

Cuadernos Gestión Sanitaria

Número 7

**Fundamentos del uso de
herramientas estadísticas
aplicadas a la gestión sanitaria**



7º Cuaderno de Gestión Sanitaria

FUNDAMENTOS DEL USO DE HERRAMIENTAS ESTADÍSTICAS APLICADAS A LA GESTIÓN SANITARIA

Alicia Coduras

FUNDACIÓN GASPAR CASAL

ISBN: 978-84-09-13189-1

D. L.: M-24598-2019

Maquetación, impresión y encuadernación:

Impresión A2

Crta. M-506 – Km. 25,800 – Vía de Servicio

28320 - Pinto (Madrid)

ÍNDICE

Prólogo	5
1. Bioestadística: estadística aplicada a las ciencias de la vida.....	7
1.1. Definición de estadística y bioestadística	7
1.2. Principales funciones de la estadística	8
1.3. ¿Qué utilidad tiene estudiar bioestadística?	10
2. Conceptos básicos de conocimiento imprescindible para el estudio de la Bioestadística.....	13
2.1. Estadística descriptiva e inferencial.....	13
2.2. Población y muestra	14
2.3. Variables y su clasificación	15
3. Fundamentos de la estadística descriptiva	19
3.1. Introducción y propuesta de ejemplo aplicado.....	19
3.2. Selección de técnicas estadísticas descriptivas	20
3.3. Resultados: tablas de frecuencias simples	22
3.4. Resultados: gráficos simples unidimensionales.....	28
3.5. Resultados: medidas de tendencia central y dispersión	29
3.5.1. La moda	30
3.5.2. La mediana	31
3.5.3. La media y las medidas de dispersión	32
3.5.4. Las medidas de forma	35
4. Fundamentos de la estadística inferencial	39
4.1. Introducción.....	39
4.2. Población y muestra.....	39
4.3. Muestreo estadístico	40
4.3.1. Métodos de muestreo	40
4.3.2. El tamaño de la muestra para estimar una proporción o porcentaje	41
4.3.3. El tamaño de la muestra para estimar una media.....	46
5. Tablas bivariables o de contingencia.....	51
5.1. Definición, construcción de tablas e interpretación de los resultados mediante ejemplos aplicados.....	51
5.1.1. Ejemplo 1 y concepto de variable dependiente	51
5.2. Contraste de hipótesis sobre diferencias de proporciones	56
5.2.1. Concepto de hipótesis estadística.....	56
5.2.2. Formulación y contraste de hipótesis para tablas de contingencia	57
5.2.3. Ejemplos complementarios de contraste de hipótesis dentro del mismo caso de estudio.....	61

6. Contraste de diferencia de medias o análisis de la varianza	65
6.1. Conceptualización	65
6.2. Ejemplos e interpretación de resultados	65
7. Correlación y regresión lineal simple	73
7.1. Conceptualización	73
7.1.1. Correlación lineal.....	74
7.1.2. Ejemplos de análisis de correlación lineal	76
7.1.3. Regresión lineal simple.....	77
7.1.4. Ejemplo complementario sobre regresión lineal.....	81
8. Regresión múltiple	83
8.1. Introducción.....	83
8.2. Formulación de un modelo de regresión lineal múltiple.....	83
8.2.1. Principales hipótesis y supuestos que deben cumplir las variables que intervienen en la estimación de un modelo de regresión múltiple.....	84
8.2.2. Estimación de los parámetros del modelo.....	85
8.2.3. Bondad del ajuste y predicción	95
8.2.4. Conocimiento básico de los problemas del modelo de regresión múltiple	97
9. Regresión logística: fundamentos.....	103
9.1. Introducción: utilidad de la regresión logística.....	103
9.2. La regresión logística binaria a través de un ejemplo aplicado	103
9.3. Ejemplo complementario.....	109
10. Análisis de supervivencia	113
10.1. Concepto general.....	113
10.2. Elementos básicos para la realización de un análisis de supervivencia	114
10.3. Aplicación práctica del análisis de supervivencia simple	115
10.4. Aplicación práctica del análisis de supervivencia comparando dos curvas.....	121
Bibliografía complementaria recomendada.....	126

PRÓLOGO

La estadística es una disciplina que se ha revelado como de uso imprescindible para el colectivo de profesionales que desarrollan su trabajo en el ámbito sociosanitario. Así, es un hecho que la gestión de cualquier sistema de salud se basa en la disposición de información en forma de datos elaborados acerca de la población a atender, así como su evolución. Pero, además de esto, hay muchos otros aspectos en que interviene, como son la gestión de recursos, la investigación, la evaluación de resultados, la de tecnologías y procesos sanitarios y muchos otros. La epidemiología, por su parte, es una disciplina específica dentro de las que forman el cuerpo de la estadística y que estudia la distribución, frecuencia y factores determinantes de las enfermedades existentes en poblaciones humanas definidas.

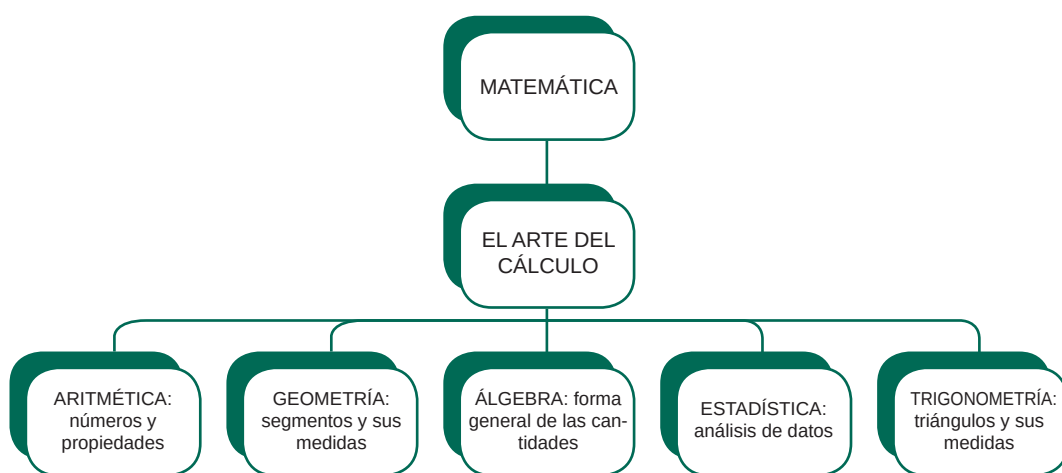
Este cuaderno proporciona un resumen en forma de apuntes básicos sobre las principales técnicas estadísticas y de análisis epidemiológico que cualquier profesional del ámbito de la gestión sociosanitaria puede necesitar en el desarrollo de su profesión. Salpicado de ejemplos que se acercan a realidades de diversos ámbitos de gestión y especialidades clínicas, esperamos que sea de utilidad práctica y que sirva de guía y fuente de consulta al afrontar diversas situaciones de gestión y de realización de estudios e informes por parte de los profesionales sociosanitarios.

1. Bioestadística: estadística aplicada a las ciencias de la vida

1.1. Definición de estadística y bioestadística

La Estadística es una rama de las Matemáticas (ver Figura 1). Consta de un conjunto de métodos, normas, reglas y principios que sirven para observar, agrupar, describir, cuantificar y analizar el comportamiento de un grupo o grupos de sujetos de cualquier tipo, acerca de los cuales se dispone de información.

Figura 1. La estadística y su lugar en las ciencias matemáticas



Fuente: elaboración de los autores

La Estadística está considerada actualmente como el vehículo que permite llevar a cabo cualquier proceso relacionado con la recopilación, descripción y análisis de datos, así como cualquier investigación científica. Es una disciplina transversal aplicable en múltiples ámbitos, desde la física a las ciencias sociales, desde las ciencias de la salud a la psicología. Sus cualidades la convierten en una herramienta idónea para la toma de decisiones, el control de calidad, la evaluación y la gestión de todo tipo de procesos y en diversas áreas.

La presencia de las matemáticas en el desarrollo científico de cualquier disciplina es esencial y, la estadística, como epígrafe de estas tiene, a su vez, sus propias ramificaciones (ver Figura 2). Así, esta materia se puede dividir en estadística paramétrica y no paramétrica. La estadística paramétrica, a su vez, se divide en estadística descriptiva e inferencial. La estadística descriptiva aglutina diversos apartados que van desde la confección de listados y tablas de frecuencias simples y de contingencia a la de gráficos de muy diversos tipos. La estadística inferencial aglutina la estimación de parámetros en poblaciones concretas, así como los contrastes de hipótesis sobre tablas de contingencia, cálculo de correlaciones, análisis de la varianza, regresiones de diversos tipos, realización de pronósticos y predicciones y muchos otros temas, todo ello sustentado especialmente por las técnicas de reco-

gida y organización de datos y las de muestreo. Asimismo, la estadística, es una disciplina que permite abordar temas de control de calidad, toma de decisiones, diseño experimental, diseño y gestión de proyectos, minería de datos, análisis demográfico, análisis actuarial y de supervivencia y otros temas que dan una idea de su gran amplitud e importancia.

Figura 2. Ramificación de la estadística



Fuente: elaboración de los autores

La estadística se aplica a múltiples ámbitos, tales como la economía, la sociología, la psicología y otros. Aunque la disciplina y sus técnicas constituyen un solo cuerpo, desde hace años, se ha establecido la costumbre de identificar la materia según el ámbito de aplicación, lo cual ha dado lugar a la aparición de nombres específicos como: econometría, sociometría, psicometría y otros, que se adaptan a cada área. Por consiguiente, la bioestadística es la aplicación de las técnicas estadísticas al ámbito de las ciencias de la vida, entre las que se encuentran la biología y la medicina.

1.2. Principales funciones de la estadística

En esencia, la estadística, estudia los métodos y procedimientos para: recoger datos; clasificarlos; resumirlos; hallar regularidades y modelos de comportamiento y, en definitiva, elaborar información de interés, que sea útil para diversos agentes de los cuales depende una gestión, un proyecto, un estudio, cierta toma de decisiones, un control de calidad, la elaboración de informes, la realización de una investigación y temas similares (ver Figura 3). Las funciones de la estadística abarcan desde simples descripciones hasta análisis complejos de datos, especialmente cuando la variabilidad y la incertidumbre sean causa in-

trínseca de los mismos o, en otras palabras, cuando éstos proceden de la extracción de muestras de poblaciones acerca de las cuales hay que realizar inferencias. Esta es la parte cuya comprensión resulta más difícil para las personas que estudian esta disciplina, ya que, cuesta asimilar que a partir de un subconjunto de información extraído de una población sea posible estimar con un margen de error muy pequeño y con una alta confianza, el verdadero valor de determinadas características de dicha población (ver Figura 4). Este tipo de metodología estadística es el que se entiende por inferencial y es uno de los más representativos de la utilidad de esta disciplina en cuanto a investigación, control de calidad, eficiencia de procesos y tecnologías, así como en la provisión de información para el diseño de políticas públicas y toma de decisiones.

Figura 3. Funciones básicas de la estadística



Fuente: elaboración de los autores

En el ámbito sanitario, los procesos inferenciales son especialmente relevantes en la realización de estudios epidemiológicos, en el estudio de cohortes, en la evaluación de efectividad de tratamientos y muchos otros. Como se verá en los siguientes apartados, el uso de la inferencia conlleva la elección concreta de técnicas de análisis según tengan que aplicarse al estudio de una variable respuesta en una población, en dos o más poblaciones o al estudio de relaciones entre variables.

Figura 4. El proceso de inferencia estadística



Fuente: elaboración de los autores

1.3. ¿Qué utilidad tiene estudiar bioestadística?

Aunque las funciones de la estadística se han descrito en los apartados anteriores, es importante que el profesional del ámbito de la salud identifique los tipos de situaciones en que tener conocimiento de esta materia le va a resultar de ayuda. Durante la realización de una carrera de medicina, enfermería, farmacia, psicología o cualquier otra especialidad relacionada con las ciencias de la salud, todos los estudiantes reciben instrucción en estadística. No obstante, además de parecer una materia ardua que no agrada a una elevada proporción de los estudiantes es difícil que éstos vean la dimensión de su utilidad y aplicabilidad hasta que no finalizan los estudios y comienzan a desarrollar su profesión en un entorno real.

Entre las diversas situaciones que enfrentan los profesionales del ámbito de la salud, comencemos por mencionar aquella en la cual tienen que trabajar con hallazgos, procedimientos y conceptos que se basan en análisis estadísticos que previamente han realizado otros.

Un primer ejemplo corresponde a la lectura e interpretación de un análisis de sangre. Este tipo de lectura, así como la de cualquier informe clínico, requiere de una madurez estadística suficiente para poder evaluarlo de forma adecuada. El análisis habrá sido realizado por otros y los valores de corte de los distintos indicadores que lo integren también habrán sido interpretados por otros dándoles una evaluación concreta sobre si son bajos, medios o altos. No obstante, el profesional que lleva cada caso es el que, bajo su conocimiento específico de cada situación, termina por valorar el cuadro que se le presenta para tomar las decisiones que correspondan.

Un segundo ejemplo muy común se refiere a los casos en que un profesional del ámbito de la salud lee, consulta, o toma referencias sobre artículos científicos o informes redactados por otros. En estos casos, hay que tener presente que la estadística es la disciplina que abre la puerta a la literatura científica. Todas las disciplinas en las que se realizan estudios cuantitativos han de justificar sus hallazgos en términos estadísticos y la validez de sus afirmaciones depende del juicio que la estadística les otorgue. Por ello la literatura científica, y la relacionada con las ciencias de la salud, está plagada de conceptos y términos estadísticos que hay que conocer para poder trabajar adecuadamente con esta documentación y, sin duda alguna, para producir nueva literatura.

Un tercer caso es considerar que la estadística supone una herramienta para el análisis de situaciones con componente aleatoria en que hay que tomar decisiones fundamentadas en información objetiva. Tal es el caso, por ejemplo, en que hay que abordar campañas de información, contención o atención extraordinaria a la población cuando se dan incrementos de demanda inesperados ocasionados por algún motivo en un territorio en concreto.

Un caso muy común y en continua expansión es el que se refiere al diseño de estudios de corte experimental, o de evolución de cohortes, efectividad de tratamientos o de cualquier

otro tipo y que suele implicar la toma de muestras representativas de poblaciones concretas. En este tipo de situación no se puede llevar a cabo ninguna fase del proyecto a menos que se sepa estadística. Todos los pasos, desde la identificación del tema de estudio o investigación hasta la recogida de datos y su análisis requieren de un conocimiento estadístico asentado y consciente: la extracción de conclusiones en temas de salud es delicada y puede tener un tipo de consecuencias que no se dan en otros ámbitos como el económico o el histórico.

Otros casos que evidencian la utilidad de saber estadística en el ámbito de las ciencias de la salud son los que se refieren a gestión de listas de espera, evaluación de tecnologías sanitarias, evaluación de políticas públicas, funcionamiento de servicios, análisis de efectividad de procesos, elaboración de información necesaria para el desarrollo de acciones dentro del marco de una triple meta y otros.

Una dificultad a la que se enfrentan los docentes en esta materia es la de poder ofrecer ejemplos que se ajusten a todas las especialidades de los profesionales del ámbito de la salud. Desde el punto de vista de los estudiantes, sería ideal el poder estudiar sobre casos que guarden relación con lo que hacen en su puesto de trabajo habitualmente. Se trata de un problema al que ambos colectivos, docentes por un lado y estudiantes por el otro, deben enfrentarse con flexibilidad. Así, por un lado, muchos docentes que escriben manuales estadísticos no proceden del entorno sanitario y les resulta muy difícil crear ejemplos afines a esta materia y, por otro lado, muchos estudiantes no se encuentran cómodos estudiando sobre textos cuyos ejemplos se basan en situaciones más propias de las ciencias sociales que de la salud, por lo que no terminan de asimilar la materia. Actualmente, ya existen muchos textos adaptados y en los cuales se explica la materia sobre ejemplos relacionados. No obstante, el estudiante debe comprender la dificultad que estriba el crear un texto para especialidades médicas concretas como es el caso de la pediatría, la hematología, la oncología, la endocrinología, la traumatología y muchas más.

Los docentes de la Fundación Gaspar Casal han realizado este esfuerzo de forma progresiva, habiendo cubierto en torno a 7 especialidades, pero es un camino lento y que se aborda sobre demanda, de forma que sea útil a un colectivo suficiente de estudiantes que provengan de una especialidad concreta. Aun así, los ejemplos son sencillos ya que deben adaptarse al propósito final que es enseñar las técnicas básicas de la bioestadística y su funcionamiento desde un punto de vista práctico y fácilmente comprensible.

2. Conceptos básicos de conocimiento imprescindible para el estudio de la Bioestadística

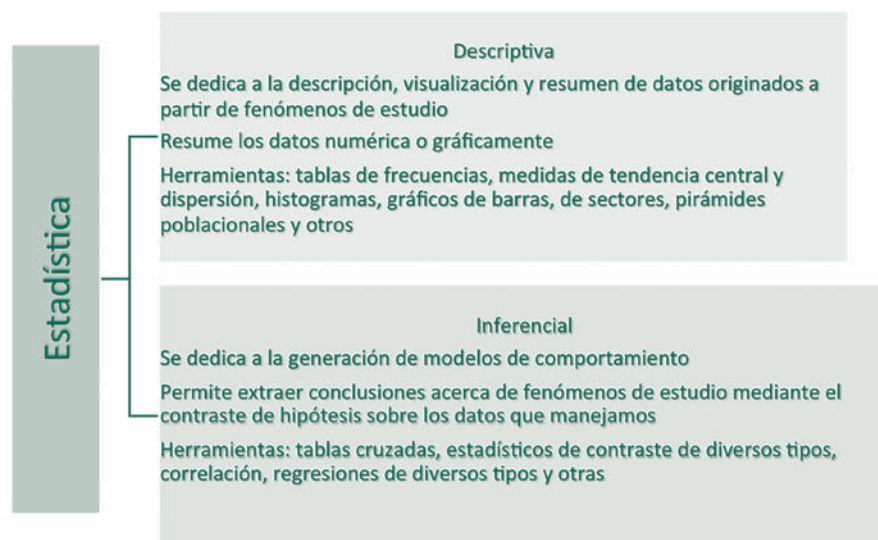
2.1. Estadística descriptiva e inferencial

La estadística se divide en dos grandes ramas: la descriptiva y la inferencial. Ambas cumplen diferentes cometidos y tienen sus propias características y herramientas (ver Figura 5).

La estadística descriptiva es más simple en cuanto a comprensión y requerimientos matemáticos, a la vez que sus resultados e interpretación son más inmediatos. Su misión es la de resumir gran cantidad de datos originados por el estudio de un fenómeno en forma de tablas y gráficos sintéticos y fáciles de interpretar. Sin embargo, la simplicidad no le resta importancia en absoluto, ya que esta sección proporciona la base de la mayor parte de informes que se confeccionan en cualquier ámbito, a la par que la información resumida acerca de las características de cualquier población o fenómeno en estudio. Así, un buen analista de resultados estadísticos descriptivos puede resumir cualquier situación de forma clara, visual y eficiente utilizando sus herramientas adecuadamente. Estas herramientas se estudian en detalle en próximos apartados y comprenden: listados, tablas de frecuencias simples y cruzadas (técnicamente denominadas de contingencia), medidas o indicadores de tendencia central, dispersión y forma y, finalmente, gráficos de muy diversos tipos como son los de sectores, barras, histogramas, pirámides, radiales y otros.

La estadística inferencial se basa en la extracción de una muestra representativa de datos de una población para, aplicando alguna técnica estadística de esta rama, proporcionar estimaciones fiables de indicadores (técnicamente denominados parámetros) de la población objetivo a partir de los que arroja la muestra (técnicamente denominados estimadores). Al igual que la descriptiva, la inferencial es una rama de la estadística en la que se trabaja con un conjunto de herramientas. Entre las más habituales disponemos de: las técnicas de muestreo, la formulación de hipótesis sobre parámetros poblacionales y los estadísticos de contraste que se pueden aplicar sobre tablas de contingencia, medias de más de una población, coeficientes de correlación, coeficientes de regresión y otros casos.

Figura 5. Estadística descriptiva y estadística inferencial



Fuente: elaboración de los autores

2.2. Población y muestra

Un estudio en que se apliquen técnicas estadísticas puede referirse a una población, también llamada universo, que es el conjunto total de elementos que son objeto de nuestro interés o estudio y acerca de los cuáles intentaremos extraer descripciones y/o conclusiones.

Los indicadores que se pueden calcular numéricamente sobre una población se denominan parámetros y suelen representarse mediante letras o caracteres griegos. Si la descripción de la población es inabordable por su gran tamaño, se dice que los parámetros poblacionales son desconocidos.

Dado que, a menudo es imposible, poco práctico o excesivamente costoso observar la totalidad de los individuos de una población, en lugar de examinar a toda la población, con frecuencia se examina una pequeña parte del grupo llamada muestra. Una muestra estadística es un subconjunto de la población que nos interesa estudiar.

Los indicadores numéricos que se obtienen a partir de los datos de una muestra se denominan estimadores o estimaciones muestrales y se representan habitualmente con letras o caracteres arábigos para distinguirlos de los parámetros poblacionales. Usualmente, las estimaciones muestrales son aproximaciones a los parámetros poblacionales. Si la muestra es suficientemente representativa, los estimadores también lo serán y se podrán utilizar para hacer inferencias sobre la población a la que representan.

Desde un punto de vista aplicado, en ciencias de la salud se aplican técnicas estadísticas sobre numerosos tipos de ámbitos: poblaciones y/o muestras los residentes en países, re-

giones, ciudades, áreas sanitarias o cualquier unidad territorial; poblaciones y/o muestras de pacientes de patologías concretas; recursos sanitarios tales como centros sanitarios y sus elementos; poblaciones /o muestras de historias clínicas; usuarios de atención primaria, especializada o sociosanitaria; información económica relacionada con procesos sanitarios tales como utilización de recursos, inversiones, costes, gastos, tiempos (de espera, estancia y otros), presupuestos, campañas, proyectos y similares. Las posibilidades son tan numerosas como los temas de estudio y análisis que se puedan proponer.

La definición del ámbito de cualquier estudio sobre el que se pueden aplicar técnicas estadísticas es uno de los puntos más importantes que hay que tener en cuenta al abordar su diseño ya que de él se desprende, entre otras cosas, si va a constituir un estudio únicamente descriptivo o si va a tener, además, alguna componente inferencial.

2.3. Variables y su clasificación

Los datos que se utilizan en cualquier estudio en que intervienen técnicas estadísticas se disponen en forma de variables. Una variable estadística es cada una de las características o cualidades que poseen los individuos u objetos que forman parte de una población. El sexo, la edad, el peso, la altura, el lugar de nacimiento, la temperatura corporal, la presión arterial y muchas otras, todas son características o, en términos estadísticos, variables que podemos observar en cualquier individuo.

Leyendo la lista anterior de ejemplos, enseguida se puede ver que hay distintas clases de variables, por lo que es imprescindible dominar su clasificación (ver Figura 6), punto clave para comprender la utilidad y forma de aplicar las diferentes técnicas estadísticas de las que vamos a hablar.

Las variables se clasifican en dos grandes grupos como son el de variables cualitativas o categóricas y el de variables cuantitativas o escala. La distinción entre ambos grupos es clara: las variables cualitativas o categóricas tienen valores que son códigos y con los cuales no es posible realizar operaciones matemáticas, aunque dichos códigos sean números, mientras que las variables cuantitativas o escala se expresan en números que responden a algún tipo de unidad de medida y con los cuales se pueden realizar todo tipo de operaciones matemáticas.

A su vez, el grupo de variables cualitativas se subdivide en variables nominales y ordinales. En las primeras, los códigos representan una mera convención para identificar categorías. Por ejemplo, la variable sexo, es una variable cualitativa nominal a la cual, por ejemplo, se puede asignar el código 1 para los casos en que los individuos son hombres y el 0 para los casos en que los individuos son mujeres. Asimismo, también se podrían haber escogido los códigos 1 y 2 para esas categorías, o cualquier otro. Ello no implica en ningún caso que el caso de un hombre valga el doble que el de una mujer. Los números sólo son códigos y no se pueden realizar operaciones con ellos. En cambio, en el caso de las variables ordinales, los códigos tienen un significado que les

otorga algún tipo de orden. Por ejemplo, si una variable recoge el grado de dolor sobre un conjunto de pacientes con una patología concreta, y los códigos que se utilizan son: 1 = muy leve, 2 = leve, 3 = intenso y 4 = muy intenso, esta variable es una de tipo ordinal, en la cual el valor 1 es el más bajo y el 4 el más alto. Estos números no se pueden utilizar para realizar operaciones matemáticas, es decir, no se sabe cuál es la distancia exacta entre una categoría y la siguiente, y el cálculo de una media no se podría tomar como una estimación fiable. En cambio, como veremos, con este tipo de variables sí que se puede averiguar la proporción de pacientes del grupo analizado que tienen un menor grado de dolor frente a los que tienen un grado intenso o muy intenso.

En definitiva, las variables cualitativas son susceptibles de los análisis descriptivos más simples, especialmente, las de tipo nominal. Entre los ejemplos más clásicos de variables nominales tenemos el sexo, el lugar de nacimiento, el padecer una enfermedad o no hacerlo, el nombre de patologías, el nombre de centros sanitarios, de servicios, el lugar de atención (primaria, especializada u otras), el tipo de ingreso (urgente, programado u otro), y muchos otros. Entre los ejemplos más clásicos de variables ordinales tenemos la categoría profesional, el grupo de edad, grados de dolor, satisfacción, gravedad y muchos otros.

Entre las variables ordinales, las que son medidas en escalas de Likert de 5, 7 o 9 puntos, constituyen un caso especial en el cual, a efectos de análisis, se las suele considerar como cuantitativas dentro del rango de la escala. Teóricamente son ordinales, pero se admite su uso en técnicas estadísticas tales como el cálculo de medias, desviaciones estándar, coeficientes de correlación, regresiones y otras que requieren la realización de operaciones matemáticas.

Las variables cuantitativas se subdividen a su vez en discretas y continuas. Todas ellas se expresan en unidades, pero en las primeras, una unidad es indivisible o, en otras palabras, se expresan siempre en números enteros, sin decimales. Son variables discretas el número de pacientes con una patología determinada, el número de camas de un hospital, el número de centros de atención primaria de un área sanitaria, el número de médicos de un país, el número de reingresos en un trimestre y muchas otras.

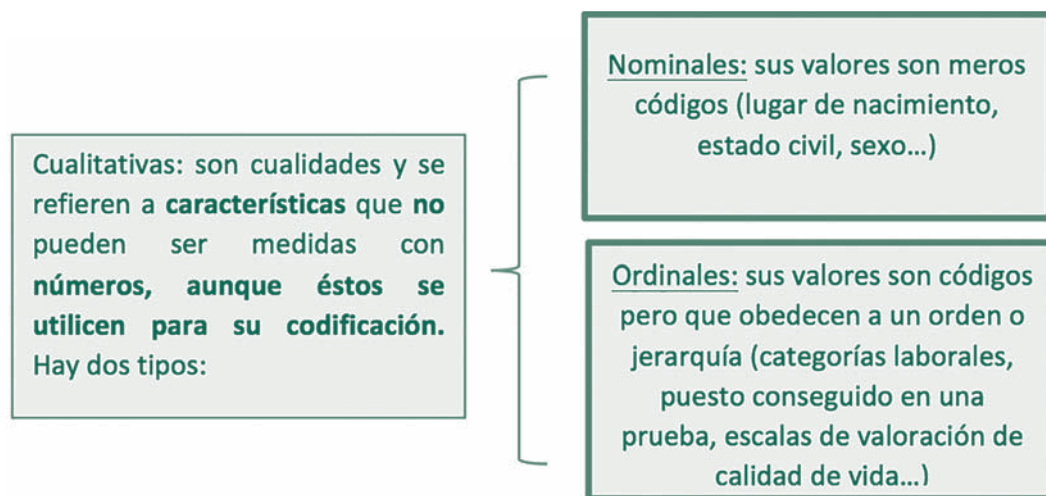
En cambio, las variables continuas se expresan como valores fijos dentro de un intervalo determinado. En estas variables, entre dos valores observables siempre va a existir un tercer valor intermedio que también podría tomar la variable continua. Una variable es continua porque toma valores a lo largo de un continuo, esto es, en todo un intervalo de valores. Por eso, una característica esencial de una variable continua es que, a diferencia de una variable discreta, nunca puede ser medida con exactitud; el valor observado depende en gran medida de la precisión de los instrumentos de medición. Con una variable continua hay inevitablemente un error de medida. Entre los ejemplos más típicos podemos mencionar, la estatura de una persona (1,72m, 1,719m, 1,7186m), el tiempo que toma un equipo quirúrgico en llevar a cabo una operación determinada (45 minutos y 15 segundos, 43 minutos y 40 segundos, 52 minutos y 17 segundos), el gasto por paciente protegido en cada Comunidad Autónoma (1543,67€,

1763,29€, 1833,21€), y así sucesivamente. La presión arterial, la mayoría de los indicadores que aparecen en análisis de sangre y de otros tipos, y muchas otras variables son cuantitativas continuas. Especial mención merecen variables como la edad, las que se expresan en unidades monetarias y otras que suelen ofrecerse sin decimales para facilitar su manejo y comprensión. Para distinguir correctamente entre variables discretas y continuas hay que centrarse en el concepto que miden y distinguir que se trata de unidades, al menos en teoría, indivisibles, tales como personas, camas, centros y similares.

Finalmente, un caso que merece especial mención entre las variables cuantitativas en general, ya sean discretas o continuas es el de las series temporales. Se trata, en realidad, de dos variables que van unidas: la medición de un concepto discreto o continuo, y un momento del tiempo asociado a la medición. En sanidad, un ejemplo muy típico es la serie temporal que se construye diariamente al tomar, por ejemplo, la temperatura o la presión arterial de un paciente en vigilancia a intervalos regulares de tiempo a lo largo de uno o más días. Así, si se toma la temperatura cada hora, tendríamos una serie temporal de 24 observaciones cada día, acerca de la cual podríamos observar su evolución, calcular los incrementos y disminuciones sufridos, la temperatura media de un día, y así sucesivamente. También son series temporales las tasas de todo tipo (mortalidad, morbilidad y muchas otras) que se ofrecen con carácter anual, semestral, trimestral o con otras periodificaciones.

Saber clasificar las variables en estas cinco categorías es imprescindible para poder escoger las técnicas estadísticas que les son aplicables.

Figura 6. Sistema de clasificación de variables



Cuantitativas: son cantidades, se expresan mediante **números y unidades** y **son** susceptibles de todas las operaciones matemáticas

Discretas: toman valores enteros, sin decimales (número de hijos, número de consultorios, número de médicos...)

Continuas: toman valores con decimales (altura, peso, presión arterial...)

Fuente: elaboración de los autores

3. Fundamentos de la estadística descriptiva

3.1. Introducción y propuesta de ejemplo aplicado

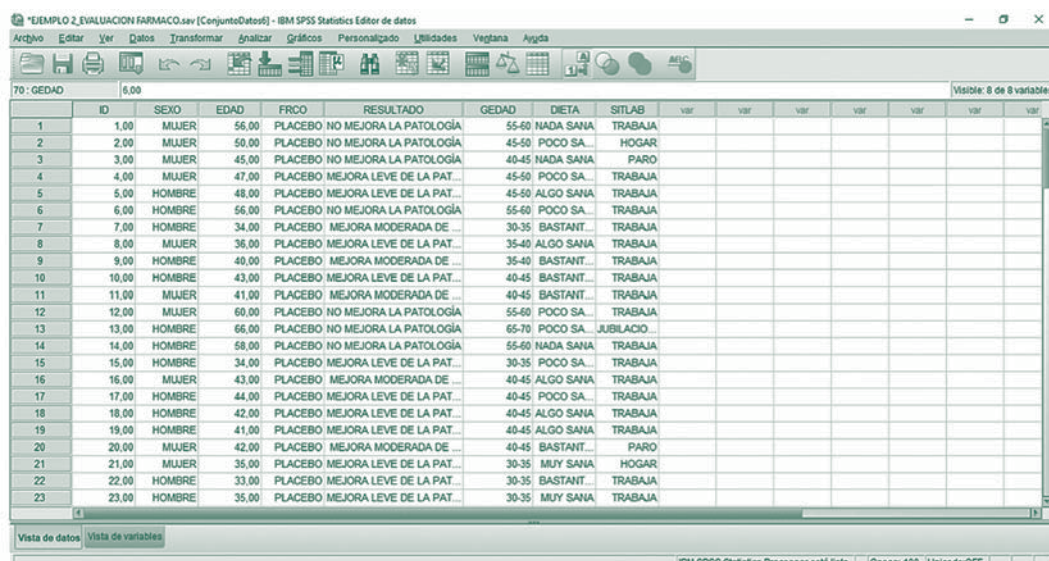
La Estadística Descriptiva proporciona un amplio conjunto de técnicas y herramientas que nos permiten estructurar información para elaborar informes, diagnósticos de situaciones, tomar decisiones, asignar recursos y otras tareas propias de cualquier área de salud y especialidad.

A lo largo de esta sección, vamos a presentar y conocer sus posibilidades y herramientas de forma dinámica y mediante ejercicios sencillos relacionados con un ejemplo clásico: la prueba de un fármaco nuevo para paliar una patología derivada de tratamientos quimioterápicos.

Imaginemos que, para llevar a cabo esta evaluación, se ha diseñado un estudio con pacientes tratados con el fármaco en evaluación y pacientes tratados con un placebo. Los datos recogidos por el equipo que realiza el estudio son dispuestos en una base de datos. Lo más recomendable cuando se está realizando un estudio que requiera llevar a cabo explotaciones estadísticas es trabajar con un paquete o software estadístico. El más popular y extendido es el denominado SPSS (Statistical Package for Social Sciences, ofrecido bajo la marca IBM en estos momentos).

En la Figura 7 se muestra la imagen de la base de datos en formato SPSS que contiene la información básica recopilada para el desarrollo de este ejemplo (los datos son ficticios).

