'a priori')*
- $p(\bar{E})$ *Probabilidad de no estar enfermo*

Las probabilidades recogidas en cada una de las columnas representan probabilidades condicionales o globales de sucesos diferentes. Por una parte tenemos los resultados de la prueba (+, -), cuyas probabilidades deben depender de si un individuo posee o no la enfermedad. Idealmente **p(+ / E)**, o sensibilidad de la prueba, y $\mathbf{p(- / \bar{E})}$, o especificidad de la prueba, deberían valer 1, es decir, si un individuo está realmente enfermo la prueba debería detectarlo con seguridad, mientras que si no posee la enfermedad también debería detectarlo con seguridad. Una prueba con estas características se denomina *gold standard*. Sin embargo, esto no es así para un gran número de pruebas diagnósticas, en cuyo caso, **p(- / E)** y $\mathbf{p(+ / \bar{E})}$, que son los complementarios de la sensibilidad y especificidad, respectivamente, son una medida del error o falta de acierto de la prueba en la detección de enfermedad o de ausencia de enfermedad.
Por otra parte, las probabilidades recogidas en la segunda columna se refieren a los resultados en enfermedad $\{\mathbf{E, \bar{E}}\}$, de forma condicional a los resultados de la prueba o de forma global. Si se piensa en un individuo cuyo estado en enfermedad es incierto, y al que se le aplica la prueba diagnóstica para disminuir esa incertidumbre, **p(E /+)**, o valor predictivo positivo, y $\mathbf{p(\bar{E} / -)}$ miden la probabilidad de que realmente esté enfermo cuando el resultado de la prueba ha sido positivo, y de que realmente no esté enfermo si el resultado de la prueba es negativo. Idealmente estas probabilidades deberían valer 1, es decir sea cual sea el resultado de la prueba sería deseable saber con seguridad si el individuo está o no enfermo. Sin embargo, este resultado sólo se produce si la sensibilidad y especificidad de la prueba toman el valor 1 (caso de una prueba *gold standard*. Además, y como se verá a continuación, existirá una relación entre las probabilidades de los resultados de la prueba y las de los resultados en enfermedad, dependiente de p(E) o prevalencia de la enfermedad.

Obtención de los valores predictivos en función de la sensibilidad y especificidad de la prueba

Si se conoce la sensibilidad y especificidad de la prueba, es posible conocer los valores predictivos a través de los teoremas de la probabilidad total y de Bayes. Para ello, baste considerar que la pareja de sucesos $\{\mathbf{E, \bar{E}}\}$ son una partición del espacio muestral asociado y que los sucesos {+, -} son otros sucesos posibles del espacio muestral, en cuyo caso se tendrá, por el teorema de Bayes:

$$\mathbf{p(E / +) = \frac{p(+ / E) \cdot p(E)}{p(+)} = \frac{p(+ / E) \cdot p(E)}{p(+ / E) \cdot p(E) + p(+ / \bar{E}) \cdot p(\bar{E})}}$$

o, equivalentemente:

$$\mathbf{Valor\ predictivo\ positivo = \frac{Sensibilidad \times Prevalencia}{Sensibilidad \times Prevalencia + (1 - Especificidad) \times (1 - Prevalencia)}}$$

y

$$\mathbf{p(\bar{E} / -) = \frac{p(- / \bar{E}) \cdot p(\bar{E})}{p(-)} = \frac{p(- / \bar{E}) \cdot p(\bar{E})}{p(- / \bar{E}) \cdot p(\bar{E}) + p(- / E) \cdot p(E)}}$$

o, equivalentemente:

$$\mathbf{Valor\ predictivo\ negativo = \frac{Especificidad \times (1 - Prevalencia)}{Especificidad \times (1 - Prevalencia) + (1 - Sensibilidad) \times Prevalencia}}$$

Como se observa, es posible obtener los valores predictivos condicionales a los resultados de la prueba si se conoce la prevalencia (o probabilidad *a priori*), p(E), de la enfermedad. Este hecho hace que una misma

prueba puede presentar diferentes valores predictivos en función del valor de la prevalencia de la enfermedad.

Como ejemplo, considérese la información suministrada anteriormente, habiendo aplicado la prueba en la detección precoz de la enfermedad en una población sobre la que se sabe que la enfermedad se produce con frecuencia del 0,5% de los individuos, se desea cuantificar la sensibilidad, especificidad, falsos positivos y negativos de la prueba, y los valores predictivos positivo y negativo.

Recordemos que los datos disponibles eran:

ESPACIO MUESTRAL ENFERMEDAD-PRUEBA		**Resultados en enfermedad**		Total
		E	_	
Resultados	+	190	120	310
de la prueba	-	10	680	690
diagnóstica	Total	200	800	1000

por lo que tendremos:

$$\textbf{Sensibilidad} = p(+ / E) = \frac{190}{200} = 0{,}95$$

$$\textbf{Especificidad} = p(- / \bar{E}) = \frac{680}{800} = 0{,}85$$

$$\textbf{Falso positivo} = p(+ / \bar{E}) = \frac{120}{800} = 0{,}15$$

$$\textbf{Falso negativo} = p(- / E) = \frac{10}{200} = 0{,}05$$

Los valores predictivos dependerán de la prevalencia, que en este caso se estima a partir del dato frecuencial 0,5%, como p(E) = 0,005. En ese caso tendremos:

$$\textbf{Valor predictivo} + = p(E / +) = \frac{0{,}95 \,.\, 0{,}005}{0{,}95 \,.\, 0{,}005 + 0{,}15 \,.\, 0{,}995} = 0{,}0308$$

$$\textbf{Valor predictivo} - = p(\bar{E} / -) = \frac{0{,}85 \,.\, 0{,}995}{0{,}85 \,.\, 0{,}995 + 0{,}05 \,.\, 0{,}005} = 0{,}9997$$

Como se observa, el valor predictivo negativo resulta altamente satisfactorio, pues la probabilidad de 'acierto' cuando el resultado es negativo es muy elevada. Sin embargo, el valor predictivo positivo no resulta aparentemente satisfactorio, puesto que la probabilidad de acierto es muy baja.

Supóngase ahora que la prueba descrita va a ser aplicada en un servicio hospitalario en el que la enfermedad considerada se presenta en el 30% de los individuos que acuden a él. Se desea calcular los valores predictivos de la prueba.

Puesto que la prueba es la misma, la probabilidad de sus resultados sólo depende del estado en enfermedad o no, por lo que su sensibilidad y especificidad son las mismas.

Por otra parte, la prevalencia será de p(E) = 0,30, interpretable más bien como la probabilidad *a priori* de

que un individuo presente la enfermedad. Esta interpretación parece más razonable dado que el término prevalencia suele ser reservado a situaciones poblacionales. Se tendrá entonces:

$$\text{Valor predictivo} + = p(E / +) = \frac{0,95 \,.\, 0,30}{0,95 \,.\, 0,30 + 0,15 \,.\, 0,70} = 0,7308$$

$$\text{Valor predictivo} - = p(\bar{E} / -) = \frac{0,85 \,.\, 0,70}{0,85 \,.\, 0,70 + 0,05 \,.\, 0,30} = 0,9754$$

En comparación con los resultados obtenidos en el ejemplo anterior, el valor predictivo positivo es aparentemente sustancialmente mejor al ser mucho más elevado, mientras que el valor predictivo negativo disminuye.

Los resultados obtenidos en estos ejemplos ponen en evidencia la forma en que la prevalencia de la enfermedad afecta a los valores predictivos, si la prevalencia se incrementa, también lo hace el valor predictivo positivo, disminuyendo el valor predictivo negativo, mientras que se invierte el resultado al disminuir la prevalencia.

Evaluación de la bondad de una prueba diagnóstica

Si se atiende al objetivo fundamental de la prueba, la disminución de la incertidumbre en el estado de enfermedad, la bondad de una prueba diagnóstica dependerá de su *reproducibilidad*, o capacidad para reproducir los mismos resultados (+ o -) en condiciones y sujetos semejantes, y de su *validez,* o grado en el que de verdad mide lo que se desea medir. Con los elementos de probabilidad descritos se puede discutir la validez de la prueba, aceptando que ésta sea reproducible.
Así, la sensibilidad y la especificidad pueden ser medidas de su validez cuando la prueba es comparada con otras pruebas y ambas medidas se dirigen en el mismo sentido. En efecto, una prueba con mayor sensibilidad y especificidad que otra producirá mejores valores predictivos sea cual sea la prevalencia de la enfermedad. Sin embargo, en situaciones cruzadas (sensibilidad mayor en una prueba que en otra y lo contrario respecto a la especificidad) o cuando la evaluación de la bondad de la prueba no se realiza por comparación con ninguna otra, es necesario calcular los valores predictivos para asegurarnos del valor de los resultados de la prueba.

Un problema aparece cuando la prevalencia es desconocida, o sólo se tiene una idea del intervalo en el que podría situarse. En este caso, suele recurrirse a la construcción de las curvas para el valor predictivo positivo y el complemento del valor predictivo negativo o falsa predicción negativa.

2.6 VARIABLES ALEATORIAS

Desde un punto de vista formal una variable aleatoria se define como una función que asigna a cada uno de los posibles resultados de un fenómeno aleatorio un valor numérico. Por ejemplo, en el caso de variables cuantitativas como el nivel de colesterol, nivel de ácido úrico o, número de ingresos en un servicio de urgencias, la variable aleatoria vendría definida por cada una de las posibilidades de las variables consideradas, puesto que en este caso ya son valores numéricos. Cuando las variables son de tipo cualitativo como el sexo o el nivel de estudios, será necesario asignar valores numéricos a cada una de las posibilidades (1 Hombre 2 Mujer para el sexo, 1 Sin estudios 2 Primaria 3 Secundaria 4 Universitarios para el nivel de estudios). Las variables aleatorias se clasifican en variables aleatorias *discretas* y variables aleatorias *continuas*. Las variables aleatorias discretas pueden tomar un número finito o infinito numerable de valores, mientras que las continuas pueden alcanzar un número infinito no numerable de posibles valores, es decir, pueden tomar cualquier valor en un intervalo.
Desde un punto de vista más práctico, las variables aleatorias podrían considerarse como variables cuyos resultados se rigen por el azar. En este sentido es importante tener en cuenta que, como se ha mencionado con anterioridad, no todos los posibles valores de una variable aleatoria tienen la misma probabilidad de ser observados. Por tanto, sería útil contar con herramientas que proporcionen información sobre la

probabilidad asociada a cada uno de los valores de una variable aleatoria.

Función de probabilidad

En el caso de las variables aleatorias discretas, se define como *función de probabilidad* de la variable aleatoria X a una función *p*(x) que para cada uno de los valores de la variable, por ejemplo x_0, le asigna su probabilidad, esto es:

$$p(x_0) = P(X = x_0)$$

Función de densidad de probabilidad

Para variables aleatorias continuas se define como *función de densidad de probabilidad f*(x) a una función no negativa que verifica:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

y donde la probabilidad de que la variable X tome valores entre x_1 y x_2 puede calcularse de la forma:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

Función de distribución

Tanto para variables aleatorias discretas como continuas puede definirse la *función de distribución* de la variable X, como una función F(x) que asigna, para cada valor concreto, la probabilidad de que la variable tome un valor menor o igual a él, es decir:

$$F(x_0) = p(X \leq x_0)$$

Para variables aleatorias discretas puede expresarse la función de distribución en función de la función de probabilidad de la forma:

$$F(x_0) = P(X \leq x_0) = \sum_{x=0}^{x_0} p(x)$$

Para variables aleatorias continuas, la función de distribución podrá expresarse en función de la función de densidad de probabilidad tal y como sigue:

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)dx$$

Características de una variable aleatoria

En el capítulo de exploración de datos se definieron diversas medidas descriptivas (medidas de tendencia central, dispersión y forma) calculables a partir de un determinado conjunto de datos. Cuando se trabaja con una variable aleatoria es posible obtener expresiones para el cálculo de este tipo de medidas basándose en las probabilidades asociadas a cada uno de los posibles valores de dicha variable, aunque hay que tener en cuenta que el sentido de estas medidas será diferente. Mientras que en el caso descriptivo representan las características de *lo que ha sucedido* (los datos observados), en el caso que nos ocupa, representarán las características *esperadas o esperables* de nuestras variables en la población a estudio.

En el cuadro 2.1 se presenta la *media* o *esperanza* de una variable aleatoria y la *varianza*, distinguiendo entre variables aleatorias discretas y continuas.

Cuadro 2.1.- Características de una variable aleatoria

Característica	Nomenclatura	Discretas	Continuas
Media o Esperanza	$\mu = E(X)$	$\sum x_i p(x_i)$	$\int_{-\infty}^{+\infty} xf(x)dx$
Varianza	$\sigma^2 = Var(X)$	$\sum(x_i - \mu)^2 p(x_i)$	$\int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$

La desviación típica se obtendría como la raíz cuadrada de la varianza de la variable aleatoria.

Para ejemplificar estas medidas, suponga que la información recogida por un equipo de investigación sobre el número de ingresos hospitalarios diarios debidos a determinada patología permitió obtener la función de probabilidad que aparece reflejada en el cuadro 2.2

Cuadro 2.2.- Probabilidad del número de ingresos/día

Número de ingresos/día x	Función de Probabilidad p(x)=P(X=x)	Función de disribución F(x)=P(X≤x)
0	0,10	0,10
1	0,15	0,25
2	0,20	0,45
3	0,26	0,71
4	0,21	0,92
5	0,08	1,00
Total	1,00	1,00

Como puede observarse, es posible calcular cualquier tipo de probabilidad sobre la variable aleatoria X=número de ingresos diarios. Por ejemplo, la probabilidad de que se produzcan al menos 3 ingresos diarios podría calcularse:

$$p(X \geq 3) = p(3) + p(4) + p(5) = p(X = 3) + p(X = 4) + p(X = 5) = 0{,}26 + 0{,}21 + 0{,}08 = 0{,}55$$

Obsérvese que también podría calcularse a partir de la función de distribución P(X≥3)=1-P(X≤2)=1-F(2)=1-0,45=0,55. Para la probabilidad de que el número de ingresos no supere los 2 diarios se obtendría

$$P(X \leq 2) = p(0)+p(1)+p(2) = 0.10+0.15+ 0.20 = 0.45 = F(2)$$

Por otra parte el número de ingresos esperado en un día determinado vendría determinado por la media o esperanza de la variable aleatoria

$$E(X) = \sum x_i p(x_i) = 0.0{,}10 + 1.0{,}15 + 2.0{,}20 + 3.0{,}26 + 4.0{,}21 + 5.0{,}08 = 2{,}57 \quad \textbf{ingresos/día}$$

con una varianza:

$$Var(X) = 0{,}10(0-2{,}57)^2+0{,}15(1-2{,}57)^2+0{,}20(2-2{,}57)^2+0{,}26(3-2{,}57)^2 +$$

$$+0{,}21(4-2{,}57)^2 +0{,}08(5-2{,}57)^2=4{,}14 \ (\text{ingresos/día})^2$$

y una desviación típica

$$D(X)=2{,}03 \ \text{ingresos/día}$$

Luego, tal y como ha podido apreciarse en el ejemplo, el conocimiento de la función de probabilidad *p*(x) o la función de distribución de una variable aleatoria discreta, posibilita el cálculo de la probabilidad de cualquiera de los sucesos relacionados con los posibles valores de dicha variable. Análogamente, cualquier probabilidad relacionada con los posibles valores de una variable aleatoria continua podrá ser obtenida a partir de la función de densidad de probabilidad *f*(x) o la función de distribución. En definitiva, los *modelos de probabilidad* no serán más que expresiones concretas para estas funciones que describirán el comportamiento probabilístico de la variable.

El modelo Binomial

Algunos de los fenómenos aleatorios objeto de estudio pueden dar lugar únicamente a dos posibles resultados, por ejemplo, *salud/enfermedad*, *positivo/negativo*, *fumador/no fumador,* etc. Esta situación dicotómica da lugar al planteamiento de diversos modelos de probabilidad, según la variable aleatoria que se defina, y entre ellos al del modelo de probabilidad binomial. Para comprender los fundamentos en la obtención del modelo binomial y, en general, de la importancia de los modelos de probabilidad se plantea el siguiente problema:

Suponga que se conoce que el 65% de los pacientes afectados por determinada patología responde positivamente al tratamiento. Si consideramos un grupo de 3 pacientes afectados, ¿cuál será la probabilidad de que 2 de ellos respondan de forma positiva al tratamiento?

Sea el suceso A ={2 pacientes responden positivamente de entre los 3 observados}.Tres son los sucesos que darían como resultado que 2 de los pacientes evolucionasen de forma favorable y que aparecen reflejados en el cuadro 2.3

Cuadro 2.3.-. Posibilidades favorables al caso en que 2 de los 3

individuos analizados respondan positivamente al tratamiento

Suceso	Paciente 1	Paciente 2	Paciente 3
A1	+	+	-
A2	+	-	+
A3	-	+	+

En realidad el número de sucesos podría obtenerse a partir del cálculo del número de combinaciones de 3 elementos de orden 2. Si se considera que el hecho de que un paciente responda de forma positiva al tratamiento es independiente de lo que ocurra con otro paciente, la probabilidad de cada uno de estos 3 sucesos puede obtenerse como el producto de probabilidades. Así, se tiene que:

$$P(A1) = P(+)\,P(+)\,P(-) = 0{,}65.0{,}65.(1\text{-}0{,}65) = 0{,}1478$$

$$P(A2) = P(+)\,P(-)\,P(+) = 0{,}65.(1\text{-}0{,}65).0{,}65 = 0{,}1478$$

$$P(A3) = P(-)\,P(+)\,P(+) = (1\text{-}0{,}65).0{,}65.0{,}65 = 0{,}1478$$

De esta forma, la probabilidad del suceso A podría calcularse como la suma de las probabilidades de los tres sucesos descritos con anterioridad A1, A2, y A3.

$$p(A) = p(A1) + p(A2) + p(A3) = 3\,.\,0{,}65^{2}\,.\,(1-0{,}65)^{1} = 0{,}4436$$

En general, supóngase que un fenómeno aleatorio únicamente puede dar lugar a dos posibles resultados (positivo/negativo, curación/no curación, supervivencia/muerte, enfermo/no enfermo, etc) y que la probabilidad de uno de los dos sucesos es *p* (por tanto la probabilidad del otro será 1-*p*). Si se obtienen *n* observaciones independientes del fenómeno aleatorio correspondientes a un grupo de *n* individuos, la probabilidad de que se observe la ocurrencia del suceso en k de los n individuos podrá determinarse a partir de la función de probabilidad para el modelo binomial:

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

donde X =número de individuos en los que se observa el suceso y

$$\text{media o esperanza} = E(X) = np$$
$$\text{varianza} = V(X) = np(1-p)$$

Si se hubiera utilizado el modelo de probabilidad binomial para resolver el ejercicio anterior se obtendría:

$$p(2) = P(X = 2) = \binom{3}{2} 0{,}65^{2} (1- 0{,}65) = 0{,}4436$$

La media o esperanza será

$$E(X) = np = 3.0{,}65 = 1{,}95 \text{ individuos}$$

mientras que la varianza quedará

$$Var(X) = np(1-p) = 3.0{,}65.(1-0{,}65) = 0{,}6825 \text{ individuos}$$

El modelo Poisson

Otro modelo discreto que se presenta con frecuencia en el ámbito de las ciencias de la salud es el modelo de Poisson. Sea λ el promedio de ocurrencias de un determinado suceso en un intervalo de tiempo o espacio. Además supóngase que se verifican las siguientes condiciones:

1. Las ocurrencias del suceso son independientes
2. Es posible observar un número infinito de ocurrencias en cada intervalo
3. La probabilidad de ocurrencia del suceso en un intervalo es proporcional a su amplitud

La variable aleatoria X=*número de ocurrencias en ese intervalo de tiempo o espacio* se distribuye según un modelo de Poisson, siendo su función de probabilidad:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}; \quad k = 0, 1, 2, 3, \ldots$$

donde:

$$\text{media o esperanza} = E(X) = \lambda = V(X) = \text{Varianza}$$

Para ejemplificar el modelo, suponga que es conocido el número promedio de ingresos en el servicio de Ginecología y Obstetricia de un Hospital y es de 4 al día. Si se supone que el número de ingresos se distribuye según un modelo de Poisson, ¿cuál sería la probabilidad de que se produjeran más de 4 ingresos en un día determinado?

$$P(X > 4) = 1 - P(X \leq 4) = 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)]$$
$$= 1 - \left[\frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!} + \frac{\lambda^4 e^{-\lambda}}{4!} \right]$$

con $\lambda=4$

El modelo normal

El modelo normal constituye la distribución de probabilidad para variables aleatorias continuas más importante de toda la estadística, debido a que en la naturaleza muchas de las variables, y en particular las relacionadas con procesos de medición, se comportarían de forma aproximada según este modelo de probabilidad y, sobre todo, porque un resultado como el *teorema central del límite* asignará al modelo normal un papel destacado en el ámbito de la estadística inferencial. La función de densidad de probabilidad que describe el modelo normal es:

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

donde:

Media o esperanza = E(X) = μ
Varianza = V(X) = σ^2

Como puede observarse en la figura 2.1, la distribución de probabilidad normal es una distribución simétrica respecto de la media de la variable que es μ que coincide además con el valor donde se alcanza el máximo de la función de densidad *f*(x), resultando ser también mediana y moda.

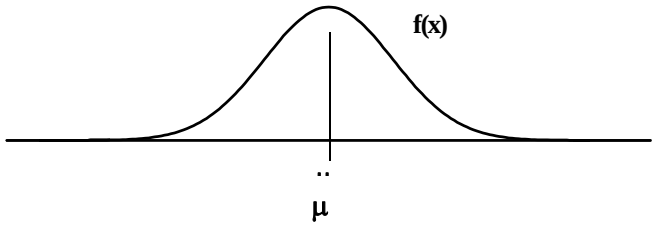


Figura 2.1.- Función de densidad del modelo normal

Como ejemplo, suponga que el nivel de colesterol en cierta población se distribuye según un modelo aproximadamente normal con media 210 mg/100 ml y una desviación típica de 15 mg/100 ml, ¿cuál es la probabilidad de que un individuo de esta población, seleccionado al azar presente un nivel de colesterol inferior a 225 mg/100 ml?

Tendríamos que resolver:

$$p(X < 225) = \int_{-\infty}^{225} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx$$

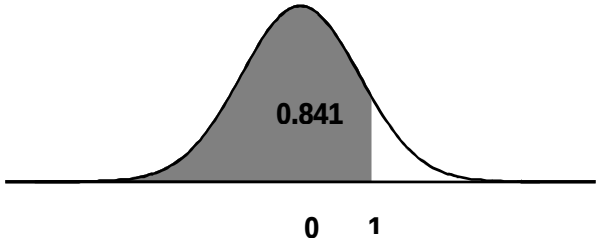
pero en la práctica no es necesario resolver las integrales, puesto que éstas se encuentran en tablas de probabilidad (ver tablas modelo normal).Dada la imposibilidad de contar con una tabla para cada una de las distribuciones normales con cualquier media y cualquier desviación típica, el cálculo de probabilidades sobre este tipo de distribuciones se realiza a partir de una única tabla correspondiente a la distribución normal con media 0 y desviación típica 1 que recibe el nombre de normal estándar. Para ello será necesario transformar la variable objeto de estudio en una variable con media 0 y desviación típica 1, proceso que recibe el nombre de tipificación, y que se consigue restando la media y dividiendo por la desviación típica. Así se tiene que:

$$P(X \leq 225) = P\left(\frac{X - \mu}{\sigma} \leq \frac{225 - 210}{15}\right)$$

$$= P\left(Z \leq \frac{225 - 210}{15}\right)$$

$$= P(Z \leq 1{,}00) = 0{,}8413$$

(ver tabla 1 en Anexo de tablas)



Otros modelos continuos

Para representar situaciones de la realidad cuyo ajuste a un modelo normal es dudoso o claramente divergente, la estadística matemática ha ido elaborando otros modelos para variables continuas. Así, algunos de los que cuentan con mayor visibilidad serían los modelos *t de Student, F de Snedecor* o *Ji-cuadrado* (ver tablas 3, 4 y 5 en anexo de tablas). Estos tres modelos serán utilizados en las aplicaciones de inferencia, en las que aparecen como herramientas de valoración probabilística del error aleatorio cometido. No obstante existen muchos otros modelos continuos (*Gamma, Beta, Exponencial,...*) que proponen a través de sus funciones de densidad otras posibles representaciones de la realidad.

C A P I T U L O 3

FUNDAMENTOS PARA EXTRAER CONCLUSIONES DE LOS DATOS: LA INFERENCIA ESTADISTICA

Hasta el momento se han descrito técnicas de organización, presentación y resumen de datos que proporcionan información sobre determinadas características de las variables y datos observados. Sin embargo, uno de los objetivos básicos en la investigación de cualquier fenómeno aleatorio es extraer conclusiones acerca de una característica de interés sobre la *población* objeto de estudio, cuando solamente se ha podido observar una pequeña parte de dicha población (*muestra*). Observe las siguientes situaciones:

- *Un investigador recopila información sobre el tiempo de supervivencia correspondiente a 41 pacientes intervenidos quirúrgicamente de determinada afección. La media de supervivencia en los 41 pacientes observados es de 17 años con una desviación típica de 1,5 años. Su intención es comprobar, a partir de los datos disponibles, que el tiempo de supervivencia medio desde el momento de la intervención y en este tipo de pacientes es superior a 15 años.*

- *En un estudio se ha recogido información sobre el nivel de colesterol correspondiente a cada uno de los 250 pacientes analizados. El nivel promedio de colesterol observado en los 250 pacientes es de 200 mg/100 ml. El objetivo del estudio es estimar el nivel promedio de colesterol en toda la población de pacientes de la que partió la muestra.*

En ambos casos, no se está interesado en comprobar si el tiempo de supervivencia medio en los 41 pacientes observados es o no superior a 15 años o si el promedio de nivel de colesterol en los 250 pacientes es 200 mg/100 ml, sino si esto ocurre en toda la población de pacientes afectados por esa patología y de la que los 41 y los 250 individuos observados no son más que una pequeña parte. Este proceso de generalización de resultados de la muestra a la población, conocido como *inferencia estadística,* constituirá el objetivo fundamental del presente capítulo.

