Metodología cuantitativa: Bioestadística básica

PID_00265798

Manuel Lozano Relaño

Tiempo mínimo de dedicación recomendado: 3 horas

Universitat Oberta de Catalunya

Manuel Lozano Relaño

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por las profesoras: Anna Bach, Marina Bosque

Primera edición: septiembre 2019

Av. Tibidabo, 39-43, 08035 Barcelona
Realización editorial: FUOC

Índice

Introducción

La investigación científica trata de explicar ciertos comportamientos en la población o el efecto sobre un parámetro de salud. Estos comportamientos de la población se definen por medio de una serie de características «medibles», como, por ejemplo, el sexo, la edad, el índice de masa corporal (IMC), la zona de residencia o el consumo de nutrientes. Unas características podrían afectar al parámetro de salud de interés y otras no. Y las que lo hacen lo harán en sentido positivo o negativo, en mayor o menor medida.

La herramienta para medir cómo, cuánto y en qué sentido se producen estas asociaciones se conoce como **investigación o metodología cuantitativa**. Esta metodología es el procedimiento de identificar y cuantificar, de entre todas las características de la población, cuáles resultan significativas en la explicación del patrón de comportamiento del parámetro de salud que nos interesa. De esta manera, la investigación cuantitativa se basa en el análisis de la causa y el efecto. Para que exista metodología cuantitativa se requiere que entre estas características y el parámetro estudiado exista una relación cuya naturaleza sea representable por algún modelo numérico. Para ello se usan magnitudes numéricas que pueden ser tratadas mediante herramientas del campo de la estadística. En el ámbito de la salud, esto se conoce como **bioestadística**.

1. Planificación de la investigación

Toda investigación comienza con la idea de que ciertos aspectos de la población afectan a un parámetro de salud que es de nuestro interés.

Por ejemplo, cómo afecta el tipo de dieta a la incidencia de cierta enfermedad, o cómo afectan las características sociodemográficas de una población al correcto cumplimiento de una dieta saludable. Este efecto se mide en un **estudio**, que puede ser de diferentes tipos (retrospectivo, prospectivo, cohortes, casos y controles...). El tipo de estudio condiciona el **diseño de la investigación**.

En estadística, fundamentalmente tratamos de valorar si el efecto observado en un estudio se debe al azar o si existe un patrón determinado por ciertos factores asociados al parámetro o a la población que también observamos y tratamos de relacionar con el efecto.

Tanto los efectos que queremos estudiar como los factores que pretendemos relacionar con estos se miden a través de lo que llamamos **variables estadísticas**.

Por tanto, dependiendo de qué queremos investigar, con qué objetivos y con qué hipótesis, deberemos plantear correctamente la investigación. Ello resulta crucial para observar correctamente tanto el efecto como estos factores. ¿Cuál es la forma correcta de plantearlo? Contestando a cinco preguntas básicas:

1.1. ¿Qué?

Para empezar, debemos tener claro qué deseamos investigar: identificar las **variables respuesta o dependientes** (el efecto estudiado) y las **variables predictivas o independientes** (posibles factores asociados). En este caso, la variable respuesta sería el nivel de triglicéridos medido antes y después del cambio dietético, y el cambio o no en la dieta sería el factor predictivo.

Podríamos valorar, por ejemplo, otros factores que pensemos que puedan influir en la variable respuesta, como el sexo, la edad, el índice de masa corporal, la dieta o la medicación.

1.2. ¿Cómo?

Dependiendo del efecto y los factores que hayamos de medir, debemos elegir las mejores herramientas para hacerlo. En este caso podríamos optar por entrevistas con cuestionarios de frecuencia de consumo alimentario y la bioquímica de los individuos. El cómo es muy importante, pues de él depende el correcto diseño *a priori* de la investigación.