Figura 7. Base de datos SPSS sobre un conjunto de pacientes en tratamiento quimioterápico divididos en un grupo de tratamiento y otro de control para la prueba de un nuevo fármaco



ID	SEXO	EDAD	FRCO	RESULTADO	GEDAD	DIETA	SITLAB
1	MUJER	56.00	PLACEBO	NO MEJORA LA PATOLOGÍA	55-60	NADA SANA	TRABAJA
2	MUJER	50.00	PLACEBO	NO MEJORA LA PATOLOGÍA	45-50	POCO SA...	HOGAR
3	MUJER	45.00	PLACEBO	NO MEJORA LA PATOLOGÍA	40-45	NADA SANA	PARO
4	MUJER	47.00	PLACEBO	MEJORA LEVE DE LA PAT...	45-50	POCO SA...	TRABAJA
5	HOMBRE	48.00	PLACEBO	MEJORA LEVE DE LA PAT...	45-50	ALGO SANA	TRABAJA
6	HOMBRE	56.00	PLACEBO	NO MEJORA LA PATOLOGÍA	55-60	POCO SA...	TRABAJA
7	HOMBRE	34.00	PLACEBO	MEJORA MODERADA DE ...	30-35	BASTANT...	TRABAJA
8	MUJER	36.00	PLACEBO	MEJORA LEVE DE LA PAT...	35-40	ALGO SANA	TRABAJA
9	HOMBRE	40.00	PLACEBO	MEJORA MODERADA DE ...	35-40	BASTANT...	TRABAJA
10	HOMBRE	43.00	PLACEBO	MEJORA LEVE DE LA PAT...	40-45	BASTANT...	TRABAJA
11	MUJER	41.00	PLACEBO	MEJORA MODERADA DE ...	40-45	BASTANT...	TRABAJA
12	MUJER	60.00	PLACEBO	NO MEJORA LA PATOLOGÍA	55-60	POCO SA...	TRABAJA
13	HOMBRE	66.00	PLACEBO	NO MEJORA LA PATOLOGÍA	65-70	POCO SA...	JUBILACIO...
14	HOMBRE	58.00	PLACEBO	NO MEJORA LA PATOLOGÍA	55-60	NADA SANA	TRABAJA
15	HOMBRE	34.00	PLACEBO	MEJORA LEVE DE LA PAT...	30-35	POCO SA...	TRABAJA
16	MUJER	43.00	PLACEBO	MEJORA MODERADA DE ...	40-45	ALGO SANA	TRABAJA
17	HOMBRE	44.00	PLACEBO	MEJORA LEVE DE LA PAT...	40-45	POCO SA...	TRABAJA
18	HOMBRE	42.00	PLACEBO	MEJORA LEVE DE LA PAT...	40-45	ALGO SANA	TRABAJA
19	HOMBRE	41.00	PLACEBO	MEJORA LEVE DE LA PAT...	40-45	ALGO SANA	TRABAJA
20	MUJER	42.00	PLACEBO	MEJORA MODERADA DE ...	40-45	BASTANT...	PARO
21	MUJER	35.00	PLACEBO	MEJORA LEVE DE LA PAT...	30-35	MUY SANA	HOGAR
22	HOMBRE	33.00	PLACEBO	MEJORA LEVE DE LA PAT...	30-35	BASTANT...	TRABAJA
23	HOMBRE	35.00	PLACEBO	MEJORA LEVE DE LA PAT...	30-35	MUY SANA	TRABAJA

Fuente: elaboración de los autores

Cada columna de la base de datos recoge los valores obtenidos acerca de las variables medidas en la muestra de pacientes. Cada fila de la base de datos representa a un paciente. Las variables que contiene la base de datos son:

ID = IDENTIFICADOR DE PACIENTE

SEXO = SEXO DEL PACIENTE

EDAD = EDAD DEL PACIENTE

FRCO = TIPO DE FÁRMACO ADMINISTRADO

RESULTADO = TIPO DE RESULTADO OBTENIDO

GEDAD = GRUPO DE EDAD

DIETA = TIPO DE DIETA HABITUAL QUE VIENE CONSUMIENDO EL PACIENTE

SITLAB = SITUACIÓN LABORAL DEL PACIENTE

Por otro lado, los valores que pueden tomar estas variables son:

ID = un número del 1 al 100, pues hay 100 pacientes realizando la prueba

SEXO = 1 (Hombre) 0 (Mujer)

EDAD = edad en años

FRCO = 1 (fármaco en pruebas) 0 (placebo)

RESULTADO = 0 (la patología no mejora), 1 (Mejora leve de la patología), 2 (Mejora moderada de la patología), 3 (Mejora sustancial de la patología), 4 (Curación de la patología)

GEDAD = 1 (30-35) 2 (35-40) 3 (40-45) 4 (45-50) 5 (50-55) 6 (55-60) 7 (60-65) 8 (66-70)

DIETA = 1 (NADA SANA) 2 (POCO SANA) 3 (ALGO SANA) 4 (BASTANTE SANA) 5 (MUY SANA)

SITLAB = 1 (TRABAJA) 2 (PARO) 3 (HOGAR) 4 (JUBILACIÓN-RETIRO)

3.2. Selección de técnicas estadísticas descriptivas

Como se ha comentado anteriormente, la estadística descriptiva ofrece un conjunto de técnicas ideales para resumir y presentar la información recogida sobre cualquier conjunto de datos. Las técnicas más comunes son:

- Tablas de frecuencias simples o unidimensionales
- Tablas de frecuencias cruzadas, de contingencia o bidimensionales
- Medidas de tendencia central, dispersión y forma: moda, mediana, media, rango, percentiles, desviación estándar, coeficiente de asimetría y coeficiente de curtosis
- Gráficos simples o unidimensionales (de sectores, barras e histogramas)

Las tablas de frecuencias simples se pueden aplicar a todos tipo de variables, pero en el caso de las variables cuantitativas, especialmente las continuas, no son de mucha utilidad a menos que se agrupen los valores en intervalos. Este caso es muy frecuente, por ejemplo, cuando trabajamos con edades, estaturas, pesos y similares. Así, en estas ocasiones,

es mejor clasificar a los individuos por tramos, por ejemplo, de entre 18 y 25 años, 25 y 35 años, 35 y 45 años y así sucesivamente, o por tramos de 1,5 metros a 1,6 metros, de 1,6 a 1,7 metros y así sucesivamente. Nótese que los intervalos, al representar un continuo tienen el extremo superior repetido al inicio del siguiente intervalo. El analista debe explicar al interpretar los resultados si el extremo superior estaba incluido en el intervalo o si se considera que ese valor es el inicial del intervalo siguiente (criterio más habitual).

Las tablas de contingencia o bidimensionales se aplican para relacionar dos variables cualitativas, ya sean nominales u ordinales y no se recomiendan para trabajar con variables cuantitativas a menos que se trate de casos muy especiales y con pocos valores en la distribución.

En cuanto a las medidas de tendencia central, la moda se puede aplicar a todo tipo de variables y es la única medida de tendencia central que se puede calcular en el caso de las cualitativas nominales. La mediana se puede aplicar a variables cualitativas ordinales y a ambos tipos de variables cuantitativas. La media y las medidas de dispersión y forma sólo se pueden aplicar a variables cuantitativas.

En definitiva, la selección de las técnicas para llevar a cabo la explotación estadística de los datos disponibles depende de dos cuestiones:

a) El tipo de preguntas que deseamos responder acerca de los datos disponibles

b) El tipo de variables con la que vayamos a trabajar en cada pregunta a responder

Por consiguiente, antes de comenzar, es importante determinar las preguntas a las que deseamos dar respuesta, así como clasificar las variables y saber qué técnicas estadísticas es posible aplicar a cada tipo de variable.

Supongamos que, entre las muchas que se pueden plantear, hemos seleccionado esta batería de preguntas para responder:

P1. ¿Cómo se distribuyó el sexo de los pacientes en la muestra? O, en otras palabras, ¿cuántos hombres y mujeres participaron en la experiencia?

P2. ¿Cómo se distribuyó la edad de los pacientes en la muestra? O, en otras palabras, ¿cuántos pacientes de cada edad del rango contemplado hubo en la muestra?

P3. ¿Cómo se distribuyó el fármaco y el placebo entre la muestra? O, en otras palabras, ¿cuántos pacientes recibieron el fármaco y cuántos el placebo?

P4. ¿Cómo se distribuyó el resultado de la experiencia? O, en otras palabras, ¿qué porcentaje hubo de éxitos totales, parciales, leves y de fracasos en la muestra?

Una vez planteadas las preguntas, clasificamos las variables disponibles (téngase presente que un analista con experiencia realiza esta operación mentalmente y sabe qué técnicas

son aplicables a cada tipo de variable gracias a su dilatado conocimiento práctico).

ID = IDENTIFICADOR DE PACIENTE = VARIABLE CUALITATIVA NOMINAL
SEXO = SEXO DEL PACIENTE = VARIABLE CUALITATIVA NOMINAL
EDAD = EDAD DEL PACIENTE = VARIABLE CUANTITATIVA CONTINUA
FRCO = TIPO DE FÁRMACO ADMINISTRADO = VARIABLE CUALITATIVA NOMINAL
RESULTADO = TIPO DE RESULTADO OBTENIDO = VARIABLE CUALITATIVA ORDINAL
GEDAD = GRUPO DE EDAD = VARIABLE CUALITATIVA ORDINAL
DIETA = TIPO DE DIETA HABITUAL = VARIABLE CUALITATIVA ORDINAL
SITLAB = SITUACIÓN LABORAL DEL PACIENTE = VARIABLE CUALITATIVA NOMINAL

Una vez clasificadas las variables, constatamos qué técnicas de la estadística descriptiva antes mencionadas se pueden aplicar a cada tipo de variable:

ID = IDENTIFICADOR DE PACIENTE: sólo listado, porque no tiene papel en la investigación
SEXO = SEXO DEL PACIENTE: tablas y gráficos más la moda
EDAD = EDAD DEL PACIENTE: todas
FRCO = TIPO DE FÁRMACO ADMINISTRADO: tablas y gráficos más la moda
RESULTADO = TIPO DE RESULTADO OBTENIDO: tablas y gráficos más la moda y la mediana
GEDAD = GRUPO DE EDAD: tablas y gráficos más la moda y la mediana
DIETA = TIPO DE DIETA HABITUAL: tablas y gráficos más la moda y la mediana
SITLAB = SITUACIÓN LABORAL DEL PACIENTE: tablas y gráficos más la moda

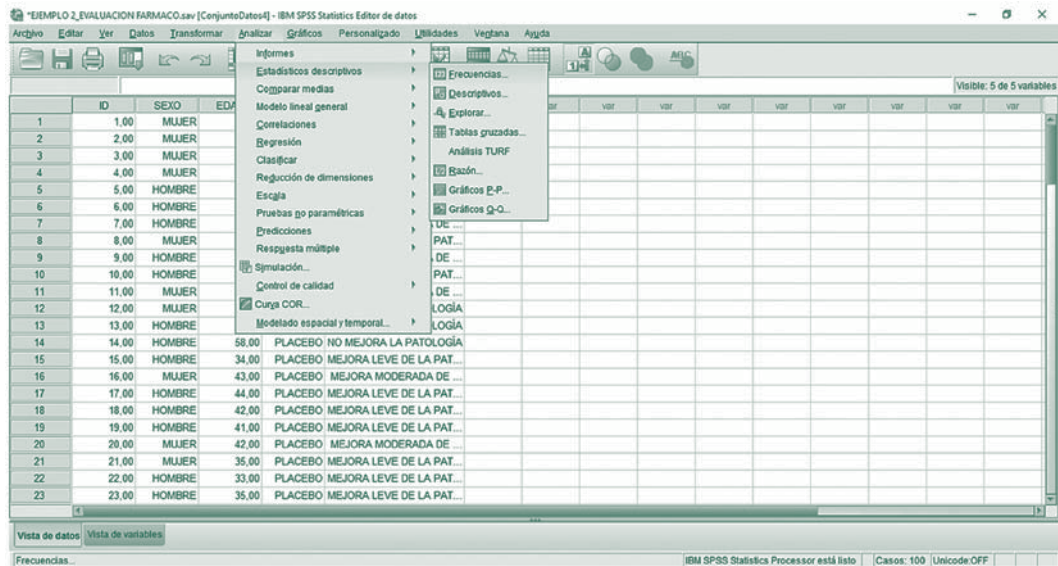
3.3. Resultados: tablas de frecuencias simples

Las tablas de frecuencias simples son herramientas de la estadística en las cuales se colocan los datos en columnas representando los distintos valores de cada variable recogidos en la muestra y las frecuencias (las veces) en que ocurre cada valor. Dado que en estadística cobra más valor trabajar con porcentajes que con frecuencias absolutas, en las tablas también aparecen y se calculan dividiendo cada frecuencia absoluta por el tamaño de la muestra y multiplicando por cien:

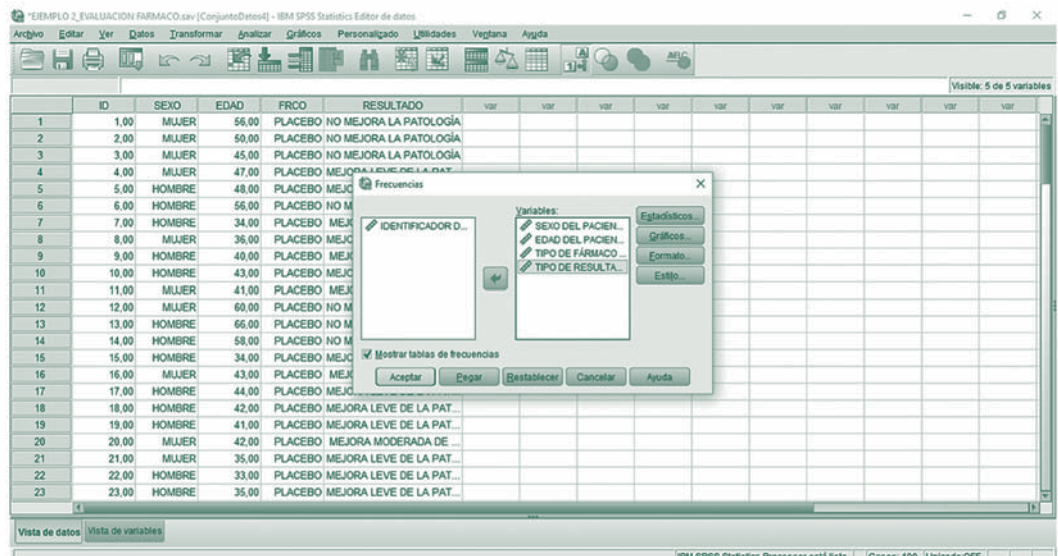
$$p_i = \frac{n_i}{n} \times 100$$

Donde p_i representa cada porcentaje, n_i cada frecuencia absoluta y n el tamaño de la muestra.

Actualmente, las tablas de frecuencias se calculan mediante programas estadísticos y el investigador debe aprender a obtenerlas e interpretarlas. Para calcular tablas de frecuencias con SPSS, basta con seguir esta secuencia de menús: Analizar, estadísticos descriptivos, frecuencias



En el cuadro de diálogo que aparece, se colocan las variables de interés a la derecha y se deja la variable ID, pues no tiene relevancia.



Se acepta en el icono de aceptar y se obtiene, en primer lugar, una tabla resumen del procesamiento de los casos que nos indica que, en todas las variables solicitadas hubo 100 casos con información y 0 casos “perdidos” o de no respuesta:

Estadísticos

		SEXO DEL PACIENTE	EDAD DEL PACIENTE	TIPO DE FÁRMACO ADMINISTRADO	TIPO DE RESULTADO
N	Válido	100	100	100	100
	Perdidos	0	0	0	0

En segundo lugar, una tabla de frecuencias simple del sexo, que nos informa de que la muestra se componía de 46 mujeres y 54 hombres, es decir, al ser la muestra de 100 casos, se tuvo un 46% de mujeres y un 54% de hombres. Gracias a esta tabla, se responde a la primera pregunta formulada, es decir: ¿Cómo se distribuyó el sexo de los pacientes en la muestra?

SEXO DEL PACIENTE

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	MUJER	46	46,0	46,0	46,0
	HOMBRE	54	54,0	54,0	100,0
	Total	100	100,0	100,0	

En esta y otras tablas similares, el porcentaje válido coincide con el porcentaje porque no ha habido casos de no respuesta. De haberlos habido, el porcentaje válido se habría recalculado sobre el número de casos válidos en lugar de sobre 100. El porcentaje acumulado, se calcula sumando los porcentajes válidos sobre el siguiente hasta alcanzar el 100% de los casos.

La tabla de frecuencias de la edad no es muy informativa porque queda demasiado desagregada al ser una variable cuantitativa continua. En variables de esta naturaleza, esto siempre sucede, por lo que sus descriptivos se suelen presentar en tablas de intervalos como la que ahora veremos. En ésta, vemos que hay un 2% de personas de 30 años, un 2% de personas de 31 años y así sucesivamente. Por consiguiente, esta tabla responde a la segunda pregunta planteada: ¿Cómo se distribuyó la edad de los pacientes en la muestra? O, en otras palabras, ¿cuántos pacientes de cada edad del rango contemplado hubo en la muestra?

EDAD DEL PACIENTE

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	30,00	2	2,0	2,0	2,0
	31,00	2	2,0	2,0	4,0
	32,00	7	7,0	7,0	11,0
	33,00	5	5,0	5,0	16,0
	34,00	8	8,0	8,0	24,0
	35,00	2	2,0	2,0	26,0
	36,00	4	4,0	4,0	30,0
	37,00	2	2,0	2,0	32,0
	40,00	1	1,0	1,0	33,0
	41,00	2	2,0	2,0	35,0
	42,00	2	2,0	2,0	37,0
	43,00	3	3,0	3,0	40,0
	44,00	5	5,0	5,0	45,0
	45,00	8	8,0	8,0	53,0
	46,00	4	4,0	4,0	57,0
	47,00	4	4,0	4,0	61,0
	48,00	3	3,0	3,0	64,0
	50,00	2	2,0	2,0	66,0
	52,00	1	1,0	1,0	67,0
	53,00	1	1,0	1,0	68,0
	54,00	2	2,0	2,0	70,0
	55,00	2	2,0	2,0	72,0
	56,00	5	5,0	5,0	77,0
	58,00	3	3,0	3,0	80,0
	59,00	1	1,0	1,0	81,0
	60,00	5	5,0	5,0	86,0
	61,00	1	1,0	1,0	87,0
	64,00	3	3,0	3,0	90,0
	65,00	5	5,0	5,0	95,0
	66,00	3	3,0	3,0	98,0
	67,00	2	2,0	2,0	100,0
	Total	100	100,0	100,0	

La siguiente tabla nos informa de que, en este experimento, la mitad de la muestra tomó el fármaco en pruebas y la otra mitad el placebo, respondiendo a la tercera pregunta: ¿Cómo se distribuyó el fármaco y el placebo entre la muestra? O, en otras palabras, ¿cuántos pacientes recibieron el fármaco y cuántos el placebo?

TIPO DE FÁRMACO ADMINISTRADO

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	PLACEBO	50	50,0	50,0	50,0
	FÁRMACO EN PRUEBAS	50	50,0	50,0	100,0
	Total	100	100,0	100,0	

La siguiente tabla nos muestra que el resultado del experimento de investigación es disperso. Un 20% no mejora, un 24% muestra una mejoría leve, un 22% mejora moderadamente, un 19% sustancialmente y un 15% se cura. Con ella se responde a la cuarta pregunta planteada: ¿Cómo se distribuyó el resultado de la experiencia? O, en otras palabras, ¿qué porcentaje hubo de éxitos totales, parciales, leves y de fracasos en la muestra?

TIPO DE RESULTADO					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	NO MEJORA LA PATOLOGÍA	20	20,0	20,0	20,0
	MEJORA LEVE DE LA PATOLOGÍA	24	24,0	24,0	44,0
	MEJORA MODERADA DE LA PATOLOGÍA	22	22,0	22,0	66,0
	MEJORA SUSTANCIAL DE LA PATOLOGÍA	19	19,0	19,0	85,0
	CURACIÓN DE LA PATOLOGÍA	15	15,0	15,0	100,0
	Total	100	100,0	100,0	

Por consiguiente, una tabla de frecuencias es una herramienta descriptiva que sintetiza y organiza la información sobre una variable observada en un conjunto de elementos o individuos que constituyen una población (o una muestra, según el caso) proporcionándonos:

- Lista de los valores o categorías que puede tomar la variable
- Número de elementos o individuos que hay para cada valor que puede tomar la variable (Frecuencia)
- Porcentaje de elementos o individuos que hay para cada valor que puede tomar la variable (Porcentaje)
- Porcentaje de elementos o individuos que tienen un registro válido de valor de la variable sobre el total de válidos (Porcentaje válido, que coincide con el Porcentaje si no ha habido casos de no respuesta o de falta de información)
- Porcentaje acumulado, que va sumando el porcentaje de elementos o individuos que hay en las sucesivas categorías. Puede ser útil en determinados casos o para resaltar detalles. En la última tabla del ejemplo, se podría comentar que un 44% de los pacientes no muestran mejoría o la tienen leve tras la experiencia

Lógicamente, quedan muchas preguntas de interés que formular, especialmente la que sería fundamental en este experimento: ¿mejoraron más los pacientes que tomaron el fármaco en pruebas que los que tomaron el placebo?, pero este tipo de preguntas no se responden con tablas de frecuencias simples...

Otras tablas que se han calculado sobre los datos son las de las variables grupo de edad, tipo de dieta y situación laboral. Como se puede apreciar, la reducción de la edad a intervalos es mucho más práctica y permite visualizar que prácticamente la cuarta parte de la muestra son personas de entre 30 y 35 años, que poco más de la mitad no supera los 45 años y otros detalles. La tabla acerca del tipo de dieta nos informa de que un 23% de los pacientes no seguían habitualmente una dieta sana, mientras que el resto sí que tendía a hacerlo en mayor o menor grado. Finalmente, la tabla acerca de la situación laboral indica que la gran mayoría de los pacientes (75%) está en situación activa, trabajando, mientras que un 7% está en el paro, un 5% realiza labores del hogar y un 13% está jubilado o retirado.

GRUPOS DE EDAD

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	30-35	26	26,0	26,0	26,0
	35-40	7	7,0	7,0	33,0
	40-45	20	20,0	20,0	53,0
	45-50	13	13,0	13,0	66,0
	50-55	6	6,0	6,0	72,0
	55-60	14	14,0	14,0	86,0
	60-65	9	9,0	9,0	95,0
	65-70	5	5,0	5,0	100,0
	Total	100	100,0	100,0	

TIPO DE DIETA

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	NADA SANA	4	4,0	4,0	4,0
	POCO SANA	19	19,0	19,0	23,0
	ALGO SANA	32	32,0	32,0	55,0
	BASTANTE SANA	32	32,0	32,0	87,0
	MUY SANA	13	13,0	13,0	100,0
	Total	100	100,0	100,0	

SITUACIÓN LABORAL					
		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	TRABAJA	75	75,0	75,0	75,0
	PARO	7	7,0	7,0	82,0
	HOGAR	5	5,0	5,0	87,0
	JUBILACION-RETIRO	13	13,0	13,0	100,0
	Total	100	100,0	100,0	

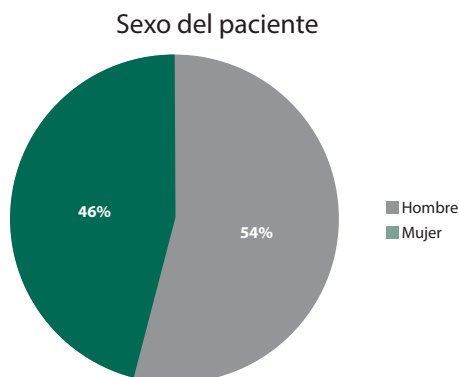
3.4. Resultados: gráficos simples unidimensionales

Todos los resultados anteriores pueden presentarse en forma de gráfico. Los gráficos simples o unidimensionales más adecuados según la clasificación de las variables son los siguientes:

- Variables cualitativas nominales con pocas categorías: gráfico de sectores
- Variables cualitativas nominales con más de 5 categorías: gráfico de barras
- Variables cualitativas ordinales: gráfico de barras
- Variables cuantitativas discretas: gráfico de barras
- Variables cuantitativas continuas: histograma

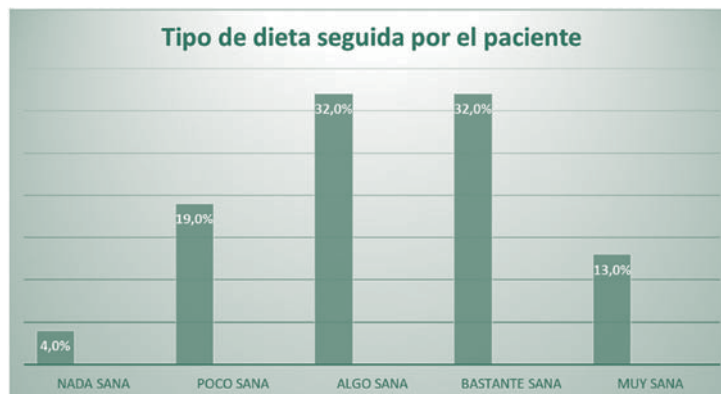
Seguidamente, se muestran ejemplos de todos estos tipos de gráficos, contruidos utilizando Power Point-Excel o Word-Excel. El motivo para utilizar este software en lugar de los gráficos que construye SPSS es que, tanto desde el punto de vista estético como de edición, si se está confeccionando un informe para publicar, los gráficos en Excel ofrecen más opciones de formato y son fácilmente gestionables para los autores y los editores. En el menú de Word y de Power Point se halla la opción Insertar, que permite insertar gráficos de todo tipo en un documento o en una presentación sin necesidad de hacerlo en Excel separadamente. Mediante este recurso, al insertar un gráfico, se abre una ventana Excel que permite la introducción de los datos del gráfico. Una vez confeccionado el modelo básico, se puede proceder a su refinamiento estético o a añadir detalles informativos relevantes como pueden ser las etiquetas, leyendas y otros.

Figura 8. Gráfico de sectores representando la distribución de sexo de los pacientes



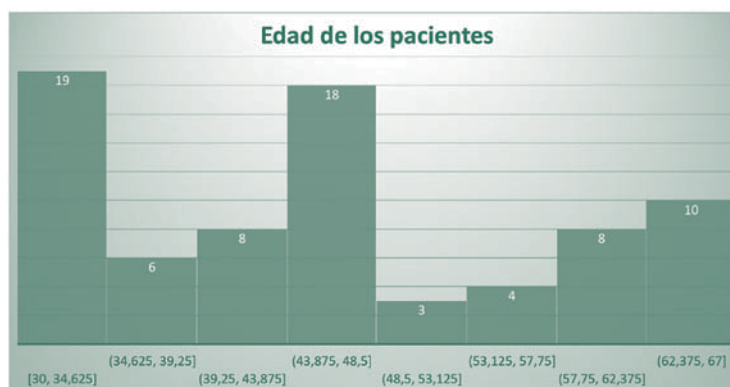
Nótese que en este tipo de gráfico el área de cada sector es proporcional a la frecuencia con que aparece cada categoría o valor de la variable sexo.

Figura 9. Gráfico de barras representando la distribución del tipo de dieta de los pacientes



Nótese que en este tipo de gráfico la altura de la barra es proporcional a la frecuencia con que aparece cada categoría o valor de la variable tipo de dieta.

Figura 10. Histograma representando la distribución de la edad de los pacientes



Nótese que, en un histograma, los valores de las áreas de los rectángulos equivalen a la probabilidad de hallar individuos de cada intervalo de edad en la muestra y que las barras están pegadas porque están representando a una variable cuantitativa continua, que toma valores en un intervalo sin interrupción.

3.5. Resultados: medidas de tendencia central y dispersión

Las medidas de tendencia central son medidas estadísticas que pretenden resumir en un solo valor a un conjunto de valores. Representan un centro en torno al cual se encuentra ubicado el conjunto de los datos. Las medidas de tendencia central más utilizadas son: moda, mediana y media.

Las medidas de dispersión en cambio miden el grado de dispersión de los valores de la variable o dicho, en otros términos, pretenden evaluar en qué medida los datos difieren entre sí. Las medidas de dispersión más utilizadas son: rango, percentiles y desviación típica o estándar.

Ambos tipos de medidas usadas conjuntamente permiten describir un conjunto de datos entregando información sintética acerca de su posición y su dispersión.

A lo largo de esta sección, vamos a familiarizarnos con estas herramientas y la utilidad que proporcionan en el ámbito de la sanidad utilizando el mismo ejemplo acerca del cual hemos calculado las tablas de frecuencias en la sección anterior.

3.5.1. La moda

La moda de una distribución se define como el valor de la variable que más se repite o que presenta la frecuencia más elevada (o el porcentaje más elevado) en la distribución de una variable. Es la única medida de tendencia central que se puede calcular para variables cualitativas nominales. Su determinación se hace por simple inspección visual de una tabla o gráfico de frecuencias. El valor de la variable que tenga la frecuencia más elevada es la moda. Pueden darse casos de variables bimodales (cuando hay más de un valor con la frecuencia más elevada) o multimodales (cuando hay más de dos valores que presentan la misma frecuencia máxima).

Siguiendo con el ejemplo anterior, teniendo en cuenta que en nuestra lista de variables hay tres que son de tipo nominal (el sexo de los pacientes, el tipo de fármaco que se les ha administrado y su situación laboral), se ha pedido a SPSS que calculase la moda para las mismas, que es la única medida de tendencia central que se puede calcular sobre ellas. El resultado es una tabla de este tipo:

Estadísticos				
		SEXO DEL PACIENTE	TIPO DE FARMACO ADMINISTRADO	SITUACIÓN LABORAL
N	Válido	100	100	100
	Perdidos	0	0	0
Moda		1,00	,00 ^a	1,00

a. Existen múltiples modos. Se muestra el valor más pequeño.

La tabla muestra el número de casos válidos (100) y de no respuesta o perdidos (0) para cada variable. La última fila muestra el valor de la moda para cada variable. En el caso del sexo, la moda es el valor 1, que representa a los hombres. En el caso del tipo de fármaco, la moda es un 0 con una nota indicando que para esta variable hay más de una moda y que se muestra la que tiene el valor más pequeño, en este caso el 0, que representa al placebo. Finalmente, la moda de la situación laboral toma el valor 1, que representa la categoría “trabaja”.

Por consiguiente, la moda de estas variables nos permite deducir que en la muestra de pacientes hay más hombres que mujeres, que ha habido tantos casos de administración de placebo como del fármaco en pruebas y que la mayoría de los pacientes está trabajando.

3.5.2. La mediana

Cuando trabajamos con datos individualizados (o desagregados), la mediana es el valor de la variable que ocupa la posición central si el número de observaciones es impar y la media de los dos valores centrales si el número de observaciones es par. En una tabla de frecuencias, como trabajamos con datos agregados, la mediana es el valor que recoge el 50% de las observaciones. Es decir, el 50% de las observaciones tiene valores iguales o inferiores a la mediana y el otro 50% tiene valores iguales o superiores a la mediana.

Siguiendo con el ejemplo anterior, en nuestra lista de variables hay tres que son de tipo ordinal (el tipo de resultado obtenido, el grupo de edad y el tipo de dieta habitual de los pacientes), se ha pedido a SPSS que calculase la moda y la mediana para las mismas, ya que son las únicas medidas de tendencia central que se puede calcular sobre ellas. El resultado es una tabla de este tipo:

Estadísticos				
		TIPO DE RESULTADO	GRUPOS DE EDAD	TIPO DE DIETA
N	Válido	100	100	100
	Perdidos	0	0	0
Mediana		2,0000	3,0000	3,0000
Moda		1,00	1,00	3,00 ^a

a. Existen múltiples modos. Se muestra el valor más pequeño.

La tabla muestra el número de casos válidos (100) y de no respuesta o perdidos (0) para cada variable. Las dos últimas filas muestran el valor de la mediana y de la moda para cada variable. En el caso del tipo de resultado, la moda o caso más frecuente es el valor 1, que representa un resultado de “Mejora leve de la patología”, mientras que la mediana es el valor 2, que representa un resultado de “Mejora moderada de la patología”. Por consiguiente, tras el experimento, la mayoría de los pacientes presentó una leve mejora de la patología, teniendo en cuenta al mismo tiempo que, según la mediana, un 50% de los casos no superó ese estadio y otro 50% sí que lo hizo. En el caso del grupo de edad, la moda tomó el valor 1, indicando que la mayoría de los pacientes tenía entre 30 y 35 años, mientras que la mediana tomó el valor 3, indicando que el 50% de los pacientes pertenecía a un grupo de edad menor que el de 40-45 años y el otro 50% a un grupo de mayor edad. Finalmente, acerca del tipo de dieta, la moda y la mediana tomaron el valor 3, que representa una dieta “algo sana”. El resultado indica que éste es el caso más frecuente entre los pacientes y que la mitad de la muestra sigue dietas por debajo de este nivel y la otra mitad por encima de éste. Sin embargo, la distribución de esta variable es bimodal (ver la tabla correspondiente

en el apartado 3.3), siendo la segunda moda el valor 4 que representa una dieta bastante sana. De ello se deduce que más de un 50% de la muestra tiende a seguir una dieta más sana que lo contrario.

3.5.3. La media y las medidas de dispersión

En el caso de trabajar con variables cuantitativas, como por ejemplo la edad, la estadística descriptiva permite calcular todas las medidas de tendencia central (moda, mediana y media) y todas las de dispersión (rango, percentiles, desviación típica y otras), ya que se pueden realizar todo tipo de operaciones matemáticas con ellas.

La medida de tendencia central más conocida y utilizada es la media aritmética o promedio aritmético. Se representa por la letra griega μ cuando se trata del promedio del universo o población (es decir, de un parámetro) y $\bar{x}$ cuando se trata del promedio de la muestra (o estimador del parámetro). Es importante destacar que μ es una cantidad fija (la mayoría de las veces desconocida), mientras que el promedio de la muestra es variable puesto que si se extrajesen diferentes muestras de la misma población estas tenderían a mostrar ciertas diferencias entre sí. La media se expresa en la misma unidad que los datos originales: años, centímetros, horas, gramos, etc.

La media se calcula sumando todos los valores de los casos que componen la distribución de una variable y dividiendo este resultado por el número total de observaciones. Esta operación se expresa mediante la fórmula:

$$\bar{x} = \frac{\sum x_i}{n}$$

Sin embargo, actualmente, tanto SPSS como Excel o cualquier calculadora de mano nos permiten efectuar los cálculos de esta y todas las medidas estadísticas con comodidad y facilidad, por lo que en la práctica lo más importante es comprender el concepto, su significado y aprender a interpretarlo cuando aparece en forma de resultado.

La media es un valor cuya aspiración es la de proporcionar el valor central representativo de una distribución mediante un solo número. Por eso, para poder evaluar si ese número es representativo o no de dicha distribución, la media debe presentarse junto a una medida de dispersión, un valor que aporta al analista una visión de la amplitud de la distribución. Para comprender la importancia de este punto en el ámbito sanitario, baste con imaginar a un médico evaluando indicadores medios de determinados conceptos en un análisis de sangre. Para el médico no va a significar lo mismo una media de, por ejemplo, valor 5 con una dispersión media de 2 unidades a su alrededor que con una dispersión media de 0,3 unidades a su alrededor. La media es la misma, pero la precisión del segundo caso es mucho mayor que la del primero.

Aunque la medida de dispersión por excelencia que debe acompañar a una media es la desviación estándar, es importante conocer otras más elementales y que también resultan de utilidad en la interpretación de resultados descriptivos de variables cuantitativas. Así, la más sencilla de las medidas de dispersión, el rango, se define como la diferencia entre el mayor valor de la variable y el menor valor de la variable. A mayor rango, mayor dispersión en la distribución de una variable cuantitativa. Por otro lado, hay que considerar a los percentiles, que son los valores de la variable que aglutinan el 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% y 90% de su distribución. Los percentiles resultan de gran utilidad para resumir distribuciones de variables con elevado rango de valores.

Como se ha apuntado antes, la mejor medida de dispersión, y la más generalizada es la varianza, o su raíz cuadrada, la desviación estándar. La varianza se representa con el símbolo σ^2 (sigma cuadrado) para el universo o población y con el símbolo s^2 (s cuadrado), cuando se trata de la muestra. La desviación estándar, que es la raíz cuadrada de la varianza, se representa por σ (sigma) cuando pertenece al universo o población y por "s", cuando pertenece a la muestra. σ^2 y s^2 son parámetros, constantes para una población particular; s^2 y s son estadígrafos o estimadores de los parámetros, valores que cambian de muestra en muestra dentro de una misma población. La varianza se expresa en unidades de variable al cuadrado y la desviación estándar simplemente en unidades de la variable en estudio.

La fórmula de cálculo de la desviación estándar es:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$

Interpretación de desviación estándar: un alto valor de la desviación típica en las unidades que estemos estudiando, indicará que los datos están alejados del promedio. A veces es difícil hacer una interpretación de la desviación típica teniendo un solo valor de ella. La situación es más clara si se comparan las dispersiones de dos muestras. Por ejemplo, si la desviación típica de una muestra A en que se ha medido la edad es igual a 18 años y la de una muestra B igual a 25 años. En ese caso diremos que los datos de la muestra B tienen mayor dispersión que los datos de la muestra A. Esto significa que en la muestra A las edades de los sujetos están más cerca del promedio y que en cambio en la muestra B los datos están más alejados del promedio, siendo la edad media de la muestra A más representativa del conjunto muestral que la edad media de la muestra B.