3.1 LA ESTIMACIÓN

En la introducción anterior se presentaba un ejemplo en el que el investigador o investigadores pretendían estimar el valor del nivel promedio de colesterol μ en una población en la que tan sólo habían sido capaces de observar a 250 pacientes y en los que el nivel promedio era de 200 mg/100 ml. Para responder a esta cuestión existen dos alternativas:

-proporcionar un único valor para la media poblacional μ

-proporcionar un intervalo que contendrá al verdadero valor de la media poblacional μ con una determinada probabilidad de error.

Estimación puntual y estimación por intervalos

La diferencia principal entre estos dos métodos estriba en que en el primero de los casos (estimación *puntual*) no se proporciona ningún tipo de información sobre la magnitud probable de μ, ni sobre el error que rodea a la estimación, algo que si ocurre en el segundo de los casos (estimación por *intervalos*: *intervalos de confianza*) y que constituirá una de las dos técnicas básicas para la realización de inferencias. Sin embargo, en el proceso de estimación por intervalos y como se comprobará más tarde, será necesario contar con las estimaciones puntuales correspondientes a cada uno de los parámetros de interés. En el ejemplo, el nivel medio de colesterol en la población sería el parámetro que se pretende estimar, pero ¿cuál sería el estimador puntual que debería utilizarse para aproximar el valor de μ? Según el criterio de *máxima verosimilitud*, basado en la información que proporcionan los datos observados y en el cálculo de probabilidades, la media muestral sería, en este caso, el mejor estimador que podría utilizarse para aproximar el valor de la media desconocida de la población μ. En el cuadro 3.1 se presenta algunos de los parámetros habituales objeto de estudio y sus estimadores puntuales máximo verosímiles.
Los estimadores son características cuantitativas calculadas a partir de los datos de la muestra observada que, por su construcción, intentan acercarse al verdadero valor del parámetro desconocido de la población.

Cuadro 3.1.- Estimadores puntuales máximo verosímiles

Parámetro		**Estimador puntual**	
Media	μ	Media muestral	$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Proporción	p	Proporción muestral	$\hat{p} = \frac{r}{n}$
Diferencia de medias	$\mu_1-\mu_2$	Diferencia de medias muestrales	$\hat{\mu}_1 - \hat{\mu}_2 = \bar{x}_1 - \bar{x}_2 = \frac{\sum x_{1i}}{n_1} - \frac{\sum x_{2i}}{n_2}$
Diferencia de proporciones	p_1-p_2	Diferencia de proporciones muestrales	$\hat{p}_1 - \hat{p}_2 = \frac{r_1}{n_1} - \frac{r_2}{n_2}$
Varianza	σ^2	Varianza muestral corregida	$\hat{\sigma}^2 = \frac{n}{n-1} S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

Un intervalo de confianza al nivel de confianza 1-α para un determinado parámetro θ podría definirse como un intervalo [a,b] que contendrá al verdadero valor de θ con una confianza 1-α. La cantidad 1-α se denomina nivel de confianza, y es frecuente que se exprese como porcentaje, es decir 100(1-α)%.

Aunque será discutido con posterioridad, cabe mencionar que la utilización del término confianza en lugar del de probabilidad se produce como consecuencia de que el parámetro toma un valor, que aunque es desconocido, es fijo y no tendría sentido plantear probabilidades sobre algo que no causa variabilidad. Sin embargo, en la práctica, la interpretación es que el intervalo contendrá al valor del parámetro con una probabilidad 1-α.
La cuestión que se plantea es de qué forma pueden ser construidos estos intervalos de confianza, bajo qué condiciones y cuál será el papel que desempeñarán los estimadores en todo este proceso.
Hay que recordar que se trata de realizar afirmaciones sobre determinada característica de la población cuando únicamente se dispone de una muestra de dicha población y, sin duda, la forma en que sean seleccionados los individuos que formarán parte de la misma influirá enormemente sobre los resultados que se puedan obtener. Lo deseable es que la muestra seleccionada sea *representativa* de la población sobre la que se quiere realizar algún tipo de inferencia, pues de lo contrario, las estimaciones se alejarán de los verdaderos valores de los parámetros poblacionales produciéndose un error denominado *error sistemático* o *sesgo.* Por tanto, será necesario contar con métodos o técnicas de muestreo en el proceso de selección de los elementos de la muestra que intenten reducir al máximo el error sistemático que pudiera cometerse.

El muestreo

Las diferentes técnicas de muestreo se clasifican en *probabilísticas* o *aleatorias* y *no probabilísticas*. La diferencia radica fundamentalmente en que, en el caso de las técnicas probabilísticas, cada uno de los elementos de la población tiene una probabilidad conocida y distinta de cero de ser incluido en la muestra, mientras que en las no probabilísticas esta cuestión se desconoce. La principal consecuencia es que las técnicas probabilísticas, al ser cuantificable el error muestral, posibilitarán la aplicación de las diferentes técnicas de inferencia estadística (determinación de tamaños muestrales, estimación por intervalos, contrastes de hipótesis,..). Por este motivo se describe brevemente y a continuación algunas de las técnicas probabilísticas básicas de muestreo.

Muestreo aleatorio simple

Este método de muestreo selecciona de forma aleatoria los *n* elementos de la muestra de manera que cada uno de los *N* individuos de la población tiene la misma probabilidad de ser incluido en la misma. Puede realizarse con o sin reemplazamiento, dependiendo de que un mismo elemento pueda ser seleccionado en más de una ocasión. En este proceso, suele ser habitual utilizar una tabla de números aleatorios como la que se presenta en el cuadro 3.2, donde cada individuo de la población tiene asignado un número. Por ejemplo, si el tamaño de la población es N=500 y se quiere obtener una muestra de tamaño n=10, se seleccionarían los 10 primeros números comprendidos entre 1 y 500 comenzando al azar por cualquier parte de la tabla. Si se comienza por la parte superior izquierda, de izquierda a derecha a grupos de tres dígitos, se obtendrían: 18, 95, 107, 67, 98, 426, 454, 266, 467 y 428, seleccionando los individuos de la población que, una vez enumerados, se correspondieran con estas cifras.

Cuadro 3.2.- Tabla de números aleatorios

018095	845107	670067	981098
426815	454815	734839	707786
856719	266467	428853	846511
914832	326989	100817	045268
577309	926758	664998	731061
147361	383121	966477	303795
117559	559584	405976	042077

Muestreo aleatorio sistemático

En el muestreo aleatorio sistemático no es necesario contar con un listado enumerado de todos los individuos de la población sino que será suficiente con que los individuos se encuentren ordenados según

algún tipo de clasificación que nada tenga que ver con la variable que se pretende estudiar, por ejemplo, por orden alfabético. Si se cuenta con una población de tamaño N y se quiere obtener una muestra de tamaño n se procederá de la forma siguiente:

1) Calcular el paso de selección k como el número entero más próximo a N/n (por defecto)
2) Seleccionar mediante muestreo aleatorio simple un número aleatorio n_0 entre 1 y k
3) Seleccionar los individuos que ocupen el lugar n_0 ,n_0+k, n_0+2k, n_0+3k, ...,n_0+(n-1)k

Este método resulta del todo inapropiado cuando los individuos presentan oscilaciones periódicas en su ordenación.

Muestreo aleatorio estratificado

Cuando la población puede ser clasificada en diferentes categorías o estratos que pueden influir directamente sobre el resultado de las estimaciones de los parámetros de interés suele ser habitual realizar un muestreo del tipo estratificado. La idea fundamental es construir muestras específicas para cada estrato, generalmente de tamaños proporcionales a la población en cada uno de los estratos o categorías. Por ejemplo, si se pretende obtener una muestra de tamaño n de una población de tamaño N en la que se consideran 3 categorías o estratos con tamaños N_1, N_2, N_3, se seleccionarán por muestreo aleatorio simple, en cada una de las 3 subpoblaciones una muestra de tamaño $(N_1/N)n$, $(N_2/N)n$ y $(N_3/N)n$ respectivamente. De esta forma se asegura que los diferentes grupos de población estén representados en la muestra que será analizada con posterioridad. Además, esta técnica de muestreo permite obtener estimaciones del parámetro objeto de estudio en cada una de las categorías o estratos de la población.

Distribuciones en el muestreo

Una vez recogida la muestra aleatoria por medio de cualquiera de las técnicas de muestreo aleatorio, puede construirse un estimador puntual tal y como ha sido descrito en el apartado anterior. Pero, dada la aleatoriedad con que se seleccionan los elementos de la muestra ¿qué ocurriría si en lugar de haberse obtenido exactamente esos elementos muestrales, se hubieran obtenido otros totalmente distintos? En realidad el número de muestras aleatorias distintas de tamaño n que podrían haberse obtenido de una población de tamaño N ascendería, en el caso de muestreo aleatorio simple sin reemplazamiento a

$$\binom{N}{n} = \text{Combinaciones de N individuos tomados de n en n}$$

y para cada una de estas muestras se dispondría de un estimador puntual del parámetro poblacional. En consecuencia, dado que el valor del estimador varía de muestra a muestra, se obtiene un resultado de gran relevancia en la estadística inferencial: un estimador es una variable aleatoria.

Distribución de la media en el muestreo

Dado que un estimador es una variable aleatoria, tiene sentido preguntarse si la distribución de probabilidad asociada a cada uno de los estimadores del tipo descrito en el cuadro 3.1 puede ser conocida. En caso afirmativo se estaría en condiciones de plantear cualquier probabilidad sobre los valores del estimador y se habrían sentado las bases para la construcción de intervalos de confianza.
Supóngase que a partir de los datos del segundo ejemplo de la introducción, se considera que la variable nivel de colesterol en la población objeto de estudio se distribuye según una distribución de probabilidad normal con media μ y una desviación típica σ. Se sabe que la suma de variables aleatorias normales sigue a su vez una distribución también normal con la misma media μ pero con una varianza distinta.

Propiedad . Sean $X_1,X_2,...,X_n$ n variables aleatorias independientes, entonces la varianza de la suma podrá expresarse como la suma de las varianzas:

$$\mathbf{Var(X_1 + X_2 + ... + X_n) = Var(X_1) + Var(X_2) + ... + Var(X_n)}$$

Propiedad . Sea X una variable aleatoria y *a* una constante, entonces la variable Y=*a*X tendrá una varianza proporcional a la varianza de la variable X

$$\mathbf{Var(Y) = Var(aX) = a^2 Var(X)}$$

La variable aleatoria media muestral se construye como la suma de n variables aleatorias independientes, donde cada una de ellas sigue una distribución normal con media μ y varianza:

$$\bar{X} = \frac{1}{n} X_1 + \frac{1}{n} X_2 + ... + \frac{1}{n} X_n$$

$$\mathbf{Var}\left(\frac{1}{n} X\right) = \frac{1}{n^2} \mathbf{Var(X)} = \frac{\sigma^2}{n}$$

En conclusión, la distribución de probabilidad asociada a la media muestral, cuando la variable se distribuye según una distribución normal de media μ y desviación típica σ, será una distribución normal con las siguientes características:

$$\mathbf{E}(\bar{X}) = \mu \quad ; \quad \mathbf{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

$$\begin{aligned} \mathbf{Var}(\bar{X}) &= \mathbf{Var}\left(\frac{1}{n} X_1 + \frac{1}{n} X_2 + ... + \frac{1}{n} X_n\right) \\ &= \mathbf{Var}\left(\frac{1}{n} X_1\right) + \mathbf{Var}\left(\frac{1}{n} X_2\right) + ... + \mathbf{Var}\left(\frac{1}{n} X_n\right) \\ &= \frac{1}{n^2} \mathbf{Var}(X_1) + \frac{1}{n^2} \mathbf{Var}(X_2) + ... + \frac{1}{n^2} \mathbf{Var}(X_n) \\ &= \frac{\sigma^2}{n^2} + \frac{\sigma^2}{n^2} + ... + \frac{\sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

En el ejemplo se cuenta con un tamaño muestral de 250. La media muestral seguirá por tanto una distribución normal con media μ y varianza:

$$\frac{\sigma^2}{n} = \frac{\sigma^2}{250}$$

Este resultado se basa, como fue señalado con anterioridad, en la suposición de que la variable (en el ejemplo el nivel de colesterol) se distribuye según un modelo de probabilidad normal. Sin embargo, en la mayoría de las ocasiones no se conoce la distribución de probabilidad de la variable objeto de estudio. La pregunta es: ¿existe alguna forma de conocer la distribución de la media muestral cuando se trabaja con variables que no se comportan según el modelo de probabilidad normal o para las que esta cuestión se desconoce?

Teorema central del límite

Sean X_1, X_2, ..., X_n n variables aleatorias independientes con media μ_i y varianza σ_i^2 que se distribuyen según un determinado modelo de probabilidad cualquiera y sea la variable aleatoria Y, una variable construida de la forma

$$Y=X_1+ X_2+ ... + X_n$$

entonces, cuando n tiende a infinito, la variable Y sigue asintóticamente una distribución normal con media y varianza

$$E(Y) = \sum_{i=1}^{n} X_i \quad ; \quad Var(Y) = \sum_{i=1}^{n} \sigma_i^2$$

Esto querría decir que, en el caso concreto de la media muestral y cuando el tamaño de la muestra fuera lo suficientemente grande (cuanto más grande mejor será la aproximación de la distribución normal), la distribución de probabilidad asociada sería aproximadamente una normal con media y varianza

$$E(\bar{X}) = \mu \quad ; \quad Var(\bar{X}) = \frac{\sigma^2}{n}$$

Habitualmente se considera que la aproximación normal es lo suficientemente buena a partir de $n \geq 30$.

Error estándar

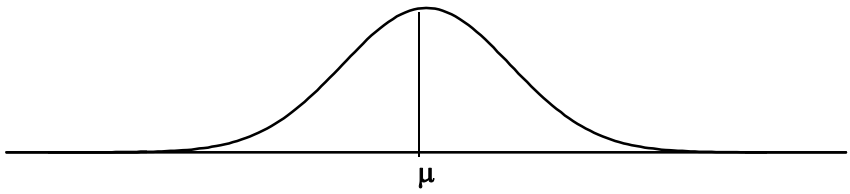
A la desviación típica de un estimador en el muestreo se la denomina *error estándar.* Por tanto, el error estándar será una medida de la variabilidad del estimador en el proceso de muestreo.

En el caso de la media de una población, el estimador es la media muestral y el error estándar sería la desviación típica de la media en el muestreo, es decir, la desviación típica de la variable aleatoria media muestral, que en este caso es

$$EE = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Intervalo de confianza para una media

El objetivo será construir un intervalo de confianza para la media desconocida de una población a un determinado nivel de confianza 1-α. Sin pérdida de generalidad se considerará el nivel de confianza 0,95.
Se sabe, como fue descrito con anterioridad, que si los datos que componen la muestra observada han sido obtenidos por muestreo aleatorio, el estimador máximo verosímil de la media (estimador puntual) es la media muestral y que éste es, a su vez, una variable aleatoria que se distribuye según el modelo de probabilidad normal si la variable objeto de estudio se distribuía según el modelo normal, o



aproximadamente normal si el

tamaño de la muestra era igual o superior a 30 e independientemente de la distribución de probabilidad de la variable. Puede apreciarse que en el modelo de probabilidad normal no todos los valores de la variable tienen la misma probabilidad de ser observados. De hecho, los valores con mayores probabilidades se concentrarían en la zona central de la distribución, mientras que los valores situados en cualquiera de los dos extremos de la distribución tendrían escasa probabilidad de ser observados.
Si se pretendiera construir un intervalo que contuviera al 95% de los valores con mayor probabilidad de ser observados correspondientes a una variable aleatoria que se distribuye según el modelo de probabilidad

normal, debería situarse en el centro de la distribución y bastaría con calcular los valores de z_1 y z_2 que verifiquen:

$$P(z_1 \leq X \leq z_2) = 0{,}95$$

En el caso de la media muestral:

$$P(z_1 \leq \overline{X} \leq z_2) = 0{,}95$$

Para calcular los valores de z_1 y z_2 a partir de la distribución normal estándar, tipificamos la variable restándole la media y dividiendo por la desviación típica

$$P(z_1 \leq \overline{X} \leq z_2) = P\left(\frac{z_1 - \mu}{\sigma/\sqrt{n}} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq \frac{z_2 - \mu}{\sigma/\sqrt{n}}\right) = P\left(\frac{z_1 - \mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{z_2 - \mu}{\sigma/\sqrt{n}}\right) = 0{,}95$$

y donde la variable Z se distribuirá según una normal estándar. En una distribución normal estándar la probabilidad de que la variable tome valores entre -1,96 y 1,96 (intervalo centrado) es 0,95 (ver tabla 2). Por tanto, los valores de z_1 y z_2 podrán obtenerse despejando de las expresiones:

$$\frac{z_1 - \mu}{\sigma/\sqrt{n}} = -1{,}96 \quad ; \quad \frac{z_2 - \mu}{\sigma/\sqrt{n}} = 1{,}96$$

de donde se obtiene:

$$z_1 = \mu - 1{,}96\,\frac{\sigma}{\sqrt{n}} \quad ; \quad z_2 = \mu + 1{,}96\,\frac{\sigma}{\sqrt{n}}$$

En definitiva:

$$P\left(\mu - 1{,}96\,\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right) = 0{,}95$$

Si se construye un intervalo para el parámetro desconocido de la población μ ,y a partir de los datos de la muestra observada, de la forma que se describe a continuación, el 95% de los intervalos que podrían construirse a partir de las diferentes muestras que podrían haberse obtenido por muestreo aleatorio, contendrían al verdadero valor del parámetro poblacional:

$$I_{0,95}(\mu) = \left[\overline{x} - 1{,}96\,\frac{\sigma}{\sqrt{n}}, \overline{x} + 1{,}96\,\frac{\sigma}{\sqrt{n}}\right]$$

Intervalo de confianza para un parámetro poblacional

Puede observarse, a partir del resultado obtenido en el apartado anterior, que los elementos necesarios para la construcción de un intervalo de confianza a un determinado nivel de confianza para la media desconocida de una población son: la media muestral, el error estándar de la media muestral y dos coeficientes. En general, la construcción de un intervalo de confianza al nivel 1-α para un parámetro desconocido cualquiera de la población, digamos θ podrá realizarse, en la mayoría de los casos (cuando la distribución muestral asociada es simétrica), de la forma:

$I_{1-\alpha}(\theta)$ = [estimador puntual-$c_{1-\alpha/2}$ x error estándar, estimador puntual + $c_{1-\alpha/2}$ x error estándar]

donde los elementos que necesarios son:

- el estimador puntual del parámetro poblacional θ
- el error estándar del estimador
- el coeficiente $c_{1-\alpha/2}$ calculado sobre la distribución de probabilidad asociada al estimador puntual

y donde el nivel de confianza 1-α significa que el 100(1-α)% de los intervalos que podrían construirse de esta forma para cada uno de los posibles resultados del estimador puntual calculados a partir de las diferentes muestras aleatorias que podrían obtenerse por muestreo aleatorio contendrán al verdadero valor del parámetro poblacional.

Otras distribuciones muestrales

La utilización del muestreo aleatorio en el proceso de selección de los datos de la muestra y el conocimiento de las distribuciones muestrales asociadas a los diferentes estimadores han sido cruciales en el proceso de estimación de parámetros desconocidos de la población. En el cuadro 3.3 aparecen reflejadas las distribuciones muestrales de diferentes estimadores y que serán utilizadas y analizadas con mayor detalle en el siguiente capítulo.

Cuadro 3.3.- Distribuciones muestrales

Parámetro	**Distribución muestral**
μ, σ conocida	Normal
μ, σ desconocida	t student con n-1 g.l
p	Aproximadamente normal
$\mu_1-\mu_2$ σ_1, σ_2 conocidas	Normal
$\mu_1-\mu_2$ $\sigma_{1,}$, σ_2 desconocidas $\sigma_1=\sigma_2$	t student con n_1+n_2-2 grados de libertad
$\mu_1-\mu_2$ $\sigma_{1,}$, σ_2 desconocidas $\sigma_1\neq\sigma_2$	t student con gl grados de libertad $$gl = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1}+\frac{(s_2^2/n_2)^2}{n_2}}$$
p_1-p_2	Aproximadamente normal
σ^2	Ji-cuadrado con n-1 g.l

3.2 CONTRASTES DE HIPÓTESIS

Al igual que ocurre con la estimación de parámetros desconocidos de la población por medio de intervalos de confianza, mediante los contrastes de hipótesis también se pretende realizar inferencias sobre la población objeto de estudio cuando únicamente se dispone de una muestra observada de dicha población. El funcionamiento de esta segunda técnica inferencial se basa en la realización de una afirmación acerca de las características de una o más variables en una o más poblaciones (*hipótesis*) y en el estudio de la compatibilidad entre esta afirmación y lo observado en la muestra. Cuanto mayor sea la discrepancia entre la hipótesis realizada y la información proporcionada por la muestra observada, mayor será la evidencia en contra de dicha hipótesis.

En el primer ejemplo del presente capítulo el investigador pretendía comprobar, a partir de la información sobre el tiempo de supervivencia correspondiente a 41 pacientes intervenidos quirúrgicamente de determinada afección, si el tiempo medio de supervivencia desde el momento de la intervención y en este tipo de pacientes era superior a 15 años (hipótesis). Si el tiempo medio de supervivencia observado en la muestra fuera de 10 años, la discrepancia entre lo supuesto en la hipótesis y lo observado en la muestra sería superior a 5 años. La cuestión es si la discrepancia observada podría ser explicada por el azar, al haber observado sólo una parte pequeña de la población, o por el contrario es consecuencia de la falsedad de la hipótesis realizada.

Hipótesis nula e hipótesis alternativa

La mecánica de los contrastes de hipótesis se basa en la definición de dos hipótesis enfrentadas, la *hipótesis nula* H_0 y la *hipótesis alternativa* H_a. La hipótesis nula es la hipótesis que se pretende contrastar y será mantenida a menos que los datos observados en la muestra indiquen una fuerte evidencia de que no es cierta, siendo ésta la razón de que, a pesar de que el contraste conduzca a aceptarla nunca se considere probada. Por el contrario si el contraste de hipótesis se decide por la hipótesis alternativa y, por tanto, rechaza la hipótesis nula, será porque la evidencia en contra de dicha hipótesis es manifiesta. En este sentido, algunos autores prefieren afirman que una hipótesis nula nunca puede ser aceptada, sino simplemente rechazada o no rechazada, si bien lo que quiere expresarse es exactamente lo mismo.

Hipótesis simples y compuestas

Las hipótesis estadísticas pueden clasificarse en dos grupos, dependiendo de si especifican un valor concreto para el parámetro o parámetros de la población (*hipótesis simples*) o si consideran varios valores, habitualmente un intervalo, como posibles (*hipótesis compuestas*). En el ejemplo anterior la hipótesis de que el tiempo medio de supervivencia sea superior a 15 años constituiría una hipótesis compuesta, puesto que cualquier valor del parámetro incluido en el intervalo]15, +∞] sería favorable a dicha hipótesis. Si el investigador hubiera estado interesado en comprobar si el tiempo medio de supervivencia es de exactamente 15 años, la hipótesis, en este caso, sería simple.

Etapas en la realización de un contraste de hipótesis

A continuación se describen los pasos necesarios para la realización de un contraste de hipótesis. Para ilustrar mejor el procedimiento así como los elementos que intervienen en el mismo se utilizarán los datos del ejemplo. Se cuenta con información sobre el tiempo medio de supervivencia de 41 pacientes en los que la media de supervivencia es de 17 años con una desviación típica de 1,5 años. A efectos de simplicidad se supondrá , en un primer momento, que lo que el investigador quiere demostrar es si el tiempo medio de supervivencia es o no distinto de 15 años.

1) Definición de las hipótesis del contraste. Hipótesis nula e hipótesis alternativa.

El parámetro sobre el que se pretende realizar un contraste de hipótesis es, en este caso, una media. Las hipótesis quedarán de la siguiente forma

$$H_0 : \mu = 15$$
$$H_a : \mu \neq 15$$

2) Definir una medida de discrepancia o *estadístico de contraste* entre lo que se afirma en la hipótesis nula y la información que proporcionan los datos de la muestra observada.