Por ejemplo, pensemos en otro estudio que pretende valorar la eficacia de una prueba diagnóstica respecto a otra para determinada patología o parámetro clínico. En ese caso debemos prever la inclusión en la muestra de individuos cuyo resultado con la nueva prueba diagnóstica resultó tanto positivo como negativo, pues hemos de analizar los falsos positivos y los falsos negativos.

1.3. ¿Cuándo?

Nuestro estudio puede ser **retrospectivo** (por ejemplo, estudiando el comportamiento dietético en el pasado de esta población y contrastarlo con su nivel de triglicéridos) o **prospectivo** (por ejemplo, estudiando la evolución del nivel de triglicéridos durante un tiempo y contrastarla con la dieta seguida durante ese periodo, con intervención o sin ella).

Pero, además, el momento de recogida en sí es relevante. Así, por ejemplo, el registro alimentario en verano variará respecto al registro en invierno.

1.4. ¿Dónde?

El lugar de realización de la investigación define ciertas características de la población que pretendemos estudiar.

No es lo mismo estudiar el nivel de triglicéridos en España que en Senegal, o en personas mayores que viven en su domicilio particular que en centros residenciales donde se dispensan dietas estandarizadas.

1.5. ¿Por qué?

En el fondo, nos estamos preguntando cuál es nuestra **hipótesis**. Si vamos a estudiar la relación entre el nivel de triglicéridos (variable respuesta) y el cambio dietético (variable predictiva), estamos hipotetizando sobre una posible relación entre ambos factores en un sentido u otro. A este respecto, es un error común buscar siempre una «asociación significativa», puesto que la falta de asociación también es un resultado válido, incluso satisfactorio.

Pensad en un estudio que trate de relacionar el consumo de agua potable y el cáncer de colon.

A continuación se describirán y abordarán los procesos que permiten describir nuestros datos, así como explorarlos y analizarlos de una manera muy básica.

Es lo que se conoce como **análisis exploratorio de datos**, y permite enfocar un análisis estadístico completo posterior, acorde con los objetivos de nuestro estudio.

Técnicas de análisis estadístico avanzadas

Quedan fuera de este breve texto las técnicas de análisis estadístico más avanzadas, debido a la gran cantidad de posibilidades y a su complejidad. Al final del documento se encuentran varios recursos que pueden ser útiles en el desarrollo de las más comunes.

2. Estadística básica Conceptos generales

Los datos recopilados en un estudio epidemiológico se organizan en una **base de datos** (casos en filas y variables en columnas), pero ¿cómo extraemos la información que buscamos a partir de ellas?

2.1. Organización y tipos de datos

La figura 1 representa un extracto de la base de datos. En ella se observan los datos obtenidos para seis de las variables de los nueve primeros participantes (de un total de cincuenta). El objetivo del estudio podría ser el análisis de los cambios en el nivel de triglicéridos entre la semana 0 y la 24 habiendo cambiado o no la dieta, considerando la edad y el género.

Figura 1. Extracto de un conjunto de datos de un estudio epidemiológico

	A	B	C	D	E	F
1	Age	Gender	Triglycerides_w0	Triglycerides_w24	Change_med	Adverse_effect
caso → 2	53	Mujer	206,2	133	si	si
3	53	Mujer	218,2	164,6	si	no
4	55	Mujer	205	145,8	no	no
5	52	Mujer	200,2	138	si	si
6	50	Mujer	212,3	171,8	si	si
7	51	Mujer	211,1	148,3	si	no
8	55	Hombre	207,9	134,4	si	si
9	50	Mujer	210,6	123,8	si	no
10	46	Mujer	217,2	188,4	si	no

variable cuantitativa continua — variable cualitativa nominal

Fuente: INDESTAP. Departamento de Estadística e Investigación Operativa de la Universitat de València.

Las variables se llaman así porque cambian entre caso y caso (de lo contrario estaríamos ante constantes), y pueden ser:

1) Cualitativas: presentan sus resultados en forma de categorías (como en el caso del sexo, hombres-mujeres) y pueden ser nominales (clasificaciones independientes, como tipos específicos de dieta) u ordinales (con un orden, como bajo, medio, alto). Suelen llamarse variables categóricas o factores.