A partir de estos conocimientos, veamos la información descriptiva que se obtiene acerca de la edad de los pacientes entrevistados en una tabla calculada por SPSS que ofrece todas las medidas de tendencia central y dispersión:

Estadísticos

EDAD DEL PACIENTE

N	Válido	100
	Perdidos	0
Media		46,4800
Mediana		45,0000
Moda		34,00 ^a
Desviación estándar		11,40307
Varianza		130,030
Rango		37,00
Mínimo		30,00
Máximo		67,00
Percentiles	10	32,0000
	20	34,0000
	30	36,3000
	40	43,4000
	50	45,0000
	60	47,0000
	70	54,7000
	80	58,8000
	90	64,9000

a. Existen múltiples modos. Se muestra el valor más pequeño.

La muestra está compuesta por 100 pacientes. Todos ellos proporcionaron su edad y no hay casos de no respuesta o “perdidos”. La edad media de esta muestra es de 46,48 años con una desviación típica de 11,4 años, un rango de 37 años (la edad mínima es de 30 años y la máxima de 67). El 10% de la muestra tiene hasta 32 años, el 20% hasta 34 años y así sucesivamente hasta el percentil 90% que indica que el 90% de los casos tiene como máximo 64 años.

La moda o caso más frecuente es tener 34 años (si bien, en este caso hay varias modas, es decir, varios valores de edad con la máxima frecuencia de aparición, y se muestra el valor más pequeño) y la mediana es de 45 años. La dispersión de la media es elevada (11,4 años) y el rango de 37 años es amplio en el contexto en el que estamos. En casos de gran dispersión como el presente, se toma la mediana como indicador de tendencia central, pues se trata de una medida “robusta”, que no se ve afectada por valores extremos (muy bajos o muy altos) de la variable y que responde a un valor situado en el centro de la distribución y no a una operación matemática como en el caso de la media. Cuando en una distribución se dan diferencias llamativas entre los valores de la moda, la mediana y la media, este hecho debe interpretarse como un síntoma inequívoco de que la media es poco representativa y que es mejor trabajar con la mediana: este valor, como se determina por el orden de los valores de los casos y no mediante operación matemática, no se ve afectado por casos de valores muy bajos o muy altos que pueda haber en los extremos de la distribución. La mediana es, verdaderamente el valor central de una distribución cuantitativa en el caso de

trabajar con un número impar de casos y la media de los dos valores centrales en el caso de trabajar con un número par de casos.

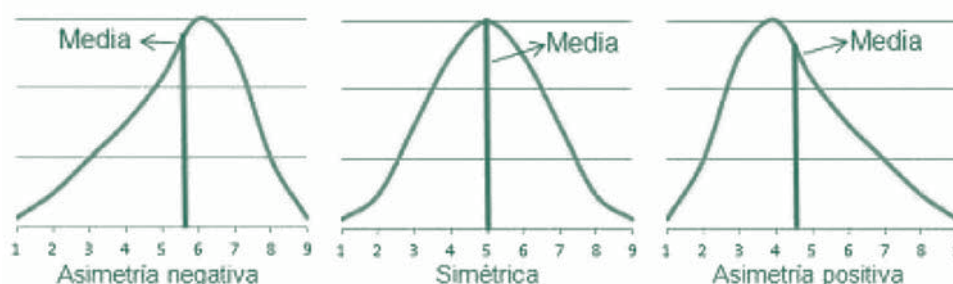
La varianza es el valor de la desviación estándar al cuadrado y no tiene utilidad práctica, ya que no resulta eficaz leer cantidades en unidades al cuadrado. Habitualmente no se pide entre el conjunto de resultados a menos que sea necesaria para algún cálculo adicional.

3.5.4. Las medidas de forma

Las medidas de forma sólo son aplicables a variables cuantitativas continuas. Las más comunes son el coeficiente de asimetría y el coeficiente de curtosis. Se trata de dos indicadores que determinan qué tipo de comportamiento muestra un histograma en cuanto a su simetría y en cuanto a su apuntamiento respectivamente.

Así, en general, la distribución de una variable cuantitativa continua puede mostrar cualquier forma al ser representada en forma de histograma, pero en muchos casos, se aproxima a un modelo de tipo simétrico, asimétrico a la derecha o asimétrico a la izquierda (ver Figura 11).

Figura 11. Asimetría: principales modelos de histogramas

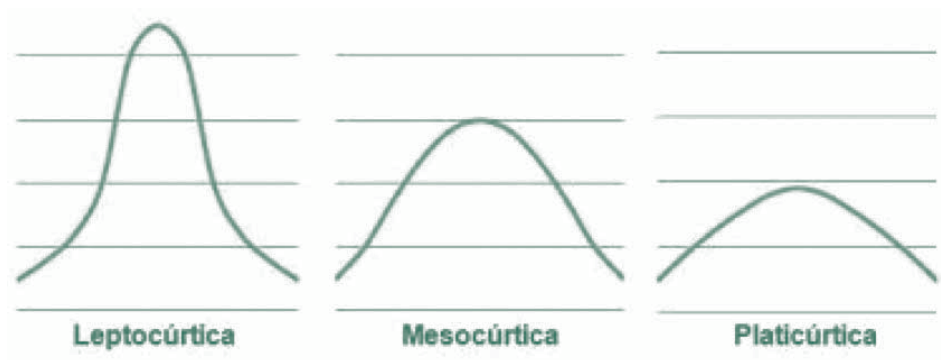


Fuente: Universo Fórmulas

Estos modelos son importantes en estadística, ya que, el simétrico indica que la variable analizada tiene lo que se denomina un comportamiento Normal y los otros dos, que se aproxima al comportamiento de otras distribuciones emblemáticas como son la Chi Cuadrado, la F de Snedecor u otras. Más adelante veremos que, en relación con el contraste de hipótesis las distribuciones Chi Cuadrado y F de Snedecor son importantes y que, en la construcción de modelos de regresión y otras técnicas, una de las condiciones requeridas a los datos de las variables implicadas es la de tener un comportamiento Normal.

Por otro lado, si una distribución sigue un comportamiento simétrico o aproximadamente simétrico, es importante saber si la distribución está muy concentrada en torno a la media (leptocúrtica), si tiene una dispersión intermedia de tipo Normal (mesocúrtica) o si la distribución está bastante dispersa en torno a la media (platicúrtica). La Figura 12 muestra estos tres tipos de modelos.

Figura 12. Curtosis: principales modelos de apuntamiento en histogramas simétricos



Fuente: Universo Fórmulas

El coeficiente de asimetría es un indicador cuyo valor nos permite saber si un histograma es simétrico o asimétrico a la derecha o a la izquierda. La fórmula de este coeficiente es:

$$CAF = \frac{\sum (x_i - \bar{x})^3}{n \cdot s_x^3}$$

Si $CAF < 0$: la distribución tiene una asimetría negativa y muestra altas frecuencias para los valores que quedan por encima de la media

Si $CAF = 0$: la distribución es simétrica y las frecuencias de los valores van disminuyendo progresivamente por encima y por debajo de la media

Si $CAF > 0$: la distribución tiene una asimetría positiva y muestra altas frecuencias para los valores que quedan por debajo de la media

El coeficiente de curtosis o apuntamiento es un indicador que nos permite clasificar un histograma simétrico en función de su grado de apuntamiento en leptocúrtico, mesocúrtico o leptocúrtico. La fórmula de este coeficiente es:

$$CAP = \frac{\sum (x_i - \bar{x})^4}{n \cdot s_x^4} - 3$$

En la fórmula se resta 3 porque es el valor de la curtosis de una distribución Normal. Por consiguiente, el coeficiente de apuntamiento valdrá 0 para la Normal, tomándose a ésta como referencia.

Si $CAP < 0$: la distribución es platicúrtica y más dispersa en torno a la media que una Normal

Si $CAP = 0$: la distribución es mesocúrtica o Normal

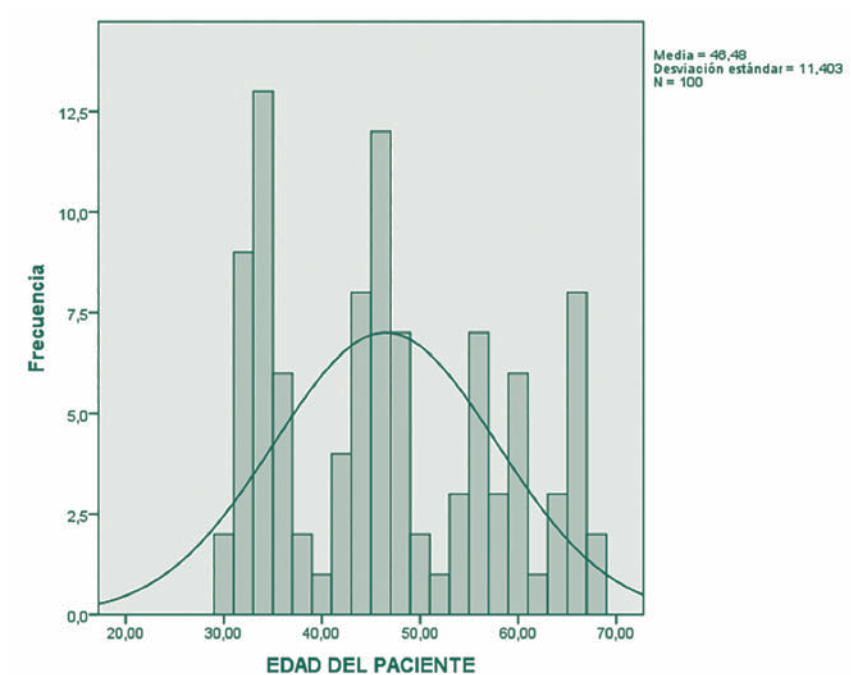
Si $CAP > 0$: la distribución es leptocúrtica y menos dispersa en torno a la media que una Normal

Actualmente, SPSS y otros programas estadísticos calculan estos coeficientes de forma automática como parte de los indicadores que se incluyen en los descriptivos de una variable cuantitativa continua. Por ejemplo, en el caso de la variable edad del ejemplo que venimos trabajando, SPSS nos proporciona el siguiente resultado:

Estadísticos		
EDAD DEL PACIENTE		
N	Válido	100
	Perdidos	0
Media		46,4800
Desviación estándar		11,40307
Asimetría		,282
Error estándar de asimetría		,241
Curtosis		-1,161
Error estándar de curtosis		,478

Como podemos observar en esta tabla, el coeficiente de asimetría es 0,282, un valor positivo pero alejado de 1, por lo que podemos decir que el histograma de esta variable es ligeramente asimétrico a la derecha o positivo, por lo que la cola es más larga a la derecha y la joroba más alta a la izquierda, lo que indica que hay un mayor volumen de personas jóvenes (valores por debajo de la media de 46,48 años) que mayores. Dado que la asimetría es ligera, la distribución es más simétrica que asimétrica, por lo que tiene sentido mirar el valor del coeficiente de curtosis. En este caso vale -1,161, valor por debajo de cero que indica que la distribución es más dispersa que una Normal. El valor de la desviación estándar es de 11,4 años, elevado, lo cual ratifica el resultado (ver Figura 13).

Figura 13. Histograma sobre la edad de los pacientes del ejemplo y su comparación con una curva Normal de igual media y desviación estándar



4. Fundamentos de la estadística inferencial

4.1. Introducción

El principal objetivo de la Estadística consiste en poder decir algo con respecto a un gran conjunto de personas, mediciones u otros entes (población) con base en las observaciones hechas sobre sólo una parte (muestra) de dicho gran conjunto. La capacidad para “decir algo” sobre poblaciones con base en muestras está basada en supuestos con respecto a algún modelo de probabilidad que permite explicar las características del fenómeno bajo observación.

Al conjunto de procedimientos estadísticos en los que interviene la aplicación de modelos de probabilidad y mediante los cuales se realiza alguna afirmación sobre poblaciones con base en la información producida por muestras se le llama Inferencia Estadística o Estadística Inferencial.

Por consiguiente, la Estadística Inferencial es el conjunto de métodos utilizados para determinar algún atributo medible acerca de una población en base a una muestra. Asimismo, elabora modelos de comportamiento y contrasta hipótesis acerca de relaciones entre más de una variable.

Su uso tiene diversas ventajas:

- El uso de muestras representativas reduce costes y posibilita estudios que no podrían abordarse al estudiar grandes poblaciones
- Si la muestra es representativa pueden incluso cometerse menos errores que si se tuviese que recoger información censal (de todos los individuos que componen una población)

En definitiva, las técnicas de estadística inferencial sirven para obtener conclusiones que van más allá de la mera descripción de los datos.

4.2. Población y muestra

Recordemos que una población es:

El conjunto total de sujetos de cualquier tipo acerca del cual interesa estudiar determinadas características o comportamientos mediante la disposición de información de estos, presentada en forma de variables.

Un ejemplo de población en el ámbito de la salud puede ser la de habitantes en cualquier ciudad o territorio; la atendida por un Centro de Atención primaria; la de en-

fermos de cualquier patología (Cáncer de piel, Alzheimer, atrofia muscular...) de un país; la de asegurados en una Compañía; la de dermatólogos de un país...la lista sería interminable.

En definitiva, en el ámbito estadístico, población es el conjunto total de elementos, individuos o entes sujetos a estudio y de los cuales queremos obtener un resultado.

Por otra parte, una muestra es:

Aplicada a la estadística se refiere a un subconjunto representativo de elementos que forman parte de población que se desea estudiar.

4.3. Muestreo estadístico

En Estadística, se conoce como muestreo a la técnica para la selección de una muestra a partir de una población.

Al elegir una muestra se espera que sus propiedades sean extrapolables a la población. Este proceso permite ahorrar recursos, obteniendo resultados parecidos a los que se obtendrían si se realizase un estudio de toda la población. Para que el muestreo sea válido y se pueda realizar un estudio fiable (que realmente represente a la población), debe cumplir ciertos requisitos.

En el muestreo, desde una perspectiva teórica, se puede extraer un elevado número de muestras diferentes de la misma población. Al conjunto de muestras que se pueden obtener de la población se le denomina espacio muestral. La variable que asocia a cada muestra su probabilidad de extracción sigue la llamada distribución muestral.

En cambio, en la práctica, lo habitual es trabajar únicamente con una de esas múltiples muestras que sería posible extraer en teoría. Por consiguiente, conocer los métodos de muestreo o extracción de la muestra con la que se trabaja realmente es imprescindible para hacerlo acertada y eficientemente.

4.3.1. Métodos de muestreo

Dado que la estadística inferencial se basa en el uso de muestras, es importante tener algunas nociones acerca de los métodos más comunes que se consideran para su extracción.

- Muestra aleatoria: es una muestra obtenida al azar en una población determinada, por ejemplo, calcular 20 números aleatorios mediante un programa informático y extraer los 20 hospitales que se correspondan con estos números en la lista total o

población de los hospitales del país ordenada previamente.

- Muestra sistemática, es una muestra seleccionada en unas determinadas condiciones, por ejemplo, una muestra de la lista total de hospitales del país de las cuales tomamos un nombre de cada cinco comenzando a partir de un número seleccionado de forma aleatoria y hasta completar una ronda completa sobre dicha lista.
- Muestra estratificada, es aquella en la que se definen características claras y precisas en la muestra que reproducen las características de la población en su conjunto, por ejemplo, muestra de hospitales del país teniendo en cuenta su figura jurídica, dependencia administrativa, región u otros estratos de interés.
- Dependiendo del tipo de investigación en ocasiones será necesario seleccionar un grupo control con similares características al grupo problema (o caso), con el fin de comparar los resultados y obtener información contrastada sobre las características diferenciales o que son propias de ambos grupos. Este planteamiento es muy común en la investigación epidemiológica, por ejemplo.

4.3.2. El tamaño de la muestra para estimar una proporción o porcentaje

Trabajar con una muestra implica trabajar con un número de sujetos menor a la población total, objeto de la investigación, pero que sea una cantidad suficiente y razonable para que los resultados obtenidos puedan ser significativos de ese total. Por lo tanto, sería la cantidad mínima necesaria para que los resultados obtenidos en la muestra sean representativos y fiables del conjunto de la población.

Existen varias formas de calcular el tamaño muestral, pero en general se recomienda aplicar una fórmula de poblaciones finitas (una población acerca de la cual se sabe el número total de integrantes) por dos razones: en primer lugar, porque la mayoría de las investigaciones relacionadas con la gestión sanitaria se realizan sobre poblaciones concretas y determinadas y, en segundo lugar, porque posee una mayor precisión.

Sin embargo, antes de mostrar la fórmula de cálculo del tamaño muestral es necesario tener en cuenta el tipo de pregunta clave o más importante que queremos resolver acerca de la población en estudio.

Supongamos que la pregunta de investigación es la siguiente:

¿Cuál fue el porcentaje de pacientes que entraron en los hospitales españoles con un diagnóstico de entrada urgente en el año 2011?

Para responder a esta pregunta, podríamos utilizar la encuesta de ámbito español de morbilidad hospitalaria para obtener un resultado y, lo que queremos estimar es un **porcentaje o proporción de casos de un determinado tipo** que se dan en la población

total. Por este motivo, la fórmula de cálculo que emplearíamos es la que ha sido especialmente diseñada para estimar porcentajes o proporciones:

$$n = \frac{z^2 \cdot N \cdot p \cdot (1 - p)}{e^2(N - 1) + z^2 \cdot p(1 - p)}$$

Donde:

- Tamaño muestral = n
- Constante del intervalo de confianza (de valor 2,64 para un nivel del 99%; 1,96 para un nivel del 95%; 1,64 para un nivel del 90%) = z
- Tamaño de la población objetivo = N
- Prevalencia estimada a priori del proceso de nuestro interés = p
- Error muestral que asumimos = e , se coloca en la fórmula dividido por 100. Es decir, 0,01 representaría un error del 1%; 0,05 del 5% y así sucesivamente

En esta fórmula hay conceptos importantes que debemos conocer antes de seguir adelante, como son el nivel de confianza, la prevalencia estimada a priori y el error muestral. Vamos a explicarlos a continuación.

4.3.2.1. El nivel de confianza

Siguiendo con la pregunta de ejemplo (¿Cuál fue el porcentaje de pacientes que entraron en los hospitales españoles con un diagnóstico de entrada urgente en el año 2011?), denominemos **P** al porcentaje total poblacional de pacientes que ingresaron de urgencia en el total de hospitales de España en el 2011. Para responder a la pregunta de investigación, sabemos que contamos con los resultados de la encuesta de morbilidad hospitalaria realizada por el INE a una **muestra representativa de hospitales**. El INE nos proporciona **una** de las muestras posibles entre todas las que habría podido extraer para elaborar la encuesta de morbilidad hospitalaria. Denominemos **p** al porcentaje de pacientes ingresados de urgencia que aparece en esa muestra que ha tomado el INE. Dado que lo que vamos a obtener es una estimación de **P** a través de **p**, desde un punto de vista estadístico nos interesa saber cuán precisa es esta estimación o, dicho de otra forma, cuánto se aproxima la estimación **p** al verdadero **P** que es desconocido. En resumen: queremos poder decir qué tan fiable es la estimación que obtiene el INE del parámetro poblacional.

El nivel de confianza es la herramienta estadística que nos proporciona información acerca de esa fiabilidad. Sin embargo, es importante comprender que el nivel de confianza lo escogemos nosotros, o en este ejemplo, el propio INE, antes de calcular el tamaño de una muestra. Así, el INE, al diseñar la muestra de la encuesta de morbilidad hospitalaria y su tamaño, tuvo que escoger con qué nivel de confianza quería trabajar. El nivel de confianza es una expresión del grado de aproximación que deseamos obtener entre la estimación muestral **p** y el verdadero **P** en la población. Los niveles más

utilizados en la práctica son los de 95% (el más habitual), el 99% (el que proporciona mayor precisión) y el de 90% (el más coste-efectivo). Se podrían estimar parámetros a cualquier otro nivel (70%, 25%, 89%.....), pero los estimadores obtenidos se alejarían mucho del verdadero y no interesan. De manera que cuando se invierte en la extracción de una muestra, no es lógico usar niveles de confianza menores al 90%.

Una vez seleccionado el nivel de confianza, y sin entrar en el marco teórico matemático acerca de su justificación y formulación, lo importante es comprender que nos permite calcular un **intervalo de confianza** acerca de cuánto se aproxima **p** al verdadero **P**. Es decir, que, a partir de la muestra, el nivel de confianza se representa como $(1-\alpha)$ siendo α la diferencia entre 1 y el nivel de confianza dividido por 100. Este número $(1-\alpha)$, nos indica la probabilidad de que los extremos superior e inferior del intervalo de confianza en torno a **p** contengan al verdadero valor **P**. La interpretación práctica es sencilla, por ejemplo, si el nivel de confianza es del 95% (entonces, $1 - \alpha = 1 - 0,05 = 0,95$), significa que en el 95% de las veces que repitiéramos el experimento, el intervalo de confianza calculado contendría al verdadero valor del parámetro y en el 5% restante el intervalo no contendría el verdadero valor. La representación gráfica del nivel de confianza se ve a continuación, así como la formulación del intervalo de confianza para una proporción o porcentaje:

Figura 14. Representación del nivel de confianza del 95%



Fuente: Emaze

Fórmula del intervalo de confianza para estimar una proporción poblacional:

$$P \left(p - z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \leq P \leq p + z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right) = 1 - \alpha$$

Una vez que el intervalo de confianza ha sido calculado para una muestra concreta, sabemos que, el intervalo obtenido contiene o no contiene al verdadero valor del parámetro **P** con probabilidad 1, por esa razón, cuando ya tenemos un valor concreto **p** hablamos de confianza y no de probabilidad. **Confiamos** en que se cumple que para los límites del in-

tervalo que hemos calculado haya un 95% (90%, 99% u otro nivel) de posibilidades de que contengan el verdadero valor **P** (que seguirá siendo desconocido, por supuesto).

De cara a la práctica, conviene precisar que:

Para un nivel de confianza del 95%, el valor de **z** en el gráfico de la Figura 14 es 1,96. En cambio, para un nivel de confianza del 99%, la zona azul se ampliaría y el valor de **z** sería de 2,58 y, finalmente, para un nivel de confianza del 90%, la zona azul se haría más pequeña y el valor de **z** sería de 1,64. Por consiguiente, al calcular un tamaño muestral, nosotros somos los que ponemos el valor de **z** en la fórmula como expresión de la precisión con que vamos a trabajar.

4.3.2.2. El error muestral

Además del nivel de confianza, también nosotros somos los que escogemos el error muestral máximo que queremos afrontar en las estimaciones que va a proporcionar la muestra. En el caso de la estimación de un porcentaje o proporción, el error muestral se expresa en forma de porcentaje. Así, podemos decir que queremos cometer un error muestral del 1%, del 2%, del 5%, etc., no teniendo sentido afrontar errores que vayan más allá del 10% y siendo éste un error ya muy elevado. El error muestral es la expresión de cuánto, como máximo, para un nivel de confianza establecido por nosotros, estamos dispuestos a admitir que nuestra estimación **p** se aleje del verdadero parámetro **P**.

El error muestral y el nivel de confianza **son conceptos diferentes** y sus cantidades no son complementarias ni depende nuestra elección de cada una de ellas del otro valor. Esta confusión se da frecuentemente porque el nivel de confianza y error que se usan habitualmente son 95% y $\pm 5\%$, pero también se podría trabajar al 95% con un error de $\pm 2\%$ u otro cualquiera que sea razonable.

De la fórmula del intervalo de confianza se desprende que el error muestra equivale a:

$$z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Pues representa el porcentaje que sustraemos y añadimos a la estimación muestral **p** de la proporción poblacional **P** para calcular entre qué límites debe hallarse dicho valor con un nivel de confianza establecido.

4.3.2.3. La prevalencia a priori

En la fórmula de cálculo del tamaño muestral también interviene una medida de la dispersión que estimamos hay en la población acerca del concepto que queremos medir: la prevalencia de un fenómeno a priori. Por ejemplo, para nuestra pregunta de investigación, tendríamos que hacer un supuesto o comprobar mediante una prueba piloto, qué proporción de pacientes entra por urgencias y qué proporción lo hace con un diagnóstico ordinario en una pequeña muestra de hospitales. Si no tenemos elementos de juicio acerca de esta cuestión, entonces se recomienda poner como estimación de dispersión o prevalencia el supuesto más extremo, que es aquél en que el 50% de los pacientes entra por urgencias y el 50% por diagnóstico ordinario.

Esto se expresaría como: $p = 0,5$ y $1-p = 0,5$

Si tenemos información, por ejemplo, de un hospital representativo en que se da la proporción 20% urgencias y 80% ordinario, entonces pondríamos que: $p = 0,2$ y $1-p = 0,8$. Si disponemos de encuestas de años anteriores, se pueden utilizar las estimaciones obtenidas en las mismas, y así sucesivamente.

4.3.2.4. Ejemplo real sobre aspectos del cálculo de tamaño muestral para estimar una proporción

Tomando la pregunta de investigación concreta acerca de la cual venimos trabajando como ejemplo:

¿Cuál fue el porcentaje de pacientes que entraron en los hospitales españoles con un diagnóstico de entrada urgente en el año 2011?

Explotando los micro datos de la encuesta de morbilidad hospitalaria de ese año, obtenemos la siguiente tabla que nos da la respuesta a la pregunta de investigación planteada:

Diagnóstico de entrada:

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Ordinario	1811897	39,3	39,3	39,3
	Urgente	2795771	60,7	60,7	100,0
	Total	4607668	100,0	100,0	

Con estos datos reales de la Encuesta de Morbilidad Hospitalaria España 2011 en que se estimó que un 39,3% de los pacientes ingresaron de forma ordinaria y un 60,7% de forma urgente, vamos a analizar los elementos que permitieron lograr este resultado.

Comenzando por el tamaño de la muestra, la tabla arroja un total de 4.607.668 casos. Es decir, el INE se basó en una muestra de este tamaño para hacer la estimación.

La pregunta que nos hacemos lógicamente es: ¿cuán fiable es esta muestra?

Para responder a la pregunta, podemos fijar un nivel de confianza, por ejemplo, del 95% y calcular el error muestral que se ha cometido:

$$e = z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 1,96 \sqrt{\frac{0,393 \cdot 0,607}{4607668}} = 0,000446$$

Multiplicando esta cantidad por 100, vemos que se ha cometido un error muestral de $\pm 0,04466\%$, un error muy bajo y alejado del habitual del $\pm 5\%$. Este error, permite concluir que el INE estimó con gran precisión el porcentaje de pacientes que ingresaron por vía urgente, ya que con un 95% de confianza, el valor poblacional de este parámetro estaba entre los límites:

$$P(60,7 - 0,000446 = 60,699 \quad \text{y} \quad 60,7 + 0,000446 = 60,700) = 0,95$$

Valores que ratifican que se pudo tomar como válida y fiable la estimación de 60,7%. Lógicamente, el INE, al ofrecer indicadores de ámbito nacional en muchos apartados clave que ofrecen información sobre el sistema sanitario, tiene que manejar muestras grandes y fiables. No obstante, el tamaño muestral no es el factor más crítico para obtener buenas estimaciones, sino que lo es la dispersión de los valores de la variable analizada en la población. Esta muestra tiene un tamaño tal que, para esta pregunta resulta incluso excesiva. Ahora bien, la encuesta de morbilidad debe trabajar con ese tamaño, no sólo para estimar esta proporción de modalidad de ingreso, sino muchos otros indicadores de altas por patologías y otras variables que están mucho más segmentadas.

4.3.3. El tamaño de la muestra para estimar una media

Para calcular el tamaño muestra necesario (n) para estimar una media, es necesario disponer de una muestra piloto que nos permita calcular la dispersión o desviación típica a priori de la muestra o bien tener información que nos permita hacer un supuesto acerca de su magnitud gracias a estudios realizados con anterioridad sobre el mismo parámetro o, incluso disponer de una estimación razonable hecha por algún experto en el tema de investigación.

Una vez en disposición de esta información, escogemos el nivel de confianza deseado y el error muestral, esta vez como dato **en las mismas unidades que la media que queremos estimar** y que tenga sentido respecto del fenómeno analizado. Cuando se tiene toda la información, se aplican los datos a la siguiente fórmula para poblaciones finitas:

$$n = \frac{z^2 \cdot N \cdot s^2}{e^2(N - 1) + z^2 \cdot s^2}$$

Donde:

- Tamaño muestral = n
- Constante del intervalo de confianza (de valor 2,64 para un nivel del 99%; 1,96 para un nivel del 95%; 1,64 para un nivel del 90%) = z
- Tamaño de la población objetivo = N
- Desviación estándar piloto de la variable objeto de nuestro interés = s
- Error muestral que asumimos = e, es un valor razonable y expresado en las mismas unidades que la variable objeto de interés

Supongamos ahora que tuviésemos la siguiente pregunta de investigación:

¿Cuál fue la media de días de estancia en el hospital de los pacientes atendidos en los hospitales españoles en el año 2011?

En este caso, como queremos estimar una cantidad media y no un porcentaje, deberíamos calcular el tamaño de muestra necesario para aproximarnos a este dato en la población española aplicando lo siguiente:

N = supongamos que la población total de pacientes ingresados fue de 10.000.000 (no toda la población de un país sufre un ingreso hospitalario cuando menos anual) y, si no tenemos ese dato, se puede aplicar la fórmula para poblaciones infinitas, que sirve igual cuando la población es muy grande

z = 1,96 supongamos que queremos una estimación al 95% de confianza

s = supongamos que en un hospital representativo se calculó que la desviación típica de los días de estancia era de 20,7 días

e = supongamos que escogemos como error muestral máximo 0,3 días (es una elección nuestra, también podemos poner 0,1 días si queremos más precisión o 1 día si queremos menos precisión....)

Con todos estos datos, la muestra que necesitaríamos para estimar la media de días de estancia sería:

$$n = \frac{1,96^2 \cdot 10000000 \cdot 20,7^2}{0,3^2(10000000 - 1) + 1,96^2 \cdot 20,7^2} = 18.289,3 \text{ pacientes o casos de ingreso}$$

Por consiguiente, entrevistado a una muestra representativa de 18.289 pacientes de toda España, podríamos obtener una estimación de la media de días de estancia en un hospital en el año 2011 para un nivel de confianza del 95% y un error de 0,3 días por encima y por debajo. Es habitual que a algunas personas les parezca una muestra pequeña sobre una población de 10 millones de pacientes, pero sin embargo sería suficiente con los niveles de fiabilidad requeridos: recordemos que la clave de un muestreo representativo es más dependiente de la variabilidad o dispersión que haya en la población acerca del tema en estudio que del tamaño de la muestra. En este caso, la dispersión es notable con 20,7 días en torno a la media, por lo que una muestra de ese tamaño, si está bien implementada en términos de aleatoriedad, reflejará correctamente la realidad.

Por último, se muestra el intervalo de confianza para la estimación de una media poblacional (μ). Téngase en cuenta que la desviación estándar poblacional (σ) no es un parámetro que suela conocerse a priori, por lo que casi siempre se pone el valor muestral (s) de la muestra piloto o uno supuesto o basado en información fiable.

$$P\left(\bar{x} - z_{\alpha/2} \sqrt{\frac{s^2}{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \sqrt{\frac{s^2}{n}}\right) = 1 - \alpha$$

4.3.3.1. Ejemplo real sobre aspectos del cálculo de tamaño muestral para estimar una media

En el plano real, el INE utiliza una muestra mucho más amplia para realizar la encuesta de morbilidad hospitalaria, como ya hemos visto anteriormente, de forma que, para responder a la pregunta:

¿Cuál fue la media de días de estancia en el hospital de los pacientes atendidos en los hospitales españoles en el año 2011?

Cuenta con un número de casos muy superior al que nos ha proporcionado el ejemplo anterior. En este caso, la estimación de la media de días de estancia arrojada por el INE al explotar estadísticamente los datos mediante SPSS es la siguiente:

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
Días de estancia	4607668	1	14851	6,67	29,184
N válido (según lista)	4607668				

El INE estimó una media de 6,67 días a partir de una muestra de 4.607.668 casos. La desviación típica es algo más amplia comparada con el supuesto que hemos tomado en el ejemplo anterior (29,184 días frente a los 20,7 días que hemos utilizado antes), lo que da

una idea de lo importante que es hacer una buena muestra piloto para poner este dato en la fórmula de cálculo. Con esta muestra, el INE comete un error muestral bajo:

$$e = z_{\alpha/2} \sqrt{\frac{s^2}{n}} = 1,96 \sqrt{\frac{29,184^2}{4607668}} = 0,027 \text{ días}$$

Es decir, el error es directamente de 0,027 días en este caso, por lo que el intervalo de confianza al 95% queda así:

$$P(6,67 - 0,027 = 6,643 \quad \text{y} \quad 6,67 + 0,027 = 6,697) = 0,95$$

Ratificando que 6,67 días es una buena estimación.

Una última consideración acerca del muestreo es que, en diseños complejos en que el muestro debe realizarse en varias etapas, caso frecuente en el ámbito sanitario, las fórmulas anteriores sobre cálculo de tamaño muestral pueden tener que ser aplicadas más de una vez para extraer muestras de cada una de las poblaciones objetivo de la etapa: por ejemplo, selección de un conjunto de centros (etapa 1) para, posteriormente seleccionar unidades de servicio (etapa 2), para seguir con una selección de personal sanitario (etapa 3), o de pacientes (etapa 3 alternativa). Lo mismo sucede cuando hay que analizar poblaciones estratificadas por sexo, grupo de edad, sintomatología y muchos otros casos. En este apartado se ha utilizado como trasfondo de los ejemplos la Encuesta de Morbilidad por ser una herramienta tradicional y cuya metodología se puede consultar en la web del INE. Una gran parte de estudios relacionados con especialidades médicas concretas como la hematología, la oncología, la dermatología y muchas otras, se basan en muestras de pacientes que sufren enfermedades de la sangre, cánceres específicos, problemas dermatológicos, etc. Las muestras en que se sustentan algunos de estos estudios no suelen ser probabilísticas y se depuran a base de criterios de inclusión en los estudios, selección de cohortes y otras técnicas, pero no son siempre adecuadas para establecer ejemplos generales de cálculos de tamaño muestral.

En cualquier caso, incluir la extensa materia que se refiere al muestreo excedería el ámbito de un cuaderno que resume las técnicas estadísticas esenciales, por lo que, lógicamente, aquellos que necesiten profundizar o hallar más información sobre cómo proceder a diseños de muestras más sofisticados deben consultar manuales dedicados específicamente a este apartado.

5. Tablas bivariantes o de contingencia

5.1. Definición, construcción de tablas e interpretación de los resultados mediante ejemplos aplicados

Las tablas de cruzamientos, o más técnicamente bivariantes o de contingencia, constituyen una herramienta útil para:

- Describir la relación existente entre dos variables cualitativas
- Contrastar hipótesis simples acerca de dicha relación

Por consiguiente, se pueden utilizar como parte de la estadística descriptiva y como parte de la inferencial dependiendo del objetivo que se persigue con su construcción.

Para elaborarlas, se coloca una variable en las filas y otra en las columnas de la tabla y se calculan los contenidos de las celdas que relacionan los valores de una y otra variable, así como los resultados de los márgenes o marginales y el total.

Veamos su funcionamiento e interpretación mediante unos ejemplos aplicados en el ámbito de la salud.

5.1.1. Ejemplo 1 y concepto de variable dependiente

Este ejemplo se basa en un caso hipotético cuyo enunciado en el siguiente: en un centro integral de oncología clínica donde se llevan a cabo tratamientos con quimioterapia, se realizó un estudio de evaluación sobre los pacientes acerca de aspectos de la enseñanza del autocuidado dispensada por el personal del centro desde el inicio del tratamiento. El objetivo del estudio fue la posible detección de carencias o necesidades por parte de los pacientes que pudiesen no resultar plenamente satisfactorias desde su óptica.