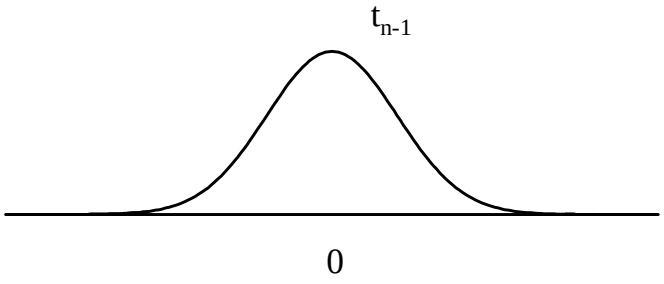
En el caso del contraste de una media, cuando se desconoce la desviación típica de la población objeto de estudio, el estadístico de contraste utilizado es

$$t = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}}$$

$$\hat{s} = \frac{n}{n-1} s = \frac{41}{41-1} 1{,}5 = 1{,}5375$$

donde μ_0 es el valor de la media que se especifica en la hipótesis nula, $\bar{x}$ es la media muestral, s el estimador puntual de la desviación típica poblacional y n el tamaño de la muestra. Puede observarse que, dado que la desviación típica es siempre una cantidad positiva, la medida de discrepancia o estadístico de contraste tomará el valor cero únicamente cuando la media observada en la muestra coincida exactamente con la que se propone en la hipótesis nula. Además, cuanto mayor sea la discrepancia entre la media observada y la media a la que se refiere la hipótesis nula, mayor será el numerador y, por tanto, el valor del estadístico de contraste. De este modo puede concluirse que valores del estadístico de contraste próximos a cero favorecerían a la hipótesis nula, mientras que valores muy distantes de cero indicarían la existencia de evidencia en contra de dicha hipótesis. En definitiva, será necesario contar con herramientas que permitan decidir cuándo la discrepancia es lo suficientemente grande como para rechazar la hipótesis nula. En este sentido será muy útil la consideración del siguiente paso.

3) Conocer la distribución de probabilidad asociada (distribución muestral) a la medida de discrepancia o estadístico de contraste. El conocimiento de la distribución de probabilidad que gobierna el comportamiento del estadístico de contraste será vital para el desarrollo final del contraste de hipótesis. En el ejemplo se cuenta con 41 datos. El teorema central del límite permitiría considerar que la distribución de probabilidad asociada a la media es aproximadamente una normal, pero como se desconoce el valor de la desviación



típica de la población la distribución de probabilidad asociada será, tal y como se mostraba en el cuadro 3.3, una t de student con n-1 grados de libertad.
La distribución es simétrica y está centrada en cero. Si la hipótesis nula fuera cierta el estadístico de contraste debería tomar valores en la zona central de la distribución, siendo muy improbable observar valores en cualquiera de sus dos extremos.

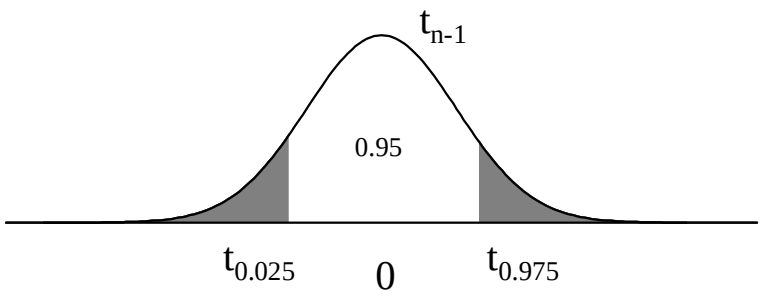
4) Establecimiento del nivel de significación del contraste.

Para decidir exactamente qué valores del estadístico de contraste tendrían una probabilidad de observarse prácticamente despreciable si la hipótesis nula fuera cierta, se define el que se denomina *nivel de significación* del contraste o nivel de significación *a priori* α. Habitualmente se consideran valores de α del

tipo 0,05, 0,01 ó 0,001. En este ejemplo se decide utilizar un nivel α=0,05.

5) Construcción de una regla de decisión.

Si el nivel de significación elegido es α=0,05 deberían despreciarse valores del estadístico de contraste con una probabilidad inferior a 0,05. Como fue discutido con anterioridad, los valores menos probables del estadístico de contraste si la hipótesis nula fuera cierta se concentran en ambos extremos de la distribución, por tanto, se define como región crítica de contraste o región de rechazo de la hipótesis nula a la región $]-\infty, t_{0,025}[\cup]t_{0,975}, +\infty[$ que aparece sombreada en la figura siguiente:



De forma complementaria, se define como región de aceptación de la hipótesis nula a la región comprendida entre $t_{0,025}$ y $t_{0,975}$ $[t_{0,025}, t_{0,975}]$. Los valores $t_{0,025}$ y $t_{0,975}$ sobre una distribución t de student con n-1=41-1=40 grados de libertad que determinan la región de aceptación y de rechazo de la hipótesis nula son aquellos que verifican:

$$\mathbf{P(t \leq t_{0,025}) = 0,025}$$
$$\mathbf{P(t \leq t_{0,975}) = 0,95 + 0,025 = 0,975}$$

Los valores son $t_{0,025}$=- 2,021 y $t_{0,975}$=2,021 (ver tabla 3 de una t de student)

La regla de decisión quedará en este caso de la siguiente forma:

- Si t >2,021 ó t<-2,021 entonces se rechazará la hipótesis nula
- Si -2,021≤ t ≤ 2,021 entonces se aceptará la hipótesis nula

6) Cálculo del estadístico de contraste, aplicación de la regla de decisión y conclusiones

Por último será necesario calcular el valor del estadístico de contraste y aplicar la regla de decisión. En el ejemplo se tendrá que:

$$\mathbf{t = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} = \frac{17 - 15}{1,5375/\sqrt{41}} = \frac{2}{1,5375/\sqrt{41}} = 8,329}$$

Como t=8,329 es mayor que 2,021 queda situado en la región crítica de contraste o región de rechazo de la hipótesis nula y la regla de decisión conducirá a rechazar la hipótesis nula, considerándose que existe evidencia suficiente de que el tiempo medio de supervivencia es *significativamente* distinto de 15 años. En otro caso, la conclusión hubiera sido que no existiría evidencia para rechazar la hipótesis nula, pudiendo ser cierta.

Contrastes bilaterales y unilaterales

Los contrastes de hipótesis pueden ser *unilaterales* o *de una cola* y *bilaterales* o *de dos colas* dependiendo de la forma en que se planteen las hipótesis. Así, cuando se contrasta la hipótesis de que determinado

parámetro de la población tome exactamente un valor dado frente a la hipótesis de que el parámetro tome un valor distinto al propuesto, el contraste será bilateral o de dos colas.

$$H_0 : \theta = \theta_0$$
$$H_a : \theta \neq \theta_0$$

Obsérvese que la región de rechazo definida por un contraste de este tipo quedaría situada a ambos extremos de la distribución, puesto que debería rechazarse la hipótesis nula tanto cuando $\theta>\theta_0$ como cuando $\theta<\theta_0$. Por otra parte, si la hipótesis se plantea de forma que se atiende únicamente al hecho de que ese mismo parámetro de la población tome un valor superior (análogamente inferior) a un valor dado, el contraste será unilateral o de una cola.

$$H_0 : \theta \leq \theta_0 \quad ; \quad H_0 : \theta \geq \theta_0$$
$$H_a : \theta > \theta_0 \quad ; \quad H_a : \theta < \theta_0$$

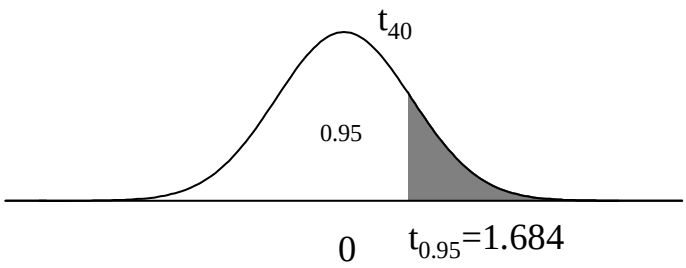
En este caso la región de rechazo definida por el contraste se situaría bien a la derecha de la distribución (se rechaza H_0 cuando $\theta>\theta_0$), o bien a la izquierda de la distribución (se rechaza H_0 cuando $\theta<\theta_0$).

En el ejemplo anterior se pretendía demostrar si el tiempo medio de supervivencia era superior a 15 años ($\mu>15$). Podría considerarse un contraste unilateral o de una cola de la siguiente forma

$$H_0 : \mu \leq 15$$
$$H_a : \mu > 15$$

Dado que la hipótesis nula nunca se considera probada, la única posibilidad a la hora de demostrar si el tiempo medio de supervivencia será significativamente superior a 15 años es que esta hipótesis se considere como hipótesis alternativa. El estadístico de contraste y la distribución de probabilidad asociada coinciden exactamente con los expuestos en el contraste bilateral realizado con anterioridad.

Obsérvese que el estadístico t tomará valores grandes y positivos cuando la media observada sea mayor que 15, mientras que en otro caso tomará valores próximos a cero o negativos. Si se utiliza el mismo nivel de significación del contraste $\alpha=0,05$, la región crítica de contraste y la regla de decisión quedarán:



Regla de decisión

Si t>1,684 entonces se rechaza la hipótesis nula
Si t≤ 1,684 entonces se acepta la hipótesis nula

Como t=8,329>1,684 se rechaza la hipótesis nula y se considera que el tiempo medio de supervivencia en este tipo de pacientes es significativamente superior a 15 años. Si se comparan las regiones críticas de contraste obtenidas para el contraste bilateral y el contraste unilateral se observa que la cola de la derecha es mayor en el contraste unilateral (a partir de 1,684 se rechazaría la hipótesis nula) que en el bilateral (a partir de 2,021 se rechazaría la hipótesis). Por tanto, existe mayor probabilidad de rechazar la hipótesis nula bajo el planteamiento de contrastes unilaterales. Probablemente el investigador del ejemplo anterior planteó el contraste unilateral una vez calculada la media de la muestra y después de haber observado que ésta era superior a 15, pretendiendo entonces demostrar si esa diferencia a favor de una media superior era significativa. Dado que previamente a la observación de los datos de la muestra se desconoce si la media será o no superior a la que se propone en la hipótesis nula, no tendría sentido plantearse si la media de la

población será o no significativamente superior a determinado valor sino, en todo caso, si sería significativamente distinta de una valor dado (mayor o menor). Por tanto, el planteamiento bilateral del contraste sería el más indicado.

Error tipo I y tipo II

En un contraste de hipótesis, donde dos son los posibles resultados: Rechazar la hipótesis nula o no rechazarla, también dos van a ser los posibles errores que puedan cometerse y que están relacionados con cada una de las decisiones anteriores. En el cuadro 3.4 ser refleja esta situación.

Cuadro 3.4.- Tipos de error en un contraste de hipótesis

		Realidad	
		H_0 falsa	H_0 cierta
Decisión del contraste	Rechazo de H_0	Ausencia de error	Error tipo I
	No rechazo de H_0	Error tipo II	Ausencia de error

El error de tipo I se define como el error que se comete al rechazar la hipótesis nula cuando ésta es cierta. El contraste conducirá al rechazo de la hipótesis nula únicamente cuando el estadístico de contraste se sitúe en la región crítica y esto sólo puede ocurrir con una probabilidad α. Luego, al establecer un nivel de significación para el contraste se está controlando la probabilidad de cometer un error de tipo I.

$$\alpha = p \text{ (cometer error tipo I)} = p(\text{Rechazar } H_0 | H_0 \text{ cierta})$$

Por otra parte el error de tipo II es el que se comete cuando no se rechaza la hipótesis nula a pesar de que es falsa y suele denotarse por β.

$$\beta = p \text{ (cometer error tipo II)} = p(\text{Aceptar } H_0 | H_0 \text{ falsa})$$

Potencia de un contraste

La potencia de un contraste de hipótesis se define como la probabilidad de rechazar H_0 cuando es falsa o, equivalentemente, la probabilidad de aceptar la hipótesis alternativa cuando ésta es cierta.

$$P(\text{Aceptar } H_a | H_a \text{ cierta}) = P(\text{Rechazar } H_0 | H_0 \text{ falsa})$$

$$= 1 - P(\text{No rechazar } H_0 | H_0 \text{ falsa})$$

$$= 1 - \beta$$

En algunos estudios, las características de la muestra seleccionada pueden impedir la detección de evidencia significativa en contra de la hipótesis nula que se plantea aunque ésta sea falsa. En este sentido, la potencia del contraste podría interpretarse como la probabilidad de encontrar en el estudio evidencia significativa en contra de la hipótesis nula en el caso de que efectivamente la hipótesis nula fuera falsa. Cuando la hipótesis alternativa es una hipótesis simple (especifica un único valor para el parámetro) el valor de β es único. Sin embargo, cuando la hipótesis alternativa es compuesta existirá una valor de β asociado a cada una de las posibilidades. Esto sugiere la definición de la que se denomina *función de potencia* de un contraste de la forma:

$$\Pi(\theta) = p \text{ (rechazar } H_0|\theta)$$

donde θ representa todas las posibilidades del parámetro sobre el que se pretende contrastar alguna hipótesis.
Cuando sustituimos θ por el valor que se especifica en la hipótesis nula θ_0, la función de potencia toma el valor α. Por otra parte, cuando θ toma cualquier otro valor, la función de potencia quedará:

$$\Pi(\theta) = p \text{ (rechazar } H_0|\theta) = 1 - \beta(\theta)$$

El contraste de hipótesis y el nivel de significación α

La posibilidad de rechazar una hipótesis nula depende en gran parte de la magnitud de la región crítica de contraste. El tamaño de esta región crítica está determinado por el valor del nivel de significación del contraste α. Por tanto, el resultado del contraste de hipótesis será dependiente del nivel de significación elegido, pudiéndose dar el caso en que se rechace la hipótesis nula trabajando con un nivel de significación α=0,05 y no rechazarse al nivel α=0,045. Por otra parte, la aplicación de la regla de decisión del contraste tan sólo permite concluir si se consigue o no se consigue rechazar la hipótesis nula pero no informa sobre la magnitud de la evidencia en contra de dicha hipótesis. Ambos problemas pueden solucionarse definiendo el que se denomina *nivel crítico p*, *p-valor* o nivel de significación *a posteriori*.

El valor *p* o nivel de significación a posteriori

El *valor p* se define como la probabilidad de observar, bajo la suposición de que la hipótesis nula es cierta, un valor del estadístico de contraste o medida de discrepancia igual o más extremo que el observado en la muestra. Por tanto, el valor de *p* no se fija a priori sino que es calculado a partir de los datos de la muestra (a posteriori).

Si se tiene en consideración que en el ejemplo 1 los valores más probables del estadístico de contraste, si la hipótesis nula fuera cierta, se encontrarían situados en la zona central de la distribución y que los más improbables se encontrarían en los extremos, el valor de *p* estaría informando sobre la situación del valor del estadístico de contraste con respecto a la distribución. Así, valores muy pequeños de *p* estarían indicando que el estadístico de contraste se encuentra situado en cualquiera de los dos extremos de la distribución y la evidencia en contra de la hipótesis nula sería patente. Además, cuanto más pequeño sea el valor de *p* mayor será la evidencia en contra de la hipótesis nula.

En el ejemplo el valor del estadístico de contraste era t=8,329. El valor de p será la probabilidad de observar un valor del estadístico o medida de discrepancia igual o más extremo al observado. En el caso del contraste unilateral se tendrá:

$$p = p\ (t \geq 8{,}329) < 0{,}025$$

En el caso del contraste bilateral debe considerarse tanto la probabilidad de que el estadístico de contraste sea mayor o igual que 8,329 como la de que sea menor o igual que -8,329, ya que las discrepancia entre lo que se afirma en la hipótesis nula y lo observado en la muestra puede ser positiva o negativa, dependiendo del extremo de la distribución en que se sitúe el estadístico de contraste. Entonces el valor de *p* deberá calcularse:

$$p = p\ (t \leq -8{,}329) + p\ (t \geq 8{,}329) = 2\ .\ p\ (t \geq 8{,}329) < 2\ .\ 0{,}025 = 0{,}05$$

C A P I T U L O 4

¿QUE PRUEBA ESTADISTICA UTILIZAMOS? SELECCION DE PRUEBAS ESTADISTICAS DE INFERENCIA: APLICACIONES BASICAS

En el capítulo 3 se ha descrito los fundamentos y elementos necesarios para la aplicación de los dos procedimientos inferenciales básicos: la estimación, especialmente la construcción de intervalos de confianza, como procedimiento para la cuantificación de parámetros desconocidos, y el contraste de hipótesis, como procedimiento para la toma de decisiones acerca de enunciados o relaciones entre parámetros o características de las variables a estudio. El objetivo de éste capítulo es resumir los procedimientos estadísticos inferenciales básicos, especialmente los relativos a pruebas o contrastes de hipótesis, subrayando las situaciones que dan lugar a su utilización y estructurando los elementos necesarios para su aplicación. Aspectos tales como la naturaleza de las variables estudiadas, el objetivo del estudio, o las limitaciones derivadas de los propios requerimientos estadísticos, pueden ayudar a seleccionar la prueba oportuna.

4.1 VARIABLE RESPUESTA Y VARIABLES EXPLICATIVAS

La aplicación de los procedimientos inferenciales se realizará sobre la información que los diferentes individuos de la muestra nos aportan a través de las observaciones de las variables a estudio. En general, el interés se centrará en el estudio de características de alguna variable, de forma general, o sea, en toda la muestra, en subgrupos de ésta, es decir, en diferentes submuestras, o en relación a otra variable. La variable a estudio suele ser identificada como variable *dependiente*, mientras que, cuando se analiza ésta en diferentes submuestras, o en relación a otra variable, la variable que define los grupos o submuestras o con la que se quiere relacionar recibe el nombre de *independiente*. Los términos dependiente e independiente son utilizados para describir el papel de las variables en el análisis y no hacen referencia a la condición real de dependencia o relación entre las variables. Por ello, en un lenguaje más actual, es preferible identificar las variables a estudio de acuerdo con su papel real. Así, la variable a estudio es denominada *variable respuesta*, mientras que la variable independiente pasará a llamarse *variable explicativa*. Para comprender el papel de las variables, las situaciones descritas, y diferentes situaciones a resolver a lo largo de este capítulo se presenta el siguiente ejemplo:

En un estudio realizado en un centro de salud sobre la eficacia de un programa educativo para promocionar la lactancia materna se seleccionaron de forma aleatoria 62 mujeres, embarazadas y primíparas. Se registró la edad, la actitud hacia la lactancia materna antes y después del parto, clasificada como positiva y negativa, y el tiempo que estuvieron amamantando a sus hijos. De las 62 mujeres, 31 de ellas fueron asignadas aleatoriamente a un programa educativo de promoción de la lactancia materna.

Los datos registrados se presentan en la base que recoge el cuadro 4.1. Algunas preguntas de interés sobre estos datos podrían ser:

0	E	Ac1	Pr	Ac2	T	0	E	Ac1	Pr	Ac2	T
1	28	1	1	1	6.5	32	31	1	2	1	2.0
2	30	2	1	2	3.5	33	32	1	2	1	3.5
3	22	1	1	1	8.5	34	17	1	2	1	7.5
4	25	1	1	1	2.5	35	28	2	2	1	4.0
5	32	2	1	2	0.0	36	26	1	2	1	3.5
6	21	1	1	1	6.0	37	26	1	2	1	3.5
7	18	1	1	2	0.0	38	23	1	2	1	6.0
8	25	1	1	2	0.0	39	28	2	2	2	7.0
9	36	2	1	2	1.5	40	22	1	2	1	6.0
10	30	1	1	2	0.0	41	22	1	2	1	7.0
11	23	1	1	1	6.0	42	33	2	2	2	0.0
12	24	1	1	1	3.0	43	34	2	2	1	0.5
13	28	2	1	2	1.5	44	25	1	2	2	0.0
14	39	2	1	2	0.0	45	24	1	2	1	8.0
15	25	1	1	1	4.5	46	26	2	2	1	5.0
16	25	1	1	2	0.0	47	27	2	2	1	4.0
17	22	1	1	1	6.5	48	29	2	2	1	3.0
18	27	2	1	2	2.0	49	31	2	2	1	0.0
19	29	2	1	2	0.0	50	34	1	2	1	2.5
20	33	2	1	2	1.0	51	25	1	2	1	3.0
21	34	1	1	1	4.5	52	38	2	2	1	2.0
22	30	1	1	2	0.5	53	32	2	2	1	2.0
23	21	1	1	1	4.5	54	21	1	2	1	5.0
24	20	1	1	1	6.0	55	20	1	2	1	7.0
25	19	1	1	1	8.5	56	27	2	2	1	3.0
26	23	1	1	1	9.0	57	26	1	2	1	3.5
27	23	2	1	1	9.0	58	25	1	2	1	1.0
28	27	2	1	2	0.5	59	24	1	2	1	7.0
29	29	2	1	1	3.0	60	22	1	2	1	6.5
30	22	1	1	2	0.0	61	32	2	2	1	2.5
31	40	2	1	2	0.0	62	31	2	2	1	2.5

Cuadro 4.1.- Datos(simulados) de un estudio sobre lactancia materna. Variables: O=N° de orden, E=Edad, Ac1=Actitud hacia la lactancia 1 positiva 2 negativa, antes del parto, Pr=Intervención en un programa educativo sobre lactancia materna 1 no, 2 si, Ac2=Idem Ac1 pero tras el parto, T=Tiempo de lactancia

- ¿Cuál es la edad (variable respuesta) media en la población de la que proceden las mujeres estudiadas?

- ¿Podemos decir que hay diferencias significativas en la edad (variable respuesta) media entre las mujeres que presentan una actitud ante la lactancia antes del parto (variable explicativa) positiva y las que la presentan negativa?

- Las diferencias observadas en la proporción de mujeres que presentan una actitud negativa hacia la lactancia antes del parto (variable respuesta) entre las sometidas al programa y las que no lo siguen (variable explicativa), ¿podemos decir que son debido al azar?

- El tiempo de lactancia (variable respuesta), ¿se ve influido por la edad (variable explicativa) de la madre?

- ¿Ha habido variación significativa en la actitud hacia la lactancia (variable respuesta) por el hecho de haber intervenido a través del programa educativo?

4.2 SELECCION DE PRUEBAS ESTADISTICAS

La selección de la prueba estadística a utilizar en una situación concreta de análisis dependerá, en primera instancia, del objetivo perseguido en el análisis (estimación de parámetros, comparación del comportamiento de una variable, etc.). No obstante, la naturaleza de los datos y el tipo de variables estudiadas sugerirá las pruebas estadísticas más oportunas.
Los cuadros 4.2 y 4.3 presentan los procedimientos básicos más utilizados en situaciones inferenciales, dependiendo de la naturaleza de las variables estudiadas. Así, el cuadro 4.2 hace referencia a aquellas situaciones en las que se está interesado en inferir acerca del comportamiento de una variable respuesta según una variable explicativa categórica que define el número de grupos. Incluye las situaciones comparativas habituales, y, en caso de probar diferencias en el comportamiento de la variable respuesta según el número de grupos, se podrá hablar de relación o asociación entre ésta y la explicativa.
El cuadro 4.3 recoge las situaciones de estudio de la relación entre dos variables para los casos en que ambas sean cualitativas o categóricas, cuantitativas ordinales o transformadas en rangos, o cuantitativas continuas. En estos casos, los procedimientos habituales permiten probar la existencia de asociación o relación a través de alguna prueba de hipótesis global (como la prueba basada en el estadístico Ji-cuadrado para variables cualitativas o categóricas) o a través de coeficientes o modelos que detecten la magnitud y/o forma de la relación entre las variables (caso del coeficiente de correlación de Spearman o el de Pearson y el modelo de regresión lineal simple). A continuación pasamos a describir los diferentes procesos de inferencia sugeridos en los cuadros mencionados.

4.3 ESTUDIO DE UNA VARIABLE RESPUESTA EN UNA POBLACION

En este apartado se describe las características de las pruebas inferenciales para el estudio de una variable en un único grupo. La selección de los procedimientos a utilizar depende de si la variable es cualitativa o cuantitativa. Cuando es cualitativa, la proporción de individuos que verifican una de sus categorías es el parámetro de interés más frecuente. En el caso de que sea cuantitativa, la media de la variable es el parámetro de interés más frecuente. Tanto en uno como en otro caso, el objetivo suele ser estimar o contrastar hipótesis sobre sus valores. A continuación se desarrolla la solución a la construcción de intervalos de confianza y contraste de hipótesis para las situaciones descritas, bajo un esquema que recorre los puntos clave para obtener la solución adecuada (ver figura 4.1).