2) Cuantitativas: presentan sus resultados en forma de valores numéricos (como en el caso de la edad, la ingesta de lípidos o el nivel de triglicéridos). Las variables cuantitativas pueden ser continuas, cuando pueden tomar cualquier valor numérico entre dos valores establecidos, y pueden alcanzar un número infinito de valores distintos, o discretas, cuando sus resultados proceden por ejemplo de un recuento, en cuyo caso pueden tomar un número finito o infinito numerable de valores.

Es frecuente convertir variables cuantitativas en cualitativas agrupando los valores numéricos en diferentes categorías. Esto suele ser útil para diferenciar el efecto medido respecto a estos grupos, como, por ejemplo, grupos de edad o rangos de índice de masa corporal.

Hojas de cálculo

Las hojas de cálculo (Excel, Open Office) son ideales para recoger nuestra información y se pueden cargar fácilmente en una aplicación estadística como R-Commander, SPSS, PSPP, Stata, etc.

2.2. Población y muestreo

Debemos prestar especial atención a los conceptos de **población (N)** y **muestra (n)**.

El conjunto de los individuos objeto de nuestro estudio es lo que se denomina **población**. Si pretendemos estudiar el nivel de triglicéridos en una ciudad, la población serían todos los habitantes de la ciudad.

Así, si el objetivo fuese estudiar el nivel de triglicéridos solo en mayores de 65 años de esa ciudad, la población estaría formada por todos los mayores de 65 años que residen en ella. En la mayoría de las ocasiones esto es inabarcable, pues deberíamos recoger los datos de todas y cada una de estas personas. Por ello suele estudiarse solo una parte representativa: una muestra. Para ello, se selecciona una muestra de personas de entre el conjunto de la población objeto del estudio. Este subconjunto, mucho menor y manejable, es sobre el que estudiaremos la característica que nos interesa (en este caso, el nivel de triglicéridos y los factores asociados).

Así, una muestra será cualquier subconjunto de individuos de una población; a la selección de individuos de entre la población se le llama **muestreo**.

El muestreo es correcto cuando garantiza que las características de la población quedan reflejadas apropiadamente en la muestra. Preservar la **representatividad** es el atributo más importante que debe reunir el muestreo, lo que nos permitirá extrapolar a la población los resultados obtenidos en nuestro estudio.

En caso de que la muestra sea poco representativa de la población, diremos que está sesgada (o que tiene **sesgo**). Este sesgo suele darse cuando algunos sectores de la población están más representados dentro de la muestra que otros.

Pero ¿cómo se selecciona la muestra y qué tamaño es el adecuado? Para que sea representativa de la población, el proceso de selección de la muestra se debe basar en el principio de **aleatorización**, es decir, eligiendo al azar, de entre la población, a los individuos que formarán parte de la muestra. Hay diferentes técnicas de muestreo. Veamos algunas de las más relevantes utilizadas en salud pública:

1) Muestreo aleatorio simple. Esta técnica asegura que todos los individuos de la población tienen la misma probabilidad de ser incluidos y que los individuos se seleccionen de manera independiente unos de otros.

2) Muestreo aleatorio estratificado. Si nuestra población objetivo es especialmente heterogénea, pudiendo diferenciarse en varios grupos (lo que se conoce como estratos) que constituyen categorías importantes para la investigación (por ejemplo, por sexos), la elección de la muestra no debe hacerse de la misma manera en todos los estratos, ya que unos podrían acabar más representados que otros. Para determinar los estratos se suele recurrir a variables espaciales (comunidades, municipios, entre otros) u otras que nos interesen en función del objetivo del estudio (la capacidad económica, la etnia, etc.). Este muestreo pretende asegurar la representación de cada grupo en la muestra. Una vez definidos los estratos, se realiza un muestreo aleatorio simple dentro de cada uno de ellos.