Para llevar a cabo el estudio, se diseñó una encuesta que se pasó a una muestra aleatoria de 100 pacientes de la población de 200 pacientes que estaban en tratamiento quimioterápico en el momento de iniciar la recogida de datos y que cumplían los criterios de inclusión. Estos criterios consistieron en incluir sólo pacientes de 20 y más años, conscientes y con buena orientación temporal y espacial. Al final de la encuesta se creó una base de datos en formato SPSS que contenía las siguientes variables:

- Código de identificación del paciente
- Sexo del paciente
- Edad del paciente en años
- Grado de satisfacción con la enseñanza de autocuidado recibida (Medida en una escala de 5 puntos donde 1 = muy insatisfecho y 5 = muy satisfecho)

- Grado de orientación percibido por los pacientes sobre los cuidados a seguir para prevenir complicaciones en el tratamiento de quimioterapia que brinda enfermería (Valores: 1 = ninguno, 2 = poco, 3 = mucho)
- Grado de enseñanza que brinda enfermería sobre los efectos adversos que produce la quimioterapia (Medido en una escala de 5 puntos donde 1 = ninguno y 5 = mucho)
- Atención rápida que brinda enfermería ante la presencia de efectos adversos (Medida en una escala de 5 puntos donde 1 = nunca y 5 = siempre)
- Enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva (Valores: 1 = insuficiente, 2 = suficiente)
- Frecuencia del cepillado dental que realizan los pacientes sometidos a tratamiento de quimioterapia (Valores: 1 vez al día, 2 veces al día, 3 veces al día)
- Elementos que utilizan para el cuidado bucal, los pacientes sometidos a tratamiento de quimioterapia (Valores: 1 = cepillo blando 2 = cepillo medio, 3 = gasas)
- Tiempo que dedican para el cepillado dental, los pacientes sometidos a tratamiento de quimioterapia (Valores: 1 = menos de 1 minuto, 2 = entre 1 y 2 minutos, 3 = entre 2 y 3 minutos)
- Utilización de pañuelos o pelucas por parte de los pacientes sometidos a tratamiento de quimioterapia (Valores: nunca, con mucha frecuencia, siempre)
- Hidratación de la piel con cremas, según opinión de los pacientes sometidos a tratamiento de quimioterapia (Nunca, con mucha frecuencia, siempre)
- Ingesta abundante de líquidos antes de asistir a un tratamiento con citostáticos, según opinión de los pacientes sometidos a tratamiento de quimioterapia (Sí, por decisión propia; Sí, por recomendación del personal; No)

El aspecto de la base de datos en formato SPSS es el siguiente:

ID	SEXO	EDAD	AUTOC	ORIENTACION	ENEFADV	ATRENF	ERCC	CEPDEN	ELCDEN	TIEMPOCO	ESTET_1	HIPEL	LIQUIDO	var
1	1	Hombre	34	Muy satisf...	Mucho	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...	
2	2	Hombre	65	Ni insatisfe...	Poco	Algo	Depende d...	Insuficiente	2 veces al...	Cepillo me...	Entre 1 y 2...	Con much...	Siempre	SI, por rec...
3	3	Hombre	45	Satisfecho	Mucho	Bastante	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Nunca	SI, por dec...
4	4	Hombre	76	Insatisfecho	Ninguno	Poco	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Siempre	SI, por rec...
5	5	Hombre	29	Muy satisf...	Mucho	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Siempre	SI, por dec...
6	6	Hombre	33	Muy satisf...	Mucho	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...
7	7	Hombre	48	Ni insatisfe...	Poco	Algo	Depende d...	Insuficiente	1 vez al día	Cepillo bla...	Menos de...	Con much...	Siempre	SI, por dec...
8	8	Hombre	46	Satisfecho	Mucho	Mucho	Casi siempre	Suficiente	3 veces al...	Cepillo me...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...
9	9	Hombre	53	Ni insatisfe...	Poco	Poco	Depende d...	Insuficiente	2 veces al...	Cepillo bla...	Entre 1 y 2...	Con much...	Con much...	SI, por dec...
10	10	Hombre	59	Satisfecho	Mucho	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Siempre	Con much...	SI, por dec...
11	11	Hombre	61	Insatisfecho	Poco	Algo	A veces	Insuficiente	1 vez al día	Cepillo me...	Entre 1 y 2...	Nunca	Con much...	SI, por dec...
12	12	Hombre	74	Ni insatisfe...	Poco	Algo	Depende d...	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Nunca	No
13	13	Hombre	30	Satisfecho	Mucho	Bastante	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Siempre	SI, por dec...
14	14	Hombre	55	Satisfecho	Mucho	Mucho	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...
15	15	Hombre	47	Ni insatisfe...	Poco	Algo	Depende d...	Insuficiente	2 veces al...	Cepillo me...	Entre 1 y 2...	Siempre	Siempre	SI, por dec...
16	16	Hombre	38	Muy satisf...	Mucho	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Siempre	No
17	17	Hombre	58	Satisfecho	Mucho	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Con much...	Siempre	SI, por dec...
18	18	Hombre	67	Ni insatisfe...	Poco	Algo	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Con much...	Siempre	SI, por dec...
19	19	Hombre	59	Insatisfecho	Poco	Bastante	Siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...
20	20	Hombre	68	Ni insatisfe...	Poco	Algo	Depende d...	Insuficiente	2 veces al...	Gasas	Menos de...	Nunca	Con much...	SI, por dec...
21	21	Hombre	44	Ni insatisfe...	Poco	Poco	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Con much...	Siempre	SI, por rec...
22	22	Hombre	69	Ni insatisfe...	Poco	Algo	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 2 y 3...	Nunca	Con much...	SI, por dec...
23	23	Hombre	78	Ni insatisfe...	Poco	Bastante	Casi siempre	Suficiente	3 veces al...	Cepillo bla...	Entre 1 y 2...	Nunca	Siempre	SI, por rec...

Ahora, en el marco de esta encuesta de pacientes en tratamiento quimioterápico, supongamos que el equipo investigador desea saber si la enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva es percibida de forma diferente por razón de sexo. En este caso, estaríamos relacionando las siguientes dos variables:

Percepción de la enseñanza sobre rasurado del cabello: cualitativa ordinal, que toma los valores 1 = insuficiente y 2 = suficiente

Sexo del paciente: cualitativa nominal, que toma los valores 1 = Hombre, 0 = mujer

Para confeccionar una tabla de frecuencias de doble entrada o cruzada, se utiliza el software estadístico SPSS, que directamente nos ofrece el siguiente resultado:

Tabla cruzada Sexo del paciente*Enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva

Recuento

		Enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva		Total
		Insuficiente	Suficiente	
Sexo del paciente	Mujer	12	32	44
	Hombre	13	43	56
Total		25	75	100

Los datos de las casillas de esta tabla se leen de la siguiente forma:

- Del total de mujeres (44) en la muestra 12 consideraron la enseñanza insuficiente, mientras que 32 la consideraron suficiente
- Del total de hombres (56) en la muestra 13 consideraron la enseñanza insuficiente, mientras que 43 la consideraron suficiente
- Del total de pacientes (100) en la muestra 25 consideraron la enseñanza insuficiente, mientras que 75 la consideraron suficiente

La fila y la columna totales se denominan marginales porque están en los márgenes y representan las distribuciones de frecuencias simples de las variables que se cruzan en la tabla.

Esta tabla proporciona información. Sin embargo, salvo casos excepcionales y descriptivos, los estadísticos nunca trabajan con cifras absolutas y menos tratándose de datos muestrales, pues no constituyen un total poblacional, por lo que hay que transformar estas cifras en porcentajes.

Dado que en una tabla cruzada se pueden ofrecer porcentajes fila y/o porcentajes columna (e incluso totales si fuese necesario), es importante determinar cuáles son los que resultan más informativos para nuestros propósitos. Ello nos lleva a considerar un nuevo concepto: el de variable dependiente y variable independiente.

Así, al establecer una relación entre dos variables, pueden producirse las siguientes situaciones:

- Que una de las variables se refiera a un fenómeno que puede ser causado por la otra variable
- Que la relación causal entre las dos variables sea recíproca y que resulte indiferente considerar a una o a otra como la dependiente
- Que no exista relación causal o de dependencia entre ninguna de las dos

Cuando se considera que existe o puede existir una relación causal unilateral, entonces a una de las variables se la denomina dependiente y a la otra independiente o explicativa, puesto que explica todo o parte del comportamiento de la otra variable.

En cuanto a la descripción que nos ocupa en el caso de los pacientes en tratamiento quimioterápico, tenemos dos variables: la percepción sobre la suficiencia de la enseñanza sobre el rasurado de cabello y el sexo de los pacientes ¿Qué tipo de relación podemos establecer entre ambas?

- El sexo depende de la percepción sobre la enseñanza
- La percepción sobre la enseñanza depende del sexo
- La dependencia es recíproca
- No tiene sentido establecer una dependencia dado que no hay base para suponer que hombres o mujeres puedan tender a valorar de forma diferente esta enseñanza

Al leer las posibilidades, el sentido común nos llevaría a considerar este caso como del segundo o del cuarto tipo. En el caso del segundo tipo, podríamos formular una hipótesis acerca de la influencia del sexo en la percepción y contrastar si es cierto que las mujeres son más propensas a valorar mejor esta enseñanza que los hombres porque, socialmente, parece que les afecta más que a los hombres el impacto estético y pueden percibir que se les da una enseñanza más extensa que a los hombres, o más insuficiente de lo que desharían, etc. En el caso del cuarto tipo, estaríamos dando por supuesto que no hay base para considerar que el sexo pueda influir en la probabilidad de tener una mejor o peor percepción sobre la enseñanza recibida.

Veamos que nos dicen los distintos tipos de porcentajes que se pueden calcular sobre esta tabla:

En esta primera tabla tenemos porcentajes fila, lo cual se indica en la esquina superior izquierda, pues la base de cálculo de estos es la variable sexo que está en las filas.

Tabla cruzada Sexo del paciente*Enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva

% dentro de Sexo del paciente

		Enseñanza que brinda enfermería sobre el rasurado del cabello para evitar la caída de forma masiva		Total
		Insuficiente	Suficiente	
Sexo del paciente	Mujer	27,3%	72,7%	100,0%
	Hombre	23,2%	76,8%	100,0%
Total		25,0%	75,0%	100,0%

Los porcentajes se leen así:

Del total de pacientes mujeres un 27,3% consideró la enseñanza insuficiente y un 72,7% suficiente

Del total de pacientes varones un 23,2% consideró la enseñanza insuficiente y un 76,8% suficiente

Del total de pacientes de la muestra un 25% consideró la enseñanza insuficiente y un 75% suficiente

Nótese que cuando utilizamos porcentajes fila, la última columna debe mostrar 100% para todas las filas.

En esta segunda tabla tenemos porcentajes columna, lo cual se indica en la esquina superior izquierda, pues la base de cálculo de estos es la variable percepción sobre la enseñanza de rasurado del cabello que está en las columnas.

Tabla cruzada Sexo del paciente*Enseñanza que brinda enfermería sobre el rasurado del cabello

% dentro de Enseñanza que brinda enfermería sobre el rasurado del cabello

		Enseñanza que brinda enfermería sobre el rasurado del cabello		Total
		Insuficiente	Suficiente	
Sexo del paciente	Mujer	48,0%	42,7%	44,0%
	Hombre	52,0%	57,3%	56,0%
Total		100,0%	100,0%	100,0%

Los porcentajes se leen así:

Del total de pacientes que opinaron que la enseñanza fue insuficiente, un 48% fueron mujeres y un 52% hombres

Del total de pacientes que opinaron que la enseñanza fue suficiente, un 42,7% fueron mujeres y un 57,3% hombres

Del total de pacientes de la muestra un 44% fueron mujeres y un 56% hombres
Cuando utilizamos porcentajes columna, la última fila debe mostrar 100% para todas las columnas.

Interpretación:

La Tabla con porcentajes fila es la que nos permite observar que la distribución de la percepción de la enseñanza en función del sexo es bastante similar: en ambos sexos, la gran mayoría de pacientes (más del 70%), consideró la enseñanza de rasurado como suficiente. La única diferencia remarcable es que, entre los hombres, la proporción de casos de percepción positiva es algo mayor. Sin embargo, para saber si esta diferencia es realmente estadísticamente significativa, es necesario plantear y contrastar una hipótesis del tipo: la proporción de mujeres con percepción positiva de la enseñanza sobre rasurado es diferente a la proporción de hombres con percepción positiva de la enseñanza sobre rasurado (contraste de diferencia de proporciones). Para poder hacerlo, hay que utilizar un estadístico de contraste denominado Chi Cuadrado y que se presenta en el próximo apartado.

En cuanto a la tabla con porcentajes columna, la información que proporciona es meramente descriptiva, pues los porcentajes muestran la composición de la muestra: por un lado hay más hombres que mujeres en la muestra, lo cual se refleja en los porcentajes y por otro la opinión acerca de la enseñanza recibida no permite pronosticar el sexo de un paciente o, en otras palabras, un paciente no es hombre o mujer por el hecho de tener una percepción negativa o positiva acerca de la enseñanza recibida sobre el rasurado del cabello. En general, en los estudios, los analistas van directamente a calcular la tabla que muestra los porcentajes de la variable independiente: si ésta está colocada en las filas, calcularán porcentajes fila y si está en las columnas calcularán porcentajes columna. El cálculo de la tabla contraria a la que debe utilizarse para contrastar la hipótesis no suele hacerse y, si se hace, suele ser con fines descriptivos, no inferenciales.

5.2. Contraste de hipótesis sobre diferencias de proporciones

5.2.1. Concepto de hipótesis estadística

Una hipótesis consiste en una explicación anticipada de lo que va a acontecer en un proceso de investigación. Incluye la suposición de las relaciones existentes entre hechos o fenómenos o entre las variables que se miden o acerca de las cuales se dispone de información. En muchas ocasiones, una hipótesis supone un eslabón entre la teoría o marco teórico y la investigación en sí misma. Lo importante de una hipótesis es que conduzca a la constatación de un supuesto, o al descubrimiento de nuevas aportaciones o que su respuesta posibilite el planteamiento de nuevas hipótesis.

Entre las características básicas de una hipótesis destacan las siguientes:

- 1.^a La hipótesis suele formularse mediante una frase que afirma o niega algo
- 2.^a La formulación de una hipótesis debe ser clara y sencilla
- 3.^a La formulación podrá ser explicativa o predictiva
- 4.^a La hipótesis debe ser susceptible de verificación
- 5.^a La hipótesis relaciona entre sí a dos o más variables
- 6.^a Es deseable que la formulación señale lo más significativo y relevante de lo que se pretende conseguir en una investigación

Para ver distintos tipos de hipótesis, resultan de utilidad estos ejemplos que pertenecen a un estudio sobre riesgo de enfermedad cardiovascular:

Hipótesis Conceptual 1: Un estilo de vida saludable previene el riesgo cardiovascular

Hipótesis Operativa 1: Media hora de ejercicio aeróbico todos los días reduce los indicadores de, al menos, dos factores de riesgo: hipertensión y obesidad medida por el IMC.

Hipótesis Conceptual 2: El abandono de malos hábitos de vida, como el sedentarismo, tabaquismo y una dieta rica en grasas, sal y azúcares, reduce el riesgo cardiovascular.

Hipótesis Operativa 2: El dejar de fumar reduce la probabilidad de volver a tener otro episodio cardíaco.

Son hipótesis de investigación muy amplias, que requieren recolección de muchos datos, un marco teórico bien estructurado, la determinación de diversas variables para medir conceptos y otras operaciones. En este cuaderno mostramos ejemplos mucho menos ambiciosos y referidos a análisis sencillos de casos hipotéticos y simulaciones como la de la encuesta de pacientes sometidos a tratamiento quimioterápico, pero hay que tener presente que los grandes progresos en investigación en ciencias de la salud suelen proceder, en muchos casos, de investigaciones en torno a hipótesis ambiciosas como las anteriores.

5.2.2. Formulación y contraste de hipótesis para tablas de contingencia

Las hipótesis para tablas cruzadas o de contingencia se formulan en sentido negativo ya que de ellas deriva el concepto de hipótesis nula.

Siguiendo con el ejemplo basado en la Encuesta a pacientes en tratamiento quimioterápico, supongamos que se desea averiguar si el uso de pañuelos, pelucas y elementos estéticos depende significativamente del sexo en la población objetivo. Para obtener una respuesta

tentativa acerca de este tema, se han seleccionado los datos que relacionan las variables sexo y frecuencia de uso de elementos estéticos y se va a contrastar la siguiente hipótesis:

H₀: la frecuencia de uso de pañuelos, pelucas y elementos estéticos durante la quimioterapia NO depende del sexo de los pacientes en la población en estudio

En otras palabras, dado que la tabla de contingencia contrasta diferencias de proporciones, estamos contrastando que las proporciones de mujeres que utilizan elementos estéticos son estadísticamente iguales que las de los hombres contra la hipótesis de que existen diferencias significativas entre las proporciones de ambos sexos.

Para contrastar este tipo de hipótesis, sabemos que las variables implicadas en su formulación son cualitativas nominales o cualitativas ordinales con pocas categorías o valores, por lo cual, lo adecuado es calcular una tabla de contingencia.

Cuando sabemos cuál es la variable dependiente y cuál la independiente, ya hemos visto que ponemos una de ellas en las filas y la otra en las columnas y seguimos la siguiente regla para calcular los porcentajes de las casillas:

- Si la variable dependiente está en las filas, hay que calcular porcentajes columna (es decir, de la variable independiente)
- Si la variable dependiente está en las columnas, hay que calcular porcentajes fila (es decir, de la variable independiente)

En nuestro caso, la variable dependiente es la frecuencia de uso de estéticos y la independiente el sexo. Hemos puesto en las filas el sexo y en las columnas la frecuencia de uso de estéticos, por lo que tenemos que trabajar con porcentajes fila.

Tabla cruzada Sexo del paciente*Utilización de pañuelos o pelucas por parte de los pacientes sometidos a tratamiento de quimioterapia

% dentro de Sexo del paciente		Utilización de pañuelos o pelucas por parte de los pacientes sometidos a tratamiento de quimioterapia			Total
		Nunca	Con mucha frecuencia	Siempre	
Sexo del paciente	Mujer	2,3%	27,3%	70,5%	100,0%
	Hombre	71,4%	19,6%	8,9%	100,0%
Total		41,0%	23,0%	36,0%	100,0%

En el resultado ya hemos visto que, a primera vista, los porcentajes de casos de frecuencia de uso de estéticos son diferentes para cada sexo (por ejemplo, un 70,5% de las mujeres los usan siempre frente a un 8,9% en el caso de los hombres).

Sin embargo, la pregunta para llegar a una conclusión acerca de si la frecuencia de uso de estéticos depende estadísticamente en algún grado del sexo del paciente, es si las diferencias entre estas proporciones o porcentajes son suficientemente amplias como para resultar estadísticamente significativas.

La respuesta acerca de esta pregunta de si las proporciones son realmente diferentes desde el punto de vista estadístico, nos la proporciona el estadístico de contraste Chi Cuadrado (X^2). Lo que nos dice la Chi Cuadrado es si los sujetos que observamos en cada casilla son los que veríamos si no hubiera diferencias ni relaciones entre los criterios de clasificación. Al estadístico Chi Cuadrado se le denomina por esta razón prueba de independencia (o lo contrario, de asociación), porque verificamos si los criterios de clasificación son independientes. La prueba de la Chi Cuadrado nos va a decir si lo que observamos:

- a) Está dentro de lo normal y probable; en ese caso afirmaremos que no hay diferencia ni relación (aceptamos la Hipótesis Nula o, en otros términos, aceptamos que los resultados están dentro de lo aleatorio).
- b) Es atípico y poco normal en el caso de no diferencia o relación; en este caso sí afirmaremos que hay relación entre los criterios de clasificación o que los grupos son distintos (rechazamos la Hipótesis Nula y, por consiguiente, concluimos que una variable depende de la otra en alguna medida).

Se trata en definitiva en verificar la probabilidad de que ocurra casualmente lo que nos hemos encontrado en el caso de que no hubiera ni diferencias ni relaciones. El estadístico Chi cuadrado lo calcula automáticamente SPSS y su valor se obtiene aplicando la siguiente fórmula:

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Donde:

- f_o son las frecuencias observadas, número de individuos que hay en cada casilla de la tabla
- f_e son las frecuencias teóricas o que tendrían que aparecer en cada casilla si los grupos de la variable independiente estuviesen distribuidos de la misma forma

A efectos prácticos, el valor de este estadístico de contraste viene acompañado de un valor de significación (bilateral en este caso) que debe interpretarse de la siguiente forma para un nivel de confianza del 95%:

- Si la significación es menor o igual que 0,025, se rechaza la hipótesis, de forma que las proporciones de los grupos comparados son diferentes
- Si la significación es mayor que 0,025, se acepta la hipótesis, de forma que las proporciones de los grupos comparados deben considerarse iguales

Para el contraste que nos ocupa, el valor del estadístico Chi Cuadrado y su significación se pueden ver en la siguiente tabla calculada por SPSS junto a la tabla cruzada:

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	55,275 ^a	2	,000
Razón de verosimilitud	66,930	2	,000
Asociación lineal por lineal	54,279	1	,000
N de casos válidos	100		

a. 0 casillas (0,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 10,12.

La significación es el área que deja el valor de la X^2 a su derecha o a su izquierda (zona $\alpha/2$). Si ese valor es menor o igual que 0,025, quiere decir que el estadístico de contraste está en la zona de rechazo de la hipótesis y, si es mayor que está en la zona de aceptación. Para una mejor comprensión, véase la Figura 15 donde se muestra la distribución Chi Cuadrado y sus características.

Figura 15. Distribución Chi Cuadrado



Fuente: Universidad Nacional de Colombia, cursos de estadística

En la figura, la zona de no rechazo o aceptación de la hipótesis es la zona central que tiene un valor de probabilidad del 95% o dividido por cien de valor 0,95 ($1 - \alpha$). La zona de rechazo de la hipótesis que tiene un valor de probabilidad del 5% o dividido por cien, de valor $0,05 = \alpha$, se presenta subdividida en dos zonas de valor 2,5% o $\alpha/2 = 0,025$: por eso se dice que el contraste es bilateral o a dos colas. La significación es el área que deja el valor de la X^2 a su derecha o a su izquierda (zona $\alpha/2$). La significación es el área que deja el valor de la X^2 a su derecha o a su izquierda (zona $\alpha/2$). Si ese valor es menor o igual que 0,025, quiere decir que el estadístico de contraste está en la zona de rechazo de la hipótesis y, si es mayor que está en la zona de aceptación. En nuestro caso, el valor de la Chi cuadrado es de 55,275 y el valor de la significación es de 0,000. Como 0,000 es menor que 0,025, debemos rechazar la hipótesis formulada:

H₀: La frecuencia de uso de estéticos NO depende del sexo de los pacientes de la población en estudio

El rechazo de la hipótesis, dado que esta formulada en sentido negativo, nos lleva a concluir que la frecuencia de uso de estéticos SÍ que depende en algún grado del sexo de los pacientes en esta población, por lo que, podemos interpretar las diferencias de proporciones que aparecen en la tabla y ver en qué sentido se da dicha dependencia:

Las pacientes femeninas muestran una proporción del 70,5% de uso continuado de estéticos, frente a un 8,9% entre los pacientes masculinos. Y, éstos muestran una proporción del 71,4% de uso nulo de estéticos frente a un 2,3% entre las pacientes femeninas. En conclusión, en esta población de pacientes, el uso de estéticos está significativamente mucho más extendido entre la mujer.

Una vez contrastada una hipótesis, SPSS también nos proporciona una medida del grado de dependencia entre dos variables. Esta medida se llama V de Cramer y viene expresada en forma de tanto por uno, de manera que multiplicándola por 100, obtenemos el grado de dependencia en porcentaje. Este valor es importante para calibrar la intensidad de la dependencia, de forma que podamos decir si los grupos comparados son realmente muy diferentes, moderadamente diferentes, etc. Téngase en cuenta que, aunque las variables sean independientes, la V de Cramer siempre nos da un valor mayor que cero, pues siempre existe un grado de dependencia, aunque sea irrelevante. Lo importante es si es significativo o no. Así, en las tablas del ejemplo, como se ve aquí debajo, obtenemos una dependencia del 74,3%, que es muy notable en el contexto del análisis. Como referencia, los grados de dependencia inferiores al 15% son más bien despreciables.

Medidas simétricas

		Valor	Significación aproximada
Nominal por Nominal	Phi	,743	,000
	V de Cramer	,743	,000
N de casos válidos		100	

5.2.3. Ejemplos complementarios de contraste de hipótesis dentro del mismo caso de estudio

El siguiente ejemplo contrasta la hipótesis nula:

H₀: El tipo de elemento de cuidado bucal que utilizan los pacientes NO depende del número de cepillados dentales diario

La tabla bivariable y los estadísticos de contraste que se obtienen son:

Tabla cruzada Frecuencia del cepillado dental que realizan los pacientes*Elementos que utilizan para el cuidado bucal, los pacientes

% dentro de Frecuencia del cepillado dental que realizan los pacientes

		Elementos que utilizan para el cuidado bucal, los pacientes			Total
		Cepillo blando	Cepillo medio	Gasas	
Frecuencia del cepillado dental que realizan los pacientes	1 vez al día	33,3%	55,6%	11,1%	100,0%
	2 veces al día	34,8%	52,2%	13,0%	100,0%
	3 veces al día	88,2%	10,3%	1,5%	100,0%
Total		71,0%	24,0%	5,0%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	30,899 ^a	4	,000
Razón de verosimilitud	30,207	4	,000
Asociación lineal por lineal	23,593	1	,000
N de casos válidos	100		

a. 4 casillas (44,4%) han esperado un recuento menor que 5. El recuento mínimo esperado es ,45.

Medidas simétricas

		Valor	Significación aproximada
Nominal por Nominal	Phi	,556	,000
	V de Cramer	,393	,000
N de casos válidos		100	

Según los resultados, la hipótesis nula se rechaza porque la significación de la Chi cuadrado vale 0,000 que es un valor menor que 0,025. La V de Cramer nos indica que el tipo de elemento que usan los pacientes para el cepillado dental depende en un 39,2% del número de cepillados diario. El sentido de la dependencia es el siguiente: los pacientes que llevan a cabo uno o dos cepillados tienden a utilizar en mayor medida un cepillo de tipo medio (55,6% y 52,2% de los casos), mientras que los que llevan a cabo tres cepillados tienden a utilizar uno blando (88,2% de los casos).

Finalmente, el siguiente ejemplo, contrasta la hipótesis nula:

Ho: El grado de satisfacción con la enseñanza recibida sobre el autocuidado NO depende del sexo de los pacientes

La tabla bivariable y los estadísticos de contraste que se obtienen son:

Tabla cruzada Sexo del paciente*Grado de satisfacción con la enseñanza de auto-cuidado recibida

% dentro de Sexo del paciente

		Grado de satisfacción con la enseñanza de auto-cuidado recibida				Total
		Insatisfecho	Ni insatisfecho ni satisfecho	Satisfecho	Muy satisfecho	
Sexo del paciente	Mujer	4,5%	25,0%	45,5%	25,0%	100,0%
	Hombre	5,4%	30,4%	39,3%	25,0%	100,0%
Total		5,0%	28,0%	42,0%	25,0%	100,0%

Pruebas de chi-cuadrado				Medidas simétricas			
	Valor	gl	Significación asintótica (bilateral)		Valor	Significación aproximada	
Chi-cuadrado de Pearson	,508 ^a	3	,917	Nominal por Nominal	Phi	,071	,917
Razón de verosimilitud	,510	3	,917		V de Cramer	,071	,917
Asociación lineal por lineal	,167	1	,683				
N de casos válidos	100					100	

a. 2 casillas (25,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 2,20.

Según los resultados, la hipótesis nula se acepta porque la significación de la Chi cuadrado vale 0,917 que es un valor mayor que 0,025. La V de Cramer nos indica que la satisfacción depende en un 7% del sexo, un porcentaje despreciable. En consecuencia, el grado de satisfacción con la enseñanza de autocuidados se distribuye de la misma forma entre mujeres y hombres, no teniendo lugar ninguna diferencia significativa en esta población a este respecto.

6. Contraste de diferencia de medias o análisis de la varianza

6.1. Conceptualización

Los contrastes de diferencias de medias se aplican cuando hay que:

- Describir la relación existente entre una variable dependiente cuantitativa y una variable independiente cualitativa que clasifica a los individuos de la muestra en distintos grupos
- Contrastar hipótesis simples acerca de dicha relación en que el argumento es que las medias de la variable dependiente son iguales entre los grupos establecidos

Por consiguiente, se utilizan mayormente en el marco de la estadística inferencial con el objetivo de comparar el comportamiento de la variable dependiente entre los grupos que se establezcan.

Veamos su funcionamiento e interpretación mediante ejemplos.

6.2. Ejemplos e interpretación de resultados

Una gran residencia de mayores está en período de realización de revisión médica bimensual de su conjunto de usuarios, por lo que estos son sometidos, entre otras pruebas a un análisis de sangre y revisión cutánea que forma parte del procedimiento regular de control. Aprovechando la revisión, el personal sanitario ha realizado y recopilado los resultados de algunas pruebas especiales para una amplia muestra (100 casos) del conjunto de usuarios en una base de datos para proceder a realizar un estudio sobre la prevalencia de melanomas malignos en el colectivo y las posibles causas que lo originan en los casos positivos.

Por consiguiente, se dispone de una base de datos (ficticia en este caso) acerca de una muestra de 100 usuarios de una gran residencia de mayores que han sido sometidos a diversas pruebas en su revisión médica bimensual. La base de datos contiene resultados de análisis de sangre (nivel de hematíes, de hemoglobina y hematocrito) además de algunas características de los usuarios (sexo, edad, tiempo en la residencia, antecedentes de tabaquismo, realización regular de ejercicio) y la presencia o no de algún tipo de melanoma maligno, y su aspecto en formato SPSS es el siguiente:

	IgE	HEMATIES	RHEMATIES	Hb	Hto	SEXO	EDAD	TRESI	TRESI_I	TABACO	EJERCICIO	HORASES	MELANOMA
1	10	5554342.00	Nivel normal	16.30	48.00	Hombre	75.00	8.00	Menos de 12 y	No	No	.00	No
2	16	4900213.00	Nivel normal	16.40	47.60	Hombre	80.00	12.00	Entre 12 y 24	No	Si	6.00	No
3	18	5089765.00	Nivel normal	15.10	48.40	Hombre	80.00	17.00	Entre 12 y 24	No	Si	7.00	No
4	19	5976543.00	PoliGlobulia	19.30	55.30	Hombre	88.00	18.00	Entre 12 y 24	Ex-fumador	No	.00	No
5	20	3453456.00	Anemia	9.60	33.00	Hombre	73.00	13.00	Entre 12 y 24	No	Si	2.00	No
6	20	5243521.00	Nivel normal	17.20	50.30	Hombre	80.00	9.00	Menos de 12 y	No	Si	5.80	No
7	23	4665347.00	Nivel normal	17.10	46.80	Hombre	84.00	26.00	Entre 12 y 24	No	Si	8.00	No
8	24	6100000.00	PoliGlobulia	19.60	55.30	Hombre	90.00	20.00	Entre 12 y 24	Si	No	.00	No
9	25	4500000.00	Nivel normal	13.50	41.00	Hombre	83.00	20.00	Entre 12 y 24	No	Si	6.00	No
10	30	3467524.00	Anemia	11.30	34.00	Hombre	76.00	12.00	Entre 12 y 24	No	Si	6.00	No
11	40	3476565.00	Anemia	10.30	28.97	Hombre	70.00	8.00	Menos de 12 y	No	Si	5.50	No
12	40	5900000.00	Nivel normal	17.50	53.00	Hombre	81.00	7.00	Menos de 12 y	No	Si	7.00	No
13	50	5856432.00	Nivel normal	15.30	46.30	Hombre	81.00	4.00	Menos de 12 y	No	Si	6.00	No
14	55	5346476.00	Nivel normal	14.90	50.80	Hombre	84.00	12.00	Entre 12 y 24	No	No	.00	No
15	60	5400876.00	PoliGlobulia	17.80	48.20	Mujer	84.00	23.00	Entre 12 y 24	Si	No	.00	No
16	60	6466751.00	Nivel normal	14.80	44.30	Hombre	75.00	14.00	Entre 12 y 24	No	Si	6.00	No
17	60	4677345.00	Nivel normal	16.20	46.00	Hombre	82.00	18.00	Entre 12 y 24	No	Si	7.00	No
18	65	6012344.00	PoliGlobulia	18.30	54.70	Hombre	93.00	26.00	Entre 12 y 24	Si	No	6.50	No
19	66	5987633.00	PoliGlobulia	19.30	54.90	Hombre	69.00	3.00	Menos de 12 y	Ex-fumador	No	.00	No
20	66	5976543.00	PoliGlobulia	14.90	56.70	Hombre	83.00	14.00	Entre 12 y 24	Ex-fumador	No	.00	No
21	67	5644534.00	Nivel normal	16.80	47.60	Hombre	83.00	18.00	Entre 12 y 24	No	Si	6.00	No
22	67	5354356.00	Nivel normal	14.20	61.20	Hombre	85.00	20.00	Entre 12 y 24	No	Si	7.00	No
23	78	4355621.00	Nivel normal	13.40	36.90	Mujer	78.00	8.00	Menos de 12 y	No	Si	4.00	No

Los investigadores, en primer lugar, se han interesado por averiguar si el nivel de IgE (inmoglobulina E), habitualmente utilizado como posible marcador de determinadas de alergias, prurito y otras afecciones cutáneas, depende en algún grado del intervalo de tiempo que llevan los usuarios en la residencia.

Para responder a esta pregunta, ha llevado a cabo un contraste de diferencia de medias en que la variable dependiente es el nivel de IgE y la independiente el tiempo que llevan los usuarios en la residencia establecido en 3 intervalos (menos de 12 meses, entre 12 y 24 y más de 24 meses).

En este caso, estaríamos relacionando las siguientes dos variables:

- Nivel de IgE: variable cuantitativa continua medida en kU/L
- Tiempo en la residencia: cualitativa ordinal, que toma los valores 1 = menos de 12 meses, 2 = entre 12 y 24 meses, 3 = más de 24 meses

Para confeccionar una descripción de las medias de nivel de IgE en cada grupo de usuarios clasificados según el tiempo de estancia, se utiliza el software estadístico SPSS, que directamente nos ofrece el siguiente resultado tras llevar a cabo el siguiente programa de instrucciones para programar un ANOVA o análisis de la varianza:

ONEWAY IgE BY TRESI_I /STATISTICS DESCRIPTIVES/MISSING ANALYSIS.

Esta programación se escribe utilizando las ventanas de los menús SPSS pulsando: Analizar, Comparar medias, ANOVA de un factor, variable dependiente IgE variable independiente TRESI_I, Opciones descriptivos, continuar, aceptar.