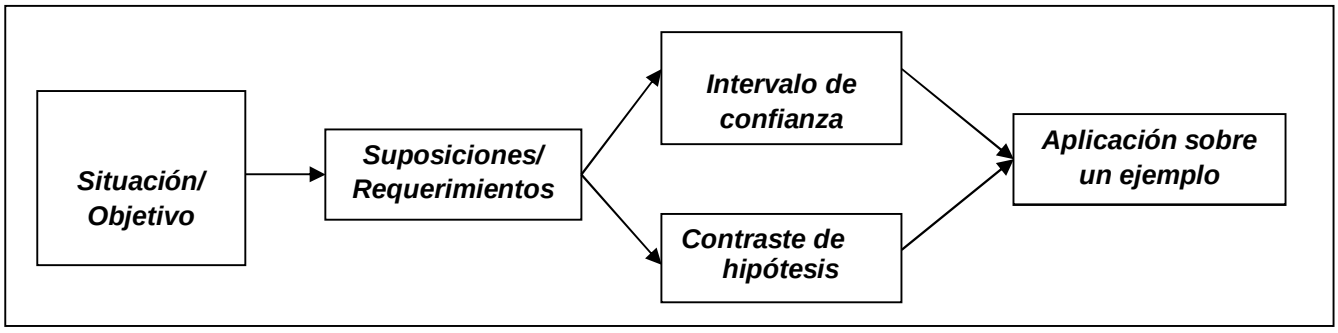


Figura 4.1.- Esquema de presentación de los procedimientos de inferencia

Cuadro 4.2.- Clasificación de métodos estadísticos de inferencia según la naturaleza de la variable respuesta, para una variable explicativa categórica, y según su número de grupos

Variable respuesta	Nº de grupos (Variable explicativa)	Diseño de los grupos	Objetivo	Pruebas estadísticas
Cualitativa	1 grupo	----------	Estimar la proporción de una categoría de la variable. Probar si la proporción puede ser igual a algún(os) valor(es) predeterminado(s)	- Intervalo de confianza y prueba z sobre proporciones
	2 grupos	Independientes	Probar si hay diferencia en las proporciones de alguna de las categorías entre los grupos	- Intervalo de confianza para la diferencia de proporciones y prueba z de comparación de proporciones
		Apareados	Probar si la proporción de una categoría de la variable es la misma en los dos grupos apareados	- Prueba de McNemar
	> 2 grupos	Independientes	Probar si hay diferencia en las proporciones de las categorías de la variable entre los grupos	- Prueba Ji-cuadrado
		Apareados	Probar si las proporciones de las categorías de la variable son las mismas en los grupos apareados	- Prueba de Cochran (variable respuesta dicotómica) - Coeficiente Kappa*
Cuantitativa	1 grupo	----------	Estimar la media de la variable. Probar si la media puede ser igual a algún(os) valor(es) predeterminado(s)	- Intervalo de confianza y prueba t sobre medias
	2 grupos	Independientes	Probar si hay diferencia en el comportamiento de la variable entre los grupos	- Intervalo de confianza para la diferencia de medias y prueba t de comparación de medias - Prueba U de Mann-Whitney
		Apareados	Probar si hay diferencia en el comportamiento de la variable entre los grupos apareados	- Intervalo de confianza para la media de las diferencias y prueba t para datos apareados - Prueba de Wilcoxon por rangos
	> 2 grupos	Independientes	Probar si hay diferencia en el comportamiento de la variable entre los grupos	- Análisis de la varianza de una vía* - Prueba de Kruskal-Wallis
		Apareados	Probar si hay diferencia en el comportamiento de la variable entre los grupos apareados	- Análisis de la varianza de medidas repetidas* - Prueba de Friedman

Cuadro 4.3.- Procedimientos estadísticos para el estudio de la relación entre dos variables, según su tipo

Tipo de variables	Objetivo	Prueba estadística
Cualitativas	Establecer si existe asociación entre las categorías de la variable	- Prueba Ji-cuadrado
Cuantitativas transformadas en rangos u ordinales	Establecer si existe relación entre las variables, una vez transformadas en rangos o posiciones ordinales	- Coeficiente de correlación de Spearman
Cuantitativas	Establecer si existe relación lineal entre las variables. Construir un modelo que dé forma lineal a la relación	- Coeficiente de correlación lineal de Pearson - Modelo de regresión lineal simple

Intervalo de confianza y contraste de hipótesis sobre una proporción

Cuando la variable es cualitativa, el interés suele residir en conocer aspectos relativos a la proporción de individuos que verifican alguna de sus categorías.

Situación

Se dispone de una muestra aleatoria de tamaño n, sobre la que se observa una variable cualitativa dicotómica, con proporción poblacional p en una de sus categorías, siendo $\hat{p}$ la proporción muestral de esa categoría

Suposiciones/Requerimientos

Se requiere que np y n(1-p) sean cantidades mayores que 5

Intervalo de confianza

El intervalo de confianza (IC) de nivel 1-α para p se obtiene como:

$$IC_{1-\alpha}(p) = \left[\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

con $z_{1-\alpha/2}$ coeficiente de un modelo normal para el nivel de confianza exigido

Contraste de hipótesis

Para contrastar la hipótesis nula de que la proporción p es igual a un valor especificado p_0, el estadístico de contraste es:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

cuya distribución muestral es normal

Un ejemplo

Sobre los datos recogidos en el cuadro 4.1, interesa conocer cuál es el intervalo de confianza para la proporción poblacional de mujeres primíparas con actitud positiva hacia la lactancia materna, antes del parto, y antes de que ninguna de ellas sea sometida al programa educativo que se describe. Adicionalmente, interesa conocer si puede aceptarse o debe rechazarse que la proporción descrita sea de 0,75.

Se dispone de una muestra aleatoria de n=62 mujeres, de las que 37 (59,7%) manifiestan tener una actitud positiva hacia la lactancia materna. El requerimiento para aceptar la normalidad de las distribuciones muestrales utilizadas puede ser evaluado de forma aproximada a través de:

$$n\hat{p} = 37 \quad y \quad n(1-\hat{p}) = 25$$

cantidades superiores a 5.

El intervalo de confianza, de nivel 95%, para la proporción poblacional será:

$$IC_{0,95}(p) = \left[0,597 \pm 1,96 \sqrt{\frac{0,597\,(1-0,597)}{62}} \right]$$
$$= [\,0,597 \pm 0,122\,] = [\,0,475\,,\,0,719\,]$$

por lo que la proporción poblacional se encontrará entre 0,475 (47,5%) y 0,719 (71,9%) con seguridad 95%.

Para decidir sobre la segunda pregunta, contrastaremos la hipótesis nula de que la proporción poblacional pueda ser p_0=0,75, resolviendo el contraste a nivel de significación α=0,05:

$$z = \frac{0,597 - 0,75}{\sqrt{\frac{0,75\,(1-0,75)}{62}}} = -2,78$$

valor que queda fuera de los valores críticos -1,96 y 1,96 que delimitan la región de aceptación de la hipótesis nula, por lo que debemos rechazar ésta, concluyendo que la proporción estudiada debe ser distinta a 0.75, con p (probabilidad *a posteriori* de error tipo I) $\leq$ 0.05.

Intervalo de confianza y contraste de hipótesis sobre una media

Cuando la variable es cuantitativa, la media es el parámetro de localización de interés más frecuente.

Situación

Se dispone de una muestra aleatoria de tamaño n, y observaciones independientes de una variable cuantitativa con media μ y desviación típica σ. La solución tanto para la construcción del intervalo como del contraste de hipótesis depende de si σ es conocida o desconocida. No obstante, puesto que σ suele ser desconocida en la práctica totalidad de las ocasiones, sólo se abordará este caso. Se tendrá,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{y} \quad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

como estimadores muestrales de la media y la desviación típica de la variable, respectivamente (en adelante, las desviaciones típicas serán siempre las correspondientes a las estimaciones muestrales descritas en el capítulo precedente)

Suposiciones/Requerimientos

Se requiere que la variable respuesta siga un modelo de probabilidad normal, o, en su defecto, que n sea mayor que 30.

Intervalo de confianza

El intervalo de confianza (IC) de nivel 1-α para μ se obtiene como:

$$IC_{1-\alpha}(\mu) = \left[\bar{x} \pm t^{n-1}_{1-\alpha/2} \frac{s}{\sqrt{n}} \right]$$

con $t^{n-1}_{1-\alpha/2}$ coeficiente de un modelo *t de student* con n-1 grados de libertad, para el nivel de confianza exigido.

Contraste de hipótesis

Para contrastar la hipótesis nula de que la media μ es igual a un valor especificado μ_0, el estadístico de contraste es:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

cuya distribución muestral es t de student con n-1 grados de libertad.

Un ejemplo

Sobre los datos recogidos en el cuadro 4.1, interesa conocer cúal es el intervalo de confianza, de nivel 95% para la media poblacional de la edad de las mujeres estudiadas. Adicionalmente, interesa conocer si puede aceptarse o debe rechazarse que la edad media es de 26 años.

Se dispone de una muestra aleatoria de n=62 mujeres, para las que la edad alcanza en la muestra una media $\bar{x} = 26{,}95$ años, y una desviación típica $s = 5{,}20$. El requerimiento para construir intervalos de confianza o contrastar hipótesis se cumple, pues n es mayor que 30.
El intervalo de confianza de nivel 95%, para la media poblacional de edad será:

$$IC_{0,95}(\mu) = [26{,}95 \pm 2{,}0\frac{5{,}2}{\sqrt{62}}] = [26{,}95 \pm 1{,}32] = [25{,}63\ ;\ 28{,}27]$$

por lo que la edad media se encontrará entre 25,63 años y 28,27 años, con confianza 95%.

Para decidir sobre si la edad media puede ser de 26 años o no, contrastaremos la hipótesis nula de que la media de edad pueda ser μ_0=26 años, resolviendo el contraste a nivel de significación α=0,05, el estadístico de contraste toma el valor:

$$t = \frac{26{,}95 - 26}{5{,}2/\sqrt{62}} = 1{,}44$$

valor que queda incluido entre los valores críticos -2 y 2, que delimitan la región de aceptación en un modelo t de student con n-1=62 grados de libertad, para el nivel de significación exigido. Se concluye que el valor 26 para la media de la edad no puede ser rechazado, no hay suficiente evidencia en los datos para hacerlo.

4.4 ESTUDIO DE UNA VARIABLE RESPUESTA EN DOS POBLACIONES

El estudio de una variable en dos poblaciones suele perseguir como objetivo habitual la comparación de sus características entre los grupos estudiados. Cuando la variable es cualitativa, el interés se centra en comparar las proporciones de alguna de sus categorías, mientras que si es cualitativa, la media es el

parámetro que habitualmente se compara entre los grupos estudiados. Sin embargo, el diseño utilizado para generar las muestras de las respectivas poblaciones influye de forma decisiva sobre las pruebas a utilizar. Los posibles diseños dependen de si entre los sujetos de las muestras estudiadas existen vínculos o nexos que relacionen, por parejas, a cada individuo de una de las muestra con otro de la otra muestra, hablando entonces de *muestras apareadas o emparejadas,* o si, por el contrario, los sujetos de cada una de las muestras son seleccionados de las respectivas poblaciones de forma independiente, hablando entonces de *muestras independientes*. Como ejemplos de estos posibles diseños, se puede considerar los datos incluidos en el cuadro 4.1. Si quisiéramos comparar la edad según la actitud inicial hacia la lactancia materna, las mujeres que inicialmente presentan actitud positiva y las que presentan actitud negativa pueden considerarse muestras independientes, cada una de ellas representativa de la correspondiente población, puesto que no existe ningún vínculo entre las que verifican un resultado y las que verifican otro. Sin embargo, si queremos comparar la proporción de actitud positiva antes y después del programa educativo, las muestras a utilizar serán las de las 31 mujeres asignadas al programa educativo, antes y después de recibirlo, es decir, las mismas mujeres, existiendo un vínculo obvio entre los grupos estudiados. Esta última situación es un ejemplo de apareamiento. El apareamiento persigue como objetivo hacer que los grupos estudiados sean lo más comparables entre sí, eliminando la posibilidad de que variables extrañas influyan sobre la nitidez de los resultados de las comparaciones.
Al igual que en el apartado anterior, la selección de procedimientos de inferencia dependerá de si la variable respuesta es cualitativa o cuantitativa. La exposición de los procedimientos básicos se guiará por el esquema descrito en la figura 4.1.

Comparación de proporciones de poblaciones independientes

Se supondrá grupos generados de forma independiente. El interés se centra en comparar las proporciones de alguna de las categorías de una variable cualitativa entre las poblaciones estudiadas.

Situación

Se dispone de dos muestras aleatorias de tamaños n_1 y n_2, de dos poblaciones independientes, y una variable cualitativa dicotómica, con proporciones $\mathbf{p_1}$ y $\mathbf{p_2}$ para una de las categorías de la variable, siendo $\hat{\mathbf{p}}_1$ y $\hat{\mathbf{p}}_2$ las correspondientes proporciones muestrales.

Suposiciones/Requerimientos

Si el tamaño total de la muestra, $n=n_1 + n_2$, está comprendido entre 20 y 40, las cantidades np_1, $n(1-p_1)$, np_2 y $n(1-p_2)$ deben ser todas ellas superiores a 5. Para n inferior o igual a 20, el procedimiento no es adecuado.

Intervalo de confianza para la diferencia de proporciones

El intervalo de confianza de nivel 1-α para $p_1 - p_2$ se obtiene como:

$$IC_{1-\alpha}(p_1 - p_2) = \left[(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

con $z_{1-\alpha/2}$ coeficiente de un modelo normal para el nivel de confianza exigido.

Contraste de hipótesis para comparar las proporciones

Para contrastar la hipótesis nula de que las proporciones p_1 y p_2 son iguales ($p_1 - p_2 = 0$), el estadístico de contraste es:

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

con $\hat{p}$ la proporción muestral de la categoría estudiada combinando ambos grupos. El estadístico z sigue una distribución muestral normal.

Un ejemplo

Sobre los datos recogidos en el cuadro 4.1, queremos saber si hay diferencias en la proporción inicial de actitud positiva hacia la lactancia materna entre las mujeres que han sido asignadas al programa educativo y las que no lo han sido. Se supone que no debe haberlas, puesto que la asignación ha sido al azar. Para discutir este objetivo construiremos el intervalo de confianza de nivel 95% para la diferencia de proporciones, y contrastaremos la hipótesis nula de igualdad de proporciones.

Se dispone de dos grupos independientes (la asignación se ha realizado al azar), con muestras de tamaños n_1=31 (grupo que no es intervenido) y n_2=31 (grupo que es intervenido). El tamaño total es n=62. Los casos observados de actitud positiva son 19 y 18, respectivamente, y, por consiguiente, las proporciones muestrales observadas son $\hat{p}_1 = 0{,}613$ y $\hat{p}_2 = 0{,}581$ y la proporción conjunta $\hat{p} = 0{,}596$.

El intervalo de confianza para la diferencia de proporciones será:

$$IC_{0,95}(p_1 - p_2) = \left[(0{,}613 - 0{,}581) \pm 1{,}96 \sqrt{\frac{0{,}613\,(1-0{,}613)}{31} + \frac{0{,}581\,(1-0{,}581)}{31}} \right]$$

$$= [\,0{,}032 \pm 0{,}244\,] = [\,-0{,}212\ ,\ 0{,}276\,]$$

en el que se observa que la diferencia puede ser a favor de una de las proporciones o de la otra, incluyendo el valor 0.

Para decidir sobre la igualdad de proporciones se contrastará la hipótesis nula de igualdad de proporciones a través del estadístico:

$$z = \frac{0{,}613 - 0{,}581}{\sqrt{0{,}596\,(1-0{,}596)\left(\frac{1}{31} + \frac{1}{31}\right)}} = 0{,}257$$

valor que queda incluido en la región de aceptación de la hipótesis nula para α=0,05, delimitada por los valores -1,96 y 1,96 de la curva normal. Concluimos que no podemos rechazar la igualdad de las proporciones estudiadas.

Comparación de proporciones de poblaciones apareadas

Si la selección de los sujetos de los grupos a estudio se realiza de forma apareada, es necesario utilizar procedimientos inferenciales que contemplen esta situación. En otro caso, los procedimientos para grupos independientes no serán satisfactorios. Presentamos en este punto la prueba más utilizada para comparar proporciones entre grupos apareados.

Situación

Se dispone de dos muestras apareadas, con tamaño n para cada una de ellas (n pares de observaciones), y una variable cualitativa dicotómica con proporciones poblacionales p_1 y p_2 para una de sus categorías. Los datos muestrales deben ser estructurados en la forma que refleja el cuadro 3.4. De los n pares de observaciones, en a de ellos tanto el individuo del grupo 1 como su pareja del grupo 2 responden a la categoría + (representación de una de las categorías de la variable a estudio), en b ocasiones el individuo del grupo 1 es - y el del grupo 2 es +. En c ocasiones el individuo del grupo 1 es + y el del 2 es -, mientras que en d ocasiones ambos son -. La información relevante es la contenida en b y en c, puesto que la de a y b es irrelevante, dado que tanto el individuo de un grupo

	Grupo 1 **+**	**-**
Grupo 2 **+**	a	b
-	c	d

Cuadro 4.4.- Estructura para datos apareados, variables cualitativas

como su pareja del otro se comportan igual.

Suposiciones/requerimientos

Es necesario que b+c sea mayor de 10.

Contraste de hipótesis para comparar las proporciones apareadas. Prueba de McNemar

Para contrastar la hipótesis nula de igualdad de proporciones de respuesta a una de las categorías el estadístico de contraste de McNemar es:

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

que sigue una distribución ji-cuadrado con 1 grado de libertad.

Un ejemplo

Sobre los datos recogidos en el cuadro 4.1, queremos comparar las proporciones de actitud positiva hacia la lactancia materna, antes y después del programa educativo, para evaluar su efecto.

Se dispone de dos grupos apareados (en realidad son el mismo), 31 mujeres antes y después del programa. Los datos observados, estructurados de acuerdo con la situación apareada son:

	Grupo 1	
	+	**-**
Grupo 2 **+**	17	11
-	1	2

donde el grupo 1 son las 31 mujeres antes del programa de intervención, y el grupo 2 después del programa, existiendo 17 casos en los que la mujer tiene actitud positiva antes y después, 11 casos en los que tiene actitud negativa antes y positiva después, 1 caso al contrario y 2 casos con actitud negativa antes y después.

Puesto que b + c = 12, los requerimientos se cumplen, y el estadístico del contraste de McNemar será:

$$\chi^2 = \frac{(11 - 1)^2}{11 + 1} = 8{,}33$$

cuyo valor, 8,33, debe ser comparado con el de un modelo ji-cuadrado con 1 grado de libertad, obteniendo que, para un contraste bilateral, se encuentra fuera de la región de aceptación de la hipótesis nula, delimitada por los valores 0,00098 y 5,024.

Comparación de medias de poblaciones independientes

Se supondrá grupos generados de forma independiente. El interés reside en comparar las medias de una variable cuantitativa entre las poblaciones estudiadas.

Situación

Se dispone de dos muestras aleatorias independientes, de tamaños n_1 y n_2, y una variable cuantitativa con medias μ_1 y μ_2 y desviaciones típicas σ_1 y σ_2 en las respectivas poblaciones.

Suposiciones/requerimientos

- La variable respuesta es normal en las dos poblaciones estudiadas o los tamaños muestrales de ambos grupos, n_1 y n_2, son mayores de 30.

La solución a los intervalos de confianza y contraste de hipótesis depende de si las desviaciones típicas poblacionales, σ_1 y σ_2 son conocidas o desconocidas, y en este último caso si pueden suponerse iguales o se rechaza la igualdad, siendo diferentes. En este texto consideraremos únicamente la posible disyuntiva sobre si las desviaciones típicas son iguales o diferentes, siendo desconocidas, puesto que la situación conocidas se presenta con escasa frecuencia. Para discutir la igualdad o diferencia de σ_1 y σ_2 recurriremos a un contraste de hipótesis de comparación de varianzas. La comparación de medias se realizará en función de su resultado. Una consideración es que el contraste de comparación de varianzas requiere normalidad de la variable a estudio en cada uno de las poblaciones estudiadas.

Contraste de comparación de varianzas

Se contrastará la hipótesis nula de igualdad de varianzas a través del estadístico de contraste:

$$F = \frac{s_1^2}{s_2^2}$$

Para un contraste bilateral, se situará en el numerador la mayor de las varianzas muestrales, y el resultado del estadístico se comparará con el valor crítico $F_{1-\alpha/2}$ de un modelo de probabilidad F de Snedecor con n_1-1 y n_2-1 grados de libertad (del numerador y del denominador), respectivamente.

Intervalo de confianza para la diferencia de medias

El intervalo de confianza para la diferencia de medias se obtiene, según el contraste de igualdad de varianzas poblacionales conduzca a aceptar la hipótesis nula (igualdad de varianzas) o a rechazarla (varianzas distintas) será:

1. Caso de varianzas poblacionales desconocidas y supuestamente iguales

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{1\bullet\alpha/2} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right]$$

con

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

y $t_{1-\alpha/2}$ coeficiente de una t de student con $n_1 + n_2 - 2$ grados de libertad

2. Caso de varianzas poblacionales desconocidas y significativamente distintas

$$IC_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x}_1 - \bar{x}_2) \pm t_{1\bullet\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

En este caso, $t_{1-\alpha/2}$ debe ser obtenido de una distribución t de student con grados de libertad f, aproximados a través de la expresión:

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

Contraste de hipótesis para comparar las medias

Podemos contrastar la hipótesis de que las medias poblacionales μ_1 y μ_2 son iguales, ($\mu_1 - \mu_2 = 0$), según las varianzas poblacionales se supongan iguales o sean diferentes a través de los estadísticos de contraste correspondientes:

1. Caso de varianzas poblacionales desconocidas y supuestamente iguales

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

con s_p^2 definido anteriormente y el estadístico t sigue un modelo t de student con $n_1 + n_2 - 2$ grados de libertad

2. Caso de varianzas poblacionales desconocidas y significativamente distintas

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

El estadístico t sigue aproximadamente una distribución t de student con f grados de libertad como se definió previamente.

Un ejemplo

Una de las variables incluida en los datos del cuadro 4.1 es la edad de las mujeres estudiadas. Interesa averiguar si existe diferencia significativa en las medias de edad de las mujeres según su actitud hacia la lactancia materna, tras el parto, sea positiva o negativa. Se desea cuantificar la diferencia promedio de edad entre uno y otro grupo.

Se dispone de dos muestras aleatorias e independientes de tamaños n_1=43 (mujeres que tras el parto tienen actitud positiva) y n_2=19 (mujeres que tras el parto presentan actitud negativa), con medias y desviaciones típicas muestrales de:

$$\bar{x}_1 = 25{,}90, \bar{x}_2 = 29{,}31, s_1 = 4{,}77, s_2 = 5{,}47.$$

Para decidir si las varianzas muestrales se suponen iguales o son diferentes se realizará el contraste de hipótesis de comparación de varianzas. Se contrastará la hipótesis nula de igualdad de varianzas poblacionales a través del estadístico:

$$F = \frac{5{,}47^2}{4{,}77^2} = 1{,}32$$

valor que no resulta significativo en una F de Snedecor con 42 y 18 grados de libertad, para α = 0,05, suponiendo entonces que las varianzas poblacionales son iguales. Para que el resultado obtenido sea riguroso es necesario suponer que la variable edad sigue un modelo normal en ambos grupos. El valor de la varianza común será $s_p^2 = 24,90$.

El intervalo de confianza, de nivel 95%, será entonces:

$$I_{0,95}(\mu_1 - \mu_2) = \left[(25,90 - 29,31) \pm 2,00 \sqrt{24,90\left(\frac{1}{43} + \frac{1}{19}\right)} \right]$$
$$= [- 6,16 , - 0,66]$$

habiendo obtenido el coeficiente t en un modelo t de student con 60 grados de libertad, y siendo la media de edad del grupo de respuesta positiva desde 0,66 hasta 6,16 años inferior al de respuesta negativa, con seguridad 95%. En promedio, las madres primíparas con actitud positiva hacia la lactancia materna son más jóvenes.

El contraste de la hipótesis nula de igualdad de medias, $\mu_1 - \mu_2 = 0$, se basará en el estadístico:

$$t = \frac{(25,90 - 29,31)}{\sqrt{24,90\left(\frac{1}{43}\right)}} = - 2,48$$

que queda fuera de los límites que delimitan la región de aceptación de la hipótesis nula, valores -2,00, 2,00, de un modelo t de student con 60 grados de libertad, rechazando la hipótesis de igualdad de medias. Para la validez de los resultados es necesario suponer normalidad de la variable en ambos grupos, pues en uno de ellos el número de efectivos es 19, inferior a 30.

Comparación de medias de poblaciones apareadas

En el caso de grupos apareados, la comparación de medias se realizará a través de las diferencias entre cada valor de la variable sobre cada individuo de uno de los grupos y el valor de la variable en su pareja.

Situación

Se dispone de una muestra de n pares de observaciones , (x_i^1, x_i^2) , i=1,...,n, de una variable cuantitativa con medias μ_1 y μ_2 respectivamente. Se dispondrá la variable creada a partir de las diferencias : $d_i = x_i^1 - x_i^2$, variable con media poblacional $\mu_d = \mu_1 - \mu_2$, y media y desviación típica muestrales $\bar{x}_d , s_d$.

Suposiciones/Requerimientos

El modelo de probabilidad seguido por la variable de las diferencias es normal o n es mayor que 30.

Intervalo de confianza para la diferencia de medias

Podemos construir el intervalo de confianza para la diferencia de medias a través del intervalo para la media de las diferencias:

$$IC_{1-\alpha}(\mu_d) = \left[\bar{x}_d \pm t_{1-\alpha/2} \frac{s_d}{\sqrt{n}} \right]$$

con $t_{1-\alpha/2}$ coeficiente de una t de student con n-1 grados de libertad

Contraste de hipótesis para la diferencia de medias

Se contrastará la hipótesis nula de igualdad de medias a través de la hipótesis nula $\mu_d = 0$. El estadístico de contraste es:

$$t = \frac{\bar{x}_d}{s_d / \sqrt{n}}$$

cuya distribución muestral es la de una t de student con n-1 grados de libertad

Un ejemplo

Los datos que se presenta en el cuadro 3.5 corresponden a observaciones obtenidas sobre 20 individuos de la variable nivel de ácido úrico (mg/100ml) y la variable respuesta positiva a un tratamiento, en tres momentos distintos del tiempo. Las variables UR1, UR2 y UR3 contienen los datos sobre ácido úrico en los tiempos que genéricamente se identificará como 1, 2 y 3. Se pretende comparar los niveles medios de ácido úrico entre los tiempos 1 y 3, para saber si se ha producido una disminución significativa.