3) Muestreo por conveniencia. Este tipo de muestreo no es aleatorio y se basa en elegir la muestra según la facilidad de acceso a ella, la disponibilidad de los participantes para formar parte del estudio, o cualquier requisito estricto de la población de estudio que restrinja el tamaño de la población. Cuando se utiliza esta técnica, se pueden observar hábitos, opiniones, y puntos de vista de manera más fácil, por lo que es muy utilizada en salud pública, aunque debe justificarse suficientemente su uso.

Todo estudio lleva implícito en la fase de diseño la determinación del tamaño de muestra necesario. Es muy importante que la muestra sea representativa de la población, y el tamaño necesario depende del error que estemos dispuestos a asumir en nuestros resultados. Veamos a continuación a qué nos referimos.

2.3. Contrastes de hipótesis

Existe una relación entre el tamaño de la muestra y el **error estadístico**. De nuestra población de estudio nos planteamos una serie de preguntas. Estas preguntas se traducen en **hipótesis** estadísticas. Se plantean dos hipótesis y se debe elegir entre la llamada **hipótesis nula (H_0)** y la **hipótesis alternativa (H_a)**. Esto es lo que se conoce como contraste de hipótesis.

Se suele considerar la hipótesis nula como aquella que determinará que no hay relación entre el efecto medido y la variable estudiada, y la alternativa estipula que sí lo hay, de manera «estadísticamente significativa». Veremos qué quiere decir esto último.

En este proceso se pueden cometer dos errores posibles:

1) Error de tipo I: rechazar H_0 cuando es cierta.

2) Error de tipo II: aceptar H_0 cuando es falsa.

Sus probabilidades, frecuentemente denominadas riesgos, vienen dadas por:

1) α = Probabilidad de rechazar H_0 cuando es cierta (error de tipo I). El investigador llega a la concusión que hay diferencias entre las hipótesis (o variables de estudio) cuando en realidad no las hay. Por lo tanto el error tipo I da falsos positivos.

2) β = Probabilidad de aceptar H_0 cuando no es cierta (error de tipo II). El investigador es incapaz de encontrar diferencias cuando en realidad existen. Por lo tanto equivale a un falso negativo.

Normalmente se establece por consenso el valor del riesgo α en 0,05 (5 %) y se denomina **nivel de significación**, mientras que 1 – β se conoce como **potencia** de la prueba, que debe ser al menos de 0,8 (80 %). Cuanto más incrementemos *n* (tamaño de la muestra), más decrecerán α y β y aumentará la potencia de la prueba, con lo que disminuirán los errores.

El grado de riesgo de cometer estos errores lo podemos cuantificar con el nivel crítico *p* (o **p-valor**).

El p-valor es la probabilidad de obtener una discrepancia mayor o igual a la observada en la muestra cuando *H*0 es cierta, y puede considerarse como el error de tipo I que cometemos con los datos observados en la muestra.

Dicho de otra manera, el **p-valor** es la probabilidad (medida en escala de 0 a 1) de que, al repetir el experimento en idénticas condiciones en otra muestra aleatoria de la misma población, encontremos resultados discrepantes.

Por tanto, un p-valor = 0,05 se entiende como un 5 % de probabilidad de que al repetir el experimento obtengamos resultados diferentes. α (nivel de significación) será el límite aceptable que imponemos a esa probabilidad, que suele ser 0,05 (5 %) por consenso entre la comunidad científica. Así:

si p-valor < α → Rechazamos H_0

si p-valor ≥ α → No rechazamos H_0

2.4. Estadística descriptiva e inferencial

La estadística tiene como objetivo último extraer información de los datos recogidos. Y lo aborda de dos maneras:

1) Estadística descriptiva: se encarga de describir, analizar y representar un grupo de datos utilizando métodos numéricos y gráficos que resumen y presentan información contenida en ellos. Solo describe la muestra.