Los resultados que se obtienen son:

Descriptivos

Nivel de IgE en kU/L

	N	Media	Desviación estándar	Error estándar	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
Menos de 12 meses	18	6,4989	9,45686	2,22900	1,7961	11,2017	,10	30,66
Entre 12 y 24 meses	49	5,4131	9,40478	1,34354	2,7117	8,1144	,16	35,76
Más de 24 meses	33	22,0276	7,67720	1,33643	19,3054	24,7498	,88	37,45
Total	100	11,0913	11,70271	1,17027	8,7692	13,4134	,10	37,45

ANOVA

Nivel de IgE en kU/L

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	5906,372	2	2953,186	37,436	,000
Dentro de grupos	7652,005	97	78,887		
Total	13558,377	99			

Veamos cómo se leen e interpretan estos resultados. En la primera tabla, de izquierda a derecha tenemos la siguiente información:

- Tiempo en la residencia: tres grupos de usuarios (Menos de 12 meses, de 12 a 24 y más de 24 meses)
- Tamaño de cada grupo (N = 18 casos de menos de 12 meses, 49 entre 12 y 24 meses y 33 casos de más de 24 meses)
- Media de nivel de IgE en kU/L de cada grupo
- Desviación típica o estándar de cada media de IgE en kU/L
- Error típico de la media (sirve para calcular los intervalos de confianza de la media)
- Intervalo de confianza de cada media al 95% (límites inferior y superior)
- Valor mínimo y máximo de nivel de IgE en cada grupo de usuarios

Interpretación:

En el grupo de usuarios que residen menos de 12 meses, la media de nivel de IgE fue de 6,4989 kU/L con una dispersión de 9,45686 kU/L en torno a la media. El intervalo de confianza de la media de este grupo (1,7961 – 11,2017) presenta unos límites que están a relativamente a cierta distancia de la media (con un rango de 9,4056), por lo que se concluye que los usuarios son bastante heterogéneos en cuanto a nivel de IgE. Dado que los niveles normales de IgE son valores de 0,35 kU/L o menores (aunque la referencia depende de la edad), se puede interpretar que una parte sustancial de este grupo de usuarios tiene el nivel de IgE elevado indicando presencia de problemas de alergias tóxicas (o de otros tipos), así como posible presencia de pruritos y otros problemas dermatológicos. El valor mínimo

de la variable en este grupo queda por debajo del nivel normal ($0,10 < 0,35$), indicando que también hay usuarios sin problemas aparentes, mientras que el valor máximo es de 30,66, delatando la presencia de casos claramente candidatos a desarrollo de problemas cutáneos.

En el grupo de usuarios que residen entre 12 y 24 meses, la media de nivel de IgE fue de 5,4131 kU/L con una dispersión de 9,40478 kU/L en torno a la media. El intervalo de confianza de la media de este grupo (2,7117 – 8,1144) es más estrecho que el anterior (con un rango de 5,4027), por lo que, aunque los valores mínimo y máximo son superiores a los del primer grupo, el nivel de IgE de estos usuarios es algo más homogéneo. En este contexto, dado que los parámetros normales de IgE son valores de 0,35 kU/L o menores (aunque la referencia depende de la edad), también se puede interpretar que una parte sustancial de este grupo de usuarios tiene el nivel de IgE elevado indicando presencia de alergias tóxicas (o de otros tipos), así como posible presencia de pruritos y otros problemas dermatológicos, si bien, la media del nivel de IgE es inferior a la del primer grupo.

En el grupo de usuarios que residen más de 24 meses, la media de nivel de IgE fue de 22,0276 kU/L con una dispersión de 7,67720 kU/L en torno a la media. El intervalo de confianza de la media de este grupo (19,3054 – 24,7498) también es más estrecho que el primero (con un rango de 5,4444), pero los valores mínimo y máximo superan a los del primer y segundo grupo. Por consiguiente, el nivel de IgE de estos usuarios es más homogéneo que en el primer grupo, pero dentro de unos niveles del marcador sustancialmente más elevados, cabiendo esperar más casos positivos de alergias y otros problemas cutáneos dentro del mismo.

En definitiva: a la vista de esta tabla, de forma intuitiva, parece que en esta residencia los usuarios que llevan más tiempo muestran un cuadro más preocupante que los que llevan menos tiempo acerca del nivel de IgE y su papel como marcador de problemas dermatológicos.

En otras palabras, de la información descriptiva de la tabla, se desprende que las medias de nivel de IgE de cada grupo de usuarios son aparentemente diferentes:

6,4989 kU/L en el grupo de residentes de menos de 12 meses

5,4131 kU/L en el grupo de residentes de entre 12 y 24 meses

22,0276 kU/L en el grupo de residentes de más de 24 meses

La media global del colectivo es de 11,0913 kU/L

Sin embargo, para poder afirmar con rigor que estas diferencias son o no estadísticamente significativas y que la media más elevada corresponde al grupo de residentes de más de 24 meses y la más baja al grupo de residentes de entre 12 y 24 meses, debemos contrastar

una hipótesis. Así, en casos como este en que se contrasta una diferencia de medias de una variable para diferentes grupos, la hipótesis se formula de la siguiente forma:

Ho: media de nivel de IgE del grupo de residentes de menos de 12 meses = media de nivel de IgE del grupo de residentes de entre 12 y 24 meses = media de nivel de IgE del grupo de residentes de más de 24 meses

Al plantear la hipótesis, estamos utilizando la parte inferencial en lugar de la descriptiva, por lo que el supuesto se refiere a lo que sucede en la población total de usuarios de la residencia. Lo que nosotros vemos en la tabla son las medias de cada grupo en la muestra de usuarios y las medias poblacionales son desconocidas.

El estadístico de contraste se denomina F de Snedecor (su valor se muestra en la quinta columna de la siguiente tabla) y no vamos a entrar en su formulación, sino en la interpretación del resultado que aparece en la tabla que ha proporcionado SPSS junto a la descriptiva anteriormente comentada.

ANOVA

Nivel de IgE en kU/L

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	5906,372	2	2953,186	37,436	,000
Dentro de grupos	7652,005	97	78,887		
Total	13558,377	99			

Desde un punto de vista práctico si el valor de la significación (Sig.) es mayor o igual que 0,05, para un nivel de confianza del 95% aceptaríamos la hipótesis nula, mientras que para un valor menor que 0,05 la rechazaríamos. En nuestro caso, la significación (0,000) es menor que 0,05, por lo que rechazamos la hipótesis.

La técnica general que se emplea para contrastar esta hipótesis se denomina ANOVA (Análisis de la Varianza) y los cálculos del estadístico de contraste los realiza SPSS automáticamente, como hemos visto.

Al rechazar esta hipótesis:

Ho: media de nivel de IgE del grupo de residentes de menos de 12 meses = media de nivel de IgE del grupo de residentes de entre 12 y 24 meses = media de nivel de IgE del grupo de residentes de más de 24 meses

Estamos diciendo que las medias son estadísticamente diferentes, lo que nos lleva a concluir que, efectivamente, en la población de usuarios, los que llevan más tiempo tienen niveles sustancialmente más elevados que los que llevan entre 12 y 24 meses y los que llevan menos de 12 meses, si bien, éstos también superan al segundo grupo. Sin embargo, dado que estamos comparando las medias de tres grupos y hay dos que se parecen bastante como son el grupo de menos de 12 meses y de entre 12 y 24 meses, estaría bien poder comparar cada

grupo con el resto y ver si las diferencias son significativas en general, o lo son grupo a grupo. Para ello, SPSS nos ofrece una tabla adicional que nos permite realizar esta operación:

Comparaciones múltiples

Variable dependiente: Nivel de IgE en kU/L

Tamhane

(I) Intervalo de tiempo en la residencia	(J) Intervalo de tiempo en la residencia	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Menos de 12 meses	Entre 12 y 24 meses	1,08583	2,60261	,967	-5,4924	7,6641
	Más de 24 meses	-15,52869*	2,59894	,000	-22,1079	-8,9495
Entre 12 y 24 meses	Menos de 12 meses	-1,08583	2,60261	,967	-7,6641	5,4924
	Más de 24 meses	-16,61451*	1,89503	,000	-21,2399	-11,9891
Más de 24 meses	Menos de 12 meses	15,52869*	2,59894	,000	8,9495	22,1079
	Entre 12 y 24 meses	16,61451*	1,89503	,000	11,9891	21,2399

*. La diferencia de medias es significativa en el nivel 0.05.

En esta tabla de comparaciones múltiples, podemos ver en la primera columna y casilla, el valor de las diferencias de medias de nivel de IgE del grupo de residentes de menos de 12 meses con cada uno de los otros dos y que la diferencia respecto de la del grupo de residentes de más de 24 meses es significativa ($\text{Sig } 0,000 < 0,05$), mientras que la diferencia respecto de la del grupo de residentes de entre 12 y 24 meses no es significativa ($\text{Sig } 0,967 > 0,05$). Por consiguiente, tras este análisis complementario podemos afirmar que la media de nivel de IgE de residentes de menos de 12 meses y de residentes de entre 12 y 24 meses es estadísticamente igual y ambas diferentes de la de residentes de más de 24 meses y que la diferencia indica que los residentes de hasta 24 meses tienen un nivel más bajo, mientras que los residentes de más de 24 meses tienden a presentar mayor nivel de IgE y, por consiguiente, mayor probabilidad de desarrollo de problemas cutáneos.

Seguidamente, los investigadores realizaron el mismo tipo de análisis tomando como variable independiente el tipo de resultado obtenido en el análisis de sangre, especialmente en cuanto a la presencia de anemia, otro factor que se asocia al problema en estudio. Los resultados proporcionados por SPSS fueron los siguientes:

Descriptivos

Nivel de IgE en kU/L

	N	Media	Desviación estándar	Error estándar	95% del intervalo de confianza para la media		Mínimo	Máximo
					Límite inferior	Límite superior		
Nivel normal	57	8,6712	10,96275	1,45205	5,7624	11,5800	,10	37,45
Anemia	24	20,8096	9,46364	1,93176	16,8134	24,8057	,20	35,76
Poliglobulia	19	6,0758	9,68913	2,22284	1,4058	10,7458	,19	28,97
Total	100	11,0913	11,70271	1,17027	8,7692	13,4134	,10	37,45

ANOVA^a

Nivel de hematíes o glóbulos rojos/ml

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Entre grupos	9,829E+12	2	4,914E+12	38,721	,000
Dentro de grupos	5,965E+12	47	1,269E+11		
Total	1,579E+13	49			

a. Sexo de los trabajadores = Mujer

Comparaciones múltiples

Variable dependiente: Nivel de IgE en KU/L
Tamhane

(I) Resultado nivel hematíes	(J) Resultado nivel hematíes	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Nivel normal	Anemia	-12,13836*	2,41664	,000	-18,1091	-6,1676
	Poliglobulia	2,59544	2,65508	,706	-4,0657	9,2566
Anemia	Nivel normal	12,13836*	2,41664	,000	6,1676	18,1091
	Poliglobulia	14,73379*	2,94495	,000	7,3816	22,0860
Poliglobulia	Nivel normal	-2,59544	2,65508	,706	-9,2566	4,0657
	Anemia	-14,73379*	2,94495	,000	-22,0860	-7,3816

*. La diferencia de medias es significativa en el nivel 0.05.

Como se aprecia en las tablas, los usuarios están ahora clasificados según el resultado del análisis en: casos con anemia, con nivel normal de hematíes o con poliglobulia. Por consiguiente, la hipótesis que se contrasta en este caso es:

Ho: La media de nivel de IgE de los usuarios con nivel normal de hematíes es igual a la media de nivel de IgE de los que tienen anemia y de los que tienen poliglobulia.

Según el resultado del contraste ANOVA, la hipótesis general se rechaza porque la significación de la F vale 0,000, valor que es menor que 0,05. Al rechazar la hipótesis, concluimos que, en principio, las medias son diferentes entre sí, siendo muy superior la correspondiente a los casos con anemia comparada con las otras dos. Por otro lado, al estar trabajando con tres grupos, es necesario atender a los resultados de comparaciones múltiples antes de concluir que las tres medias son diferentes entre sí como sugiere el resultado general. Así, observando la tabla de comparaciones múltiples, concluimos que la media de nivel de IgE de los que tienen nivel normal de hematíes es estadísticamente igual que la de los usuarios que tienen poliglobulia, ya que, en este caso, la significación de la comparación entre estas dos medias vale 0,706, valor superior a 0,05 y que conduce a aceptar la hipótesis parcial de que estas dos medias son iguales.

Para finalizar esta sección, es importante comentar que la prueba o test ANOVA es la técnica general que se puede utilizar para cualquier caso en que haya que comparar dos o más medias, cubriendo el espectro de este problema. Sin embargo, cuando se comparan dos medias, muchos analistas utilizan otra prueba denominada T, que ofrece algunos detalles adicionales de información que pueden resultar de interés. La inclusión de estas técnicas en este cuaderno excede los objetivos de resumen planteados acerca de su confección, por lo que se remite al lector a los manuales generales sobre inferencia estadística donde hallará ésta y otras opciones de análisis.

7. CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

7.1. Conceptualización

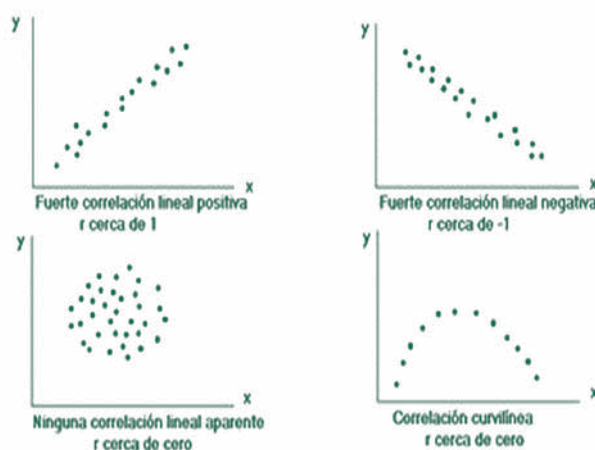
Muchos de los trabajos de investigación que se realizan en las diferentes disciplinas de las ciencias médicas y de la salud, centran su atención en la determinación de la relación existente entre dos variables cuantitativas.

Por ejemplo, en estudios de oncología se puede desear calcular cuál es la relación existente entre la edad y diversos indicadores cancerígenos; entre diversos indicadores de cáncer de pulmón y tiempo como fumador; entre indicadores genéticos y de características de un tipo de cáncer, etc. En el ámbito epidemiológico, es constante la aplicación de estas técnicas para analizar cómo varía en el laboratorio la respuesta de un animal a un fármaco a medida que varía la dosis; o cuál es la relación que existe entre la presión sanguínea y la edad; o si existe una relación entre dos determinaciones fisiológicas y bioquímicas obtenidas en los mismos pacientes.

La lista de ejemplos que servirían para ilustrar problemas comunes para investigadores que trabajan en este campo es interminable y se plasma en la dilata literatura científica y académica que ofrece estudios y artículos en que se emplean estas técnicas estadísticas.

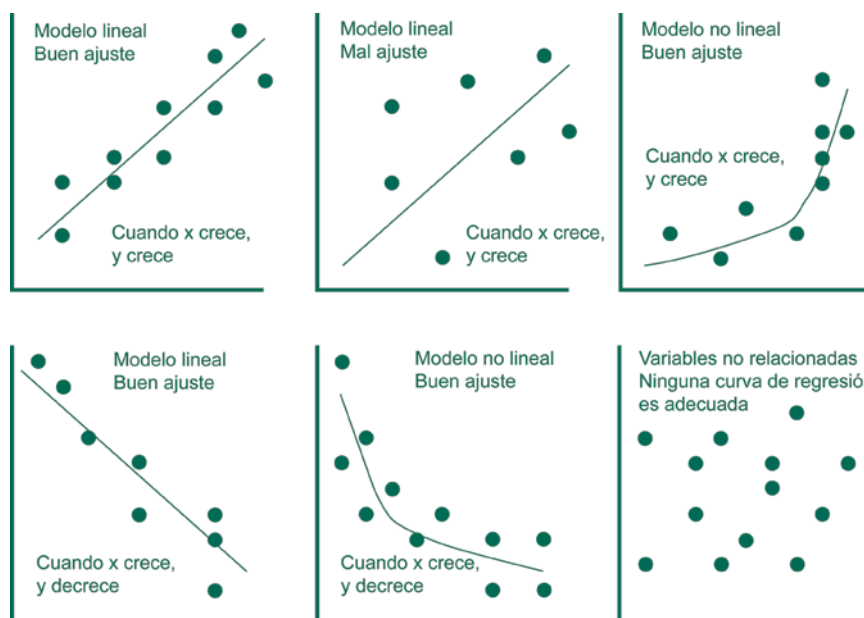
La correlación y la regresión constituyen las técnicas estadísticas utilizadas para investigar el tipo de relaciones antes descritas. Ambas técnicas están relacionadas, pero tienen propósitos diferentes: la correlación es útil para medir si la relación que existe entre las variables es aproximadamente lineal o no y con qué intensidad y signo, mientras que la regresión sirve para buscar la ecuación matemática que mejor describe la relación entre las variables. Las Figuras 16 y 17 ilustran ambos conceptos y propósitos.

Figura 16. Correlación lineal: patrones de comportamiento de relaciones entre los valores de dos variables cuantitativas



Fuente: <http://inferencialuic.blogspot.com/2017/11/regresion-lineal.html>

Figura 17. Ajustes de regresión más habituales sobre relaciones entre los valores de dos variables cuantitativas



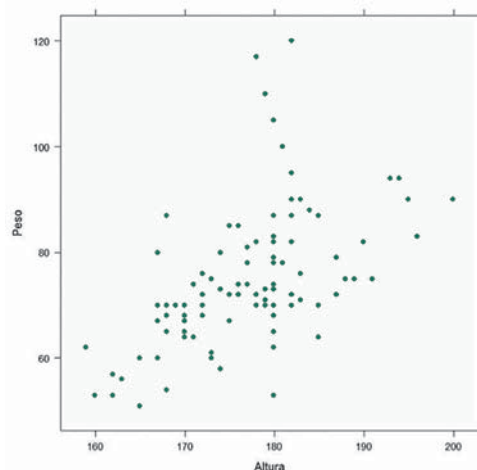
Fuente: <http://inferencialuic.blogspot.com/2017/11/regresion-lineal.html>

7.1.1. Correlación lineal

Se dice que dos variables están correlacionadas cuando los cambios de una de ellas se acompañan de cambios en la otra variable en una misma unidad de observación. La correlación lineal es un método que permite medir si la relación que existe entre dos variables es aproximadamente lineal o no, es decir, si su representación gráfica da lugar, aunque sea aproximadamente, a una línea recta. Si los resultados de la medición de una variable se incrementan con el incremento de la otra, entonces la correlación es positiva; mientras que, si la relación es inversa, entonces la correlación es negativa.

Para estudiar una relación entre dos variables cuantitativas, lo primero que se hace es representar gráficamente a través de un diagrama de dispersión (ver Figura 18) tal relación, llamado así porque lo que hace es mostrar la dispersión de los pares de valores de las dos variables.

Figura 18. Ejemplo de diagrama de dispersión que muestra la relación entre la talla (eje de abscisas) y el peso (eje de ordenadas)



Fuente: Universidad de Granada

Posteriormente, se calcula el coeficiente de correlación lineal (r), estadístico que cuantifica el grado de relación lineal entre dos variables cuantitativas y que puede tomar valores entre -1 y +1, mediante esta fórmula:

$$r = \frac{S_{XY}}{S_X S_Y}$$

El numerador de esta expresión representa la covarianza entre dos variables y el denominador es el producto de sus desviaciones estándar. El coeficiente r no tiene unidades. La interpretación de r , un coeficiente que no tiene unidades ya que en su cálculo se anulan las de las variables cuya relación describe, depende principalmente de los detalles de la investigación y de la experiencia propia en el tema de estudio. No obstante, es bastante común aceptar el siguiente criterio de interpretación:

Valores de 0 – 0,25: Se considera la relación lineal escasa o nula

Valores de 0,25 – 0,50: Relación lineal positiva débil

Valores de 0,50 – 0,75: Relación lineal positiva entre moderada y fuerte

Valores de 0,75 – 1,00: Relación lineal positiva entre fuerte y perfecta

En caso de coeficientes de correlación con signo negativo, la interpretación de la intensidad es la misma, variando sólo su signo.

Sin entrar en la justificación de la fórmula del coeficiente de correlación lineal de Pearson, es importante recordar que:

- El coeficiente sólo nos informa acerca de la existencia o no de una relación lineal, de su signo y de su intensidad
- En ningún caso su valor se puede interpretar como una medida del grado de dependencia entre dos variables cuantitativas
- En caso de existir una base para el análisis de esa dependencia, los estadísticos que miden esa relación nos proporciona el análisis de regresión lineal, no el coeficiente de correlación
- Finalmente, si el coeficiente se calcula por medio de datos muestrales, se puede contrastar la hipótesis de que su valor es distinto de cero. Si la significación que acompaña al valor del coeficiente es menor que 0,05 (o que 0,025 si el contraste es bilateral), se rechazará la hipótesis de que r sea igual a cero y existirá una relación lineal entre dos variables por débil que esta sea

7.1.2. Ejemplos de análisis de correlación lineal

Siguiendo con el análisis de los datos referentes a los usuarios de la residencia utilizado en el apartado 6, los analistas han hecho un estudio de correlación lineal entre el nivel de IgE rojos y el tiempo en meses que llevan los usuarios en el centro. Para pedir un análisis de correlación lineal entre estas dos variables, han ejecutado la siguiente sintaxis en SPSS, que se obtiene haciendo: Analizar, Correlaciones, Bivariadas, Variables colocar IgE Y TRESI, marcar Pearson y aceptar:

CORRELATIONS/VARIABLES=IgE TRESI/PRINT=TWOTAIL NOSIG/MISSING=PAIRWISE.

El resultado para el conjunto de usuarios que componen la muestra es el siguiente:

Correlaciones			
		Nivel de IgE en kU/L	Tiempo en la residencia en meses
Nivel de IgE en kU/L	Correlación de Pearson	1	,466**
	Sig. (bilateral)		,000
	N	100	100
Tiempo en la residencia en meses	Correlación de Pearson	,466**	1
	Sig. (bilateral)	,000	
	N	100	100

** La correlación es significativa en el nivel 0,01 (bilateral).

SPSS nos ha proporcionado una tabla simétrica donde se puede ver la correlación entre cada variable y ella misma (casilla superior izquierda y casilla inferior derecha) y la correlación entre las dos variables solicitadas: nivel de IgE y tiempo de residencia en meses. En el primer caso, la correlación vale 1, ya que estamos viendo la relación de una variable con ella misma. En el segundo caso, la correlación vale 0,466 indicando que la relación lineal es poco intensa y positiva: el nivel de IgE aumenta a medida que lo hace el tiempo de residencia de los usuarios en estudio. El coeficiente es distinto de cero, pues se rechaza la hipótesis de que r es igual a cero con mucha contundencia. Dos asteriscos indican que se rechaza al 99% de confianza. Como la significación es bilateral, esta hipótesis se rechaza siempre que la significación es menor o igual que 0,025 y, en este caso vale $0,000 < 0,025$.

Otro ejemplo dentro del mismo caso proporcionó esta tabla al explorar la correlación lineal entre el nivel de hematíes y el de hemoglobina:

Correlaciones			
		Nivel de hematíes o globulos rojos/ml	Nivel de hemoglobina en g/dl
Nivel de hematíes o globulos rojos/ml	Correlación de Pearson	1	,857**
	Sig. (bilateral)		,000
	N	100	100
Nivel de hemoglobina en g/dl	Correlación de Pearson	,857**	1
	Sig. (bilateral)	,000	
	N	100	100

** La correlación es significativa en el nivel 0,01 (bilateral).

En este segundo ejemplo, la correlación lineal es mucho más elevada e intensa que en el ejemplo anterior, de signo positivo también y con un contraste de hipótesis potente acerca del valor obtenido para el coeficiente, ya que es significativo al 99% de confianza.

7.1.3. Regresión lineal simple

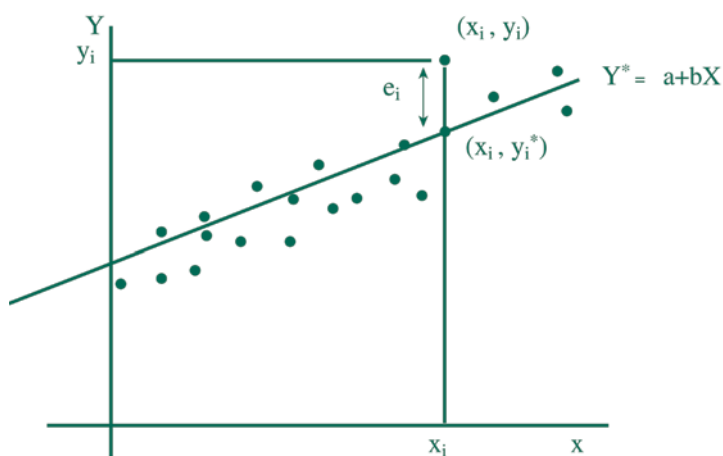
El concepto que presenta la regresión se basa en encontrar la relación que puede existir entre una o varias variables independientes con una dependiente todas ellas cuantitativas. En nuestro caso, vamos a tratar acerca de la regresión lineal, es decir, la que ajusta una línea recta al diagrama de dispersión, pero también existen otros modelos no lineales como el exponencial, el logístico y otros.

Una recta de regresión lineal se calcula a través de la técnica denominada de mínimos cuadrados, que se basa en encontrar la distancia mínima entre los puntos reales obtenidos de mediciones históricas y una recta estimada que represente todos los puntos (ver Figura 19). Por consiguiente, la regresión es una técnica estadística que nos permite buscar la ecuación matemática que mejor describe la relación que existe entre las variables impli-

cadass. La regresión lineal consiste en cuantificar la relación que existe entre variables con el fin de predecir cuáles serán los valores de una variable cuando se conocen los valores de las otras.

En la regresión lineal simple sólo se consideran dos variables: X e Y. Habitualmente se utiliza la Y para identificar a la variable dependiente y la X para identificar a la independiente o explicativa. La ecuación que buscamos estimar es del tipo: $Y^* = a + bX$ a partir de los datos de la nube de puntos representados por la fórmula: $Y = a + bX + E$.

Figura 19. Descripción del fundamento de la técnica de los mínimos cuadrados ordinarios para estimar una recta de regresión a partir de un diagrama de dispersión sobre dos variables cuantitativas



Fuente: <http://biplot.usal.es/problemas/regresion/teoria/regsimple.htm>

En las ecuaciones mencionadas:

- “a” y “b” son coeficientes estimados de la función desconocida, de forma que:
- “a” es la intersección con el eje Y cuando X toma el valor 0
- “b” simboliza la pendiente en la población, y es el coeficiente de regresión lineal simple, que mide la unidad de cambio en promedio de la variable Y por cada unidad de cambio de la variable independiente
- E es el error aleatorio, representa el promedio de las distancias entre los puntos de la recta ideal y las predicciones que se obtienen mediante la recta estimada

Todos estos estadísticos los estima actualmente el software SPSS e incluso Excel y otros programas informáticos. Lo importante a efectos prácticos es saber interpretar los resultados y utilizarlos para nuestros propósitos de toma de decisiones, realización de informes, investigación epidemiológica o cualquier otro proceso de nuestro interés.

Para familiarizarnos con los resultados proporcionados por esta técnica, siguiendo con el ejemplo de usuarios de la residencia, analicemos los que se obtienen, por ejemplo, al estimar la recta de regresión simple que relaciona el nivel de IgE con el tiempo de residencia en meses de los usuarios de la residencia en estudio. Para ello, partimos del supuesto de que existe esta ecuación, puesto que, como hemos visto anteriormente, la correlación lineal de Pearson ha salido medianamente intensa y positiva:

$$\text{Nivel de IgE} = a + b (\text{Tiempo de residencia en meses})$$

Por consiguiente, solicitamos a SPSS que haga el análisis de regresión simple para la muestra de usuarios. La sintaxis para realizar la regresión en SPSS, que se obtiene al programar: analizar, regresiones, lineales, colocar las variables y aceptar, es la siguiente:

REGRESSION/MISSING LISTWISE/STATISTICS COEFF OUTS R ANOVA/CRITERIA=PIN (.05) POUT(.10)/NOORIGIN/DEPENDENT IgE/METHOD=ENTER TRESI.

En la primera tabla que proporciona, el programa nos informa de que la variable dependiente es la que recoge el nivel de IgE, y que ha introducido como variable independiente el tiempo de estancia:

Variables entradas/eliminadas^a

Modelo	Variables entradas	Variables eliminadas	Método
1	Tiempo en la residencia en meses ^b	.	Entrar

a. Variable dependiente: Nivel de IgE en kU/L

b. Todas las variables solicitadas introducidas.

Seguidamente, el programa nos proporciona un resumen del modelo de regresión. La R es el coeficiente de correlación (0,466) y la R cuadrado es, ahora sí, el estadístico que nos informa del grado de dependencia de las variables, concretamente, nos indica que el nivel de IgE depende en un 21,7% del tiempo de residencia en esta población de mujeres usuarios. La R cuadrado-corregida sirve para tener esta misma información cuando el modelo de regresión es múltiple, es decir, cuando hay más de una variable explicativa, que no es nuestro caso. El error típico de la estimación representa a la E del modelo y nos indica que, por término medio los puntos del diagrama de dispersión se separan de la recta estimada en 10,40904 kU/L, valor lógico en este contexto y, relativamente elevado, ya que el ajuste a la recta no es bueno (la R cuadrado no está próxima a 1 o 100%).

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,466 ^a	,217	,209	10,40904

a. Predictores: (Constante), Tiempo en la residencia en meses

El siguiente resultado importante a efectos prácticos es la estimación de la recta. En esta tabla podemos identificar el valor estimado de “a” y de “b” y si estos valores son significativos:

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	-1,283	2,593		-,495	,622
	Tiempo en la residencia en meses	,625	,120	,466	5,209	,000

a. Variable dependiente: Nivel de IgE en kU/L

La recta estimada es: $IgE^* = -1,283 + 0,625(\text{tiempo de residencia en meses})$

El coeficiente “a” no es significativo ($Sig = 0,622 > 0,05$), lo que significa que estadísticamente hablando es cero. En cambio “b” es significativo ($Sig = 0,000 < 0,05$) y distinto de cero.

Así, en este contexto, cuando el tiempo de residencia es de cero meses, se estima que por término medio el nivel estimado de IgE es estadísticamente igual a cero, lo cual es lógico. Por consiguiente, en el contexto de este caso de estudio, es muy natural que “a” no sea estadísticamente significativo.

Por otro lado, según los resultados, como la pendiente “b” es significativa, a medida que aumenta el tiempo de residencia en 1 unidad (un mes en este caso), el nivel de IgE aumenta por término medio en 0,625 kU/L.

El modelo obtenido, aunque no es excelente para efectuar pronósticos, ya que su capacidad explicativa es bastante limitada ($R^2 = 21,7\%$), nos permite presentar la mecánica de su realización. Así, por ejemplo, para un usuario que lleva 8 meses en la residencia, cabe esperar que en promedio tenga este nivel de IgE:

$$IgE \text{ pronosticado}^* = -1,283 + 0,625(8) = 3,717 \text{ kU/L}$$

Comentarios adicionales:

Este es un caso muy discreto de regresión lineal simple, pero así suelen ser la mayoría. En general, se dice que una regresión es “buena” o interesante desde el punto de vista de resultados cuando se alcanza una R^2 del 0,75 (75% en términos porcentuales) o superior, ya que ello indica que la variable independiente es capaz de explicar un 75% o más del comportamiento de la dependiente. Sin embargo, en la práctica, pocas veces se alcanza dicho nivel al analizar la relación entre dos variables. Por consiguiente, hay que saber interpretar lo que nos brindan los datos y resultados. Cuando una regresión simple tiene poca capacidad de pronóstico, la razón más habitual para ello suele ser la falta de otras variables explicativas que son importantes en el estudio que se está llevando a cabo, por lo que, lo más recomendable es buscar datos sobre variables que contribuyan a explicar el fenómeno en estudio y ampliar el modelo simple a uno múltiple.

En nuestro ejemplo, la ecuación obtenida indica que existe cierta relación o influencia del tiempo de estancia en la residencia sobre el nivel de IgE que, aunque débil, es significativa. Para los gestores de la residencia este dato es importante, porque hay que esperar que los problemas cutáneos se puedan acentuar a medida que pasa el tiempo en una proporción significativa de usuarios. Por otro lado, si el tiempo explica un 21,7% del comportamiento del nivel de IgE, eso quiere decir que hay otras variables no contempladas en este análisis que pueden ayudar a explicar dicho nivel. Queda por explicar un 78,3% del nivel de IgE, por lo que, los investigadores pueden buscar y añadir otras variables a la ecuación desarrollando un modelo múltiple que les proporcione una estimación de la importancia de cada nuevo indicador que añaden en la explicación del comportamiento del IgE.

7.1.4. Ejemplo complementario sobre regresión lineal

A partir de los datos del estudio de usuarios de la residencia ha interesado calcular las regresiones para hombres y mujeres por separado entre el nivel de IgE y el hematocrito. Los principales resultados obtenidos han sido:

Resumen del modelo				
Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
	Sexo de los usuarios = Hombre (Seleccionado)			
1	,494 ^a	,244	,229	8,10738

a. Predictores: (Constante), Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total

Coefficientes^{a,b}

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error estándar	Beta		
1 (Constante)	31,738	6,016		5,275	,000
Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,520	,132	-,494	-3,940	,000

a. Variable dependiente: Nivel de IgE en KU/L

b. Selección de casos sólo para los cuales Sexo de los usuarios = Hombre

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
	Sexo de los usuarios = Mujer (Seleccionado)			
1	,303 ^a	,092	,073	12,82498

a. Predictores: (Constante), Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total

Coefficientes^{a,b}

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error estándar	Beta		
1 (Constante)	43,098	13,445		3,206	,002
Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,728	,330	-,303	-2,206	,032

a. Variable dependiente: Nivel de IgE en KU/L

b. Selección de casos sólo para los cuales Sexo de los usuarios = Mujer

Interpretación de los resultados:

El nivel de IgE depende en un 24,4% del hematocrito en el caso de los hombres y en un 9,2% en el caso de las mujeres. Los coeficientes de las rectas de regresión estimadas son significativos en ambos casos, pero de forma mucho más contundente en la recta estimada sobre la muestra de hombres. Las ecuaciones obtenidas indican que existe cierta relación o influencia del hematocrito sobre el nivel de IgE que, aunque débil, es significativa, especialmente en el caso de los hombres. Observando las pendientes de las rectas estimadas, se aprecia que la relación entre ambas variables es negativa y que el nivel de IgE disminuye, en promedio en -0,520 KU/L cuando aumenta en una unidad (1% en este caso), el volumen de hematíes en sangre en los hombres y disminuye, en promedio, en -0,728 KU/L en el caso de las mujeres.

8. REGRESIÓN MÚLTIPLE

8.1. Introducción

Tras la presentación de los análisis de regresión simple, abordamos su ampliación para conocer las posibilidades de estos modelos cuando se añaden más variables explicativas. Muchos fenómenos que se explican a través de modelos de regresión lineal simple en que una variable explicativa alcanza una capacidad inferior al 50% para explicar el comportamiento de la variable dependiente, pueden completarse y alcanzar capacidades muy superiores a base de añadir más variables explicativas de forma coherente y razonada. El modelo de regresión múltiple permite así, proporcionar visiones más completas de muchos fenómenos y, una vez se le conoce, se suele utilizar mucho más que el simple, ya que enriquece ampliamente nuestras conclusiones y explicaciones.

En este apartado, vamos a conocer este tipo de modelos y sus pormenores técnicos de forma aplicada y, partiendo de la base de conocimiento que ha proporcionado el apartado anterior acerca de la regresión lineal simple. Finalmente, veremos que, en los modelos múltiples, la “convivencia” entre varias variables explicativas no es sencilla, pues muchas veces vienen medidas en distintos tipos de unidades; a veces se solapan en cuanto a la información que proporcionan y otras tienen comportamientos particulares, todo lo cual deriva en los llamados “problemas del modelo de regresión lineal múltiple”. Por consiguiente, un analista debe saber que estos problemas existen, debe saber detectarlos y también, conocer que existen soluciones para paliarlos en lo posible.