Se dispone de dos muestras apareadas, pues se trata de los mismos individuos, referidas a los instantes 1 y 3. La variable de las diferencias entre los niveles de ácido úrico del momento 1 y el 3 presenta una media y desviación típica de $\bar{x}_d = 2{,}97$, $s_d = 2{,}39$, y el intervalo de confianza para la media de las diferencias,
$\mu_d = \mu_1 - \mu_2$, de nivel 95% será:

$$IC_{0.95}(\mu_d) = \left[2{,}97 \pm 2{,}093 \frac{2{,}392}{\sqrt{20}} \right] = [\,1{,}85\,,\,4{,}09\,]$$

UR1	UR2	UR3	R1	R2	R3
7.2	6.2	6.0	0	0	1
8.4	6.4	6.0	1	1	1
9.0	9.0	8.2	0	1	1
7.0	6.9	6.4	1	0	0
6.5	6.5	7.1	0	1	1
10.2	9.1	9.0	0	1	1
9.5	8.4	8.2	1	1	1
6.4	5.2	5.2	0	0	0
7.7	6.0	4.4	0	0	1
11.4	9.1	6.4	1	1	0
15.0	6.2	5.0	0	1	1
8.5	4.2	4.1	1	1	1
9.5	6.4	6.6	0	0	0
7.3	5.2	5.4	1	0	1
12.4	9.1	6.4	0	1	0
10.0	9.0	8.0	0	0	1
11.1	6.6	6.4	0	0	1
9.4	5.2	5.0	0	1	1
8.4	4.4	4.4	1	0	0
8.2	6.0	5.5	0	1	1

Cuadro 4.5.- Observaciones sobre 20 individuos

por lo que se puede afirmar, con seguridad 95%, que se produce una disminución media entre 1,85 y 4,09 mg/100ml en el nivel de ácido úrico.

A través del contraste de hipótesis para la hipótesis nula de que la media de las diferencias es cero, $\mu_d = \mu_1 - \mu_2 = 0$, tendremos:

$$t = \frac{2{,}97}{2{,}392 / \sqrt{20}} = 5{,}55$$

valor que, para un nivel de significación $\alpha = 0{,}05$, excede los límites de la región de aceptación de la hipótesis nula, delimitada por los valores -2,093 y 2,093, obtenidos de una t de student con 19 grados de libertad, detectando diferencias significativas entre las medias. Tanto para el intervalo de confianza como

para el contraste de hipótesis es necesario suponer normalidad de la variable nivel de ácido úrico, puesto que el tamaño muestral es 20, inferior a 30.

Pruebas no paramétricas para la comparación de una variable respuesta cuantitativa en dos grupos

En los apartados precedentes se han establecidos las pruebas más habituales para la comparación de una variable cuantitativa en dos grupos, independientes o apareados, a través de las medias de la variable respuesta. Sin embargo, las suposiciones y requerimientos que se establecen pueden resultar de difícil cumplimiento en algunas ocasiones, especialmente si el tamaño de los grupos estudiados es pequeño. En estos casos puede recurrirse a la utilización de *pruebas no paramétricas* o *de libre distribución*. La ventaja de estas pruebas estriba en que no necesitan apenas requerimientos. Sin embargo, las hipótesis que permiten contrastar no pueden considerarse equivalentes a las descritas para los métodos precedentes, debiendo tener presente que las conclusiones obtenidas no son absolutamente equivalentes. En general, el mecanismo de funcionamiento de estas pruebas se basa en la transformación de los valores de la variable cuantitativa en posiciones ordinales o rangos una vez los hemos ordenado de menor a mayor. Así, un valor de 35 años para la variable edad pasaría, en una situación concreta, a ser, por ejemplo, el rango 84 de entre una muestra de 300 observaciones, si ese valor ocupa el orden 84 una vez hemos ordenado todas las edades de menor a mayor. Presentamos a continuación dos pruebas no paramétricas útiles para comparar una variable respuesta cuantitativa entre dos poblaciones, según los grupos muestrales sean independientes o apareados. La existencia de dudas sobre el cumplimiento de los requerimientos descritos para las situaciones contempladas en los apartados precedentes, o la constatación de que no se cumplen, sugiere la utilización de estas pruebas como procedimientos que ayudarán a apoyar nuestras conclusiones con el rigor estadístico necesario.

Prueba U de Mann-Whitney para grupos independientes

Esta prueba permitirá comparar las medianas de una variable en dos grupos independientes. Hay que tener en cuenta que si la distribución de la variable es aproximadamente simétrica, la comparación de medianas será aproximadamente equivalente a la de medias.

Situación

Se dispone de dos muestras aleatorias independientes de tamaños n_1 y n_2, extraídas de cada una de las poblaciones a estudio y una variable cuantitativa con medianas Md_1 y Md_2 en las respectivas poblaciones.

Suposiciones/Requerimientos

La variable a estudio es cuantitativa, al menos ordinal.

Contraste de hipótesis U de Mann-Whitney

Se contrastará la hipótesis nula de igualdad de medianas, $Md_1 = Md_2$, a través del estadístico:

$$U = S_1 - \frac{n_1(n_1+1)}{2}$$

con S_1 la suma de los rangos asignados a las observaciones de la muestra del grupo 1 (es indiferente construir el estadístico en función del grupo 2). Para valores de n_1 y n_2 inferiores a 6, el estadístico calculado debe ser consultado en tablas de percentiles adecuadas para el estadístico (ver bibliografía recomendada), obteniendo los valores críticos $w_{\alpha/2}$ y $w_{1-\alpha/2}$ que delimitan la región de aceptación de la hipótesis nula. En general, cuando n_1 y n_2 son superiores a 6, se transforma el estadístico U en:

$$z = \frac{U - (n_1\, n_2 / 2)}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}}$$

cuya distribución es aproximadamente normal.

Un ejemplo

En el apartado de comparación de medias con grupos independientes anterior se compararon las medias de edad de las mujeres, tras el parto, según la actitud hacia la lactancia, habiendo encontrado diferencias significativas, p<0.05, en las medias de edad de ambas poblaciones. Se desea comparar las medianas a través de la prueba U de Mann-Whitney.

Se dispone de dos muestras independientes de tamaños n_1=43 y n_2=19, ambos mayores de 6.

El estadístico U valdrá:

$$U = 1199 - \frac{43\,(43+1)}{2} = 253$$

tomando como referencia para calcularlo el grupo 1, para el que la suma de los rangos de las edades de las mujeres que lo forman es: S_1=1199.
La transformación normal del estadístico toma el valor:

$$z = \frac{253 - (43 \,.\, 19) / 2}{\sqrt{43 \,.\, 19\,(43 + 19 + 1)/12}} = -2{,}38$$

que, para α = 0,05, se encuentra fuera de la región de aceptación de la hipótesis nula de igualdad de medianas, delimitada por los valores -1,96 y 1,96. Se rechazará la hipótesis nula, concluyendo diferencias significativas en las medianas de la edad entre los grupos estudiados. A diferencia de la comparación de medias, no es necesaria ninguna suposición adicional para sustentar el resultado obtenido.

Prueba de Wilcoxon por rangos para grupos apareados

Prueba para comparar las medianas de poblaciones apareadas.

Situación

Se dispone de una muestra de n pares de observaciones , (x_i^1, x_i^2) , i=1,...,n, de una variable cuantitativa, con medianas Md_1 y Md_2 en las respectivas poblaciones.

Suposiciones/Requerimientos

Se presenta la prueba únicamente en su aproximación por un modelo normal. En este caso, se requiere que n > 15. Para n≤15, puede ser consultada la prueba exacta en la bibliografía recomendada. El estadístico que se obtiene en este caso debe ser contrastado en las tablas pertinentes.

Contraste de hipótesis de Wilcoxon por rangos

Se contrastará la hipótesis de igualdad de medianas, $Md_1 = Md_2$, a través del estadístico:

$$z = \frac{T}{\sqrt{\frac{k\,(k+1)\,(2k+1)}{6}}}$$

con:

$$T = r_1 + r_2 + \ldots + r_k$$

$$r_i = \text{Rango}(D_i) \text{ si } D_i > 0$$

$$= -\text{Rango}(D_i) \text{ si } D_i < 0$$

$$D_i = x_i^1 - x_i^2 \;;\; k = \text{n° de diferencias no nulas}$$

El estadístico z sigue una distribución normal.

Un ejemplo

En el ejemplo de comparación de medias con datos apareados anterior se obtenía diferencia significativa en las medias de los grupos apareados, para la variable nivel de ácido úrico entre los instantes temporales 1 y 3, siendo necesario suponer normalidad de la variable para aceptar con rigor el resultado. Puesto que n=20, se compararán las medianas de ambos grupos a través de la prueba de Wilcoxon.

Los datos para las variables U1 y U3, sus diferencias, rangos, y rangos con signo se pueden observar en la tabla adjunta

U1	U3	D_i	Rango(D_i)	r_i
7.2	6.0	1.2	5.5	5.5
8.4	6.0	2.4	10.0	10.0
9.0	8.2	.8	3.0	3.0
7.0	6.4	.6	2.0	2.0
6.5	7.1	-.6	1.0	-1.0
10.2	9.0	1.2	4.0	4.0
9.5	8.2	1.3	7.0	7.0
6.4	5.2	1.2	5.5	5.5
7.7	4.4	3.3	13.0	13.0
11.4	6.4	5.0	18.0	18.0
15.0	5.0	10.0	20.0	20.0
8.5	4.1	4.4	15.5	15.5
9.5	6.6	2.9	12.0	12.0
7.3	5.4	1.9	8.0	8.0
12.4	6.4	6.0	19.0	19.0
10.0	8.0	2.0	9.0	9.0
11.1	6.4	4.7	17.0	17.0
9.4	5.0	4.4	15.5	15.5
8.4	4.4	4.0	14.0	14.0
8.2	5.5	2.7	11.0	11.0

El estadístico de contraste tomará el valor:

$$z = \frac{208}{\sqrt{\frac{20\,(20+1)\,(2\,.\,20+1)}{6}}} = 3{,}88$$

que queda fuera de los límites de aceptación de la hipótesis nula, -1,96 y 1,96, para nivel de significación 0,05. Se rechaza la igualdad de medianas concluyendo que estas son distintas.

4.5 ESTUDIO DE UNA VARIABLE RESPUESTA EN MAS DE DOS POBLACIONES

Las situaciones descritas en el apartado anterior pueden ser generalizadas a más de dos poblaciones. El objetivo habitual, al igual que en el caso de dos poblaciones suele ser la comparación de las características de la variable respuesta entre los grupos estudiados, encontrando situaciones similares a las descritas en el apartado anterior. Así, el tipo de variable respuesta, cualitativa o cuantitativa o el diseño de las muestras estudiadas, apareadas o independientes, condicionarán las pruebas a utilizar. Existirá igualmente la posibilidad de utilizar, cuando proceda, pruebas no paramétricas, con requerimientos más débiles que las paramétricas.

Prueba Ji-cuadrado para la comparación de proporciones en más de dos poblaciones independientes. Asociación o relación entre variables

Cuando la variable respuesta es cualitativa, el interés suele residir en la comparación de las proporciones de alguna de sus categorías entre los grupos estudiados o, equivalentemente, la detección de asociación significativa entre la variable respuesta y la explicativa.

Situación

Para exponer esta situación supondremos una variable respuesta con I categorías, estudiada en J grupos independientes, con tamaños muestrales n_j, j=1,2,...,J, siendo el tamaño muestral total, n, la suma de los tamaños de los J grupos. Los datos pueden ser estructurados en una tabla de doble entrada o tabla de contingencia como:
donde o_{ij} representa la frecuencia observada para la situación conjunta {categoría i de la variable respuesta, grupo o categoría j de la variable explicativa}, y $p_{i/j}$ representa la proporción poblacional de la categoría i en el grupo j. En una situación genérica, con tabla de dimensión I x J, I filas y J columnas, se tiene:

		Variable explicativa: grupos					
		1	**2**	**...**	**j**	**...**	**J**
Variable respuesta	**1**	o_{11} $p_{1/1}$	o_{12} $p_{1/2}$				o_{1J} $p_{1/J}$
	2	o_{21} $p_{2/1}$	o_{22} $p_{2/2}$				o_{2J} $p_{2/J}$
	.				.		
	.				.		
	i			…	o_{ij} $p_{i/j}$	…	
	.				.		
	.				.		
	I	o_{I1} $p_{I/1}$	o_{I2} $p_{I/2}$				o_{IJ} $p_{I/J}$

$$\hat{p}_{i/j} = \frac{o_{ij}}{o_{+j}} \;\; ; \;\; n_j = o_{+j} = \sum_{i=1}^{I} o_{ij} \;\; ; \;\; i = 1, 2, ..., I \; ; \; j = 1, 2, ..., J$$

y, de manera similar, se podría obtener:

$$\hat{p}_{j/i} = \frac{o_{ij}}{o_{i+}} \;\; ; \;\; o_{i+} = \sum_{j=1}^{J} o_{ij} \;\; ; \;\; i = 1, 2, \ldots, I \; ; \; j = 1, 2, \ldots, J$$

La igualdad de proporciones se producirá cuando para cada categoría de la variable respuesta, por ejemplo la i, suceda que las proporciones correspondientes a cada uno de los grupos, $p_{i/j}$ sean iguales (para j=1, 2,..., J). Esta situación es equivalente a afirmar que no existe asociación entre la variable respuesta y la explicativa, puesto que las proporciones de ocurrencia de sus diferentes categorías son iguales en todos los grupos estudiados. Si esto no sucede, y existen diferencias significativas en las proporciones, se podrá de hablar de asociación o relación entre la variable respuesta y la explicativa, en el sentido de que algunas proporciones de alguna o algunas categorías difieren entre los grupos estudiados, asociándose categorías y grupos (unas combinaciones presentan mayor y otras menor proporción).

Suposiciones/Requerimientos

Se requiere que las frecuencias esperadas (ver después) sean superiores o iguales a 5. Si la tabla es de dimensión superior a 2 x 2, puede aceptarse frecuencias esperadas inferiores a 5 pero superiores a 1 como máximo en el 20% de las celdas.

Contraste de hipótesis de igualdad de proporciones o ausencia de asociación. Prueba Jicuadrado

Podemos contrastar la hipótesis nula de que las proporciones poblacionales $p_{i/j}$ son iguales, j=1,...,J, o, equivalentemente, que no existe asociación entre la variable respuesta y la explicativa a través del estadístico:

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \;\; ; \;\; e_{ij} = \frac{o_{i+}\, o_{+j}}{n}$$

donde e_{ij} representa las frecuencias esperadas en cada una de las celdas bajo la hipótesis nula de no asociación entre las variables. En el caso de tablas de dimensión 2 x 2, el contraste ji-cuadrado es equivalente a la prueba z de comparación de proporciones.

Un ejemplo

A partir de los datos del cuadro 4.1, se ha clasificado las mujeres estudiadas en 'lactancia materna baja' si el tiempo de lactancia ha sido inferior a 6 meses, y 'lactancia materna normal' si el tiempo ha sido superior o igual a 6 meses. Se desea averiguar si existe asociación entre esta nueva variable, lactancia materna, y la actitud hacia la lactancia materna, evaluada después del parto.

La tabla adjunta presenta las frecuencias observadas para las variables estudiadas:

	Lactancia materna	
	Baja	**Normal**
Actitud +	26	17
Actitud -	18	1

mientras que la siguiente tabla contiene los porcentajes de las categorías de lactancia materna según la actitud, y las frecuencias esperadas para la hipótesis nula de no asociación entre las variables (cursiva), o, lo que es lo mismo, proporciones de las categorías de lactancia materna iguales según actitud:

	Lactancia materna	
	Baja	**Normal**
Actitud +	26 30,5 60,5%	17 12,5 39,5%
Actitud -	18 13,5 94,7%	1 5,5 5,3%

habiendo obtenido las frecuencias esperadas como el producto del total de la fila por el total de la columna dividido por 62 (n total) (p. ej., en la primera celda, la frecuencia esperada será [43 x 44]/62, cuyo resultado es 30,5). Se observa diferencias muestrales importantes en los porcentajes de lactancia materna normal según la actitud sea positiva o negativa (39,5% frente a 5,3%). Para constatar si esas diferencias son significativas, o, lo que es lo mismo, si las variables se asocian, el estadístico de contraste será:

$$\chi^2 = \frac{(26-30,5)^2}{30,5} + \frac{(18-13,5)^2}{13,5} + \frac{(17-12,5)^2}{12,5} + \frac{(1-5,5)^2}{5,5}$$
$$= 7,51$$

valor que, contrastado con el de un modelo Ji-cuadrado con (2-1) x (2-1) = 1 grados de libertad, resulta significativo para $\alpha = 0,05$, puesto que el valor crítico es 3,84.

Prueba de Cochran para la comparación de proporciones en más de dos poblaciones apareadas

La situación resuelta con la prueba ji-cuadrado, para poblaciones independientes, no es aplicable cuando los grupos han sido apareados. En este caso, hay que recurrir a una prueba que contemple la información debida al apareamiento. La prueba que se presenta es válida cuando la variable respuesta es cualitativa pero dicotómica.

Situación

Se dispone de 3 o más muestras apareadas, obtenidas aleatoriamente de las correspondientes poblaciones, con n observaciones en cada una de ellas de una variable cualitativa dicotómica.
Cada grupo de tres, cuatro, o más individuos apareados recibirá el nombre de estrato de apareamiento. Así, si el número de grupos es k, en cada estrato habrá k individuos apareados.

Suposiciones/Requerimientos

Se requiere que $n'.k \geq 25$, donde n' es el número real de estratos que intervienen en el análisis (todos aquellos para los que existan diferencias de respuesta entre los individuos que lo componen), $n' \leq n$, y k es el número de grupos de apareamiento.

Contraste de hipótesis de Cochran para la comparación de proporciones apareadas

Para el contraste de la hipótesis nula de igualdad de proporciones de respuesta a una de las categorías de la variable respuesta el estadístico de contraste de Cochran es:

$$Q = \frac{k\sum_{j=1}^{k} C_j^2 - T^2}{kT - \sum_{i=1}^{n} F_i^2}(k-1)$$

C_j = Respuestas a la categoría a studio en el grupo j

F_i = Respuestas a la categoría a estudio en el estrato i

$$T = \sum_{j=1}^{k} C_j = \sum_{i=1}^{n} F_i$$

El estadístico Q sigue una distribución ji-cuadrado con k-1 grados de libertad.

Un ejemplo

En los datos del cuadro 4.5 se incluyen observaciones sobre 20 individuos de la respuesta a un tratamiento en tres momentos distintos (variables R1, R2, y R3) en las que el valor 1 representa respuesta positiva. Se pretende comparar las proporciones de respuesta positiva en los tres grupos.

Se dispone de tres grupos apareados (son los mismos individuos). Se contrastará la hipótesis nula de igualdad de proporciones de respuesta positiva a través del estadístico de Cochran:

$$Q = \frac{3(7^2 + 11^2 + 14^2) - 32^2}{3.32 - (1^2 + 3^2 + 2^2 + \ldots + 1^2)}(3-1) = 4{,}35$$

R1	R2	R3	F_i
0	0	1	1
1	1	1	3
0	1	1	2
1	0	0	1
0	1	1	2
0	1	1	2
1	1	1	3
0	0	0	0
0	0	1	1
1	1	0	2
0	1	1	2
1	1	1	3
0	0	0	0
1	0	1	2
0	1	0	1
0	0	1	1
0	0	1	1
0	1	1	2
1	0	0	1

C_1=7 C_2=11 C_3=14; T=32

siendo las cantidades C, F y T las que aparecen en la tabla adjunta:

El valor del estadístico Q, para α = 0,05, está por debajo del valor crítico de una ji-cuadrado con 3 - 1 = 2 grados de libertad, que resulta ser 5,991, no pudiendo rechazar la hipótesis nula de igualdad de proporciones de respuesta positiva.

Comparación de medias de poblaciones independientes

Cuando la variable respuesta es cuantitativa, y al igual como fue expuesto en el apartado previo de comparación, las medias suelen ser los parámetros de interés para comparar el comportamiento de una variable en 3 o más grupos. Los procedimientos que se presentan a continuación son la generalización natural de los presentados anteriormente para dos grupos. Sin embargo, algunos de ellos, no serán expuestos con detalle, sugiriendo al lector la consulta de la bibliografía recomendada, en la que podrá encontrar los detalles de tales procedimientos.

Análisis de la varianza de una vía

Es la prueba paramétrica adecuada para la comparación de medias de una variable respuesta cuantitativa en tres o más poblaciones. Permite contrastar la hipótesis nula de igualdad de medias frente a la alternativa de que alguna o algunas de las medias son diferentes. Debido al carácter básico de este texto, se presenta a continuación la situación y suposiciones/requerimientos de la prueba, recomendando al lector la consulta de la bibliografía recomendada para la aplicación de la prueba.

Situación

Se dispone de k muestras (k>2) independientes, cada una de ellas aleatoriamente seleccionada de su respectiva población, con n_i observaciones de una variable cuantitativa en el grupo i (i=1,..., k). La variable posee medias poblacionales μ_i en cada una de las poblaciones.

Suposiciones/Requerimientos

Se requiere que la variable respuesta siga un modelo de probabilidad normal en cada una de las poblaciones estudiadas, con la misma varianza en cada una de ellas. Las observaciones de la variable en cada uno de los grupos son independientes.

Prueba de Kruskal-Wallis para la comparación de tres o más poblaciones independientes

Esta prueba permitirá comparar las medianas de una variable en tres o más grupos. Puede ser considerada una generalización de la prueba U de Mann-Whitney para dos grupos.

Situación

Se dispone de k (k>2) muestras aleatorias independientes de tamaños n_i (i=1,..., k), cada una de ellas seleccionada de su respectiva población, y una variable cuantitativa con medianas Md_i, en cada una de las poblaciones.

Suposiciones/Requerimientos

Al menos en alguno de los grupos estudiados, el tamaño muestral es superior a 5.

Contraste de hipótesis de Kruskal-Wallis

Se contrastará la hipótesis nula de igualdad de medianas, $Md_1 = Md_2 = ... = Md_k$, a través del estadístico:

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1)$$

$$n = n_1 + n_2 + ... + n_k$$

R_i = **Suma de rangos del grupo i**

para el que las n observaciones deben ser previamente ordenados, asignando sus correspondientes rangos. La distribución muestral del estadístico es ji-cuadrado con k-1 grados de libertad.

Un ejemplo

A partir de los datos recogidos en el cuadro 4.1, se ha creado una variable categórica tomando como base la edad de las mujeres. Esta variable define tres grupos: edad inferior a 24 años, edad entre 24 y 30 años y edad superior o igual a 30 años, de tamaños n_1=18, n_2=25 y n_3=19, e independientes. Se desea comparar el tiempo de lactancia entre estos tres grupos.

Se dispone de tres grupos independientes, y una variable cuantitativa, el tiempo de lactancia. La comparación de medias entre los tres grupos a través de una análisis de la varianza requeriría normalidad de esta variable en cada grupo y homogeneidad de varianzas. Además, los tamaños muestrales son pequeños, dificultando la comprobación de estos requerimientos. Se procederá a contrastar la hipótesis nula de igualdad de medianas entre los tres grupos, a través del estadístico de contraste de Kruskal-Wallis:

$$H = \frac{12}{62(62+1)}\left(\frac{843^2}{18}+\frac{757^2}{25}+\frac{354^2}{19}\right)-3(62+1)=22{,}74$$

Las sumas de los rangos de cada uno de los grupos se han obtenido tras ordenar las 62 observaciones y asignar a cada tiempo su rango. El resultado obtenido excede al valor crítico de una ji-cuadrado con 2 grados de libertad para α = 0,05, que es 5,991, por lo que rechazaremos la hipótesis nula de igualdad de medianas, concluyendo tendencia en unos grupos a alcanzar valores más elevados del tiempo de lactancia que en otros.

Comparación de medias de poblaciones apareadas

Al igual que en el apartado anterior, se presenta los procedimientos habituales a este fin.

Análisis de la varianza de medidas repetidas

Cuando en una situación se observan dos o más medidas sobre los mismos sujetos estudiados puede hablarse de medidas repetidas. El análisis de la varianza antes descrito pasa a ser denominado de medidas repetidas. En realidad, se puede pensar que se trata de una generalización de la prueba t para datos apareados, pero con más de dos grupos que se corresponden con las sucesivas repeticiones de las observaciones de los sujetos. Al igual que en el caso del análisis de la varianza para poblaciones independientes, se presenta a continuación la situación y los requerimientos/suposiciones del procedimiento, dejando al lector la ampliación sobre esta prueba en la bibliografía recomendada.

Situación

Se dispone de una muestra aleatoria de n sujetos, y k observaciones repetidas, $(x_i^1, x_i^2, \ldots, x_i^k)$, i=1,...,n, para cada uno de ellos. La variable a estudio posee medias $(\mu_1, \mu_2, \ldots, \mu_k)$, y el objetivo es comparar tales medias, es decir contrastar la hipótesis nula de igualdad de medias frente a la alternativa de que al menos alguna difiere de las demás.

Suposiciones/Requerimientos

De forma similar al análisis de la varianza para poblaciones independientes, se requiere normalidad de la variable a estudio y homogeneidad de las varianzas de los grupos. Adicionalmente, se requiere que las correlaciones de la variable entre los grupos considerados sean homogéneas.