2) Estadística inferencial: a partir del cálculo de probabilidades y de datos muestrales, se encarga del cálculo de estimaciones y predicciones de la relación entre variables. Saca conclusiones sobre el posible patrón de comportamiento en una muestra.

3. Estadística descriptiva

Nuestras variables se pueden describir de manera tanto numérica como gráfica. Cuando describimos una sola variable hablamos de **análisis univariante**, cuando describimos una variable en función de otra nos referimos a **análisis bivariante**, y cuando en el análisis hay implicadas más variables (por ejemplo, una regresión lineal con una variable respuesta y varias predictivas), hablamos de **análisis multivariante o múltiple**. Tanto la descripción de variables como los análisis estadísticos se realizan fácilmente con la ayuda de diferentes programas estadísticos.

3.1. Descripción numérica de los datos

La descripción numérica de los datos se realiza por medio de lo que llamamos **estadísticos descriptivos** Cualquier software estadístico los proporciona fácilmente.

¿Cuáles son? Depende de si la variable es categórica o cuantitativa. Cuando trabajamos con variables categóricas suelen utilizarse frecuencias para su descripción:

1) Frecuencia absoluta: número de veces que aparece en el estudio un determinado valor o categoría.

2) Frecuencia relativa: cociente entre la frecuencia absoluta y el tamaño de la muestra, n.

3) Porcentaje: frecuencia relativa multiplicada por 100.

En el caso de las variables cuantitativas nos referimos a las medidas de tendencia central, dispersión y posición, que comparan cualquier individuo de la muestra en relación con el valor central y permiten comparar resultados medios obtenidos por dos o más grupos de individuos:

1) Media: es la suma de todos los valores de la muestra de la variable dividida entre n.

2) Mediana: es el valor de la variable que deja por debajo a la mitad de los datos, una vez que estos están ordenados de menor a mayor. Equivale al percentil 50 de la distribución (p50).

3) Moda: es el dato más repetido, el valor de la variable con mayor frecuencia absoluta.

4) Rango: es la diferencia entre el valor mínimo y el valor máximo de la variable.

5) Máximo: es el valor máximo de la variable.

6) Mínimo: es el valor mínimo de la variable.

7) Varianza: es una medida estadística que mide la dispersión de los valores respecto a su media.

8) Desviación típica o estándar: es la raíz cuadrada de la varianza y tiene la misma función.

9) Percentil p: es el valor que deja por debajo el p% de observaciones de la muestra. Se corresponde con los cuartiles: percentil 25 (cuartil 1), percentil 50 (cuartil 2), percentil 75 (cuartil 3). No obstante, se suelen utilizar muchos otros percentiles para comparar mediciones individuales respecto al resto de la población.

10) Intervalo de confianza (IC): es un intervalo que nos indica entre qué valores se encuentra el parámetro que estamos contrastando a nuestro nivel de confianza. Para una confianza habitual del 95 % (= 0,05), de cada 100 experimentos, en el 95 % el IC comprenderá el verdadero valor. El IC es equivalente al p-valor para hacer contrastes de hipótesis. Rechazamos H_0 cuando el IC no contenga el valor del parámetro.

Tanto la varianza como la desviación típica son sensibles a la existencia de valores extremos u ***outliers*** (valores atípicos muy distantes del resto de los datos. No hay que confundirlos con el **valor máximo** y el **valor mínimo**). Estos valores pueden parecer prescindibles porque distorsionan los estadísticos descriptivos, pero a menudo son los datos más interesantes, especialmente en investigación sanitaria, por lo que debemos valorar cómo abordarlos de la manera que mejor convenga a nuestro estudio.

3.2. Descripción gráfica de los datos

Una vez descritas las distintas variables de forma numérica mediante los estadísticos descriptivos, podemos inspeccionar su comportamiento, especialmente respecto al resto de las variables, por medio de **tabulaciones** y de la **representación gráfica**.

Este proceso será también distinto en el caso de variables categóricas o cuantitativas.