8.2. Formulación de un modelo de regresión lineal múltiple

En estadística, la regresión lineal múltiple se formula como una ampliación del modelo lineal simple, a base de añadir variables independientes o explicativas a la recta simple. Cuando un modelo tiene dos variables explicativas, ya se considera múltiple. La ecuación para expresar el modelo general toma la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_n x_{ni} + u_i$$

donde:

y_i = variable dependiente

$x_1, x_2, \dots, x_n$ = variables independientes

β_0 = punto de corte del plano (a partir de 2 variables explicativas tenemos un plano y no una recta) o constante del modelo

$\beta_1, \beta_2, \beta_3, \dots, \beta_n$ = parámetros o coeficientes que acompañan a cada variable dependiente y que miden su influencia sobre la variable dependiente

u_i = residuos del modelo o parte de la variación de la variable dependiente que no es posible llegar a explicar

Como es fácil intuir, el modelo, aunque a veces se siga denominando lineal, ya no proporciona una línea recta, sino un plano de “n” dimensiones: tantas como variables explicativas hayamos puesto en la ecuación. Por consiguiente, un primer punto a tener en cuenta

es que, a diferencia de la regresión lineal simple, ya no podemos crear un gráfico de dispersión que nos muestre la relación entre la variable dependiente y las independientes.

Desde un punto de vista matemático, el modelo así especificado cumple con un conjunto de hipótesis y supuestos rigurosos para funcionar y proporcionar estimaciones válidas. No vamos a entrar en la definición de todas ellas a fondo, porque nuestro objetivo es conocer el funcionamiento práctico del modelo, pero las mencionamos para tener una idea de lo que los matemáticos diseñaron para garantizar el buen funcionamiento del modelo múltiple y los problemas que se derivan de la “convivencia” de variables distintas para explicar un único fenómeno.

8.2.1. Principales hipótesis y supuestos que deben cumplir las variables que intervienen en la estimación de un modelo de regresión múltiple

1. La variable dependiente se relaciona linealmente con los coeficientes del modelo y el modelo se especifica correctamente, es decir, no puede tener variables elevadas a ninguna potencia.
2. Los valores de las variables independientes (x_i) no están correlacionados significativamente con los valores de los residuos o perturbaciones (u_i).
3. La esperanza matemática de los residuos es nula: $E(u_i) = 0$. Esta hipótesis indica que, para cada valor de las variables independientes, la perturbación o residuo tomará distintos valores de forma aleatoria, pero no tomará sistemáticamente valores positivos o negativos, sino que se supone tomará algunos valores mayores que cero y otros menores que cero, de tal forma que su valor esperado o media sea cero.
4. Homocedasticidad: $\text{Var}(u_i) = \sigma^2$. Esta hipótesis indica que la varianza de los residuos o perturbaciones es constante, es decir, que todos los términos de la perturbación tienen la misma varianza (σ^2) o dispersión en torno a su valor esperado y que es desconocida.
5. Ausencia de correlación: $\text{Cov}(u_i, u_j) = 0$. Esta hipótesis indica que las covarianzas entre las distintas perturbaciones o residuos son nulas, lo que quiere decir que no están correlacionadas. Esto implica que el valor de la perturbación para cualquier observación muestral no viene influenciado por los valores de las perturbaciones correspondientes a otras observaciones muestrales.
6. Que las variables independientes tengan un comportamiento estocástico (aleatorio, no sistemático).
7. Ausencia de multicolinealidad o independencia lineal. No existen relaciones lineales exactas entre las variables independientes o, en otras palabras, las correlaciones entre pares de variables independientes no son significativas, ya que, si lo fuesen, dos variables independientes estarían proporcionando la misma información y no podrían “convivir” en el mismo modelo estimado.
8. Suponemos que no existen errores de especificación en el modelo, ni errores de medida en las variables explicativas.
9. Normalidad de las perturbaciones: los residuos se comportan según una Normal de media cero y varianza constante σ^2 .
10. Que el error total sea la suma de todos los errores.

Los supuestos e hipótesis del modelo de regresión lineal son exigentes y, en la práctica, es difícil que se cumplan de forma rigurosa. Lo que importa, desde el punto de vista del analista, es que ninguno de los problemas que se puedan presentar sea realmente grave. Así, es muy habitual que se presenten problemas de multicolinealidad (violación del supuesto 7), de heteroscedasticidad (que es lo contrario a la homocedasticidad, violación del supuesto 4), de autocorrelación de los residuos (violación del supuesto 5) y de falta de comportamiento Normal en los residuos (violación del supuesto 9). El analista debe aprender a detectar estos problemas y las soluciones que se adoptan comúnmente para paliarlos con el objetivo de trabajar con modelos que proporcionen las mejores estimaciones y predicciones de comportamiento de la población. Sin embargo, el dominio de esta parcela excede los límites de este cuaderno: simplemente debemos dejar constancia y advertir al lector de que es un tema acerca del cual hay que vigilar y profundizar, si es necesario, cuando se utiliza regresión lineal múltiple.

8.2.2. Estimación de los parámetros del modelo

Vamos a abordar la estimación de los parámetros del modelo a través de un ejemplo aplicado. Seguimos utilizando una base de datos de una muestra de 100 usuarios de una gran residencia para mayores, pero esta vez adaptada a otro tema. Veamos el enunciado y la base de datos.

En una residencia de mayores se está llevando a cabo un estudio sobre alergias dermatológicas, tales como la presencia de prurito sine materiae en una muestra de 100 usuarios. Dado que se conoce que la anemia y un nivel elevado de IgE pueden ser factores asociados a dicho prurito, se han tomado datos de análisis de sangre donde aparecen indicadores de estos factores. En una primera fase, antes de analizar la presencia de prurito, se ha estudiado mediante un análisis de regresión múltiple, el nivel de IgE como variable dependiente, para ver en qué grado viene explicado por el hematocrito, el nivel de hemoglobina, la edad, el tiempo de residencia y las horas semanales de ejercicio de los usuarios.

tema1_base de datos residencia mayores.sav [ConjuntoDatos] - IBM SPSS Statistics Editor de datos

Archivo Editar Ver Datos Transformar Analizar Gráficos Personalizado Utilidades Ventana Ayuda

Visible: 13 de 13 variables

	IgE	HEMATIES	RHEMATIES	Hb	Hto	SEXO	EDAD	TRESI	TRESI_I	TABACO	EJERCICIO	HORASEJS	PRURITO	VER	VER	VER
1	23.67	3756455.00	Anemia	11.20	30.70	Mujer	84.00	26.00	Más de 24	No	Si	2.00	Si			
2	22.45	3689000.00	Anemia	10.90	30.90	Mujer	90.00	25.00	Más de 24	No	Si	3.00	Si			
3	30.33	3256437.00	Anemia	8.90	31.00	Mujer	76.00	17.00	Entre 12 y	No	No	.00	Si			
4	27.89	3077866.00	Anemia	9.76	32.10	Mujer	88.00	15.00	Entre 12 y	No	Si	2.50	Si			
5	27.65	3477765.00	Anemia	10.60	32.70	Mujer	81.00	13.00	Entre 12 y	No	Si	3.50	Si			
6	35.76	3270890.00	Anemia	9.18	32.80	Mujer	80.00	12.00	Entre 12 y	No	No	.00	Si			
7	34.32	3390655.00	Anemia	11.30	33.20	Mujer	93.00	30.00	Más de 24	No	Si	4.00	Si			
8	26.78	3500877.00	Anemia	9.12	33.60	Mujer	90.00	28.00	Más de 24	No	No	.00	Si			
9	25.56	3276554.00	Anemia	10.30	33.60	Mujer	76.00	15.00	Entre 12 y	No	Si	3.50	Si			
10	30.66	3487654.00	Anemia	10.30	34.60	Mujer	78.00	7.00	Menos de	No	Si	3.80	Si			
11	33.87	4625531.00	Nivel normal	15.50	37.30	Mujer	91.00	31.00	Más de 24	No	Si	4.80	Si			
12	27.56	5199875.00	Nivel normal	15.70	38.40	Mujer	83.00	26.00	Más de 24	Ex-fumador	No	.00	Si			
13	28.13	5073349.00	Nivel normal	16.00	40.30	Mujer	88.00	28.00	Más de 24	Ex-fumador	No	.00	Si			
14	22.80	5098765.00	Nivel normal	15.80	40.80	Mujer	74.00	5.00	Menos de	Ex-fumador	No	.00	Si			
15	30.55	5133231.00	Nivel normal	15.40	41.20	Mujer	80.00	26.00	Más de 24	Ex-fumador	No	.00	Si			
16	37.45	5344672.00	Nivel normal	12.80	43.50	Mujer	93.00	29.00	Más de 24	Si	Si	7.00	Si			
17	28.78	4176523.00	Nivel normal	14.20	43.60	Mujer	90.00	32.00	Más de 24	No	Si	8.00	Si			
18	23.67	5143782.00	Nivel normal	16.00	44.20	Mujer	83.00	28.00	Más de 24	Ex-fumador	No	.00	Si			
19	26.43	5200000.00	Nivel normal	16.00	46.00	Mujer	80.00	26.00	Más de 24	Ex-fumador	No	8.00	Si			
20	22.19	5233654.00	Poiglobulia	18.80	48.60	Mujer	91.00	38.00	Más de 24	Si	No	.00	Si			
21	25.67	5437655.00	Poiglobulia	18.40	49.30	Mujer	90.00	30.00	Más de 24	Si	No	.00	Si			
22	28.97	5234543.00	Poiglobulia	18.10	49.40	Mujer	87.00	29.00	Más de 24	Si	Si	7.00	Si			
23	17.89	3977765.00	Anemia	10.60	32.70	Hombre	80.00	39.00	Más de 24	No	Si	2.00	Si			

Vista de datos Vista de variables

IBM SPSS Statistics Processor está listo Casos: 100 Unicode OFF

A partir de estos datos, en esta ocasión, se ha efectuado un primer análisis de regresión múltiple en que la variable dependiente es el nivel de IgE y las independientes: el hematocrito, el nivel de hemoglobina, la edad, el tiempo de residencia y las horas semanales de ejercicio. Los resultados se muestran un poco más adelante. Antes de verlos, es necesario explicar una serie de consideraciones previas que cualquier analista realiza antes de programar un análisis de este tipo.

Consideraciones previas efectuadas por el analista:

Cuando un analista aborda la realización de un análisis de regresión lineal múltiple con SPSS, tiene que saber que el paquete ofrece varios métodos de estimación del modelo. Entre ellos, los más relevantes y utilizados son:

1. Intro: directo, introduce todas las variables explicativas sin tener en cuenta los problemas que pueda producir su combinación en un modelo en cuanto a cumplimiento de las hipótesis de la regresión lineal.
2. Stepwise o por pasos: introduce las variables que aportan algo al modelo, eliminando las que no aportan capacidad explicativa o aquellas que explican lo mismo que otras, eligiendo la más potente. Este método elimina el problema de multicolinealidad o violación de la hipótesis 7 entre las presentadas en el apartado anterior. Las variables explicativas que permanecen en el modelo tras la aplicación de este método son las que proporcionan el mejor modelo en cuanto a capacidad explicativa y no están correlacionadas entre sí.

La mayoría de los analistas estiman el modelo utilizando ambos: el método “Intro” para dar un primer vistazo y el método “Stepwise” para eliminar el problema de la multicolinealidad o variables redundantes si lo hubiese. Vamos a ver los resultados que se obtienen para este ejemplo, comenzando por el método “Intro”.

El primer resultado que proporciona SPSS es una tabla informativa que indica que en el modelo 1 se han introducido las variables explicativas: Horas de ejercicio semanales, tiempo en la residencia en meses, hematocrito, edad y nivel de hemoglobina mediante el método directo “intro” o entrar. Al pie de la tabla, se indica que la variable dependiente es el nivel de IgE y que se han introducido todas las variables explicativas en el modelo.

Variables entradas/eliminadas^a

Modelo	Variables entradas	Variables eliminadas	Método
1	Horas de ejercicio semanales, Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Edad, Nivel de hemoglobina en g/dl ^b	.	Entrar

a. Variable dependiente: Nivel de IgE en kU/L

b. Todas las variables solicitadas introducidas.

Seguidamente, SPSS nos proporciona la tabla resumen del modelo que nos ofrece sólo dos datos útiles en una regresión múltiple, que son: la R cuadrado-ajustada y el Error estándar de la estimación. La R cuadrado-ajustada indica que las variables independientes tienen una capacidad explicativa de la variable dependiente del 43,4%, valor discreto pero superior a los alcanzados en las regresiones simples vistas en el apartado anterior (recordemos que los coeficientes de determinación toman valores entre 0 y 1 y que cuanto más próximo a 1, mejor es el modelo), que indica que hemos estimado un modelo de capacidad explicativa moderada.

El error estándar de la estimación viene dado en las mismas unidades que la variable dependiente. Por consiguiente, por término medio, la IgE pronosticada o ajustada al plano multidimensional se separa, en promedio, de las observaciones originales de IgE en 8,80084 kU/L. Teniendo en cuenta que los valores de esta variable han oscilado entre un valor mínimo de 0,10 kU/L y un valor máximo de 37,45 kU/L, la dispersión entre los datos originales y los ajustados es más alta que baja: otro síntoma de que el modelo tiene un ajuste discreto y que hay que utilizarlo con cautela para efectuar pronósticos

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,680 ^a	,463	,434	8,80084

a. Predictores: (Constante), Horas de ejercicio semanales, Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Edad, Nivel de hemoglobina en g/dl

La siguiente tabla, denominada ANOVA, sirve para contrastar las hipótesis siguientes acerca de los coeficientes de las variables independientes:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_n = 0$$

$$H_1: \beta_1 \neq \beta_2 \neq \dots \neq \beta_n \neq 0$$

En este modelo hay 5 variables explicativas. Por consiguiente, la hipótesis sería del tipo:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1: \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq \beta_5 \neq 0$$

La hipótesis nula, H_0 , se acepta si el estadístico F de la tabla tiene una significación igual o superior a 0,05 y se rechaza cuando la significación es inferior a 0,05. En este caso, la significación es $0,000 < 0,05$, por lo que se rechaza H_0 y se acepta la H_1 que indica que los coeficientes de las variables independientes son distintos de cero, es decir, que aportan explicación acerca de la variable dependiente. Sin embargo, como veremos, hay que tener cuidado con esta evaluación global, ya que a veces enmascara un problema de multicolinealidad.

ANOVA^a

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	6277,625	5	1255,525	16,210	,000 ^b
	Residuo	7280,752	94	77,455		
	Total	13558,377	99			

a. Variable dependiente: Nivel de IgE en kU/L

b. Predictores: (Constante), Horas de ejercicio semanales, Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Edad, Nivel de hemoglobina en g/dl

Finalmente, el programa proporciona la tabla de coeficientes del modelo en que podemos leer la siguiente información de izquierda a derecha:

Coeficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	9,168	14,295		,641	,523
	Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,445	,250	-,290	-1,778	,079
	Nivel de hemoglobina en g/dl	-,782	,612	-,201	-1,277	,205
	Edad	,327	,220	,181	1,490	,140
	Tiempo en la residencia en meses	,453	,156	,338	2,901	,005
	Horas de ejercicio semanales	-,965	,290	-,254	-3,331	,001

a. Variable dependiente: Nivel de IgE en kU/L

En la primera columna, tenemos la constante y los nombres de las variables que acompañan a cada coeficiente estimado. En la columna B tenemos los coeficientes:

- La constante vale 9,168 kU/L y representa el punto de corte (b_0) del plano de 5 dimensiones que forma la estimación del modelo:

$$IgE^* = b_0 + b_1(\text{Hematocrito}) + b_2(\text{Hemoglobina}) + b_3(\text{Edad}) + b_4(\text{Tiempo}) + b_5(\text{horas de ejercicio semanales})$$

La constante viene en las mismas unidades que la variable dependiente, por lo que, teóricamente, representa la media de IgE cuando las variables independientes toman el valor de 0. La constante no siempre tiene sentido, pero es recomendable incluirla porque es un artificio matemático para lograr un mejor ajuste. Sólo en el caso en el que todas las variables explicativas pudieran tomar el valor cero tiene sentido interpretar el parámetro que acompaña a la constante como el valor de la endógena cuando no toman valor el resto de las exógenas. En estudios de salud, rara vez ocurre, pero hay situaciones en que se puede dar el caso.

- El coeficiente del Hematocrito es -0,445(%). El resultado se interpreta como el número medio de unidades que disminuye el indicador de IgE cada vez que el indicador de He-

matocrito aumenta en una unidad, manteniendo constante el nivel o valor del resto de variables independientes.

- El coeficiente del nivel de hemoglobina es -0,782 (g/dl). El resultado se interpreta como el número medio de unidades que disminuye el indicador de IgE cada vez que el nivel de hemoglobina aumenta en una unidad, manteniendo constantes los valores del resto de variables independientes.
- El coeficiente la edad es 0,327 (años). El resultado se interpreta como el número medio de unidades que aumenta el indicador de IgE cada vez que la edad aumenta en un año, manteniendo constante el nivel o valor del resto de variables independientes.
- El coeficiente del tiempo de residencia es 0,453 (meses). El resultado se interpreta como el número medio de unidades que aumenta el indicador de IgE cada vez que el tiempo de residencia aumenta en una unidad (un mes en este caso), manteniendo constantes los valores del resto de variables independientes.
- El coeficiente de las horas de ejercicio semanales es -0,965 (horas). El resultado se interpreta como el número medio de unidades que disminuye el indicador de IgE cada vez que el tiempo de ejercicio aumenta en una unidad (una hora en este caso), manteniendo constantes los valores del resto de variables independientes.

En la siguiente columna tenemos los errores estándar de los coeficientes, que se calculan internamente mediante las fórmulas:

Para el error estándar de la constante:

$$S_{bo} = S_u \sqrt{\frac{1}{n} \left(1 + \frac{\bar{x}^2}{S_x^2}\right)}$$

Para el error estándar del resto de coeficientes:

$$S_{bi} = \frac{S_u}{\sqrt{n S_{xi}}}$$

Siendo:

$$S_u = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}} = \text{error estándar de la regresión}$$

La utilidad de estos errores estándar se comenta en el siguiente apartado.

En la siguiente columna tenemos los denominados coeficientes Beta, de gran utilidad en regresiones múltiples pues, calculan la regresión utilizando los valores estandarizados de las variables explicativas, eliminando los efectos producidos por venir estas medidas en distintos tipos de unidades. El uso de unidades distintas (% , gr/L...) produce que los efectos de los coeficientes estimados originales puedan ser muy dispares, enmascarando cuál de ellos es el que tiene mayor impacto en la variable dependiente.

Así, observando los coeficientes originales, parece que el mayor efecto sobre el nivel de IgE lo tiene el tiempo semanal dedicado a ejercicio, seguido por el nivel de hemoglobina. En cambio, los coeficientes Beta, muestran que el valor absoluto más grande corresponde al tiempo de residencia, seguido del coeficiente del Hematocrito. En muchas ocasiones, en regresiones múltiples ambos coeficientes coinciden en cuando a orden de impacto de las variables, pero en otras no se mantiene, por lo que siempre hay que mirar los coeficientes Beta. Al estandarizar los valores de las variables independientes, el coeficiente Beta de la constante pasa a ser cero en todos los análisis, por lo que no aparece nunca en esta columna.

Por consiguiente: la columna Beta es clave en la interpretación de una regresión múltiple. Recordemos que una variable (x) queda estandarizada cuando se aplica a sus observaciones (xi) la fórmula:

$$z_i = \frac{x_i - \bar{x}}{S_{xi}}$$

La estandarización anula las unidades de medida y transforma los valores de cualquier variable en valores de la $N(0,1)$ que se mueven, aproximadamente, entre los valores -3 y +3.

Los coeficientes divididos por sus errores estándar proporcionan los valores de la columna t, que recoge los estadísticos que contrastan las hipótesis individuales de cada coeficiente:

Ho: $\beta_0 = 0$	contra	H1: $\beta_0 \neq 0$
Ho: $\beta_1 = 0$	contra	H1: $\beta_1 \neq 0$
Ho: $\beta_2 = 0$	contra	H1: $\beta_2 \neq 0$
Ho: $\beta_3 = 0$	contra	H1: $\beta_3 \neq 0$
Ho: $\beta_4 = 0$	contra	H1: $\beta_4 \neq 0$
Ho: $\beta_5 = 0$	contra	H1: $\beta_5 \neq 0$

El estadístico t que sirve para contrastar cualquiera de estas hipótesis obedece a la fórmula:

$$t = \frac{b_i - 0}{S_{bi}} \sim t_{n-k}$$

Donde b_i es el valor de cada coeficiente estimado y s_{b_i} es el error estándar de cada coeficiente estimado y que están en la tercera columna de la tabla de resultados.

Cualquiera de las hipótesis nulas se acepta si la significación del estadístico t que le corresponde es igual o mayor que 0,05 y, se rechaza si la significación es menor que 0,05. En nuestro ejemplo, se aceptan todas las hipótesis excepto las que atañen al tiempo de residencia en meses y a las horas de ejercicio semanales. Esto evidencia que la tabla ANOVA estaba enmascarando un problema de multicolinealidad, ya que su resultado nos decía que todos los coeficientes conjuntamente evaluados eran distintos de cero, cuando en realidad, al contrastar una hipótesis individual para cada uno, resulta que tres de ellos tienen significaciones mayores que 0,05, lo que les convierte en ceros desde el punto de vista estadístico. Por consiguiente, ni el hematocrito, ni la hemoglobina ni la edad están aportando capacidad explicativa al modelo estimado y, la constante, también es cero.

Este punto de contradicción entre el resultado de la tabla ANOVA y los contrastes individuales es muy importante, pues ilustra la existencia de un problema de multicolinealidad en el modelo: hay tres variables que, en principio, no pueden convivir en el modelo porque están correlacionadas y explican lo mismo. La detección del problema de multicolinealidad es sencilla: si la tabla ANOVA indica que todos los coeficientes son distintos de cero y en los contrastes individuales se acepta alguna hipótesis de los coeficientes (dejando aparte el caso de la constante), entonces, hay un problema de multicolinealidad o información redundante. El problema se soluciona, o se puede evitar de raíz, aplicando el método “Stepwise” o “Por pasos” en lugar del método “Intro”.

Seguidamente, vamos a ver los resultados del modelo aplicando este tipo de metodología. La sintaxis para llevar a cabo esta estimación es la siguiente:

REGRESSION/MISSING LISTWISE/STATISTICS COEFF OUTS R ANOVA/CRITERIA=PIN(.05)
POUT(.10)/NOORIGIN/DEPENDENT IgE/METHOD=STEPWISE Hto Hb EDAD TRESI HORASEJS.

La primera tabla que proporciona SPSS es informativa y, como indica su encabezamiento, nos explica qué variables han entrado en cada paso:

Variables entradas/eliminadas^a

Modelo	Variables entradas	Variables eliminadas	Método
1	Tiempo en la residencia en meses		Por pasos (Criterios: Probabilidad-de-F-para-entrar <= ,050, Probabilidad-de-F-para-eliminar >= ,100).
2	Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total		Por pasos (Criterios: Probabilidad-de-F-para-entrar <= ,050, Probabilidad-de-F-para-eliminar >= ,100).
3	Horas de ejercicio semanales		Por pasos (Criterios: Probabilidad-de-F-para-entrar <= ,050, Probabilidad-de-F-para-eliminar >= ,100).

a. Variable dependiente: Nivel de IgE en kU/L

En la columna Modelo, vemos que han tenido lugar tres pasos. En el primer paso ha entrado la variable Tiempo de residencia en meses, en el segundo paso la variable que representa el hematocrito y, en el tercer paso las horas de ejercicio semanales.

No hay variables eliminadas: a veces, se eliminan variables porque cuando entran resultan incompatibles con alguna de las que ya han entrado en pasos anteriores y, por consiguiente, el programa las hace salir. Esto puede suceder cuando hay muchas variables explicativas candidatas a entrar en el modelo. En caso de haber variables eliminadas en esta fase, aparecen en la segunda columna. No hay que confundir la eliminación de variables en esta fase con la exclusión de variables. La exclusión se produce cuando una variable no llega a entrar en el modelo: esto es lo que parece haberle pasado a la variable que representa a la Hemoglobina, así como a la Edad.

La entrada y exclusión de variables se basa en el cumplimiento de las condiciones de entrada que se explican en la columna Método. El programa calcula la F (de la Tabla ANOVA) en cada paso y deja entrar a la variable de ese paso siempre que la significación de esa F sea menor o igual a 0,05 y la elimina si la significación es mayor o igual que 0,1. Es decir, acepta la convivencia de variables hasta un límite del 90% de confianza (significación 0,1 como máximo) y lo hace una por una.

Seguidamente, en la tabla resumen del modelo tenemos las medidas de bondad del ajuste y sus valores en cada paso, lo que nos permite seguir su evolución y ver la ganancia a medida que el programa avanza. Los modelos obtenidos en cada paso se explican a pie de tabla. Tenemos tres pasos: en el primer paso, la regresión todavía es simple, pues sólo hay una variable explicativa. Por consiguiente, la bondad del ajuste viene dada por $R^2 = 21,7\%$.

En el segundo paso, la regresión ya es múltiple y la bondad del ajuste viene dada por R cuadrado ajustado = 37,0% y en el tercer paso tenemos la bondad del ajuste final: 42,4%.

El error estándar disminuye entre el primer paso y el segundo, y entre el segundo y el tercero lo cual indica que se mejora el ajuste a cada nuevo modelo. Como vemos, el modelo en el primer paso es muy discreto, pues el tiempo en la residencia explica un 21,7% del nivel de IgE. Al entrar el hematocrito, mejora el modelo y, al entrar las horas de ejercicio, también. El resto de las variables no aportarían explicación del nivel de IgE.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,466 ^a	,217	,209	10,40904
2	,619 ^b	,383	,370	9,28818
3	,663 ^c	,440	,422	8,89713

a. Predictores: (Constante), Tiempo en la residencia en meses

b. Predictores: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total

c. Predictores: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Horas de ejercicio semanales

Seguidamente, la Tabla ANOVA se calcula en cada paso del proceso. En el primer paso contrasta la hipótesis sobre el coeficiente de la primera variable que ha entrado:

$$H_0: \beta_1 = 0 \text{ contra } H_1: \beta_1 \neq 0 \text{ La } H_0 \text{ se rechaza, ya que } \text{Sig} = 0,000 < 0,05$$

En el segundo paso, ya hay dos variables, así que se contrasta la hipótesis:

$$H_0: \beta_1 = \beta_2 = 0 \text{ contra } H_1: \beta_1 \neq \beta_2 = 0 \text{ La } H_0 \text{ se rechaza, ya que } \text{Sig} = 0,000 < 0,05$$

En el tercer paso contrasta la hipótesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ contra } H_1: \beta_1 \neq \beta_2 \neq \beta_3 \neq 0 \text{ La } H_0 \text{ se rechaza, ya que } \text{Sig} = 0,000 < 0,05$$

El proceso se detiene porque ya no hay más variables que cumplan con este criterio. Los coeficientes existen y la regresión estimada tiene tres variables explicativas.

ANOVA^a

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	2940,258	1	2940,258	27,137	,000 ^b
	Residuo	10618,120	98	108,348		
	Total	13558,377	99			
2	Regresión	5190,158	2	2595,079	30,081	,000 ^c
	Residuo	8368,219	97	86,270		
	Total	13558,377	99			
3	Regresión	5959,118	3	1986,373	25,093	,000 ^d
	Residuo	7599,259	96	79,159		
	Total	13558,377	99			

a. Variable dependiente: Nivel de IgE en kU/L

b. Predictores: (Constante), Tiempo en la residencia en meses

c. Predictores: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total

d. Predictores: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Horas de ejercicio semanales

Seguidamente, la tabla de coeficientes nos proporciona el modelo estimado en cada paso:

$$\text{Primer paso: } \text{IgE}^* = -1,283 + 0,625(\text{Tiempo de residencia en meses})$$

$$\text{Segundo paso: } \text{IgE}^* = 25,390 + 0,620 (\text{Tiempo}) -0,625 (\text{Hto})$$

$$\text{Tercer paso: } \text{IgE}^* = 25,390 + 0,620 (\text{Tiempo}) -0,625 (\text{Hto}) -0,906(\text{Horas Ej.})$$

Como se puede apreciar, los valores de los coeficientes B cambian en cada paso y muestran que el Hto y el tiempo de ejercicio tienen una relación negativa con el nivel de IgE: a medida que aumentan Hto y las horas de ejercicio, la IgE disminuye.

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	-1,283	2,593		-,495	,622
	Tiempo en la residencia en meses	,625	,120	,466	5,209	,000
2	(Constante)	25,390	5,713		4,444	,000
	Tiempo en la residencia en meses	,620	,107	,462	5,793	,000
	Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,625	,122	-,407	-5,107	,000
3	(Constante)	29,061	5,598		5,192	,000
	Tiempo en la residencia en meses	,617	,103	,460	6,018	,000
	Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,629	,117	-,410	-5,367	,000
	Horas de ejercicio semanales	-,906	,291	-,238	-3,117	,002

a. Variable dependiente: Nivel de IgE en kU/L

En el resto de las columnas de la Tabla de coeficientes tenemos:

Los errores estándar de los coeficientes en cada paso, que sirven para calcular las t y poder efectuar los contrastes individuales de cada coeficiente en cada paso: como vemos, en los tres pasos todos los coeficientes son significativos, si bien, en el tercer paso, las horas de ejercicio aparecen como un poco menos significativa que las otras dos variables.

Los coeficientes Beta, que indican en cada paso el peso o importancia (en valor absoluto) de las variables del modelo. En todos los pasos el tiempo de residencia es la variable que más peso tiene en el modelo ya que, en valor absoluto muestra los coeficientes más elevados.

La última Tabla que proporciona SPSS, muestra información sobre las variables excluidas, variables que no han entrado en el modelo estimado en cada paso:

En el paso 1: no entraron ni el hematocrito, ni el nivel de hemoglobina, ni la edad ni las horas de ejercicio semanales. La columna En beta nos informa del peso que tendría cada variable si entrase en el modelo. Como se puede ver, el peso del hematocrito es mayor (siempre en valor absoluto) y, por consiguiente, es mejor candidata para entrar en el siguiente paso.

La t y su significación, tienen el mismo efecto: cuanto mayor es la t y menor es su significación, mejor es la variable para entrar en el modelo. De nuevo, el hematocrito es mejor candidata. La correlación parcial es la correlación de cada una de estas variables candidatas con la variable dependiente, el nivel de IgE: cuanto mayor es esta correlación, mejor para el modelo. De nuevo gana el hematocrito, aunque con una correlación moderada. Finalmente, tenemos la tolerancia.

La Tolerancia es un indicador que mide la colinealidad o multicolinealidad. Su fórmula es:

$$\text{Tolerancia} = 1 - R_i^2$$

Siendo R_i^2 el coeficiente de correlación múltiple al cuadrado de una variable explicativa con las restantes. Para que haya multicolinealidad dicha correlación ha de ser alta, o lo que es lo mismo la tolerancia baja: en términos “humanizados”, las variables no se toleran, no conviven bien. Por consiguiente, el hematocrito y las horas de ejercicio son las mejores candidatas para entrar en el modelo en el paso 2, al mostrar la Tolerancia más alta con las variables que ya hay en el modelo: cuanto más alta la tolerancia, menor el problema de multicolinealidad. Gana el hematocrito por tener una t más elevada en el paso 1.

Variables excluidas^a

Modelo		En beta	t	Sig.	Correlación parcial	Estadísticas de colinealidad Tolerancia
1	Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total	-,407 ^b	-5,107	,000	-,460	1,000
	Nivel de hemoglobina en g/dl	-,394 ^b	-4,893	,000	-,445	,999
	Edad	-,118 ^b	-,913	,363	-,092	,482
	Horas de ejercicio semanales	-,233 ^b	-2,694	,008	-,264	1,000
2	Nivel de hemoglobina en g/dl	-,159 ^c	-,968	,336	-,098	,235
	Edad	,175 ^c	1,374	,173	,139	,389
	Horas de ejercicio semanales	-,238 ^c	-3,117	,002	-,303	1,000
3	Nivel de hemoglobina en g/dl	-,216 ^d	-1,367	,175	-,139	,232
	Edad	,191 ^d	1,570	,120	,159	,388

a. Variable dependiente: Nivel de IgE en KU/L

b. Predictores en el modelo: (Constante), Tiempo en la residencia en meses

c. Predictores en el modelo: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total

d. Predictores en el modelo: (Constante), Tiempo en la residencia en meses, Hematocrito (%): volumen de hematíes en sangre expresado como un porcentaje sobre el volumen sanguíneo total, Horas de ejercicio semanales

Por consiguiente, en el primer paso, se pone de manifiesto que la variable que representa el hematocrito es la mejor candidata para entrar en el modelo en el segundo paso, cosa que sucede. En el segundo paso, la mejor candidata es la variable Horas de ejercicio semanales con la mayor t y la mayor tolerancia. En el tercer paso, el proceso se detiene porque ninguna t es significativa para el nivel de hemoglobina o para la edad.

8.2.3. Bondad del ajuste y predicción

El principal propósito de realización de un análisis de regresión múltiple es el de obtener una descripción lo más completa posible sobre el comportamiento de una población estimando un modelo de ese comportamiento.

Si el modelo tiene una capacidad explicativa suficiente (habitualmente 70%-75% o más, si bien dependen del contexto y de los objetivos de cada estudio), entonces, se pueden hacer predicciones y simulaciones fiables con él y obtener información útil para la toma de decisiones en múltiples ámbitos.

La capacidad explicativa de un modelo de regresión lineal múltiple se determina a través del coeficiente de determinación ajustado, que es una medida de la bondad del ajuste (cuán bien o mal ajustada está una nube de puntos al plano teórico que lo sustenta). Es lo mismo que sucedía con una recta de regresión simple, pero con más dimensiones.

La fórmula del coeficiente de determinación ajustado deriva del coeficiente de determinación sin ajustar y es la siguiente:

$$\bar{R}^2 = \frac{\sum \hat{u}_i^2 / n - k}{\sum (y_i - \bar{y})^2 / (n - 1)} = 1 - \frac{n - 1}{n - k} (1 - R^2)$$

Al igual que el coeficiente de determinación sin ajustar, éste toma valores entre 0 y 1. Suele leerse como porcentaje (multiplicado por 100) y nos dice la proporción de variación de la variable dependiente explicada por el conjunto de variables independientes o explicativas introducidas en el modelo. Si el valor del coeficiente se aproxima a 1, el modelo estimado es válido para efectuar predicciones acerca del comportamiento de la población objetivo. La pregunta es: ¿hasta qué valor se admite que el modelo es bueno para efectuar predicciones?

No hay respuesta exacta a esta pregunta. Los analistas suelen admitir un modelo como válido hasta valores próximos a 0,6 y superiores, pero a veces hacen excepciones y admiten valores más bajos. Todo depende del contexto de la investigación y de la rigurosidad que haya que aplicar. A veces, con carácter exploratorio se relaja el criterio.

En nuestro ejemplo, ya hemos visto que el modelo por pasos ha arrojado un valor de 0,422, lo que indica que las variables seleccionadas explican un 42,2% del nivel de IgE. Se trata de un ajuste discreto que no permite efectuar predicciones muy fiables, pero que se puede utilizar con carácter exploratorio. Al menos advierte de que algunos factores se asocian a dicho nivel. En este punto, falta de saber si el modelo tiene algún otro problema no explorado todavía, cosa que hacemos en el próximo apartado.