Prueba de Friedman para la comparación de tres o más poblaciones apareadas

Esta prueba permite comparar las medianas de una variable en tres o más grupos apareados.

Situación

Se dispone de una muestra de n estratos de apareamiento, y en cada uno de ellos de k observaciones de una variable cuantitativa, $(x_i^1, x_i^2, \ldots, x_i^k)$, i=1,...,n. El valor k representa el número de grupos de apareamiento. La variable posee medianas Md_i en los grupos considerados.

Suposiciones/Requerimientos

Se requiere que $n \geq 8$.

Contraste de hipótesis de Friedman

Se contrastará la hipótesis nula de igualdad de medianas, $Md_1 = Md_2 = \ldots = Md_k$, a través del estadístico:

$$Q = \frac{12}{n\,k\,(k+1)} \sum_{i=1}^{k} R_i^2 - 3\,n\,(k+1)$$

R_i = Suma de rangos del grupo i obtenidos al asignarlos dentro de cada estrato

cuya distribución de probabilidad es ji-cuadrado con k-1 grados de libertad.

Un ejemplo

A partir de los datos del cuadro 4.5, las variable U1, U2 y U3 contienen las observaciones del nivel de ácido úrico sobre los 20 sujetos estudiados en tres momentos de tiempo distintos. Interesa averiguar si existe diferencia significativa en las medianas de esta variable entre los tres grupos estudiados.

Se dispone de 20 estratos de apareamiento de tamaño 3 (hay tres grupos). Se contrastará la hipótesis nula de comparación de medianas a través del estadístico de Friedman:

$$Q = \frac{12}{20 \,.\, 3\,(3+1)} (58^2 + 37^2 + 25^2) - 3 \,.\, 20\,(3+1) = 27{,}9$$

cuyo valor excede a 5,991, valor de una distribución ji-cuadrado con 2 grados de libertad para $\alpha = 0{,}05$, rechazando la hipótesis nula de igualdad de medianas.

4.6. ESTUDIO DE LA RELACION ENTRE VARIABLES

En el cuadro 3.3 se describía los procedimientos para el estudio de la asociación o relación entre variables. A continuación se describe con mayor detalle estos procedimientos.

Asociación entre variables cualitativas

Cuando se dispone de dos variable cualitativas, la situación es la descrita a través de la solución basada en la prueba de hipótesis que contrasta la hipótesis nula de no asociación entre las variables se obtiene a partir del estadístico ji-cuadrado descrito.

Relación entre variables ordinales o cuantitativas transformadas en rangos

Si las variables son ordinales o cuantitativas pero transformadas en rangos, se trata de obtener una medida que nos permita establecer si las posiciones o rangos ocupadas por los individuos para una de las variables se corresponden con las ocupadas para la otra variable.

Situación

Se dispone de n parejas de observaciones $\{x_i, y_i\}$, de dos variables X e Y. Ambas variables son cuantitativas u ordinales. Los valores de las variables, si éstas son cuantitativas, son transformados en rangos (r_x^i, r_y^i)

Coeficiente de correlación de Spearman

El estadístico que mide la relación existente entre los rangos ocupados por los sujetos en una de las variables y los ocupados en la otra, sobre la muestra, es el de Spearman:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

$$d_i = r_y^i - r_x^i$$

cuyo valor puede oscilar entre -1 y +1, y para el que los valores -1 y +1 representan concordancia perfecta entre los rangos ocupados en ambas variables, aunque el valor -1 de sentido negativo (mayor rango en una menor en la otra), mientras que el valor +1 de sentido positivo (mayor rango en una mayor en la otra). El valor 0 representa la ausencia de relación entre los rangos.

Suposiciones/requerimientos

Se requiere que no se produzcan demasiados solapamientos (rangos idénticos) en los rangos de cada variable. Se requiere que $n \geq 10$.

Contraste de hipótesis sobre la correlación de Spearman

Puede tener interés contrastar la hipótesis nula de ausencia de correlación de Spearman en la población. Si denotamos por ρ_s al coeficiente de Spearman poblacional, se contrastará la hipótesis nula $\rho_s = 0$. El contraste de hipótesis puede ser resuelto utilizando como estadístico:

$$z = r_s \sqrt{n-1}$$

cuya distribución es normal.

Un ejemplo

En el cuadro 4.6 se presenta valores de las tasas de mortalidad (x 100000) por todas las causas para las áreas de salud de la Comunidad Valenciana, para hombres y mujeres. Se pretende averiguar si existe relación entre las tasas para hombres y las tasas para mujeres.

Se dispone de 20 parejas de observaciones (tasa hombres, tasa mujeres) de una variable cuantitativa (tasa de mortalidad) obtenidas sobre las áreas de salud. En el cuadro 3.6 figuran los rangos de las diferentes tasas, tanto en hombres como en mujeres. El coeficiente de correlación de Spearman será:

$$\sum_{i=1}^{n} d_i = 306$$

$$r_s = 1 - \frac{6 \cdot 306}{20(20^2 - 1)} = 0{,}77$$

indicando una correlación muestral elevada y positiva entre los rangos de las áreas de salud en hombres y mujeres. El contraste de la hipótesis nula de correlación poblacional igual a 0, se realizará a través del estadístico:

$$z = 0{,}77\sqrt{20 - 1} = 3{,}36$$

valor que resulta significativo cuando es contrastado con una curva normal (valores críticos para α = 0,05, -1,96, 1,96).

Area	Hombres	Rango hombres	Mujeres	Rango mujeres
1	1298.5	20	1062.6	19
2	977.5	13	887.9	14
3	1005.3	15	967.7	18
4	874.1	11	742.2	6
5	807.2	5	785.3	9
6	873.3	10	890.1	15
7	740.4	2	403.0	1
8	760.8	3	722.7	5
9-12	1042.7	18	865.8	12
13	802.0	4	753.9	7
14	1000.7	14	1123.2	20
15	1021.1	16	868.0	13
16	1022.0	17	894.4	16
17	1239.4	19	941.4	17
18	907.4	12	698.5	3
19	852.4	8	839.6	11
20	849.1	7	788.0	10
21	867.9	9	721.2	4
22	676.3	1	555.1	2
23	808.9	6	778.3	8

Cuadro 4.6.- Rangos para el coeficiente de correlación de Spearman

Relación entre variables cuantitativas

En el capítulo 1 fue descrita la idea sobre la relación entre variables cuantitativas. Como se expuso, no es posible separar las ideas de magnitud de la relación (correlación) de la forma de la relación (regresión). A continuación se resumirá los aspectos más destacables de la regresión y correlación lineal para el estudio de la relación entre dos variables cuantitativas.

Regresión lineal simple. Correlación lineal

Se partirá de dos variables cuantitativas, X e Y, para las que se desea estudiar si la relación entre ellas puede ser de tipo lineal , $y = \alpha + \beta x$, y la magnitud de ésta.

Situación

Se dispone de una muestra aleatoria de n observaciones de las variables cuantitativas X e Y, (x_i, y_i) , obtenidas sobre los n sujetos. La variable Y es considerada variable respuesta, mientras que la X es la variable explicativa.

Suposiciones/Requerimientos

Se requiere que la distribución de la variable Y para cada valor de X sea normal, con la misma varianza, que las observaciones de Y son independientes y que las medias de las distribuciones de Y se relacionan linealmente con X.

Estimación del modelo de regresión lineal simple

Se trata de obtener el modelo estimado, $\hat{y} = \hat{\alpha} + \hat{\beta} x$ a partir de la información muestral. El método de estimación más frecuente es el conocido como *método de mínimos cuadrados*. Se trata de obtener la recta de regresión para la que $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ sea mínima. Los estimadores de los coeficientes de la recta son entonces:

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Coeficiente de correlación lineal simple de Pearson

El coeficiente de correlación de Pearson fue descrito en el capítulo 1. La expresión para su estimación es:

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

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

Inferencias sobre el parámetro β de la ecuación de regresión

El coeficiente β de la ecuación representa el efecto, en escala aditiva, de la variable explicativa X sobre la variable respuesta. De hecho representa el incremento de Y por unidad de incremento de X. Puede resultar de interés calcular intervalos de confianza o resolver pruebas de hipótesis sobre él. Tanto para los intervalos de confianza como para los contrastes de hipótesis, será fundamental el error estándar del coeficiente, $\mathbf{SE(\hat{\beta})}$, cuya estimación es:

$$SE(\hat{\beta}) = \frac{s_{y.x}}{\sqrt{(n-1)}\, s_x}$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)}} \quad ; \quad s_{y.x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-2)}}$$

El intervalo de confianza de nivel 1-α será:

$$I_{1-\alpha}(\beta) = [\,\hat{\beta} \pm t_{1-\alpha/2}\, SE(\hat{\beta})\,]$$

donde $t_{1-\alpha/2}$ es el coeficiente de una t de student con n-2 grados de libertad para el nivel de confianza exigido.Se puede contrastar la hipótesis nula de que β tome un valor específico, $\beta = \beta_0$, a través del estadístico de contraste:

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$$

cuya distribución es una t de student con n-2 grados de libertad.

Inferencias sobre el coeficiente de correlación

Si se dispone del coeficiente de correlación lineal de Pearson estimado, r, es posible contrastar la hipótesis nula de ausencia de correlación lineal poblacional, a través del estadístico:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

cuya distribución es una t de student con n-2 grados de libertad. Esta prueba de hipótesis es equivalente a la descrita para β, contrastando la hipótesis nula β = 0.

Un ejemplo

En los datos del cuadro 4.1 se incluye las observaciones de las variables tiempo de lactancia tras el parto y edad de la madre. Se desea inspeccionar la relación lineal existente entre estas dos variables, tomando como variable respuesta el tiempo de lactancia y explicativa la edad. Además, se desea contrastar si existe relación lineal significativa entre las variables estudiadas, y cuantificar el efecto de la edad sobre el tiempo de lactancia.

En relación a la nomenclatura descrita, X=Edad e Y=Tiempo, sobre los datos recogidos en el cuadro, tenemos las siguientes expresiones:

$$\sum_{i=1}^{n} x_i = 1671 \; ; \; \sum_{i=1}^{n} x_i^2 = 46685 \; ;$$

$$\sum_{i=1}^{n} y_i = 216{,}5 \; ; \; \sum_{i=1}^{n} y_i^2 = 1226{,}75 \; ;$$

$$\sum_{i=1}^{n} x_i\, y_i = 5329{,}5$$

por lo que los parámetros del modelo de regresión lineal serán estimados como:

$$\hat{\beta} = \frac{62 \,.\, 5329{,}5 - 1671 \,.\, 216{,}5}{62 \,.\, 46685 - 1671^2} = -0{,}307$$

$$\hat{\alpha} = \frac{216{,}5}{62} + 0{,}307\,\frac{1671}{62} = 11{,}76$$

obteniendo el modelo de regresión estimado:

$$\hat{y} = 11{,}76 - 0{,}307\, x$$

El coeficiente de correlación lineal de Pearson estimado valdrá:

$$r = \frac{62 \,.\, 5329{,}5 - 1671 \,.\, 216{,}5}{\sqrt{62 \,.\, 46685 - 1671^2}\,\sqrt{62 \,.\, 1226{,}75 - 216{,}5^2}} = -0{,}574$$

representando una relación lineal muestral inversa (a mayor edad menor tiempo de lactancia).La hipótesis nula de ausencia de correlación lineal en la población de la que proceden los datos puede ser contrastada

a través del estadístico:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = -0{,}574\sqrt{\frac{62-2}{1-(-0{,}574)^2}} = -5{,}42$$

que, al ser comparado con una t de student con 60 grados de libertad resulta significativo (los valores críticos para $\alpha = 0{,}05$ en prueba bilateral son -2,0 y 2,0). La correlación lineal entre las variables estudiadas es significativamente distinta de 0 en la población.
El parámetro β representa el efecto de la variable explicativa sobre la respuesta, por unidad de cambio. Así, en este caso, representará la disminución (el coeficiente es negativo) en el tiempo de lactancia por incrementar un año la edad de la madre. El intervalo de confianza para este parámetro cuantificará su valor en la población. Para nivel de confianza 95%, este intervalo será:

$$I_{0,95}(\beta) = [\hat{\beta} \pm t_{0,975}\, SE(\hat{\beta})]$$

$$= [-0{,}307 \pm 2 \,.\, 0{,}056] = [-0{,}419\,,\, -0{,}195]$$

pudiendo afirmar, con seguridad 95%, que la disminución del tiempo de lactancia se encontrará entre 0,419 y 0,195 meses por un año más de la madre.

C A P I T U L O 5

PRECAUCIONES EN EL ANÁLISIS DE LOS DATOS

En capítulos precedentes se ha visto diferentes elementos y procedimientos estadísticos de uso frecuente en el análisis de la información contenida en la investigación. La utilización de estos procedimientos como herramienta en el espectro general del método científico puede dar lugar a diversos errores atribuibles a diversas causas y en diferentes etapas de la aplicación del método estadístico. Diversos autores han revisado y constatado la existencia de errores en la aplicación de las técnicas estadísticas en publicaciones de revistas científicas de prestigio.

En el presente capítulo se pretende describir algunos de los aspectos de la secuencia de análisis estadístico en los que, por su relevancia con vistas a la 'calidad' de la evaluación de la información de la investigación, bien por la calidad del proceso o bien por su efecto sobre la extracción de conclusiones correctas, debe ser prestada una atención especial.

Algunos de los puntos considerados guardan relación con la comunicación de los resultados obtenidos, pues no debe ser olvidado que el lector de los resultados y conclusiones de la investigación debe ser informado de las técnicas, criterios, etc. utilizados, a fin de que pueda evaluarlos correctamente.

5.1 LA SELECCION DE LOS SUJETOS

Es un aspecto primordial en el análisis estadístico. Su efecto sobre la calidad de las inferencias estadísticas es esencial. Una selección o asignación aleatoria es la mejor garantía para la generalización de resultados y evita sesgos que cuestionarían ésta. Formalmente, a la hora de comunicar los resultados de nuestra investigación no es suficiente garantía para el lector el incluir frases como 'el muestreo se realizó al azar', y se debe recomendar la descripción del proceso estricto seguido para la selección de los sujetos.

En algunos estudios del entorno de las ciencias de la salud, fundamentalmente clínicos, es necesario especificar los criterios de inclusión o exclusión del estudio, cuando éstos existan. Cuando requieran seguimiento de los sujetos, debe darse suficiente información acerca de los procedimientos utilizados para el seguimiento, cifras de pérdidas ocurridas, y, si es posible, las características diferenciales de los sujetos perdidos.

Cabe por último señalar que cuando existan dudas acerca de la representatividad de la muestra, los métodos estadísticos, por sofisticados que sean, no ofrecen garantía para la generalización de los resultados. Este aspecto obliga a reconocer y discutir la situación, estableciendo las limitaciones del estudio, y moderando así sus pretensiones inferenciales.

5.2 LA OBTENCION DE LAS OBSERVACIONES

Los procedimientos utilizados en la medición u observación de los datos del estudio pueden ser fuente de error. Deben ser utilizados métodos (aparatos de medida, cuestionarios, etc.) que hayan sido validados previamente, es decir cuyos resultados sólo estén sujetos a error aleatorio. En el caso de aparatos de medida o cuestionarios de uso común, éstos deben ser identificados, y ser utilizados con la seguridad de no producir mediciones u observaciones erróneas. Cuando los métodos de medición u observación no son de uso frecuente, en la comunicación de los resultados del estudio es necesario describirlos brevemente, a fin de que pueda ser evaluada la calidad de las mediciones o la pertinencia de los cuestionarios que puedan ser utilizados.
Algunos estudios requieren la evaluación subjetiva por parte de observadores externos como parte de la información necesaria. En estos casos es necesario establecer mediciones de la concordancia entre los observadores, a fin de poder asegurar que todos ellos evalúan con criterios semejantes.

5.3 ALMACENAMIENTO Y PROCESAMIENTO DE LA INFORMACION

La progresiva informatización de los centros de trabajo, estudio, etc., con la correspondiente incorporación de los recursos informáticos a la investigación hace necesaria algunas consideraciones, dado el efecto que la utilización de estos recursos puede tener sobre la producción de errores en la etapa de análisis estadístico.
Un primer aspecto se deriva del almacenamiento de la información en una base de datos. Conviene elegir para ello un programa de gestión y almacenamiento de datos contrastado y con la mayor versatilidad. Se trata de que los datos puedan ser accesibles a través de los procedimientos de análisis que se utilizarán, habitualmente programas de ordenador. El punto más delicado en el almacenamiento de los datos deriva de la propia transcripción desde el soporte original hasta un archivo informático. Así, los errores de tecleo, confusión sobre variables, etc., pueden dar lugar a una base de datos con gran cantidad de errores en las observaciones que contiene. La trascendencia de estos errores sobre la calidad de las inferencias es crucial, pues, aunque no existan sesgos, los resultados serán generalizados erróneamente a la población origen. Se debe utilizar procedimientos que garanticen la calidad de la base de datos, tales como la introducción duplicada o triplicada de los datos, con corrección de las incompatibilidades encontradas, o la selección de una muestra aleatoria de las observaciones introducidas, a fin de evaluar la cantidad de errores cometidos, y sobre que variables se producen éstos, adoptando algún criterio para la repetición de la introducción cuando superen una determinada cifra.
El segundo aspecto tiene que ver con la aplicación de los procedimientos estadísticos a través del ordenador. Conviene asegurarse de que se utiliza un programa o paquete de programas estadísticos cuya calidad en los cálculos está reconocida. Además es recomendable que el paquete sea lo más versátil posible, puesto que a menudo la aplicación de unas u otras técnicas puede depender de requerimientos que deben ser comprobados sobre los propios datos. En cualquier caso, no se debe partir nunca de la premisa de que 'como lo ha hecho el ordenador debe estar bien', pues la mayoría de procedimientos estadísticos requieren de consideraciones no incluidas en los programas informáticos de aplicación estadística.

5.4 ANALISIS ESTADISTICO

En la fase de aplicación de los métodos estadísticos cada uno de los procedimientos utilizados responde a una justificación matemático-estadística y tiene unas propiedades o características determinadas que sugerirán su utilización o no, según el fin perseguido. No obstante, se remarca a continuación algunos puntos de especial atención, tanto por su frecuencia de uso como por la posibilidad de error en su aplicación.

Desviación estándar y error estándar

Los procedimientos estadísticos de inferencia se fundamentan en la variabilidad en el muestreo de los estadísticos. Esta idea, aunque suele requerir cierta abstracción para su comprensión, pues debemos imaginar múltiples muestras extraídas de una población, puede ser comprendida sin excesiva dificultad. Si embargo, suele ser fuente de confusión con la de variabilidad de una variable, basada únicamente en una muestra, aquella de la que disponemos en nuestro estudio. Los conceptos error estándar, medida de la variabilidad en el muestreo de un estadístico y desviación estándar o típica, medida de la variabilidad en una muestra o población concreta de una variable suelen ser confundidos, especialmente en lo referente a la presentación de los resultados obtenidos en una muestra. Así, es frecuente encontrar expresiones en la forma *Media* ± *Desviación típica* , $(\bar{x} \pm s)$, dando a entender que el intervalo comprendido entre la media menos la desviación típica y la media más la desviación típica da lugar a algún resultado concreto, cuando sólo en variables normales se puede afirmar el contenido de ese intervalo (aproximadamente el 68% de las observaciones muestrales), siendo un resultado descriptivo. Sin embargo, expresiones del tipo *Media* ± *Error estándar de la media* , $(\bar{x} \pm s_{\bar{x}})$, son expresiones inferenciales, de las que se puede deducir rápidamente un intervalo de confianza para la media de la variable en la población al disponer del error estándar de la media muestral. Mientras que la primera de las expresiones tiene una lectura esencialmente descriptiva, la segunda permite la realización de inferencias.

No determinación del tamaño de la muestra

La determinación previa del tamaño de la muestra permite prever, cuando el análisis va a requerir de estimaciones por intervalos de confianza o pruebas de hipótesis, las condiciones en las que se desea resolver las aplicaciones. Así, en el caso de los intervalos de confianza, permite establecer la confianza y precisión deseadas para la estimación, mientras que en el caso de pruebas de hipótesis permite definir los errores α y β deseados para la toma de decisiones. Estos aspectos son de gran relevancia puesto que la previsión correcta del tamaño muestral permitirá rentabilizar al máximo los resultados y por tanto el esfuerzo realizado en la investigación. En otro caso, puede suceder que los resultados obtenidos carezcan de la precisión necesaria para aportarlos al entorno científico (caso de los intervalos) o no podamos concluir con suficiente evidencia a favor de la hipótesis nula (caso de las pruebas de hipótesis).
No obstante, la determinación previa del tamaño muestral no está exenta de problemas. A menudo la información necesaria para su cálculo es difícil de establecer o los objetivos del análisis son muy diversos y para cada uno de ellos el tamaño muestral es diferente. Sin embargo su previsión aproximada conduce a una reflexión obligada sobre los métodos estadísticos a utilizar y sobre las posibilidades de realización de inferencias.

No utilización de pruebas de inferencia estadística

Los resultados descriptivos de un estudio son muy escasas veces suficientemente evidentes para obviar la utilización de métodos inferenciales. Hay que tener en cuenta que las conclusiones basadas en la descripción de los resultados muestrales están sometidas a error aleatorio. La no cuantificación de éste puede dar lugar a conclusiones erróneas o contradictorias.

No identificación de las pruebas de inferencia utilizadas

Este es un problema que se observa con cierta frecuencia en la comunicación de los resultados y conclusiones de los estudios. Los autores establecen la conclusión, incluso con valoración estadística, como por ejemplo el valor de p, pero no identifican las pruebas utilizadas para obtenerlas. Esta identificación es de gran importancia para el lector, pues es la única forma de que pueda evaluar los resultados y juzgar la pertinencia de las conclusiones

Utilización incorrecta de las pruebas de inferencia estadística

Un primer aspecto se deriva de la no verificación de las suposiciones/requerimientos que poseen la mayoría de las pruebas de inferencia. Hay que tener en cuenta que estos requerimientos son necesarios para que las conclusiones que se desprendan sean aceptadas con el rigor necesario.
La confusión sobre las conclusiones que pueden desprenderse de la aplicación de una prueba estadística de inferencia puede dar lugar a utilizaciones incorrectas de éstas. Es el caso, por ejemplo de la utilización de algunas pruebas no paramétricas con conclusiones que son propias de otras pruebas paramétricas.
Un problema frecuente es el derivado de la realización de pruebas de inferencia múltiples. Una primera situación que da lugar a que se produzca esta aplicación incorrecta se da en el caso de realizar comparaciones entre múltiples grupos a través de pruebas construidas para la comparación entre dos grupos. Como ejemplo, si deseamos comparar las medias poblacionales de una variable entre 5 poblaciones, intentando concluir si son iguales o alguna o algunas difieren, la solución basada en realizar las

$$\binom{5}{2} = 10$$

comparaciones entre parejas de medias (combinaciones de 5 tomadas de dos en dos), a través de pruebas t puede producir conclusiones erróneas. Para comprenderlo, si se realiza cada prueba t con nivel de significación $\alpha = 0,05$, la posibilidad de actuar correctamente al rechazar la hipótesis nula en cada una de las pruebas será $1 - \alpha = 0,95$, y, por tanto, si aceptamos que cada prueba es independiente de las demás, la probabilidad de actuar correctamente en las 10 pruebas realizadas será $0,95^{10} = 0,599$, por lo que nivel de significación real para el contraste de hipótesis planteado será $1 - 0,599 = 0,401$, es decir, la probabilidad de error al rechazar la hipótesis nula de que todas las medias son iguales será 0,401, valor exageradamente superior al nivel de significación utilizado en cada una de las pruebas t. La solución a este problema pasa por utilizar una prueba única que permita resolver el contraste de hipótesis planteado, como es el caso del análisis de la varianza de una vía para el problema suscitado. Otras soluciones pueden estar basadas en la corrección de los niveles de significación de cada una de las pruebas, disminuyendo su valor pero su aplicación resulta menos satisfactoria que la mencionada. Una segunda situación tiene que ver con la realización de pruebas de inferencia, por ejemplo comparaciones o asociaciones, sobre un gran número de variables, por ejemplo una variable respuesta y 20 variables explicativas para las que se desea estudiar si se asocian o relacionan significativamente con la respuesta. Es lógico pensar que al examinar muchas variables crece la probabilidad de obtener resultados significativos por azar, aunque cada prueba sea independiente de las demás. Desgraciadamente sólo hay una forma de resolver este problema, y es reduciendo el nivel de significación de cada una de las pruebas.

Pruebas de hipótesis unilaterales

En la mayoría de estudios observacionales o experimentales, las pruebas de hipótesis para un nivel de significación especificado, digamos α, deben ser bilaterales o de dos colas, salvo en aquellos contrastes que por su naturaleza de construcción son siempre unilaterales (por ejemplo los de asociación entre variables cualitativas a través del estadístico ji-cuadrado), puesto que el resultado que se pueda producir tras obtener y analizar los datos, puede ir en cualquier dirección. Ello es así con independencia de los deseos o intereses del investigador. Hay que tener en cuenta que es más fácil rechazar la hipótesis nula de un contraste si éste es unilateral, lo cual no es procedente si la naturaleza no garantiza la unilateralidad. Además, la realización del contraste de forma bilateral no significa que no pueda ser direccionado el resultado de acuerdo con la dirección de los datos. Una forma fácil de corregir este problema en la lectura de resultados en los que se incluye el valor de p para un contraste unilateral es multiplicarlo por un factor de dos para obtener la probabilidad de error tipo I como contraste de dos colas.