Por consiguiente, a la práctica, podemos hacer predicciones como la siguiente:

Hacer un pronóstico de nivel de IgE para un hipotético usuario con tiempo de residencia de 14 meses, con un valor de Hto de 33,4% y 10 horas de ejercicio semanales. Para calcular el pronóstico, tomamos el modelo ajustado del tercer y último paso en la Tabla de Coeficientes:

$$\text{IgE}^* = 29,061 + 0,617(\text{Tiempo}) - 0,629(\text{Hto}) - 0,906(\text{Horas de ejercicio semanales})$$

Y, sustituimos los valores deseados o propuestos de las variables donde corresponde, resolviendo la ecuación para ellos:

$$\text{IgE}^* = 29,061 + 0,617(14) - 0,629(33,4) - 0,906(10) = 7,6304 \text{ kU/L}$$

Por consiguiente, para un usuario con estos valores, cabe esperar una IgE promedio de unos 7,63 kU/L aproximadamente. Se pueden hacer tantos pronósticos y simulaciones

como se desee, cambiando los valores de las variables independientes. No obstante, hay que tener presente la bondad del ajuste y, en este caso, es bastante discreta.

8.2.4. Conocimiento básico de los problemas del modelo de regresión múltiple

El conocimiento profundo de los problemas del modelo de regresión múltiple y sus soluciones es complejo y requiere mucho más tiempo del disponible en un resumen de estas características. Sin embargo, no se puede cerrar este tema sin, por lo menos, proporcionar una explicación general de estos problemas ya que, su existencia, puede condicionar cualquier análisis de regresión múltiple infravalorando o supervalorando sus resultados. En el ámbito de las ciencias de la salud, ello cobra gran relevancia debido a que los resultados de análisis estadísticos pueden ser críticos en determinados estudios. Por consiguiente, aunque sea a nivel teórico y básico, se proporcionan las principales características de los principales problemas y sus soluciones.

Antes de llevar a cabo predicciones, es aconsejable chequear el modelo y ver si tiene, al menos, alguno de estos problemas, que se derivan del incumplimiento de los supuestos más críticos o hipótesis clave de funcionamiento del modelo:

- 1) Ver si hay problemas de multicolinealidad: el más evidente y sencillo de solucionar
- 2) Ver si hay un problema de heteroscedasticidad
- 3) Ver si hay problemas de falta de normalidad en el comportamiento de los residuos del modelo

Se trata de los problemas más habituales. Vamos a comentarlos de forma individual.

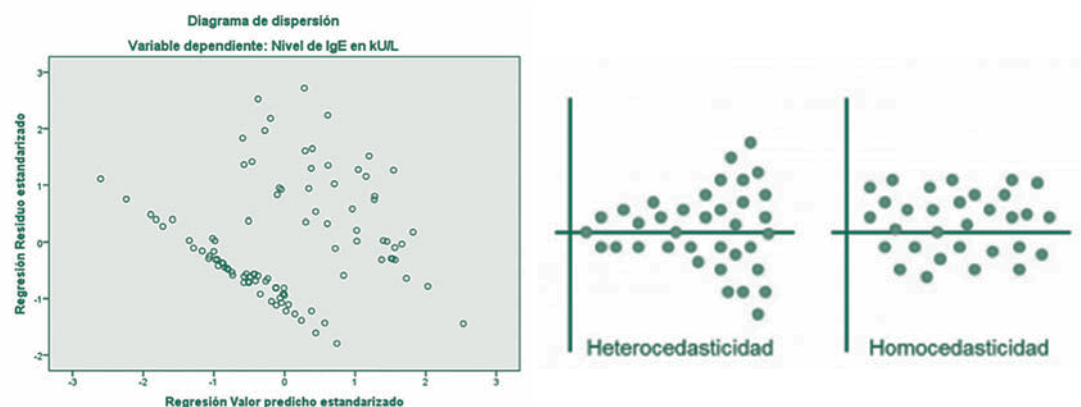
- 1) Como hemos visto, la multicolinealidad se produce cuando dos o más variables independientes del modelo están linealmente correlacionadas entre sí.

En estos casos, se dice que tenemos información redundante porque hay más de una variable explicando la variable dependiente con el mismo tipo de información y capacidad explicativa similar. Anteriormente, hemos visto que el problema se detecta cuando el contraste ANOVA general es significativo, pero alguno de los contrastes t individuales no sale significativo. La solución más rápida y práctica es aplicar el método “Stepwise o por pasos” para estimar el modelo, pues este método escoge las variables que tienen más capacidad explicativa y deja fuera las que no aportan valor al modelo.

- 2) La heteroscedasticidad se produce cuando la dispersión de los residuos o diferencias entre las observaciones originales de la variable dependiente y sus predicciones no sigue un patrón aleatorio sino sistemático.

Es decir, a priori se espera que algunos valores de los errores o diferencias entre los valores originales de la variable dependiente y sus predicciones sean positivos y otros negativos, distribuyéndose aleatoriamente alrededor de su media (que recordemos que es cero). Cuando la inspección de los residuos muestra que hay zonas con mayor dispersión y zonas con menor dispersión, entonces, hay un problema de heteroscedasticidad o falta de varianza constante entre los residuos.

La heteroscedasticidad se detecta inspeccionando un gráfico de dispersión en que se colocan los residuos en el eje vertical y las predicciones de la variable dependiente en el horizontal. SPSS construye este tipo de gráficos si se lo pedimos al estimar la regresión múltiple. Aquí se puede ver el resultado para nuestro ejemplo. Se han utilizado residuos y predicciones estandarizadas.



El modelo no presenta un gran problema de heteroscedasticidad, ya que, la nube de puntos resultante sigue un patrón bastante aleatorio: si acaso algo estructurado en los casos abajo a la izquierda. A la derecha del gráfico SPSS se muestra el típico patrón de la nube de puntos con homocedasticidad, es decir, sin problema de heteroscedasticidad comparado con el más típico cuando hay este problema, que es una nube de puntos con cierta forma de “trompeta”. Este tipo gráfico se puede construir utilizando el submenú de gráficos del menú de regresión en SPSS. En cualquier caso, la sintaxis SPSS que proporciona este tipo de gráfico es la siguiente:

```
REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=
PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT IgE /METHOD=STEPWISE Hto Hb EDAD
TRESI HORASEJS /SCATTERPLOT=(*ZRESID,*ZPRED).
```

Existen diversas soluciones a este problema, que pasan por aplicar transformaciones al modelo como las que se comentan a continuación, pero desarrollarlas excede los objetivos de este cuaderno, por lo que simplemente se sugieren.

Solución 1: cuando se asume que la varianza de los residuos es proporcional al cuadrado de una variable explicativa, es decir:

$$V(u_i) = \sigma^2 x_i^2$$

En este caso, se recomienda estimar el modelo bajo esta transformación:

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{u_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + y_i$$

donde: y_i

Es el término de perturbación transformado y cuya varianza es constante porque:

$$V(y_i) = E\left(\frac{u_i}{x_i}\right)^2 = \frac{1}{x_i^2} E(u_i)^2 = \frac{1}{x_i^2} \sigma^2 x_i^2 = \sigma^2$$

Solución 2: cuando se asume que la varianza de los residuos es proporcional a la variable explicativa, es decir:

$$V(u_i) = \sigma^2 x_i$$

En este caso, se recomienda estimar el modelo bajo este tipo de transformación:

$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1 \sqrt{x_i} + \frac{u_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1 \sqrt{x_i} + y_i$$

donde: y_i

Es el término de perturbación transformado y cuya varianza es constante porque:

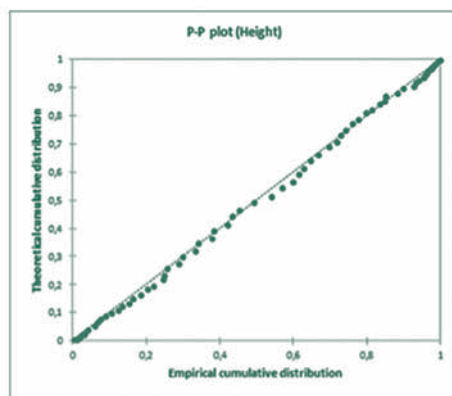
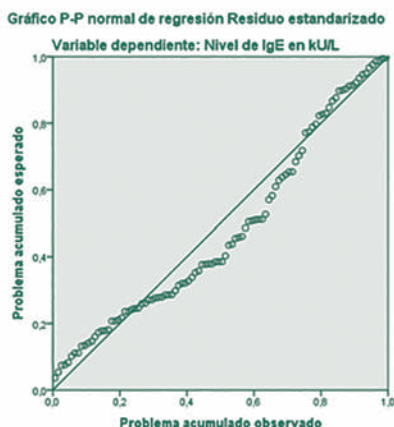
$$V(y_i) = E\left(\frac{u_i}{\sqrt{x_i}}\right)^2 = \frac{1}{x_i} E(u_i)^2 = \frac{1}{x_i} \sigma^2 x_i = \sigma^2$$

Como se puede apreciar, la solución de la heteroscedasticidad no es sencilla y requiere cierta especialización. Si el lector se pregunta cuál escoger, lo habitual es probar diversas soluciones y chequear de nuevo el comportamiento del modelo para ver si mejora. Tras las pruebas se suele decidir qué solución es la mejor. En realidad, hay muchas más posibles soluciones además de las presentadas y, antes de aplicarlas, es necesario hacer un estudio profundo del comportamiento de los residuos. Se pueden consultar manuales para profundizar en este tema o contar con el consejo de un estadístico especializado.

En cualquier caso, es importante chequear el modelo y no aceptar modelos de regresión lineal múltiple sin analizar la posible existencia de problemas de violación de los principales supuestos: en muchos artículos académicos, no se deja constancia de realización

de este chequeo y de solución de problemas si se han detectado, lo cual puede poner en entredicho los resultados y conclusiones. Se recomienda profundizar en el estudio de este apartado si se van a llevar a cabo análisis de regresión múltiple en cualquier contexto.

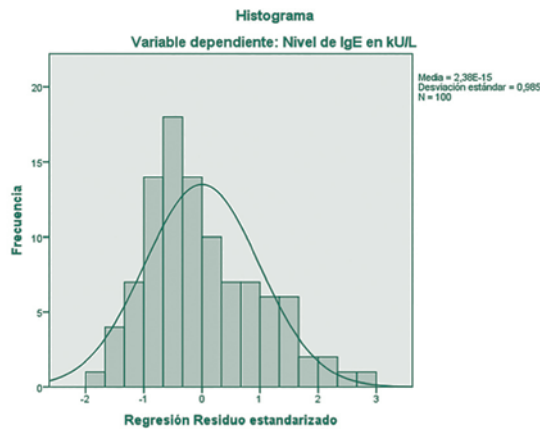
- 3) La normalidad de los residuos se chequea mediante la construcción de un gráfico denominado Q-Q o P-P Plot. Este gráfico, compara la forma de la distribución de los residuos estandarizados contra la forma de la distribución Normal estándar y también sirve para detectar casos atípicos en la misma. SPSS calcula este tipo de gráfico si se lo pedimos. A su lado, se ha colocado un gráfico que refleja una situación de normalidad.



En el calculado para este ejemplo, vemos que ambas distribuciones se parecen y que los residuos se comportan de forma bastante Normal. Cuanto más encima de la recta estén los puntos, más parecidas son ambas distribuciones y menos problema de falta de Normalidad tenemos, tal y como se va en el gráfico de la derecha. En este caso, los residuos de algunas observaciones se nos van un poco por debajo de la recta, pero no aparece como un problema grave.

Esta es la sintaxis SPSS para obtener este gráfico, e incluso un histograma donde se puede ver la distribución de los residuos comparada con una Normal estándar que ratifica lo que muestra el gráfico P-P: la distribución de los residuos se aproxima a una Normal, no habiendo violación de este supuesto en nuestro ejemplo.

```
REGRESSION /MISSING LISTWISE /STATISTICS COEFF OUTS R ANOVA /CRITERIA=
PIN(.05) POUT(.10) /NOORIGIN /DEPENDENT IgE /METHOD=STEPWISE Hto Hb EDAD
TRESI HORASEJS/RESIDUALS HISTOGRAM(ZRESID)NORMPROB(ZRESID).
```



Las violaciones del supuesto de normalidad de los residuos aparecen muchas veces porque (a) las distribuciones de la variable dependiente o de las independientes no siguen, en sí mismas, una distribución Normal o porque (b) el modelo no es lineal, sino de otro tipo.

En estos casos, la solución que se adopta suele ser llevar a cabo una transformación no lineal de las variables, la cual solventará ambos tipos de problemas. Entre las transformaciones más típicas para probar están las siguientes:

- Tomar logaritmos naturales de la variable dependiente en lugar de sus valores originales: $\ln(y_i)$
- Tomar logaritmos naturales de la variable dependiente más 1 en lugar de sus valores originales: $\ln(y_i+1)$
- Tomar las raíces cuadradas de los valores de la variable dependiente: $\sqrt{y_i}$
- Aplicar la función inversa de la variable dependiente: $1/y_i$

Todas estas transformaciones se pueden realizar utilizando el menú de SPSS dedicado a transformaciones calculando nuevas variables. Consultar manuales dedicados a problemas de la regresión si es necesario. El analista puede probar una a una estas transformaciones y chequear el gráfico Q-Q para ver si el problema desaparece. Si se aplica una transformación, es importante tener en cuenta cuál ha sido para deshacerla a la hora de efectuar predicciones. Estos son los problemas más comunes que hay que controlar cuando se estima un modelo de regresión lineal múltiple, pero hay otros como el de autocorrelación de los residuos, muy común cuando se trabaja con series temporales, o la aparición de casos atípicos, denominados outliers. Se recomienda al lector consultar manuales que traten extensamente estos problemas para ampliar sus conocimientos.

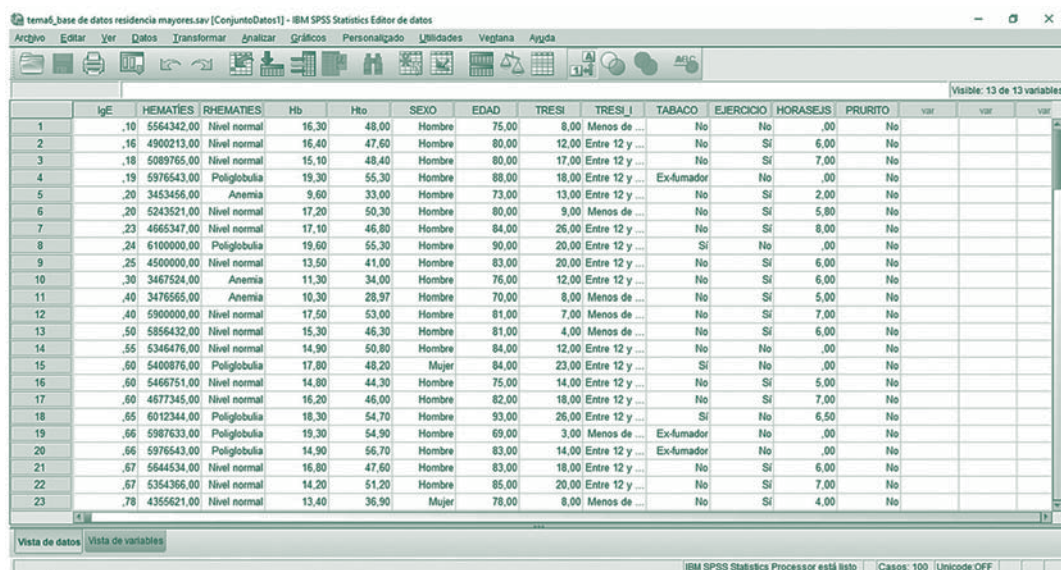
9. REGRESIÓN LOGÍSTICA: FUNDAMENTOS

9.1. Introducción: utilidad de la regresión logística

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico (es decir, que puede darse o no darse), la presencia o no de uno o más factores y el grado o nivel de influencia de estos. La metodología de la regresión lineal no es aplicable en estos casos porque ahora la variable dependiente sólo presentará dos valores (nos centraremos en el caso dicotómico), como puede ser presencia/ausencia de una patología.

Vamos a ver cómo se utiliza esta técnica sobre la base de datos de los usuarios de la residencia de mayores para explicar la presencia o ausencia del prurito sine materiae cuyos factores presumiblemente asociados hemos estudiado en apartados previos.

Recordemos el aspecto de la base de datos. Ahora una variable de nuestro interés y que va a actuar como dependiente, es la que está en la última columna: esta variable recoge la presencia o ausencia de prurito sine materiae reportada en la revisión de los usuarios.



	ID	HEMATIES	RHEMATIES	Hb	Hto	SEXO	EDAD	TRESI	TRESI_I	TABACO	EJERCICIO	HORASEJS	PRURITO			
1	,10	5564342.00	Nivel normal	16.30	48.00	Hombre	75.00	8.00	Menos de ...	No	No	,00	No			
2	,16	4900213.00	Nivel normal	16.40	47.60	Hombre	80.00	12.00	Entre 12 y ...	No	Si	6.00	No			
3	,18	5089765.00	Nivel normal	15.10	48.40	Hombre	80.00	17.00	Entre 12 y ...	No	Si	7.00	No			
4	,19	5976543.00	Poliglobulia	19.30	55.30	Hombre	88.00	18.00	Entre 12 y ...	Ex-fumador	No	,00	No			
5	,20	3453456.00	Anemia	9.60	33.00	Hombre	73.00	13.00	Entre 12 y ...	No	Si	2.00	No			
6	,20	5243521.00	Nivel normal	17.20	50.30	Hombre	80.00	9.00	Menos de ...	No	Si	5.80	No			
7	,23	4665347.00	Nivel normal	17.10	46.80	Hombre	84.00	26.00	Entre 12 y ...	No	Si	8.00	No			
8	,24	6100000.00	Poliglobulia	19.60	55.30	Hombre	90.00	20.00	Entre 12 y ...	Si	No	,00	No			
9	,25	4500000.00	Nivel normal	13.50	41.00	Hombre	83.00	20.00	Entre 12 y ...	No	Si	6.00	No			
10	,30	3467524.00	Anemia	11.30	34.00	Hombre	76.00	12.00	Entre 12 y ...	No	Si	6.00	No			
11	,40	3476565.00	Anemia	10.30	28.97	Hombre	70.00	8.00	Menos de ...	No	Si	5.00	No			
12	,40	5900000.00	Nivel normal	17.50	53.00	Hombre	81.00	7.00	Menos de ...	No	Si	7.00	No			
13	,50	5856432.00	Nivel normal	15.30	46.30	Hombre	81.00	4.00	Menos de ...	No	Si	6.00	No			
14	,55	5346476.00	Nivel normal	14.90	50.80	Hombre	84.00	12.00	Entre 12 y ...	No	No	,00	No			
15	,60	5400876.00	Poliglobulia	17.80	48.20	Mujer	84.00	23.00	Entre 12 y ...	Si	No	,00	No			
16	,60	5466751.00	Nivel normal	14.80	44.30	Hombre	75.00	14.00	Entre 12 y ...	No	Si	5.00	No			
17	,60	4677345.00	Nivel normal	16.20	46.00	Hombre	82.00	18.00	Entre 12 y ...	No	Si	7.00	No			
18	,65	6012344.00	Poliglobulia	18.30	54.70	Hombre	93.00	26.00	Entre 12 y ...	Si	No	6.50	No			
19	,66	5987633.00	Poliglobulia	19.30	54.90	Hombre	69.00	3.00	Menos de ...	Ex-fumador	No	,00	No			
20	,66	5976543.00	Poliglobulia	14.90	56.70	Hombre	83.00	14.00	Entre 12 y ...	Ex-fumador	No	,00	No			
21	,67	5644534.00	Nivel normal	16.80	47.60	Hombre	83.00	18.00	Entre 12 y ...	No	Si	6.00	No			
22	,67	5354366.00	Nivel normal	14.20	51.20	Hombre	85.00	20.00	Entre 12 y ...	No	Si	7.00	No			
23	,78	4355621.00	Nivel normal	13.40	36.90	Mujer	78.00	8.00	Menos de ...	No	Si	4.00	No			

9.2. La regresión logística binaria a través de un ejemplo aplicado

Supongamos que en una residencia de mayores se está llevando a cabo un estudio acerca de la prevalencia y causas asociadas de prurito sine materias.

Supongamos también que los investigadores disponen de una muestra de 100 usuarios acerca de los cuales han medido estas variables independientes:

- El nivel de IgE
- El nivel de hematíes, de hemoglobina
- El hematocrito
- La edad, el tiempo en la residencia en meses
- La práctica de ejercicio y las horas semanales dedicadas a esta actividad
- La presencia o no de tabaquismo

Supongamos, además, que toman como variable dependiente a la siguiente:

Prurito sine materiae. Valores: presencia de prurito = 1, ausencia de prurito = 0

Finalmente, supongamos que los investigadores desean cuantificar la posible relación entre la presencia de prurito y el nivel de IgE como posible factor asociado.

En un caso como este, el modelo de regresión lineal simple no nos sirve, ya que en la ecuación:

$$\text{Prurito sine materiae} = a + b(\text{Nivel de IgE})$$

Se ha escogido un formato cualitativo nominal para la variable dependiente, cuyos valores son 1 (presencia de prurito) y 0 (ausencia de prurito).

En estos casos hay que utilizar un modelo en el que la variable dependiente sea la probabilidad p de que un usuario padezca prurito sine materiae de la siguiente forma:

$$\text{Ln}(p/1-p)$$

El uso del logaritmo neperiano de la probabilidad de que un usuario padezca prurito nos permite que la variable dependiente pase a ser cuantitativa, es decir, que pueda tomar cualquier valor en el intervalo 0-1 (que es el rango de valores de las probabilidades) en lugar de estrictamente (0,1) de forma que, ahora sí que podemos plantear una regresión lineal de esta forma:

$$\text{Ln}(p/1-p) = a + b (\text{Nivel de IgE})$$

Este es un modelo que se puede estimar y, una vez obtenidos los valores de los coeficientes a y b , entonces se puede calcular la probabilidad “ p ” de que un usuario padezca prurito de la siguiente forma:

$$\text{Probabilidad de Prurito} = \frac{1}{1 + e^{-a-b(\text{Nivel de IgE})}}$$

Y este es precisamente el tipo de ecuación que se conoce como modelo logístico binario, donde el número de factores podría ser más de uno. Así en el exponente que figura en el denominador de la ecuación podríamos tener:

$$-a - b_1 \cdot \text{Nivel de IgE} - b_2 \cdot \text{Hematocrito} - b_3 \cdot \text{Edad} - b_4 \cdot \text{Sexo} \quad \text{etc.}$$

(nótese que en este tipo de regresión se pueden combinar variables explicativas cuantitativas y cualitativas, pero nosotros nos limitaremos a ver una regresión simple con variable independiente cuantitativa).

Al estimar la regresión logística binaria sobre los datos recogidos en el estudio utilizando SPSS y esta sintaxis, se obtienen los siguientes resultados que comentamos seguidamente:

LOGISTIC REGRESSION VARIABLES PRURITO /METHOD = ENTER IgE /CLASSPLOT /PRINT = GOODFIT CI(95) /CRITERIA = PIN(.05) POUT(.10) ITERATE(20) CUT(.5) .

En primer lugar, SPSS proporciona una tabla con el número de casos procesados. En el ejemplo son 100 personas que configuran la muestra de estudio

Resumen del procesamiento de los casos

Casos no ponderados ^a		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	100	100,0
	Casos perdidos	0	,0
	Total	100	100,0
Casos no seleccionados		0	,0
Total		100	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

En segundo lugar, la codificación de la variable dependiente:

Codificación de la variable dependiente

Valor original	Valor interno
NO	0
SI	1

Nos indica que la variable dependiente toma el valor 0 cuando no hay presencia de prurito y el valor 1 cuando hay presencia de prurito.

En tercer lugar, la tabla inicial de clasificación de los casos en que vemos 44 personas que tienen prurito y 56 que no tienen, y a las que, antes de calcular el modelo las pronostica a todas como casos de presencia de prurito, resultado que variará en función de la capacidad predictiva de la variable nivel de IgE:

Tabla de clasificación^{a,b}

Observado			Pronosticado		
			¿Padece prurito sine materiae?		Porcentaje correcto
			Sí	No	
Paso 0	¿Padece prurito sine materiae?	Sí	0	44	,0
		No	0	56	100,0
Porcentaje global					56,0

a. En el modelo se incluye una constante.

b. El valor de corte es ,500

Seguidamente, se inicia la construcción del modelo en el que aparece la constante “a”, que resulta ser no significativa (sig 0,231 > 0,05):

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0 Constante	,241	,201	1,433	1	,231	1,273

A continuación, aparece la variable independiente (nivel de IgE) que aún no está en el modelo, su puntuación para entrar y significación:

Variables que no están en la ecuación

	Puntuación	gl	Sig.
Paso 0 Variables IgE	76,538	1	,000
Estadísticos globales	76,538	1	,000

La significación menor que 0,05 ya nos indica que el Nivel de IgE va a ser una variable con capacidad predictiva y significativa para explicar la presencia de prurito en esta muestra estudiada. Seguidamente, hay que tener en cuenta que siempre que se construye un modelo de regresión es fundamental, antes de pasar a extraer conclusiones, el corroborar que el modelo calculado se ajusta efectivamente a los datos usados para estimarlo. En el caso de la regresión logística SPSS proporciona estadísticos de bondad del ajuste como el siguiente, que nos indica que la variable de Nivel de IgE aporta información y es significativa (sig. < 0,05). Así, en el caso de la regresión logística una idea bastante intuitiva es calcular la probabilidad de aparición del suceso, presencia de prurito en nuestro caso, para todos los usuarios de la muestra. Si el ajuste es bueno, es de esperar que un valor alto de probabilidad se asocie con presencia real de prurito, y viceversa, si el valor de esa probabilidad calculada es bajo, cabe esperar también ausencia de prurito. Esta idea intuitiva se lleva a cabo formalmente mediante la prueba conocida como de Hosmer-Lemeshow (1989), que básicamente consiste en dividir el recorrido de la probabilidad en deciles de riesgo (esto es probabilidad de prurito < 0.1, < 0.2, y así hasta <1) y calcular tanto la distribución de casos positivos, como de negativos prevista por la ecuación y los valores realmente observados. Ambas distribuciones, esperada y observada, se contrastan mediante una prueba de Chi². En nuestro ejemplo, la prueba es significativa ya que la significación es menor que 0,05, tal

y como se ve en esta tabla que calcula SPSS:

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	103,862	1	,000
	Bloque	103,862	1	,000
	Modelo	103,862	1	,000

Por otro lado, la R cuadrado de Cox y Snell es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por la(s) variable(s) independiente(s). En el caso de nuestro ejemplo es muy elevada, tal y como se aprecia en la siguiente tabla:

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	33,324 ^a	,646	,866

a. La estimación ha finalizado en el número de iteración 7 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Sin embargo, la R cuadrado de Cox y Snell tiene un defecto: su valor máximo queda por debajo de 1 incluso en un modelo perfecto. La R cuadrado de Nagelkerke, también mostrada en la tabla, corrige este defecto re-escalando la de Cox y Snell, de manera que sus valores se mueven entre 0 y 1. Por consiguiente, tomamos este indicador y podemos decir que, para esta muestra de personas, el Nivel de IgE explica en un 86,6% la aparición de prurito sine materiae. La bondad del ajuste se interpreta mejor en forma de porcentaje, por eso multiplicamos la R cuadrado por cien. La tabla final de clasificación que calcula SPSS se corresponde con el resultado anterior:

Tabla de clasificación^a

Observado			Pronosticado		
			¿Padece prurito sine materiae?		Porcentaje correcto
			Sí	No	
Paso 1	¿Padece prurito sine materiae?	Sí	42	2	95,5
		No	3	53	94,6
Porcentaje global					95,0

a. El valor de corte es ,500

Así, de las 44 personas con prurito, el modelo clasifica bien a 42 y pronostica que 2 no padecerían esta molestia según su nivel de IgE. Por consiguiente, en este grupo clasifica bien al 95,5% de los casos. Por otro lado, de las 56 personas sin prurito, el modelo clasifica bien a 53, lo que representa un 94,6% de este grupo. El resultado final global es que este modelo

clasifica correctamente un 95% de los casos, lo que le otorga una muy buena capacidad predictiva.

El modelo estimado final es:

Variables en la ecuación								
		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95,0% para EXP(B)
Paso	IgE	-,345	,065	27,800	1	,000	,708	Inferior Superior
	Constante	4,562	1,000	20,794	1	,000	95,756	,623 ,805

a. Variable(s) introducida(s) en el paso 1: IgE.

Por consiguiente, se obtiene la siguiente ecuación:

$$\ln(p/1-p) = 4,562 - 0,345 (\text{Nivel de IgE})$$

A partir del cual, finalmente, podemos obtener la probabilidad de tener prurito (nótese que la constante, aunque no significativa, toma un valor positivo en el modelo estimado):

$$\text{Probabilidad de Prurito} = \frac{1}{1 + e^{a-b(\text{Nivel de IgE})}}$$

A partir de los datos de la ecuación, se puede programar en una hoja Excel una simulación de probabilidades de padecer prurito en función del Nivel de IgE:

Modelo estimado			Nivel de IgE	Cálculo de (a-b*Nivel de IgE)	Probabilidad de Prurito
			0,35	4,44125	0,011644
Constante a	4,562		0,5	-0,5175	0,626563
Coefficiente b	-0,345		0,75	0,74125	0,322731
Necesario para cociente	1		1	-0,345	0,585405
			1,5	-0,5175	0,626563
			2	-0,69	0,665967
			2,5	-0,8625	0,703183
			3	-1,035	0,737884
			5	-1,725	0,848772
			10	-3,45	0,969231
			12	-4,14	0,984327
			15	-5,175	0,994376
			20	-6,9	0,998993
			30	-10,35	0,999968

Como se puede ver, se colocan los valores de la constante “a” (4,562) y del coeficiente “b” (-0,345) en unas casillas, además de un valor 1 que se necesitará para la fórmula de cálculo de la probabilidad. Seguidamente, se crean valores de Nivel de IgE: en este ejemplo, se parte del valor de referencia 0,35 kU/L y se va subiendo arbitrariamente para ver lo que pasa en valores intermedios y elevados. A continuación, se calculan los valores de la expresión (a-b*Nivel de IgE) para cada valor de nivel de IgE y, finalmente, se calcula la probabilidad de presencia de prurito, aplicando la fórmula: Probabilidad de Prurito = 1/(1+EXP(a-b*Nivel de IgE)) en la última columna del Excel.

Los resultados indican que el nivel de referencia de IgE arroja una probabilidad muy baja de padecer prurito (0,011644). En cambio, para un nivel de IgE de 0,5 kU/L la probabilidad ya se acerca al 63% y, para un nivel de 30 kU/L la probabilidad supera el 99%. Los modelos de regresión logística permiten la realización de simulaciones de este tipo que pueden resultar muy útiles para planificar actuaciones. Si el modelo es múltiple, aún es más interesante, ya que permite explorar diversos escenarios dando valores a diferentes variables para un mismo fenómeno.

El ejemplo que hemos desarrollado es muy simple y los datos han sido “algo manipulados” para obtener un resultado claro y fácilmente interpretable. El ejemplo sirve para exponer y comprender esta técnica que es una de las más utilizadas en estudios epidemiológicos. En la realidad se calculan modelos más complejos en que intervienen varias variables explicativas y no sólo una. En la realidad, el nivel de IgE puede ser un factor explicativo de la aparición de prurito sine materiae, pero combinado con muchos otros y no tan contundente como en el ejemplo. Conocer los mecanismos de la regresión logística resulta casi imprescindible para leer artículos científicos en el ámbito de las ciencias de la salud y para diseñar estudios epidemiológicos. Existe una muy amplia literatura en que se emplea esta técnica, que se puede hallar explorando entre los múltiples artículos científicos que ofrece la red.

9.3. Ejemplo complementario

En un conjunto de clínicas especializadas en dermatología se ha realizado una experiencia con 100 pacientes con dermatitis seborreica y presencia de colesterol. A 50 pacientes se les ha aplicado una dieta especial y a los otros 50 se les ha dado un fármaco (tratamiento) experimental. Ambos métodos se han aplicado con el objetivo de reducir el colesterol y ver si con ello mejora la patología. Al cabo de un mes, se calculó el porcentaje de cambio (reducción) del colesterol obtenido para cada paciente tras el tratamiento. Seguidamente, se realizó un análisis de regresión logística binaria para ver qué fue más efectivo: ¿la dieta o el fármaco experimental?

La variable dependiente fue el haber tomado el fármaco (valor 1) o el no haberlo tomado (valor 0) y, por consiguiente, haber seguido la dieta y la independiente el cambio en el nivel de colesterol (RCL) de carácter cuantitativo continuo. Los resultados que proporcionó SPSS acerca de este análisis son los que se muestran a continuación.

Resumen del procesamiento de los casos

Casos no ponderados ^a		N	Porcentaje
Casos seleccionados	Incluidos en el análisis	100	100,0
	Casos perdidos	0	,0
	Total	100	100,0
Casos no seleccionados		0	,0
Total		100	100,0

a. Si está activada la ponderación, consulte la tabla de clasificación para ver el número total de casos.

Codificación de la variable dependiente

Valor original	Valor interno
NO	0
SÍ	1

Tabla de clasificación^{a,b}

Observado			Pronosticado		
			¿Ha tomado el nuevo fármaco?		Porcentaje correcto
			NO	SÍ	
Paso 0	¿Ha tomado el nuevo fármaco?	NO	0	50	,0
		SÍ	0	50	100,0
Porcentaje global					50,0

a. En el modelo se incluye una constante.

b. El valor de corte es ,500

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 0 Constante	,000	,200	,000	1	1,000	1,000

Variables que no están en la ecuación

		RCL	Puntuación	gl	Sig.
Paso 0	Variables		24,390	1	,000
	Estadísticos globales		24,390	1	,000

Pruebas omnibus sobre los coeficientes del modelo

		Chi-cuadrado	gl	Sig.
Paso 1	Paso	27,835	1	,000
	Bloque	27,835	1	,000
	Modelo	27,835	1	,000

Resumen de los modelos

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	110,795 ^a	,243	,324

a. La estimación ha finalizado en el número de iteración 5 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla de clasificación^a

Observado			Pronosticado		
			¿Ha tomado el nuevo fármaco?		Porcentaje correcto
			NO	Sí	
Paso 1	¿Ha tomado el nuevo fármaco?	NO	41	9	82,0
		Sí	18	32	64,0
	Porcentaje global				73,0

a. El valor de corte es ,500

Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1	RCL	,491	,113	18,777	1	,000	1,634
	Constante	-6,236	1,429	19,050	1	,000	,002

a. Variable(s) introducida(s) en el paso 1: RCL.

Con todos los resultados obtenidos, se realizó una simulación acerca del comportamiento de la reducción de colesterol y su relación con haber tomado o no el fármaco experimental. La simulación es la siguiente:

Modelo estimado		Reducción del colesterol %	Cálculo de (-a-b*reducción del colesterol)	Probabilidad de haber tomado el fármaco exp.
		5	2,699	0,06303239
Constante a	-6,236	6	0,529	0,370750152
Coefficiente b	0,491	7	-1,641	0,837670962
		8	-3,811	0,978352922
		9	-5,981	0,997480067
		10	-8,151	0,999711636
		11	-10,321	0,999967067
		12	-12,491	0,99999624
		13	-14,661	0,999999571
		14	-16,831	0,999999951
		15	-19,001	0,999999994
		16	-21,171	0,999999999
		17	-23,341	1

En base a los resultados, se extraen las siguientes conclusiones acerca de este caso:

- La variable que recoge la reducción del colesterol explica un 32,4% del tipo de tratamiento aplicado al paciente, según indica la R2 de Nagelkerke.
- El modelo predictivo que hemos estimado adopta la forma:

$$\ln(p/1-p) = -6,236 + 0,491(\% \text{ reducción del colesterol})$$

- Según esta ecuación y modelo el efecto de la toma del fármaco experimental es positivo
- Según la tabla de clasificación final, el modelo tiene mayor capacidad de identificación de los pacientes que hicieron dieta que de los que tomaron el fármaco

- En función de los resultados de la simulación, los pacientes en que se registran reducciones de colesterol del 7% y más tienen una probabilidad superior al 80% de haber tomado el fármaco

Este modelo es mucho más discreto que el de la sección anterior en términos de ajuste, pero se aproxima mucho más a los que nos podemos encontrar en la realidad.

10. ANÁLISIS DE SUPERVIVENCIA

10.1. Concepto general

El análisis de supervivencia de Kaplan-Meier es una metodología estadística que se aplica cuando se investiga el tiempo hasta que ocurre un evento de nuestro interés. Para su realización, el analista precisa recoger información de al menos dos variables:

- 1) El tiempo que transcurre entre un momento cero y la producción del suceso de interés
- 2) La ocurrencia o no del evento terminal

Esta técnica es de gran ayuda en la realización de estudios quirúrgicos y de otros tipos, especialmente en relación con la oncología, donde muy habitualmente los profesionales necesitan valorar el tiempo que transcurre hasta un suceso tal como una defunción, una recidiva, una reaparición de síntomas u otros. Aunque el análisis se denomina de supervivencia, el evento de interés no tiene por qué tratarse de la defunción. De hecho, la técnica toma el nombre supervivencia del tiempo hasta que se da el evento a partir de un momento tomado como referencia. Las principales diferencias o valor añadido que aporta esta técnica respecto de otras, es que, por un lado, permite el estudio del suceso de interés cuando no está presente en todos los individuos de una muestra y, por otro lado, que no precisa de que la distribución de los datos siga una distribución Normal (requisito exigido por muchas técnicas estadísticas y que dificultaría el análisis de este tipo de datos).

Por consiguiente, en los estudios en que se aplica el análisis de supervivencia, cada paciente participante es seguido durante un determinado período y, durante el período de observación, se recoge el intervalo de tiempo entre el momento de inicio del estudio y la ocurrencia del evento de interés, o hasta que acaba el período de estudio sin que se haya producido el evento en cuestión (Rebasa, 2005).

Esta metodología, ofrece pruebas estadísticas para crear y comparar curvas de supervivencia que, como veremos, se construyen a partir de los datos recopilados sobre muestras de pacientes participantes en estudios, así como modelos basados en técnicas de regresión, adecuados para valorar cuál es el efecto de una o más variables explicativas en la ocurrencia del evento.

Téngase en cuenta que, en las enfermedades crónicas, tales como el cáncer, la supervivencia se mide como una probabilidad de permanecer vivo durante una determinada cantidad de tiempo. La supervivencia al año o a los 5 años se utilizan frecuentemente como indicadores de la severidad de una enfermedad y también como pronóstico. Así, el pronóstico del cáncer se valora usualmente determinando el porcentaje de pacientes que sobrevive al menos cinco años después del diagnóstico (Pita Fernández, 2001).

Dada la amplitud del tema, el objetivo de esta sección es el de proporcionar una pincelada acerca de las posibilidades que ofrecen estas técnicas a través de una familiarización con sus principales conceptos y terminología, aprender a estimar la proporción acumulada de supervivencia y mostrar las pruebas estadísticas que se usan habitualmente para comparar dos curvas de supervivencia.

10.2. Elementos básicos para la realización de un análisis de supervivencia

En primer lugar, se establece el suceso acerca del cual vamos a analizar la supervivencia. Supongamos que, en nuestro caso fuese el fallecimiento del paciente. Supongamos, además, que la observación de cada paciente se inicia en el momento del diagnóstico y que, en principio continua hasta que se produce el fallecimiento o hasta que se interrumpe por otro motivo. La forma de expresar este planteamiento es:

Tiempo 0 = fecha del diagnóstico

Tiempo 1 = fecha del fallecimiento

Tiempo 2 = fecha de interrupción del seguimiento por otro motivo distinto del fallecimiento

Tiempo 3 = fecha de finalización del período de observación establecido

Cuando el tiempo de seguimiento se interrumpe antes del fallecimiento o del final del período de observación, al paciente en este tipo de caso se le denomina habitualmente “censurado”. En cualquier caso, el período de seguimiento de pacientes puede terminar por las razones siguientes:

- a. El paciente decide abandonar el estudio.
- b. El paciente se pierde por algún motivo y no se dispone de más información.
- c. El estudio del paciente termina antes de producirse el suceso de interés.
- d. El paciente fallece por causas que no guardan relación con el estudio.

Cuando, por cualquier motivo, los tiempos de supervivencia no se conocen con exactitud, los datos se consideran censurados. Un tiempo de seguimiento válido siempre viene definido por una fecha de inicio y una fecha de cierre que permiten calcular el tiempo transcurrido entre ambas. Ahora bien, en este tipo de estudio, las fechas de inicio y cierre son diferentes para cada sujeto, ya que cada uno se incorpora en un momento determinado en función de sus propias circunstancias.

Una vez tenido en cuenta lo anterior, el analista debe definir perfectamente qué se entiende por estado inicial del paciente o sujeto a estudiar y que se entiende por suceso o evento a estudiar.

Entre las opciones más habituales de estado inicial destacan: fecha de diagnóstico, fecha de una intervención quirúrgica, fecha de inicio de un tratamiento y similares. En cualquier caso, se trata de una fecha que corresponde a un punto de partida que tiene asociado un estado del paciente y que, habitualmente, son diferentes para cada sujeto del estudio.

Entre las opciones más habituales de acontecimiento o suceso de interés destacan: fecha de recuperación o alta, fecha de fallecimiento, fecha de aparición de un acontecimiento adverso, fecha de recaída, fecha de recidiva y similares. En cualquier caso, se trata de una fecha que corresponde a un punto en que se produce un evento determinado y, como en el caso del estado inicial, habitualmente, son diferentes para cada sujeto del estudio.

Seguidamente, hay que planificar que, en la última observación se deben registrar dos variables fundamentales como son el estado del sujeto en ese momento y la fecha.

A partir de toda esta información, se pueden establecer y calcular:

- El tiempo de supervivencia de cada sujeto = tiempo transcurrido entre la fecha del estado inicial y la fecha del acontecimiento de estudio
- Tiempos incompletos = tiempo transcurrido entre la fecha inicial y la pérdida del paciente por cualquier motivo no relacionado con el estudio
- Duración del estudio = tiempo transcurrido entre la fecha de inicio del estudio y la fecha prevista de cierre

El analista debe tener en cuenta que en la última observación del sujeto hay que anotar su estado en ese momento y la fecha de información acerca del mismo. El período de tiempo transcurrido entre la fecha de entrada y la fecha de la última observación o contacto se conoce como tiempo de participación en el estudio.

Además de las informaciones acerca de tiempos y estados inicial y final, los analistas suelen contar con datos adicionales acerca de variables explicativas que pueden influir en la supervivencia del sujeto. Algunas de estas variables pueden ser constantes en el tiempo, como es el caso del sexo del paciente o determinadas características y factores genéticos, mientras que otras variables se irán modificando a lo largo del tiempo, como es el caso del peso, las dosis de un tratamiento, las recurrencias y otras.

10.3. Aplicación práctica del análisis de supervivencia simple

En su aplicación más simple, para realizar un estudio de supervivencia sólo se necesitan dos variables: la que recoge el tiempo de seguimiento del sujeto y una variable dicotómica que indique si el tiempo disponible es censurado (valor 0) o completo y válido (valor 1). Un análisis más elaborado puede incorporar la comparación de los sujetos en función de alguna variable que los estratifique en grupos de sexo u otras variables, o que los clasifique acerca de un aspecto relevante del estudio.

Para comprender el funcionamiento del análisis de supervivencia vamos a ilustrarlo con un ejemplo en que se estudiaron retrospectivamente una serie de historias clínicas de pacientes que fueron operadas de cáncer de ovarios en un área sanitaria en los últimos veinte años y que fueron seguidas, en algunos casos, hasta durante casi 15 años.

El programa SPSS ofrece diversos procedimientos en el entorno de los análisis de supervivencia. En este ejemplo vamos a aplicar el método de Kaplan-Meier, para obtener los estadísticos más habituales, construir gráficas de supervivencia y compararlas mediante la prueba del Log-Rank.

Supongamos que a partir de la fecha de inicio y de finalización de cada seguimiento se calcularon los tiempos de supervivencia de las pacientes. La fecha de inicio se estableció como el momento de consulta prequirúrgica y la de finalización se estableció en los 20 años que abarcaban las historias clínicas. Además, las pacientes fueron clasificadas en función de haber recibido o no tratamiento quimioterápico antes de la cirugía.

Como ya se ha comentado, durante el período de tiempo abarcado por el estudio, se hizo un seguimiento individual de cada historia clínica, de forma que los analistas se encontraron con diversos tipos de casos: fallecimientos, supervivencias y pérdida de contacto o del seguimiento por parte del equipo que estuviese llevando a la paciente. Con la información, se pudo clasificar a las pacientes en casos de muerte y censurados.

Con toda esta información, los analistas crearon una base de datos en formato SPSS que contenía 80 casos. Las variables que integran esta base de datos son: ID como identificador del paciente mediante un número exclusivo del 1 al 80; el tiempo de supervivencia (en este caso, definido como el transcurrido desde la fecha de consulta prequirúrgica hasta la fecha del último contacto con la paciente, ya fuese por fallecimiento o censura) en meses; el estado (censura con el código 0 y muerte con el código 1) y, finalmente la variable que indica si la paciente recibió sesiones quimioterápicas antes de la cirugía (valor 0 = no y valor 1 = sí).

	ID	Tiempo_S	Estado	Quimio	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR	VAR
1	1.00	14.00	Muerte	Si															
2	2.00	19.00	Muerte	Si															
3	3.00	20.00	Muerte	Si															
4	4.00	60.00	Censura	Si															
5	5.00	45.00	Censura	No															
6	6.00	38.00	Muerte	No															
7	7.00	40.00	Censura	Si															
8	8.00	58.00	Muerte	Si															
9	9.00	40.00	Muerte	No															
10	10.00	102.00	Censura	No															
11	11.00	128.00	Censura	No															
12	12.00	70.00	Censura	Si															
13	13.00	47.00	Muerte	Si															
14	14.00	40.00	Muerte	Si															
15	15.00	55.00	Censura	Si															
16	16.00	100.00	Censura	No															
17	17.00	73.00	Censura	No															
18	18.00	80.00	Censura	Si															
19	19.00	24.00	Muerte	Si															
20	20.00	26.00	Muerte	Si															
21	21.00	43.00	Censura	Si															
22	22.00	108.00	Censura	No															
23	23.00	14.00	Muerte	No															

Antes de comenzar el análisis, es importante comprobar que la variable que recoge el tiempo está clasificada dentro de la base de datos como cuantitativa (escala en SPSS) y que la variable que recoge el estado está clasificada como nominal (esto se comprueba en la pantalla de variables de SPSS, al final, a la derecha). Es también muy importante comprobar cómo está codificada la variable estado y tener presente con qué código se identifica a los sujetos que tuvieron el acontecimiento evaluado (muerte en este caso), puesto que el programa nos lo va a pedir. El ejemplo ofrece los resultados que proporciona SPSS al utilizar el método de Kaplan-Meier, el más habitual en este tipo de análisis. Se trata de un método no paramétrico, que tiene muy pocas restricciones, ya que, de hecho, el único supuesto en que se basa es el de que los sujetos censurados se habrían comportado del mismo modo que los seguidos hasta que se produjo el evento (lo que se conoce por “censura no informativa”).

Una vez introducidos los datos, al ejecutar el análisis, SPSS ordena internamente de menor a mayor todos los tiempos de supervivencia observados, anotando para cada uno de ellos el número de muertes y censuras producidas. El programa calcula para cada periodo de tiempo la probabilidad de sobrevivir, de forma que la función de Kaplan-Meier es “la probabilidad de supervivencia individual acumulada a lo largo del tiempo”. Todos estos pasos los hace automáticamente el programa estadístico.

Ahora, para llevar a cabo el procedimiento, utilizamos los menús de SPSS y escogemos la opción Analizar, Supervivencia y, finalmente Kaplan-Meier. Cuando se abre la nueva ventana de diálogo, seleccionamos la variable que mide el tiempo de seguimiento y la colocamos en la casilla correspondiente. Ahora, en la casilla Estado debe colocarse la variable categórica dicotómica que establece el estado individual durante el seguimiento (censura y muerte). El programa nos pide Definir evento (aparece un paréntesis con un signo de interrogación tras el nombre de la variable) y hay que pulsar esa pestaña y especificar el valor que tiene asignado el resultado que se evalúa (en nuestro caso el valor 1 = muerte).

Para finalizar, se pueden pedir opciones interesantes pulsando la pestaña de la parte inferior de la ventana y denominada Opciones. Así, se pueden pedir “Tablas de supervivencia” y otros estadísticos, como la “Media y mediana de supervivencia” y los “Cuartiles”. También podemos obtener diferentes gráficas siendo lo habitual señalar “Supervivencia”.

La sintaxis es la siguiente:

KM Tiempo_S /STATUS=Estado(1) /PRINT TABLE MEAN /PERCENTILES /PLOT SURVIVAL .

Y, tras ejecutarla, el resultado que se obtiene para este ejemplo es el siguiente:

Resumen del procesamiento de los casos

Nº total	Nº de eventos	Censurado	
		Nº	Porcentaje
80	43	37	46,3%

Esta tabla nos informa de que se han procesado 80 casos de pacientes, de los cuales 43 (53,7%) han mostrado el evento o muerte y 37 (46,3%) han sido censurados o, en otras palabras, han sobrevivido o, no han fallecido por causa de un cáncer de ovarios.

Tabla de supervivencia

	Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos que permanecen
			Estimación	Error típico		
1	14,000	Muerte	.	.	1	79
2	14,000	Muerte	,975	,017	2	78
3	16,000	Muerte	,963	,021	3	77
4	17,000	Muerte	,950	,024	4	76
5	17,000	Censura	.	.	4	75
6	19,000	Muerte	,937	,027	5	74
7	20,000	Muerte	.	.	6	73
8	20,000	Muerte	,912	,032	7	72
9	24,000	Muerte	,899	,034	8	71
10	25,000	Muerte	,887	,036	9	70
11	25,000	Censura	.	.	9	69
12	26,000	Muerte	,874	,037	10	68
13	27,000	Muerte	,861	,039	11	67
14	28,000	Muerte	,848	,040	12	66
15	29,000	Muerte	,835	,042	13	65
16	30,000	Muerte	,822	,043	14	64
17	34,000	Muerte	,810	,044	15	63
18	36,000	Censura	.	.	15	62
19	38,000	Muerte	,797	,045	16	61
20	39,000	Muerte	,783	,047	17	60
21	40,000	Muerte	.	.	18	59
22	40,000	Muerte	,757	,049	19	58
23	40,000	Censura	.	.	19	57
24	43,000	Muerte	,744	,049	20	56
25	43,000	Censura	.	.	20	55
26	43,000	Censura	.	.	20	54
27	44,000	Muerte	,730	,050	21	53
28	44,000	Censura	.	.	21	52
29	45,000	Censura	.	.	21	51
30	46,000	Muerte	,716	,051	22	50
31	47,000	Muerte	.	.	23	49
32	47,000	Muerte	,687	,053	24	48
33	48,000	Muerte	.	.	25	47
34	48,000	Muerte	,659	,055	26	46
35	53,000	Censura	.	.	26	45
36	55,000	Censura	.	.	26	44
37	56,000	Muerte	.	.	27	43
38	56,000	Muerte	,629	,056	28	42
39	58,000	Muerte	,614	,057	29	41
40	58,000	Censura	.	.	29	40
41	58,000	Censura	.	.	29	39
42	60,000	Censura	.	.	29	38
43	63,000	Muerte	,598	,058	30	37
44	65,000	Censura	.	.	30	36
45	68,000	Muerte	.	.	31	35
46	68,000	Muerte	,564	,059	32	34
47	70,000	Censura	.	.	32	33
48	73,000	Censura	.	.	32	32
49	74,000	Censura	.	.	32	31
50	76,000	Censura	.	.	32	30

51	78,000	Censura	.	.	32	29
52	79,000	Muerte	,545	,060	33	28
53	80,000	Muerte	,525	,061	34	27
54	80,000	Censura	.	.	34	26
55	83,000	Censura	.	.	34	25
56	84,000	Censura	.	.	34	24
57	90,000	Censura	.	.	34	23
58	100,000	Censura	.	.	34	22
59	100,000	Censura	.	.	34	21
60	102,000	Censura	.	.	34	20
61	103,000	Muerte	,499	,063	35	19
62	108,000	Censura	.	.	35	18
63	110,000	Muerte	,471	,066	36	17
64	110,000	Censura	.	.	36	16
65	112,000	Muerte	,442	,068	37	15
66	118,000	Censura	.	.	37	14
67	119,000	Censura	.	.	37	13
68	120,000	Muerte	.	.	38	12
69	120,000	Muerte	,374	,072	39	11
70	120,000	Censura	.	.	39	10
71	122,000	Muerte	,337	,074	40	9
72	126,000	Censura	.	.	40	8
73	126,000	Censura	.	.	40	7
74	128,000	Muerte	,289	,078	41	6
75	128,000	Censura	.	.	41	5
76	130,000	Muerte	,231	,081	42	4
77	145,000	Censura	.	.	42	3
78	156,000	Censura	.	.	42	2
79	167,000	Muerte	,115	,091	43	1
80	178,000	Censura	.	.	43	0

En la segunda tabla la columna de la izquierda (Tiempo) es el tiempo durante el cual las pacientes han estado en seguimiento, ordenados por tiempos crecientes. La segunda columna (Estado) nos indica si se ha producido el desenlace evaluado o no, en nuestro caso etiquetados como “muerte” (en caso contrario aparece el valor “censura”, que corresponde a las pacientes perdidas o retiradas durante el seguimiento y, también, a las que al final del periodo de estudio seguían sin tener el evento, esto es, estaban vivas). La siguiente columna (Proporción acumulada que sobrevive hasta el momento) es la supervivencia acumulada, o proporción de casos para los que no ha tenido lugar el evento en cada tiempo. La siguiente columna (Error estándar) es el error estándar correspondiente a la estimación puntual de Kaplan-Meier en cada tiempo. La quinta columna (N.º de eventos acumulados) recoge los desenlaces acumulados, esto es, las pacientes que han muerto hasta ese tiempo. La última columna (N.º de casos que permanecen) es el número de pacientes que quedan en cada momento sin que haya ocurrido en ellas el evento final o desenlace evaluado, y representan a las que “están a riesgo en el siguiente periodo”.

Seguidamente el programa SPSS nos proporciona los valores del tiempo medio de supervivencia (95,141 meses) y de la mediana de supervivencia (el tiempo en el que el 50% de los sujetos siguen sin haber desarrollado el evento, que vale 103 meses), con sus correspondientes errores estándar y los intervalos de confianza al 95%, señalándonos que el estudio está limitado por el periodo máximo de seguimiento (en nuestro ejemplo 147,617≈148 meses).

Medias y medianas del tiempo de supervivencia

Media ^a				Mediana			
Estimación	Error típico	Intervalo de confianza al 95%		Estimación	Error típico	Intervalo de confianza al 95%	
		Límite inferior	Límite superior			Límite inferior	Límite superior
95,141	7,328	80,778	109,504	103,000	22,764	58,383	147,617

a. La estimación se limita al mayor tiempo de supervivencia si se ha censurado.

Si lo hemos pedido, el programa también nos calcula los cuartiles, es decir, el tiempo en el que el 25%, el 50% y el 75% de las pacientes siguen sin haber fallecido, y los acompaña de sus errores estándar para tener una medida de la representatividad de esos valores. En este caso, el menor error estándar corresponde al tercer cuartil (5,258 meses), mientras que para el segundo cuartil hay mucha más dispersión (22,764 meses).

Percentiles

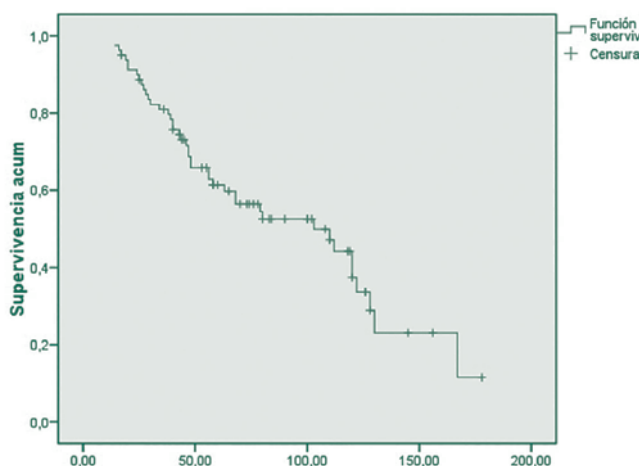
25,0%		50,0%		75,0%	
Estimación	Error típico	Estimación	Error típico	Estimación	Error típico
130,000	16,430	103,000	22,764	43,000	5,258

Finalmente, tenemos la Curva de Supervivencia. En este gráfico se sitúa en la ordenada la supervivencia acumulada en términos de probabilidad (entre 0 y 1), y el tiempo de supervivencia en el eje de abscisas, permitiéndonos responder a preguntas como, por ejemplo:

¿Cuántas pacientes operadas de cáncer de ovarios sobreviven dos años?

Para responder, trazáramos una línea vertical desde el tiempo 24 meses y en el punto donde corte a la curva leeríamos en horizontal el valor que le corresponde en el eje de ordenadas (daría aproximadamente el 90%).

Función de supervivencia



10.4. Aplicación práctica del análisis de supervivencia comparando dos curvas

Siguiendo con el mismo ejemplo, si por ejemplo deseamos comparar las curvas de supervivencia de las pacientes que fueron y no fueron sometidas a tratamiento quimioterápico antes de la cirugía, tenemos que plantear una hipótesis nula y contrastarla. La hipótesis en este ejemplo es:

Ho: la supervivencia media (o mediana) de las pacientes sometidas a quimioterapia es igual que la supervivencia media (o mediana) de las pacientes no sometidas a quimioterapia

Para contrastar la hipótesis hay varias pruebas, siendo la más empleada la del Log-Rank, que tiene en cuenta las diferencias de supervivencia entre los grupos en todos los puntos que dura el seguimiento. Por consiguiente, para llevar a cabo este análisis, seguiríamos los mismos pasos que en el anterior, pero añadiríamos la variable que separa a los dos grupos de pacientes, colocándola en la ventana “Factor”. En nuestro ejemplo, se trata de la variable “Quimio”. Hay que recordar que debemos asegurarnos de que la variable que hemos establecido como factor de comparación es cualitativa (nominal en este caso). Seguidamente, en la pestaña Comparar factor, seleccionamos el estadístico de contraste “Log rango” (y por defecto se activará la opción “Combinada sobre los estratos”. Pulsamos aceptar y obtenemos los siguientes resultados:

Resumen del procesamiento de los casos

Tratamiento pre-operatorio	Nº total	Nº de eventos	Censurado	
			Nº	Porcentaje
No	32	13	19	59,4%
Sí	48	30	18	37,5%
Global	80	43	37	46,3%

Esta primera tabla nos informa de que hemos procesado 80 casos de pacientes, de las cuales 32 no recibieron tratamiento quimioterápico y 48 sí que lo hicieron. Entre las que no lo recibieron, 13 fallecieron y 19 sobrevivieron, lo cual representa un 59,4% de este grupo. Entre las que recibieron el tratamiento, 30 fallecieron y 18 sobrevivieron, lo cual representa un 37,5%. En el conjunto total, 43 pacientes fallecieron y 37 sobrevivieron, lo cual representó un 46,3% en la muestra estudiada. Por consiguiente, intuitivamente, parece que la tasa de supervivencia es superior entre las pacientes que no fueron sometidas a quimioterapia. Sin embargo, esto no se puede comprobar hasta que se efectúa el contraste de hipótesis.

La segunda tabla es como la vista en el apartado anterior, pero para cada uno de los grupos de pacientes, por lo que la primera parte es sobre 32 casos y la segunda parte sobre 48 casos.

Por consiguiente, en la primera mitad de la tabla vemos que la columna de la izquierda (Tiempo) es el tiempo durante el cual las pacientes no sometidas a quimioterapia han estado en seguimiento, ordenados por tiempos crecientes. La segunda columna (Estado) nos indica si se ha producido el desenlace evaluado o no, en nuestro caso etiquetados como “muerte” (en caso contrario aparece el valor “censura”, que corresponde a las pacientes perdidas o retiradas durante el seguimiento y, también, a las que al final del periodo de estudio seguían sin tener el evento, esto es, estaban vivas).

La siguiente columna (Proporción acumulada que sobrevive hasta el momento) es la supervivencia acumulada, o proporción de casos para los que no ha tenido lugar el evento en cada tiempo. La siguiente columna (Error estándar) es el error estándar correspondiente a la estimación puntual de Kaplan-Meier en cada tiempo. La quinta columna (Nº de eventos acumulados) recoge los desenlaces acumulados, esto es, las pacientes que han muerto hasta ese tiempo. La última columna (Nº de casos que permanecen) es el número de pacientes que quedan en cada momento sin que haya ocurrido en ellas el evento final o desenlace evaluado, y representan a las que “están a riesgo en el siguiente periodo”.

La misma lectura se puede hacer en la segunda parte de la tabla, pero esta vez referida a las pacientes que sí que fueron sometidas a quimioterapia antes de la operación quirúrgica.

Tabla de supervivencia

Tratamiento pre-operatorio con quimioterapia		Tiempo	Estado	Proporción acumulada que sobrevive hasta el momento		Nº de eventos acumulados	Nº de casos que permanecen
				Estimación	Error típico		
No	1	14,000	Muerte	,969	,031	1	31
	2	17,000	Muerte	,938	,043	2	30
	3	17,000	Censura	.	.	2	29
	4	25,000	Censura	.	.	2	28
	5	27,000	Muerte	,904	,053	3	27
	6	28,000	Muerte	,871	,061	4	26
	7	30,000	Muerte	,837	,067	5	25
	8	38,000	Muerte	,804	,072	6	24
	9	39,000	Muerte	,770	,076	7	23
	10	40,000	Muerte	,737	,080	8	22
	11	43,000	Muerte	,703	,083	9	21
	12	43,000	Censura	.	.	9	20
	13	45,000	Censura	.	.	9	19
	14	47,000	Muerte	,666	,087	10	18
	15	73,000	Censura	.	.	10	17
	16	74,000	Censura	.	.	10	16
	17	78,000	Censura	.	.	10	15
	18	80,000	Muerte	,622	,092	11	14
	19	83,000	Censura	.	.	11	13
	20	90,000	Censura	.	.	11	12
	21	100,000	Censura	.	.	11	11
	22	100,000	Censura	.	.	11	10
	23	102,000	Censura	.	.	11	9
	24	103,000	Muerte	,553	,104	12	8
	25	108,000	Censura	.	.	12	7
	26	110,000	Censura	.	.	12	6
	27	126,000	Censura	.	.	12	5
	28	128,000	Muerte	,442	,129	13	4
	29	128,000	Censura	.	.	13	3
	30	145,000	Censura	.	.	13	2
	31	156,000	Censura	.	.	13	1
	32	178,000	Censura	.	.	13	0
Sí	1	14,000	Muerte	,979	,021	1	47
	2	16,000	Muerte	,958	,029	2	46
	3	19,000	Muerte	,938	,035	3	45
	4	20,000	Muerte	.	.	4	44
	5	20,000	Muerte	,896	,044	5	43
	6	24,000	Muerte	,875	,048	6	42
	7	25,000	Muerte	,854	,051	7	41
	8	26,000	Muerte	,833	,054	8	40
	9	29,000	Muerte	,813	,056	9	39
	10	34,000	Muerte	,792	,059	10	38

11	36,000	Censura	.	.	10	37
12	40,000	Muerte	,770	,061	11	36
13	40,000	Censura	.	.	11	35
14	43,000	Censura	.	.	11	34
15	44,000	Muerte	,748	,063	12	33
16	44,000	Censura	.	.	12	32
17	46,000	Muerte	,724	,065	13	31
18	47,000	Muerte	,701	,067	14	30
19	48,000	Muerte	.	.	15	29
20	48,000	Muerte	,654	,070	16	28
21	53,000	Censura	.	.	16	27
22	55,000	Censura	.	.	16	26
23	56,000	Muerte	.	.	17	25
24	56,000	Muerte	,604	,073	18	24
25	58,000	Muerte	,579	,075	19	23
26	58,000	Censura	.	.	19	22
27	58,000	Censura	.	.	19	21
28	60,000	Censura	.	.	19	20
29	63,000	Muerte	,550	,076	20	19
30	65,000	Censura	.	.	20	18
31	68,000	Muerte	.	.	21	17
32	68,000	Muerte	,489	,079	22	16
33	70,000	Censura	.	.	22	15
34	76,000	Censura	.	.	22	14
35	79,000	Muerte	,454	,081	23	13
36	80,000	Censura	.	.	23	12
37	84,000	Censura	.	.	23	11
38	110,000	Muerte	,413	,083	24	10
39	112,000	Muerte	,371	,085	25	9
40	118,000	Censura	.	.	25	8
41	119,000	Censura	.	.	25	7
42	120,000	Muerte	.	.	26	6
43	120,000	Muerte	,265	,088	27	5
44	120,000	Censura	.	.	27	4
45	122,000	Muerte	,199	,087	28	3
46	126,000	Censura	.	.	28	2
47	130,000	Muerte	,099	,083	29	1
48	167,000	Muerte	,000	,000	30	0

A continuación, SPSS nos ofrece la tabla con las medias y medianas por un lado y la tabla con los cuartiles por el otro, ambas divididas en resultados para las pacientes sometidas y no sometidas a quimioterapia. A primera vista, la media y la mediana de tiempo de supervivencia aparece como superior en el caso de las pacientes no sometidas a quimioterapia. Sin embargo, no se puede afirmar que exista una diferencia significativa hasta que no se evalúa el resultado de la prueba de contraste.

Medias y medianas del tiempo de supervivencia

Tratamiento pre-operatorio con quimioterapia	Media ^a				Mediana			
	Estimación	Error típico	Intervalo de confianza al 95%		Estimación	Error típico	Intervalo de confianza al 95%	
			Límite inferior	Límite superior			Límite inferior	Límite superior
No	114,421	12,950	89,040	139,802	128,000	29,251	70,668	185,332
Sí	83,518	8,018	67,803	99,234	68,000	24,743	19,504	116,496
Global	95,141	7,328	80,778	109,504	103,000	22,764	58,383	147,617

a. La estimación se limita al mayor tiempo de supervivencia si se ha censurado.

Percentiles

Tratamiento pre-operatorio	25,0%		50,0%		75,0%	
	Estimación	Error típico	Estimación	Error típico	Estimación	Error típico
No	.	.	128,000	29,251	40,000	5,244
Sí	122,000	5,061	68,000	24,743	44,000	7,572
Global	130,000	16,430	103,000	22,764	43,000	5,258

Seguidamente, el programa proporciona el resultado del contraste de hipótesis Log Rank:

Comparaciones globales

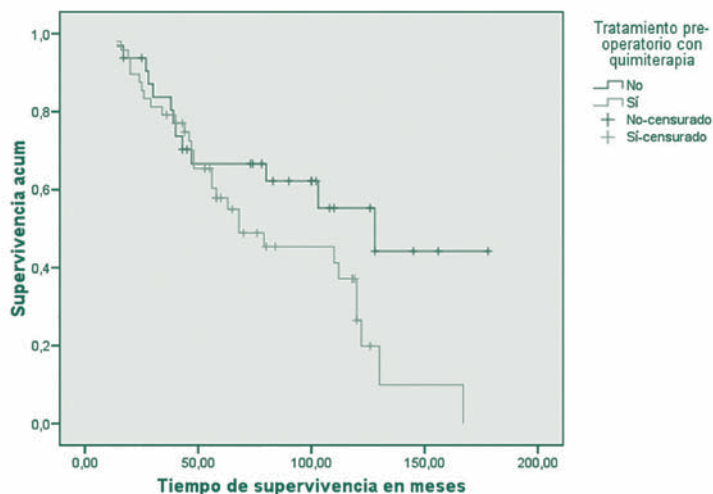
	Chi-cuadrado	gl	Sig.
Log Rank (Mantel-Cox)	2,977	1	,084

El vector de ponderaciones de tendencia es -1, 1. Éste es el valor por defecto.

En la tabla podemos ver el contraste estadístico Log Rank, con el valor del estadístico (2,977), los grados de libertad (1, en nuestro caso, al ser dos grupos los que se comparan) y el valor de significación del estadístico asociado al contraste (0,084), que corresponde a una probabilidad de la distribución que no nos permite rechazar la hipótesis nula al ser mayor que 0,05. Por consiguiente, aunque lo parecía desde un punto de vista descriptivo, la diferencia no es suficientemente relevante y hay que concluir que no hay diferencias estadísticamente significativas en la supervivencia de ambos grupos.

Finalmente, el programa nos ofrece el gráfico con las curvas de supervivencia de ambos grupos. Como puede apreciarse hay una separación entre ambas curvas de supervivencia, a favor del grupo de pacientes no tratadas con quimioterapia (“parecen sobrevivir más”), pero el test estadístico Log Rank ya hemos visto que acepta, aunque sea por poco, la H_0 de “no hay diferencias” en la supervivencia globalmente considerada.

Funciones de supervivencia



Finalmente, los analistas estudiarían en detalle el resultado, ya que un solo factor (haber sido sometidas a quimioterapia o no) es poco informativo de las causas que pueden estar tras la tendencia a una mayor supervivencia aparente entre las pacientes no sometidas a quimioterapia, aunque no se haya llegado a constatar esta diferencia estadísticamente. Sería necesario introducir la gravedad de los casos, la necesidad o no de quimioterapia tras la operación, la influencia de la edad y muchas otras variables. Como a lo largo de todo el cuaderno, los ejemplos utilizados son muy sencillos y no son realistas: su misión es simplemente la de ilustrar el funcionamiento de los métodos estadísticos y proporcionar las claves para la interpretación de los resultados que ofrecen los programas estadísticos como SPSS.

BIBLIOGRAFÍA COMPLEMENTARIA RECOMENDADA

- 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
- Daniel WW. Bioestadística. Base para el análisis de las Ciencias de las Ciencias de la Salud. México: Limusa. 2002
- Silva Ayçaguer, Luis Carlos. Excursión a la regresión logística en ciencias de la salud. Ed. Díaz de Santos, Madrid 1995
- Víctor Abraira Santos, V., Pérez de Vargas Luque, A. Métodos multivariantes en bioestadística. Ed. Centro de Estudios Ramón Areces, Madrid 1996-
- Aguayo Canela, Mariano. Cómo hacer una Regresión Logística con SPSS® “paso a paso”. (I). Servicio de Medicina Interna. Hospital Universitario Virgen Macarena. Sevilla.
- Fiuza Pérez, MD, Rodríguez Pérez, JC. La regresión logística: una herramienta versátil.
- Rebasa, P. (2005). Conceptos básicos del análisis de supervivencia (Basic concepts in survival analysis). Cirugía Española. Vol. 78. Núm. 4. octubre 2005, páginas 209-278. DOI: 10.1016/S0009 739X (05)70923-4. Elsevier
- Pita Fernández, S. Análisis de supervivencia. Unidad de Epidemiología Clínica y Bioestadística. Complejo Hospitalario-Universitario Juan Canalejo. A Coruña (España)
- Cad Aten Primaria 1995; 2: 130-135. Actualización 29/04/01.
https://www.fisterra.com/mbe/investiga/supervivencia/analisis_supervivencia2.pdf
- Aguayo Canela M, Lora Monge E. Cómo hacer “paso a paso” un Análisis de Supervivencia con SPSS para Windows. DOCUWEB FABIS. Dot. Núm 0702006. Servicio de Medicina Interna. Hospital Universitario Virgen Macarena. Sevilla.



RRII 7028