Interpretación del valor de p

Al realizar un prueba de hipótesis, sirva como ejemplo una comparación entre dos grupos, el valor p representa la probabilidad, a posteriori, de que las diferencias observadas sean atribuibles al muestreo, y observadas si la hipótesis nula fuese cierta y no hubiera diferencias. De alguna forma p refleja la credibilidad atribuida al rechazo de la hipótesis nula, pero de ninguna manera la credibilidad con la que aceptamos que la hipótesis nula es cierta, es decir, que no hay diferencias. Por ejemplo, si p = 0,48, podemos rechazar la

hipótesis nula con una probabilidad 0,48 de cometer error tipo I, es decir de error (por tanto no debemos rechazarla). Sin embargo, p no mide nada más que eso; no es la probabilidad de ningún otro tipo de error. En realidad, el otro tipo de error posible, el de tipo II, el de error al aceptar la hipótesis nula no puede ser medido a través de p, y puede ser incluso muy grande. En definitiva, un valor grande de p no excluye la posibilidad de que se esté cometiendo un error al aceptar la hipótesis nula, y no se puede apoyar la aceptación de la hipótesis nula en el valor de p.
Otra apreciación relacionada con el valor de p es el límite establecido para el rechazo de la hipótesis nula. Hay que tener en cuenta que el valor de α establecido puede depender de diferentes aspectos. Al menos no hay ninguna razón para que siempre se utilice el mismo. Si suponemos una situación concreta en la que p = 0,062, es obvio que para α = 0,05, no podremos rechazar la hipótesis nula. Sin embargo, sólo hay una probabilidad 0,062 de obtener valores tan o más extremos del estadístico de contraste utilizado suponiendo que la hipótesis nula sea cierta. A los ojos de muchos investigadores, la diferencia entre un riesgo de error de 0,05 y otro de 0,062 sería difícilmente apreciable. Si se une a esta reflexión la del valor no estadístico sino científico que pueda tener los resultados obtenidos, quizás pueda ser más razonable el rechazo de la hipótesis nula que su aceptación. En definitiva, resulta conveniente reflexionar sobre la magnitud del nivel de significación en cada situación concreta.

¿Intervalos de confianza o contrastes de hipótesis?

Algunos autores suscitan cierta polémica sobre la mayor o menor conveniencia de utilizar uno u otro procedimiento de inferencia. Lo cierto es que son procedimientos complementarios. Los intervalos de confianza, cuando su construcción es posible, son útiles y pueden complementar los resultados de una prueba de hipótesis por diferentes razones. En primer lugar poseen capacidad para matizar conclusiones obtenidas a través de una prueba de hipótesis. Si se supone una prueba de hipótesis, por ejemplo una comparación de medias, en la que la decisión nos lleva a rechazar la hipótesis nula, concluyendo que las medias de la variable deben ser diferentes, el intervalo de confianza para la diferencia de medias nos ofrecerá la magnitud de ésta, complementando los resultados obtenidos. En muchos estudios, tales como los ensayos clínicos, es necesario conocer la magnitud del efecto, diferencias o asociaciones, no siendo suficiente la comprobación de que ésta existe a través de una prueba de hipótesis. Se trata de reflexionar sobre el valor no estadístico de los hallazgos encontrados. Esto se puede comprender fácilmente si se piensa en la comparación de dos fármacos, uno de ellos de poco coste económico y escasos efectos secundarios, mientras que el otro es de coste elevado y grandes efectos secundarios. La comparación de alguna variable respuesta para los dos fármacos evidencia diferencias significativas. Parece necesario conocer la magnitud de las diferencias para recomendar el segundo fármaco, pues, si éstas son muy pequeñas tal vez no resulte aconsejable su utilización. El intervalo de confianza para las diferencias resulta imprescindible.
Por otra parte, no debe ser olvidado que un intervalo de confianza de nivel de seguridad (1 - α)% contiene todos los valores admisibles para el parámetro estudiado a la seguridad establecida. Para una prueba de hipótesis sobre el parámetro en cuestión, realizada a nivel de significación α de forma bilateral, todos aquellos valores del parámetro incluidos en el intervalo no podrán ser rechazados como valores para la hipótesis nula, mientras que los no incluidos en el intervalo serán rechazados. En definitiva existe una equivalencia entre la construcción de un intervalo de confianza para un parámetro y una prueba de hipótesis para algún valor de ese parámetro, de forma que los valores no incluidos en el intervalo son valores para los que se rechazará la hipótesis nula, con $p \leq \alpha$, siendo la prueba de hipótesis bilateral y el nivel de confianza del intervalo (1 - α)%.

A N E X O

TABLAS DE ALGUNOS MODELOS CONTINUOS DE PROBABILIDAD

Tabla 1: Probabilidad acumulada P(Z≤z) de una normal estándar Z=N(0,1)

z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00	z
-3,90	0,00003	0,00003	0,00004	0,00004	0,00004	0,00004	0,00004	0,00004	0,00005	0,00005	-3,90
-3,80	0,00005	0,00005	0,00005	0,00006	0,00006	0,00006	0,00006	0,00007	0,00007	0,00007	-3,80
-3,70	0,00008	0,00008	0,00008	0,00008	0,00009	0,00009	0,00010	0,0001	0,00010	0,00011	-3,70
-3,60	0,00011	0,00012	0,00012	0,00013	0,00013	0,00014	0,00014	0,00015	0,00015	0,00016	-3,60
-3,50	0,00017	0,00017	0,00018	0,00019	0,00019	0,00020	0,00021	0,00022	0,00022	0,00023	-3,50
-3,40	0,00024	0,00025	0,00026	0,00027	0,00028	0,00029	0,00030	0,00031	0,00032	0,00034	-3,40
-3,30	0,00035	0,00036	0,00038	0,00039	0,00040	0,00042	0,00043	0,00045	0,00047	0,00048	-3,30
-3,20	0,00050	0,00052	0,00054	0,00056	0,00058	0,00060	0,00062	0,00064	0,00066	0,00069	-3,20
-3,10	0,00071	0,00074	0,00076	0,00079	0,00082	0,00084	0,00087	0,00090	0,00094	0,00097	-3,10
-3,00	0,00100	0,00104	0,00107	0,00111	0,00114	0,00118	0,00122	0,00126	0,00131	0,00135	-3,00
-2,90	0,00139	0,00144	0,00149	0,00154	0,00159	0,00164	0,00169	0,00175	0,00181	0,00187	-2,90
-2,80	0,00193	0,00199	0,00205	0,00212	0,00219	0,00226	0,00233	0,0024	0,00248	0,00256	-2,80
-2,70	0,00264	0,00272	0,00280	0,00289	0,00298	0,00307	0,00317	0,00326	0,00336	0,00347	-2,70
-2,60	0,00357	0,00368	0,00379	0,00391	0,00402	0,00415	0,00427	0,00440	0,00453	0,00466	-2,60
-2,50	0,00480	0,00494	0,00508	0,00523	0,00539	0,00554	0,00570	0,00587	0,00604	0,00621	-2,50
-2,40	0,00639	0,00657	0,00676	0,00695	0,00714	0,00734	0,00755	0,00776	0,00798	0,00820	-2,40
-2,30	0,00842	0,00866	0,00889	0,00914	0,00939	0,00964	0,00990	0,01017	0,01044	0,01072	-2,30
-2,20	0,01101	0,01130	0,01160	0,01191	0,01222	0,01255	0,01287	0,01321	0,01355	0,01390	-2,20
-2,10	0,01426	0,01463	0,01500	0,01539	0,01578	0,01618	0,01659	0,01700	0,01743	0,01786	-2,10
-2,00	0,01831	0,01876	0,01923	0,01970	0,02018	0,02068	0,02118	0,02169	0,02222	0,02275	-2,00
-1,90	0,02330	0,02385	0,02442	0,02500	0,02559	0,02619	0,02680	0,02743	0,02807	0,02872	-1,90
-1,80	0,02938	0,03005	0,03074	0,03144	0,03216	0,03288	0,03362	0,03438	0,03515	0,03593	-1,80
-1,70	0,03673	0,03754	0,03836	0,0392	0,04006	0,04093	0,04182	0,04272	0,04363	0,04457	-1,70
-1,60	0,04551	0,04648	0,04746	0,04846	0,04947	0,05050	0,05155	0,05262	0,05370	0,05480	-1,60
-1,50	0,05592	0,05705	0,05821	0,05938	0,06057	0,06178	0,06301	0,06426	0,06552	0,06681	-1,50
-1,40	0,06811	0,06944	0,07078	0,07215	0,07353	0,07493	0,07636	0,07780	0,07927	0,08076	-1,40
-1,30	0,08226	0,08379	0,08534	0,08691	0,08851	0,09012	0,09176	0,09342	0,09510	0,09680	-1,30
-1,20	0,09853	0,10027	0,10204	0,10383	0,10565	0,10749	0,10935	0,11123	0,11314	0,11507	-1,20
-1,10	0,11702	0,11900	0,12100	0,12302	0,12507	0,12714	0,12924	0,13136	0,13350	0,13567	-1,10
-1,00	0,13786	0,14007	0,14231	0,14457	0,14686	0,14917	0,15151	0,15386	0,15625	0,15866	-1,00
-0,90	0,16109	0,16354	0,16602	0,16853	0,17106	0,17361	0,17619	0,17879	0,18141	0,18406	-0,90
-0,80	0,18673	0,18943	0,19215	0,19489	0,19766	0,20045	0,20327	0,20611	0,20897	0,21186	-0,80
-0,70	0,21476	0,21770	0,22065	0,22363	0,22663	0,22965	0,23270	0,23576	0,23885	0,24196	-0,70
-0,60	0,24510	0,24825	0,25143	0,25463	0,25785	0,26109	0,26435	0,26763	0,27093	0,27425	-0,60
-0,50	0,27760	0,28096	0,28434	0,28774	0,29116	0,2946	0,29806	0,30153	0,30503	0,30854	-0,50
-0,40	0,31207	0,31561	0,31918	0,32276	0,32636	0,32997	0,33360	0,33724	0,34090	0,34458	-0,40
-0,30	0,34827	0,35197	0,35569	0,35942	0,36317	0,36693	0,37070	0,37448	0,37828	0,38209	-0,30
-0,20	0,38591	0,38974	0,39358	0,39743	0,40129	0,40517	0,40905	0,41294	0,41683	0,42074	-0,20
-0,10	0,42465	0,42858	0,43251	0,43644	0,44038	0,44433	0,44828	0,45224	0,45620	0,46017	-0,10
0,00	0,46414	0,46812	0,47210	0,47608	0,48006	0,48405	0,48803	0,49202	0,49601	0,50000	0,00
z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00	z

Tabla 1(Cont): Probabilidad acumulada P(Z≤z) de una normal estándar Z=N(0,1)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	z
0,00	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586	0,00
0,10	.0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535	0,10
0,20	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409	0,20
0,30	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173	0,30
0,40	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793	0,40
0,50	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240	0,50
0,60	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490	0,60
0,70	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524	0,70
0,80	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327	0,80
0,90	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891	0,90
1,00	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214	1,00
1,10	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298	1,10
1,20	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147	1,20
1,30	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774	1,30
1,40	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189	1,40
1,50	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408	1,50
1,60	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449	1,60
1,70	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327	1,70
1,80	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062	1,80
1,90	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670	1,90
2,00	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169	2,00
2,10	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574	2,10
2,20	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899	2,20
2,30	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158	2,30
2,40	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361	2,40
2,50	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520	2,50
2,60	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643	2,60
2,70	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736	2,70
2,80	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807	2,80
2,90	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861	2,90
3,00	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900	3,00
3,10	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929	3,10
3,20	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950	3,20
3,30	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965	3,30
3,40	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976	3,40
3,50	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983	3,50
3,60	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989	3,60
3,70	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992	3,70
3,80	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995	3,80
3,90	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997	3,90
z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	z

Tabla 2: Percentiles z_p de la distribución Normal Z=N(0,1) P(Z≤z_p)=p

p	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	p
0,000	∞	-3,090	-2,878	-2,748	-2,652	-2,576	-2,512	-2,457	-2,409	-2,366	0,000
0,010	-2,326	-2,290	-2,257	-2,226	-2,197	-2,170	-2,144	-2,120	-2,097	-2,075	0,010
0,020	-2,054	-2,034	-2,014	-1,995	-1,977	-1,960	-1,943	-1,927	-1,911	-1,896	0,020
0,030	-1,881	-1,866	-1,852	-1,838	-1,825	-1,812	-1,799	-1,787	-1,774	-1,762	0,030
0,040	-1,751	-1,739	-1,728	-1,717	-1,706	-1,695	-1,685	-1,675	-1,665	-1,655	0,040
0,050	-1,645	-1,635	-1,626	-1,616	-1,607	-1,598	-1,589	-1,580	-1,572	-1,563	0,050
0,060	-1,555	-1,546	-1,538	-1,530	-1,522	-1,514	-1,506	-1,499	-1,491	-1,483	0,060
0,070	-1,476	-1,468	-1,461	-1,454	-1,447	-1,440	-1,433	-1,426	-1,419	-1,412	0,070
0,080	-1,405	-1,398	-1,392	-1,385	-1,379	-1,372	-1,366	-1,359	-1,353	-1,347	0,080
0,090	-1,341	-1,335	-1,329	-1,323	-1,317	-1,311	-1,305	-1,299	-1,293	-1,287	0,090
0,100	-1,282	-1,276	-1,270	-1,265	-1,259	-1,254	-1,248	-1,243	-1,237	-1,232	0,100
0,110	-1,227	-1,221	-1,216	-1,211	-1,206	-1,200	-1,195	-1,190	-1,185	-1,180	0,110
0,120	-1,175	-1,170	-1,165	-1,160	-1,155	-1,150	-1,146	-1,141	-1,136	-1,131	0,120
0,130	-1,126	-1,122	-1,117	-1,112	-1,108	-1,103	-1,098	-1,094	-1,089	-1,085	0,130
0,140	-1,080	-1,076	-1,071	-1,067	-1,063	-1,058	-1,054	-1,049	-1,045	-1,041	0,140
0,150	-1,036	-1,032	-1,028	-1,024	-1,019	-1,015	-1,011	-1,007	-1,003	-0,999	0,150
0,160	-0,994	-0,990	-0,986	-0,982	-0,978	-0,974	-0,970	-0,966	-0,962	-0,958	0,160
0,170	-0,954	-0,950	-0,946	-0,942	-0,938	-0,935	-0,931	-0,927	-0,923	-0,919	0,170
0,180	-0,915	-0,912	-0,908	-0,904	-0,900	-0,896	-0,893	-0,889	-0,885	-0,882	0,180
0,190	-0,878	-0,874	-0,871	-0,867	-0,863	-0,860	-0,856	-0,852	-0,849	-0,845	0,190
0,200	-0,842	-0,838	-0,834	-0,831	-0,827	-0,824	-0,820	-0,817	-0,813	-0,810	0,200
0,210	-0,806	-0,803	-0,800	-0,796	-0,793	-0,789	-0,786	-0,782	-0,779	-0,776	0,210
0,220	-0,772	-0,769	-0,765	-0,762	-0,759	-0,755	-0,752	-0,749	-0,745	-0,742	0,220
0,230	-0,739	-0,736	-0,732	-0,729	-0,726	-0,722	-0,719	-0,716	-0,713	-0,710	0,230
0,240	-0,706	-0,703	-0,700	-0,697	-0,693	-0,690	-0,687	-0,684	-0,681	-0,678	0,240
0,250	-0,674	-0,671	-0,668	-0,665	-0,662	-0,659	-0,656	-0,653	-0,650	-0,646	0,250
0,260	-0,643	-0,640	-0,637	-0,634	-0,631	-0,628	-0,625	-0,622	-0,619	-0,616	0,260
0,270	-0,613	-0,610	-0,607	-0,604	-0,601	-0,598	-0,595	-0,592	-0,589	-0,586	0,270
0,280	-0,583	-0,580	-0,577	-0,574	-0,571	-0,568	-0,565	-0,562	-0,559	-0,556	0,280
0,290	-0,553	-0,550	-0,548	-0,545	-0,542	-0,539	-0,536	-0,533	-0,530	-0,527	0,290
0,300	-0,524	-0,522	-0,519	-0,516	-0,513	-0,510	-0,507	-0,504	-0,502	-0,499	0,300
0,310	-0,496	-0,493	-0,490	-0,487	-0,485	-0,482	-0,479	-0,476	-0,473	-0,470	0,310
0,320	-0,468	-0,465	-0,462	-0,459	-0,457	-0,454	-0,451	-0,448	-0,445	-0,443	0,320
0,330	-0,440	-0,437	-0,434	-0,432	-0,429	-0,426	-0,423	-0,421	-0,418	-0,415	0,330
0,340	-0,412	-0,410	-0,407	-0,404	-0,402	-0,399	-0,396	-0,393	-0,391	-0,388	0,340
0,350	-0,385	-0,383	-0,38	-0,377	-0,375	-0,372	-0,369	-0,366	-0,364	-0,361	0,350
0,360	-0,358	-0,356	-0,353	-0,350	-0,348	-0,345	-0,342	-0,340	-0,337	-0,335	0,360
0,370	-0,332	-0,329	-0,327	-0,324	-0,321	-0,319	-0,316	-0,313	-0,311	-0,308	0,370
0,380	-0,305	-0,303	-0,300	-0,298	-0,295	-0,292	-0,290	-0,287	-0,285	-0,282	0,380
0,390	-0,279	-0,277	-0,274	-0,272	-0,269	-0,266	-0,264	-0,261	-0,259	-0,256	0,390
0,400	-0,253	-0,251	-0,248	-0,246	-0,243	-0,240	-0,238	-0,235	-0,233	-0,230	0,400
0,410	-0,228	-0,225	-0,222	-0,220	-0,217	-0,215	-0,212	-0,210	-0,207	-0,204	0,410
0,420	-0,202	-0,199	-0,197	-0,194	-0,192	-0,189	-0,187	-0,184	-0,181	-0,179	0,420
0,430	-0,176	-0,174	-0,171	-0,169	-0,166	-0,164	-0,161	-0,159	-0,156	-0,154	0,430
0,440	-0,151	-0,148	-0,146	-0,143	-0,141	-0,138	-0,136	-0,133	-0,131	-0,128	0,440
0,450	-0,126	-0,123	-0,121	-0,118	-0,116	-0,113	-0,111	-0,108	-0,105	-0,103	0,450
0,460	-0,100	-0,098	-0,095	-0,093	-0,090	-0,088	-0,085	-0,083	-0,080	-0,078	0,460
0,470	-0,075	-0,073	-0,070	-0,068	-0,065	-0,063	-0,060	-0,058	-0,055	-0,053	0,470
0,480	-0,050	-0,048	-0,045	-0,043	-0,040	-0,038	-0,035	-0,033	-0,030	-0,028	0,480
0,490	-0,025	-0,023	-0,02	-0,018	-0,015	-0,013	-0,010	-0,008	-0,005	-0,003	0,490
p	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	p

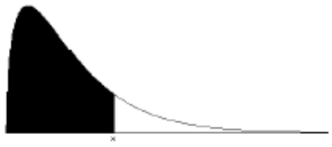
Tabla 2(Cont.): Percentiles z_p de la distribución Normal Z=N(0,1) P(Z≤z_p)=p

p	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	p
0,500	0,000	0,003	0,005	0,008	0,010	0,013	0,015	0,018	0,020	0,023	0,500
0,510	0,025	0,028	0,030	0,033	0,035	0,038	0,040	0,043	0,045	0,048	0,510
0,520	0,050	0,053	0,055	0,058	0,060	0,063	0,065	0,068	0,070	0,073	0,520
0,530	0,075	0,078	0,080	0,083	0,085	0,088	0,090	0,093	0,095	0,098	0,530
0,540	0,100	0,103	0,105	0,108	0,111	0,113	0,116	0,118	0,121	0,123	0,540
0,550	0,126	0,128	0,131	0,133	0,136	0,138	0,141	0,143	0,146	0,148	0,550
0,560	0,151	0,154	0,156	0,159	0,161	0,164	0,166	0,169	0,171	0,174	0,560
0,570	0,176	0,179	0,181	0,184	0,187	0,189	0,192	0,194	0,197	0,199	0,570
0,580	0,202	0,204	0,207	0,210	0,212	0,215	0,217	0,220	0,222	0,225	0,580
0,590	0,228	0,230	0,233	0,235	0,238	0,240	0,243	0,246	0,248	0,251	0,590
0,600	0,253	0,256	0,259	0,261	0,264	0,266	0,269	0,272	0,274	0,277	0,600
0,610	0,279	0,282	0,285	0,287	0,290	0,292	0,295	0,298	0,300	0,303	0,610
0,620	0,305	0,308	0,311	0,313	0,316	0,319	0,321	0,324	0,327	0,329	0,620
0,630	0,332	0,335	0,337	0,340	0,342	0,345	0,348	0,350	0,353	0,356	0,630
0,640	0,358	0,361	0,364	0,366	0,369	0,372	0,375	0,377	0,38	0,383	0,640
0,650	0,385	0,388	0,391	0,393	0,396	0,399	0,402	0,404	0,407	0,410	0,650
0,660	0,412	0,415	0,418	0,421	0,423	0,426	0,429	0,432	0,434	0,437	0,660
0,670	0,440	0,443	0,445	0,448	0,451	0,454	0,457	0,459	0,462	0,465	0,670
0,680	0,468	0,470	0,473	0,476	0,479	0,482	0,485	0,487	0,49	0,493	0,680
0,690	0,496	0,499	0,502	0,504	0,507	0,510	0,513	0,516	0,519	0,522	0,690
0,700	0,524	0,527	0,530	0,533	0,536	0,539	0,542	0,545	0,548	0,550	0,700
0,710	0,553	0,556	0,559	0,562	0,565	0,568	0,571	0,574	0,577	0,580	0,710
0,720	0,583	0,586	0,589	0,592	0,595	0,598	0,601	0,604	0,607	0,610	0,720
0,730	0,613	0,616	0,619	0,622	0,625	0,628	0,631	0,634	0,637	0,640	0,730
0,740	0,643	0,646	0,650	0,653	0,656	0,659	0,662	0,665	0,668	0,671	0,740
0,750	0,674	0,678	0,681	0,684	0,687	0,690	0,693	0,697	0,700	0,703	0,750
0,760	0,706	0,71	0,713	0,716	0,719	0,722	0,726	0,729	0,732	0,736	0,760
0,770	0,739	0,742	0,745	0,749	0,752	0,755	0,759	0,762	0,765	0,769	0,770
0,780	0,772	0,776	0,779	0,782	0,786	0,789	0,793	0,796	0,800	0,803	0,780
0,790	0,806	0,810	0,813	0,817	0,820	0,824	0,827	0,831	0,834	0,838	0,790
0,800	0,842	0,845	0,849	0,852	0,856	0,860	0,863	0,867	0,871	0,874	0,800
0,810	0,878	0,882	0,885	0,889	0,893	0,896	0,900	0,904	0,908	0,912	0,810
0,820	0,915	0,919	0,923	0,927	0,931	0,935	0,938	0,942	0,946	0,950	0,820
0,830	0,954	0,958	0,962	0,966	0,970	0,974	0,978	0,982	0,986	0,990	0,830
0,840	0,994	0,999	1,003	1,007	1,011	1,015	1,019	1,024	1,028	1,032	0,840
0,850	1,036	1,041	1,045	1,049	1,054	1,058	1,063	1,067	1,071	1,076	0,850
0,860	1,080	1,085	1,089	1,094	1,098	1,103	1,108	1,112	1,117	1,122	0,860
0,870	1,126	1,131	1,136	1,141	1,146	1,150	1,155	1,160	1,165	1,170	0,870
0,880	1,175	1,180	1,185	1,190	1,195	1,200	1,206	1,211	1,216	1,221	0,880
0,890	1,227	1,232	1,237	1,243	1,248	1,254	1,259	1,265	1,270	1,276	0,890
0,900	1,282	1,287	1,293	1,299	1,305	1,311	1,317	1,323	1,329	1,335	0,900
0,910	1,341	1,347	1,353	1,359	1,366	1,372	1,379	1,385	1,392	1,398	0,910
0,920	1,405	1,412	1,419	1,426	1,433	1,44	1,447	1,454	1,461	1,468	0,920
0,930	1,476	1,483	1,491	1,499	1,506	1,514	1,522	1,530	1,538	1,546	0,930
0,940	1,555	1,563	1,572	1,580	1,589	1,598	1,607	1,616	1,626	1,635	0,940
0,950	1,645	1,655	1,665	1,675	1,685	1,695	1,706	1,717	1,728	1,739	0,950
0,960	1,751	1,762	1,774	1,787	1,799	1,812	1,825	1,838	1,852	1,866	0,960
0,970	1,881	1,896	1,911	1,927	1,943	1,960	1,977	1,995	2,014	2,034	0,970
0,980	2,054	2,075	2,097	2,120	2,144	2,170	2,197	2,226	2,257	2,290	0,980
0,990	2,326	2,366	2,409	2,457	2,512	2,576	2,652	2,748	2,878	3,090	0,990
p	0,000	0,001	0,002	0,003	0	0,005	0,006	0,007	0,008	0,009	p

Tabla 3: Percentiles de la distribución t de Student $P(t_n \leq t_p)=p$

grados de libertad	$t_{0,005}$	$t_{0,01}$	$t_{0,025}$	$t_{0,05}$	$t_{0,10}$	$t_{0,25}$	$t_{0,50}$	$t_{0,75}$	$t_{0,90}$	$t_{0,95}$	$t_{0,975}$	$t_{0,99}$	$t_{0,995}$
1	-63,657	-31,821	-12,706	-6,314	-3,078	-1,000	0,000	1,000	3,078	6,314	12,706	31,821	63,657
2	-9,925	-6,965	-4,303	-2,920	-1,886	-0,816	0,000	0,816	1,886	2,920	4,303	6,965	9,925
3	-5,841	-4,541	-3,182	-2,353	-1,638	-0,765	0,000	0,765	1,638	2,353	3,182	4,541	5,841
4	-4,604	-3,747	-2,776	-2,132	-1,533	-0,741	0,000	0,741	1,533	2,132	2,776	3,747	4,604
5	-4,032	-3,365	-2,571	-2,015	-1,476	-0,727	0,000	0,727	1,476	2,015	2,571	3,365	4,032
6	-3,707	-3,143	-2,447	-1,943	-1,440	-0,718	0,000	0,718	1,440	1,943	2,447	3,143	3,707
7	-3,499	-2,998	-2,365	-1,895	-1,415	-0,711	0,000	0,711	1,415	1,895	2,365	2,998	3,499
8	-3,355	-2,896	-2,306	-1,860	-1,397	-0,706	0,000	0,706	1,397	1,860	2,306	2,896	3,355
9	-3,250	-2,821	-2,262	-1,833	-1,383	-0,703	0,000	0,703	1,383	1,833	2,262	2,821	3,250
10	-3,169	-2,764	-2,228	-1,812	-1,372	-0,700	0,000	0,700	1,372	1,812	2,228	2,764	3,169
11	-3,106	-2,718	-2,201	-1,796	-1,363	-0,697	0,000	0,697	1,363	1,796	2,201	2,718	3,106
12	-3,055	-2,681	-2,179	-1,782	-1,356	-0,695	0,000	0,695	1,356	1,782	2,179	2,681	3,055
13	-3,012	-2,650	-2,160	-1,771	-1,350	-0,694	0,000	0,694	1,350	1,771	2,160	2,650	3,012
14	-2,977	-2,624	-2,145	-1,761	-1,345	-0,692	0,000	0,692	1,345	1,761	2,145	2,624	2,977
15	-2,947	-2,602	-2,131	-1,753	-1,341	-0,691	0,000	0,691	1,341	1,753	2,131	2,602	2,947
16	-2,921	-2,583	-2,120	-1,746	-1,337	-0,690	0,000	0,690	1,337	1,746	2,120	2,583	2,921
17	-2,898	-2,567	-2,110	-1,740	-1,333	-0,689	0,000	0,689	1,333	1,740	2,110	2,567	2,898
18	-2,878	-2,552	-2,101	-1,734	-1,330	-0,688	0,000	0,688	1,330	1,734	2,101	2,552	2,878
19	-2,861	-2,539	-2,093	-1,729	-1,328	-0,688	0,000	0,688	1,328	1,729	2,093	2,539	2,861
20	-2,845	-2,528	-2,086	-1,725	-1,325	-0,687	0,000	0,687	1,325	1,725	2,086	2,528	2,845
21	-2,831	-2,518	-2,080	-1,721	-1,323	-0,686	0,000	0,686	1,323	1,721	2,080	2,518	2,831
22	-2,819	-2,508	-2,074	-1,717	-1,321	-0,686	0,000	0,686	1,321	1,717	2,074	2,508	2,819
23	-2,807	-2,500	-2,069	-1,714	-1,319	-0,685	0,000	0,685	1,319	1,714	2,069	2,500	2,807
24	-2,797	-2,492	-2,064	-1,711	-1,318	-0,685	0,000	0,685	1,318	1,711	2,064	2,492	2,797
25	-2,787	-2,485	-2,060	-1,708	-1,316	-0,684	0,000	0,684	1,316	1,708	2,060	2,485	2,787
26	-2,779	-2,479	-2,056	-1,706	-1,315	-0,684	0,000	0,684	1,315	1,706	2,056	2,479	2,779
27	-2,771	-2,473	-2,052	-1,703	-1,314	-0,684	0,000	0,684	1,314	1,703	2,052	2,473	2,771
28	-2,763	-2,467	-2,048	-1,701	-1,313	-0,683	0,000	0,683	1,313	1,701	2,048	2,467	2,763
29	-2,756	-2,462	-2,045	-1,699	-1,311	-0,683	0,000	0,683	1,311	1,699	2,045	2,462	2,756
30	-2,750	-2,457	-2,042	-1,697	-1,310	-0,683	0,000	0,683	1,310	1,697	2,042	2,457	2,750
35	-2,724	-2,438	-2,030	-1,690	-1,306	-0,682	0,000	0,682	1,306	1,690	2,030	2,438	2,724
40	-2,704	-2,423	-2,021	-1,684	-1,303	-0,681	0,000	0,681	1,303	1,684	2,021	2,423	2,704
45	-2,690	-2,412	-2,014	-1,679	-1,301	-0,680	0,000	0,680	1,301	1,679	2,014	2,412	2,690
50	-2,678	-2,403	-2,009	-1,676	-1,299	-0,679	0,000	0,679	1,299	1,676	2,009	2,403	2,678
55	-2,668	-2,396	-2,004	-1,673	-1,297	-0,679	0,000	0,679	1,297	1,673	2,004	2,396	2,668
60	-2,660	-2,390	-2,000	-1,671	-1,296	-0,679	0,000	0,679	1,296	1,671	2,000	2,390	2,660
65	-2,654	-2,385	-1,997	-1,669	-1,295	-0,678	0,000	0,678	1,295	1,669	1,997	2,385	2,654
70	-2,648	-2,381	-1,994	-1,667	-1,294	-0,678	0,000	0,678	1,294	1,667	1,994	2,381	2,648
75	-2,643	-2,377	-1,992	-1,665	-1,293	-0,678	0,000	0,678	1,293	1,665	1,992	2,377	2,643
80	-2,639	-2,374	-1,990	-1,664	-1,292	-0,678	0,000	0,678	1,292	1,664	1,990	2,374	2,639
85	-2,635	-2,371	-1,988	-1,663	-1,292	-0,677	0,000	0,677	1,292	1,663	1,988	2,371	2,635
90	-2,632	-2,368	-1,987	-1,662	-1,291	-0,677	0,000	0,677	1,291	1,662	1,987	2,368	2,632
95	-2,629	-2,366	-1,985	-1,661	-1,291	-0,677	0,000	0,677	1,291	1,661	1,985	2,366	2,629
100	-2,626	-2,364	-1,984	-1,660	-1,290	-0,677	0,000	0,677	1,290	1,660	1,984	2,364	2,626
110	-2,621	-2,361	-1,982	-1,659	-1,289	-0,677	0,000	0,677	1,289	1,659	1,982	2,361	2,621
120	-2,617	-2,358	-1,980	-1,658	-1,289	-0,677	0,000	0,677	1,289	1,658	1,980	2,358	2,617
130	-2,614	-2,355	-1,978	-1,657	-1,288	-0,676	0,000	0,676	1,288	1,657	1,978	2,355	2,614
140	-2,611	-2,353	-1,977	-1,656	-1,288	-0,676	0,000	0,676	1,288	1,656	1,977	2,353	2,611
150	-2,609	-2,351	-1,976	-1,655	-1,287	-0,676	0,000	0,676	1,287	1,655	1,976	2,351	2,609
160	-2,607	-2,350	-1,975	-1,654	-1,287	-0,676	0,000	0,676	1,287	1,654	1,975	2,350	2,607
170	-2,605	-2,348	-1,974	-1,654	-1,287	-0,676	0,000	0,676	1,287	1,654	1,974	2,348	2,605
180	-2,603	-2,347	-1,973	-1,653	-1,286	-0,676	0,000	0,676	1,286	1,653	1,973	2,347	2,603
190	-2,602	-2,346	-1,973	-1,653	-1,286	-0,676	0,000	0,676	1,286	1,653	1,973	2,346	2,602
200	-2,601	-2,345	-1,972	-1,653	-1,286	-0,676	0,000	0,676	1,286	1,653	1,972	2,345	2,601
210	-2,599	-2,344	-1,971	-1,652	-1,286	-0,676	0,000	0,676	1,286	1,652	1,971	2,344	2,599
220	-2,598	-2,343	-1,971	-1,652	-1,285	-0,676	0,000	0,676	1,285	1,652	1,971	2,343	2,598
230	-2,597	-2,343	-1,970	-1,652	-1,285	-0,676	0,000	0,676	1,285	1,652	1,970	2,343	2,597
240	-2,596	-2,342	-1,970	-1,651	-1,285	-0,676	0,000	0,676	1,285	1,651	1,970	2,342	2,596
250	-2,596	-2,341	-1,969	-1,651	-1,285	-0,675	0,000	0,675	1,285	1,651	1,969	2,341	2,596
∞	-2,576	-2,326	-1,960	-1,645	-1,282	-0,674	0,000	0,674	1,282	1,645	1,960	2,326	2,576

Tabla 4: Percentiles de la distribución Ji-cuadrada P($\chi_{n2} < \chi_p$)=p



grados de libertad	$\chi_{0,005}$	$\chi_{0,01}$	$\chi_{0,025}$	$\chi_{0,05}$	$\chi_{0,10}$	$\chi_{0,25}$	$\chi_{0,50}$	$\chi_{0,75}$	$\chi_{0,90}$	$\chi_{0,95}$	$\chi_{0,975}$	$\chi_{0,99}$	$\chi_{0,995}$
1	0,000	0,000	0,001	0,004	0,016	0,102	0,455	1,323	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,773	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	1,213	2,366	4,108	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	1,923	3,357	5,385	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	2,675	4,351	6,626	9,236	11,070	12,833	15,086	16,750
6	0,676	0,872	1,237	1,635	2,204	3,455	5,348	7,841	10,645	12,592	14,449	16,812	18,548
7	0,989	1,239	1,690	2,167	2,833	4,255	6,346	9,037	12,017	14,067	16,013	18,475	20,278
8	1,344	1,646	2,180	2,733	3,490	5,071	7,344	10,219	13,362	15,507	17,535	20,090	21,955
9	1,735	2,088	2,700	3,325	4,168	5,899	8,343	11,389	14,684	16,919	19,023	21,666	23,589
10	2,156	2,558	3,247	3,940	4,865	6,737	9,342	12,549	15,987	18,307	20,483	23,209	25,188
11	2,603	3,053	3,816	4,575	5,578	7,584	10,341	13,701	17,275	19,675	21,92	24,725	26,757
12	3,074	3,571	4,404	5,226	6,304	8,438	11,340	14,845	18,549	21,026	23,337	26,217	28,300
13	3,565	4,107	5,009	5,892	7,042	9,299	12,340	15,984	19,812	22,362	24,736	27,688	29,819
14	4,075	4,660	5,629	6,571	7,790	10,165	13,339	17,117	21,064	23,685	26,119	29,141	31,319
15	4,601	5,229	6,262	7,261	8,547	11,037	14,339	18,245	22,307	24,996	27,488	30,578	32,801
16	5,142	5,812	6,908	7,962	9,312	11,912	15,338	19,369	23,542	26,296	28,845	32,000	34,267
17	5,697	6,408	7,564	8,672	10,085	12,792	16,338	20,489	24,769	27,587	30,191	33,409	35,718
18	6,265	7,015	8,231	9,390	10,865	13,675	17,338	21,605	25,989	28,869	31,526	34,805	37,156
19	6,844	7,633	8,907	10,117	11,651	14,562	18,338	22,718	27,204	30,144	32,852	36,191	38,582
20	7,434	8,260	9,591	10,851	12,443	15,452	19,337	23,828	28,412	31,410	34,170	37,566	39,997
21	8,034	8,897	10,283	11,591	13,240	16,344	20,337	24,935	29,615	32,671	35,479	38,932	41,401
22	8,643	9,542	10,982	12,338	14,041	17,240	21,337	26,039	30,813	33,924	36,781	40,289	42,796
23	9,260	10,196	11,689	13,091	14,848	18,137	22,337	27,141	32,007	35,172	38,076	41,638	44,181
24	9,886	10,856	12,401	13,848	15,659	19,037	23,337	28,241	33,196	36,415	39,364	42,980	45,559
25	10,520	11,524	13,120	14,611	16,473	19,939	24,337	29,339	34,382	37,652	40,646	44,314	46,928
26	11,160	12,198	13,844	15,379	17,292	20,843	25,336	30,435	35,563	38,885	41,923	45,642	48,290
27	11,808	12,879	14,573	16,151	18,114	21,749	26,336	31,528	36,741	40,113	43,195	46,963	49,645
28	12,461	13,565	15,308	16,928	18,939	22,657	27,336	32,620	37,916	41,337	44,461	48,278	50,993
29	13,121	14,256	16,047	17,708	19,768	23,567	28,336	33,711	39,087	42,557	45,722	49,588	52,336
30	13,787	14,953	16,791	18,493	20,599	24,478	29,336	34,800	40,256	43,773	46,979	50,892	53,672
35	17,192	18,509	20,569	22,465	24,797	29,054	34,336	40,223	46,059	49,802	53,203	57,342	60,275
40	20,707	22,164	24,433	26,509	29,051	33,660	39,335	45,616	51,805	55,758	59,342	63,691	66,766
45	24,311	25,901	28,366	30,612	33,350	38,291	44,335	50,985	57,505	61,656	65,410	69,957	73,166
50	27,991	29,707	32,357	34,764	37,689	42,942	49,335	56,334	63,167	67,505	71,420	76,154	79,490
55	31,735	33,570	36,398	38,958	42,060	47,610	54,335	61,665	68,796	73,311	77,380	82,292	85,749
60	35,534	37,485	40,482	43,188	46,459	52,294	59,335	66,981	74,397	79,082	83,298	88,379	91,952
65	39,383	41,444	44,603	47,450	50,883	56,990	64,335	72,285	79,973	84,821	89,177	94,422	98,105
70	43,275	45,442	48,758	51,739	55,329	61,698	69,334	77,577	85,527	90,531	95,023	100,425	104,215
75	47,206	49,475	52,942	56,054	59,795	66,417	74,334	82,858	91,061	96,217	100,839	106,393	110,286
80	51,172	53,540	57,153	60,391	64,278	71,145	79,334	88,130	96,578	101,879	106,629	112,329	116,321
85	55,170	57,634	61,389	64,749	68,777	75,881	84,334	93,394	102,079	107,522	112,393	118,236	122,325
90	59,196	61,754	65,647	69,126	73,291	80,625	89,334	98,650	107,565	113,145	118,136	124,116	128,299
95	63,250	65,898	69,925	73,520	77,818	85,376	94,334	103,899	113,038	118,752	123,858	129,973	134,247
100	67,328	70,065	74,222	77,929	82,358	90,133	99,334	109,141	118,498	124,342	129,561	135,807	140,169
110	75,550	78,458	82,867	86,792	91,471	99,666	109,334	119,608	129,385	135,480	140,917	147,414	151,948
120	83,852	86,923	91,573	95,705	100,624	109,220	119,334	130,055	140,233	146,567	152,211	158,950	163,648
130	92,222	95,451	100,331	104,662	109,811	118,792	129,334	140,482	151,045	157,610	163,453	170,423	175,278
140	100,655	104,034	109,137	113,659	119,029	128,38	139,334	150,894	161,827	168,613	174,648	181,840	186,847
150	109,142	112,668	117,985	122,692	128,275	137,983	149,334	161,291	172,581	179,581	185,800	193,208	198,360
160	117,679	121,346	126,870	131,756	137,546	147,599	159,334	171,675	183,311	190,516	196,915	204,530	209,824
170	126,261	130,064	135,790	140,849	146,839	157,227	169,334	182,047	194,017	201,423	207,995	215,812	221,242
180	134,884	138,820	144,741	149,969	156,153	166,865	179,334	192,409	204,704	212,304	219,044	227,056	232,620
190	143,545	147,610	153,721	159,113	165,485	176,514	189,334	202,76	215,371	223,160	230,064	238,266	243,959
200	152,241	156,432	162,728	168,279	174,835	186,172	199,334	213,102	226,021	233,994	241,058	249,445	255,264
210	160,969	165,283	171,759	177,465	184,201	195,838	209,334	223,436	236,655	244,808	252,027	260,595	266,537
220	169,727	174,160	180,813	186,671	193,582	205,512	219,334	233,762	247,274	255,602	262,973	271,717	277,779
230	178,512	183,063	189,889	195,895	202,978	215,194	229,334	244,080	257,879	266,378	273,898	282,814	288,994
240	187,324	191,990	198,984	205,135	212,386	224,882	239,334	254,392	268,471	277,138	284,802	293,888	300,182
250	196,161	200,939	208,098	214,392	221,806	234,577	249,334	264,697	279,050	287,881	295,689	304,940	311,340

Tabla 5: Percentiles al 95% de una F de Snedecor-Fisher con n y m grados de libertad (g.l.) $P(F_{n,m}<F)=0.95$

	g.l. numerador (n)														
g.l. denomi-nador (m)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	242,98	243,91	244,69	245,36	245,95
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,4	19,4	19,41	19,42	19,42	19,43
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,71	8,70
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,87	5,86
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,6	3,57	3,55	3,53	3,51
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,24	3,22
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	3,03	3,01
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,86	2,85
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,55	2,53
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,48	2,46
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,42	2,40
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,37	2,35
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,22	2,20
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,17	2,15
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,15	2,13
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,11	2,09
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,09	2,07
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	2,08	2,06
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	2,05	2,03
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	2,04	2,01
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,99	1,95	1,92	1,89	1,87
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,89	1,86	1,84
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97	1,93	1,89	1,86	1,84	1,81
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95	1,91	1,88	1,84	1,82	1,79
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94	1,90	1,86	1,83	1,80	1,78
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,89	1,85	1,82	1,79	1,77
110	3,93	3,08	2,69	2,45	2,30	2,18	2,09	2,02	1,97	1,92	1,88	1,84	1,81	1,78	1,76
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87	1,83	1,80	1,78	1,75
130	3,91	3,07	2,67	2,44	2,28	2,17	2,08	2,01	1,95	1,90	1,86	1,83	1,80	1,77	1,74
140	3,91	3,06	2,67	2,44	2,28	2,16	2,08	2,01	1,95	1,90	1,86	1,82	1,79	1,76	1,74
150	3,90	3,06	2,66	2,43	2,27	2,16	2,07	2,00	1,94	1,89	1,85	1,82	1,79	1,76	1,73
∞	3,87	3,03	2,63	2,40	2,24	2,13	2,04	1,97	1,91	1,86	1,82	1,78	1,75	1,72	1,70

Tabla 5 (Cont): Percentiles al 95% de una F de Snedecor-Fisher con n y m grados de libertad (g.l.) $P(F_{n,m}<F)=0.95$

	g.l. numerador (n)												
g.l. denominador (m)	16	17	18	19	20	25	30	40	60	80	100	120	∞
1	246,46	246,92	247,32	247,69	248,01	249,26	250,1	251,14	252,2	252,72	253,04	253,25	253,89
2	19,43	19,44	19,44	19,44	19,45	19,46	19,46	19,47	19,48	19,48	19,49	19,49	19,49
3	8,69	8,68	8,67	8,67	8,66	8,63	8,62	8,59	8,57	8,56	8,55	8,55	8,54
4	5,84	5,83	5,82	5,81	5,80	5,77	5,75	5,72	5,69	5,67	5,66	5,66	5,64
5	4,60	4,59	4,58	4,57	4,56	4,52	4,50	4,46	4,43	4,41	4,41	4,40	4,38
6	3,92	3,91	3,9	3,88	3,87	3,83	3,81	3,77	3,74	3,72	3,71	3,70	3,68
7	3,49	3,48	3,47	3,46	3,44	3,40	3,38	3,34	3,30	3,29	3,27	3,27	3,24
8	3,20	3,19	3,17	3,16	3,15	3,11	3,08	3,04	3,01	2,99	2,97	2,97	2,94
9	2,99	2,97	2,96	2,95	2,94	2,89	2,86	2,83	2,79	2,77	2,76	2,75	2,72
10	2,83	2,81	2,80	2,79	2,77	2,73	2,70	2,66	2,62	2,60	2,59	2,58	2,55
11	2,70	2,69	2,67	2,66	2,65	2,60	2,57	2,53	2,49	2,47	2,46	2,45	2,42
12	2,60	2,58	2,57	2,56	2,54	2,50	2,47	2,43	2,38	2,36	2,35	2,34	2,31
13	2,51	2,50	2,48	2,47	2,46	2,41	2,38	2,34	2,30	2,27	2,26	2,25	2,23
14	2,44	2,43	2,41	2,40	2,39	2,34	2,31	2,27	2,22	2,20	2,19	2,18	2,15
15	2,38	2,37	2,35	2,34	2,33	2,28	2,25	2,20	2,16	2,14	2,12	2,11	2,09
16	2,33	2,32	2,3	2,29	2,28	2,23	2,19	2,15	2,11	2,08	2,07	2,06	2,03
17	2,29	2,27	2,26	2,24	2,23	2,18	2,15	2,10	2,06	2,03	2,02	2,01	1,98
18	2,25	2,23	2,22	2,20	2,19	2,14	2,11	2,06	2,02	1,99	1,98	1,97	1,94
19	2,21	2,20	2,18	2,17	2,16	2,11	2,07	2,03	1,98	1,96	1,94	1,93	1,90
20	2,18	2,17	2,15	2,14	2,12	2,07	2,04	1,99	1,95	1,92	1,91	1,90	1,86
21	2,16	2,14	2,12	2,11	2,10	2,05	2,01	1,96	1,92	1,89	1,88	1,87	1,83
22	2,13	2,11	2,10	2,08	2,07	2,02	1,98	1,94	1,89	1,86	1,85	1,84	1,81
23	2,11	2,09	2,08	2,06	2,05	2,00	1,96	1,91	1,86	1,84	1,82	1,81	1,78
24	2,09	2,07	2,05	2,04	2,03	1,97	1,94	1,89	1,84	1,82	1,80	1,79	1,76
25	2,07	2,05	2,04	2,02	2,01	1,96	1,92	1,87	1,82	1,80	1,78	1,77	1,73
26	2,05	2,03	2,02	2,00	1,99	1,94	1,90	1,85	1,80	1,78	1,76	1,75	1,71
27	2,04	2,02	2,00	1,99	1,97	1,92	1,88	1,84	1,79	1,76	1,74	1,73	1,70
28	2,02	2,00	1,99	1,97	1,96	1,91	1,87	1,82	1,77	1,74	1,73	1,71	1,68
29	2,01	1,99	1,97	1,96	1,94	1,89	1,85	1,81	1,75	1,73	1,71	1,70	1,66
30	1,99	1,98	1,96	1,95	1,93	1,88	1,84	1,79	1,74	1,71	1,70	1,68	1,65
40	1,90	1,89	1,87	1,85	1,84	1,78	1,74	1,69	1,64	1,61	1,59	1,58	1,54
50	1,85	1,83	1,81	1,8	1,78	1,73	1,69	1,63	1,58	1,54	1,52	1,51	1,47
60	1,82	1,80	1,78	1,76	1,75	1,69	1,65	1,59	1,53	1,50	1,48	1,47	1,42
70	1,79	1,77	1,75	1,74	1,72	1,66	1,62	1,57	1,50	1,47	1,45	1,44	1,39
80	1,77	1,75	1,73	1,72	1,70	1,64	1,60	1,54	1,48	1,45	1,43	1,41	1,36
90	1,76	1,74	1,72	1,70	1,69	1,63	1,59	1,53	1,46	1,43	1,41	1,39	1,34
100	1,75	1,73	1,71	1,69	1,68	1,62	1,57	1,52	1,45	1,41	1,39	1,38	1,32
110	1,74	1,72	1,7	1,68	1,67	1,61	1,56	1,50	1,44	1,40	1,38	1,36	1,31
120	1,73	1,71	1,69	1,67	1,66	1,60	1,55	1,50	1,43	1,39	1,37	1,35	1,30
130	1,72	1,70	1,68	1,67	1,65	1,59	1,55	1,49	1,42	1,38	1,36	1,34	1,29
140	1,72	1,70	1,68	1,66	1,65	1,58	1,54	1,48	1,41	1,38	1,35	1,33	1,28
150	1,71	1,69	1,67	1,66	1,64	1,58	1,54	1,48	1,41	1,37	1,34	1,33	1,27
∞	1,68	1,66	1,64	1,62	1,61	1,54	1,50	1,43	1,36	1,32	1,30	1,28	1,21

BIBLIOGRAFÍA

A continuación se incluyen algunos textos en castellano orientados hacia la aplicación de la estadística en Ciencias de la Salud que recorren los temas desarrollados en este texto. Se ha pretendido seleccionar textos completos, que aborden exhaustivamente los temas abordados en este material. Incluyen definiciones y numerosas aplicaciones con ejemplos del entorno de las Ciencias de la Salud que pueden ser útiles para complementar la adquisición de conocimientos y el trabajo del alumno.

- Daniel WW. Bioestadística. Base para el análisis de las Ciencias de las Ciencias de la Salud. México: Limusa. 2002
- Martín A, Luna JD. Bioestadística para las Ciencias de la Salud. Madrid: Capitel Ediciones. 2004
- Dawson B, Trapp R. Bioestadística médica. México: Manual moderno. 2005
- Macchi RL. Introducción a la Estadística en Ciencias de la Salud. Editorial médica Panamericana. 2013
- Moncho J. Estadística aplicada a las Ciencias de la Salud. Barcelona: Elsevier. 2015