1) Variables categóricas. Suelen tabularse mediante **tablas de frecuencias** o de contingencia, en las que se distribuyen las frecuencias de la variable respuesta, categórica en este caso, en relación con otro factor, también categóri-

co, respecto al cual queremos valorar cómo se comporta. Por ejemplo, a partir de los datos de la tabla 1, podemos describir el cambio de medicación en los cincuenta participantes en el estudio en función del género:

Tabla 1. Cambio de medicación en función del género

	Gender		
Change_med	**Hombre**	**Mujer**	
no	2	12	14
sí	11	25	36
	13	37	50

Pearson's Chi-squared test

X-squared 13.868,00, df=1, p-value = 0.2389

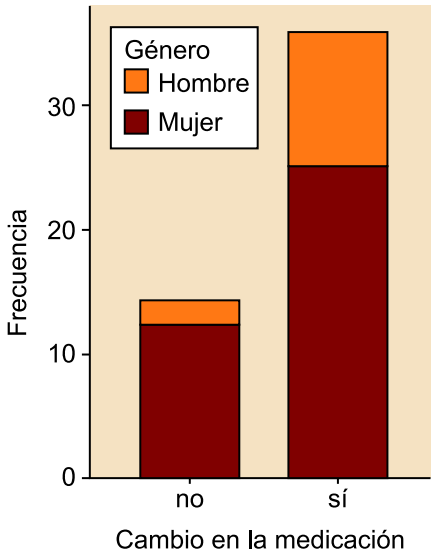
Como puede apreciarse, a menudo se asocia una prueba estadística que contrasta la hipótesis nula (no hay diferencias entre las categorías), frente a la hipótesis alternativa (sí las hay). Explicaremos este concepto más adelante.

Para representar gráficamente estas diferencias entre categorías, disponemos de varias opciones, las más comunes de las cuales son:

a) Diagramas de barras (*barplot*). Las gráficas de barras son una manera de representar frecuencias; las frecuencias están asociadas con categorías. Una gráfica de barras se presenta de dos maneras: horizontal o vertical.

La **gráfica de barras** sirve para comparar y tener una representación gráfica de la diferencia de frecuencias o de intensidad de la característica numérica de interés.

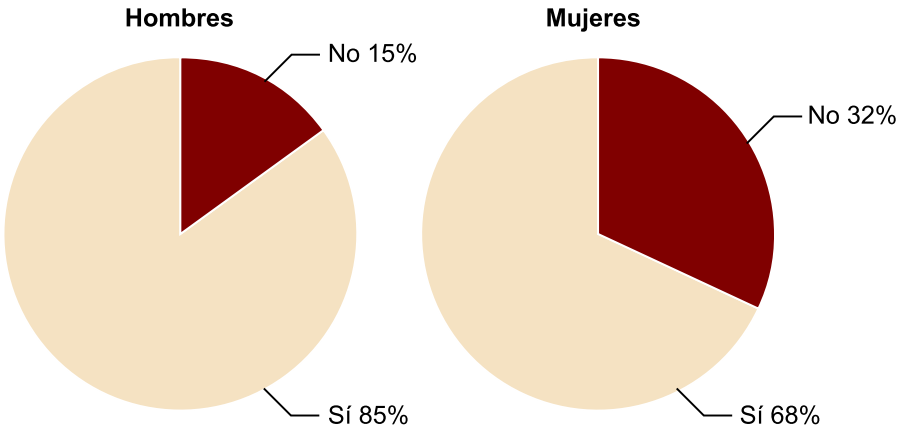
Figura 2. Diagrama de barras categórico



Participantes con cambio en la dieta en función del sexo.
Cambio en la medicación, diferenciado por género.

b) Diagramas de sectores (*pie*). Se trata de un gráfico que consiste en un círculo dividido en sectores de amplitud proporcional a la frecuencia de cada valor. Pueden añadirse etiquetas con las frecuencias de cada categoría o los porcentajes de cada una de ellas.

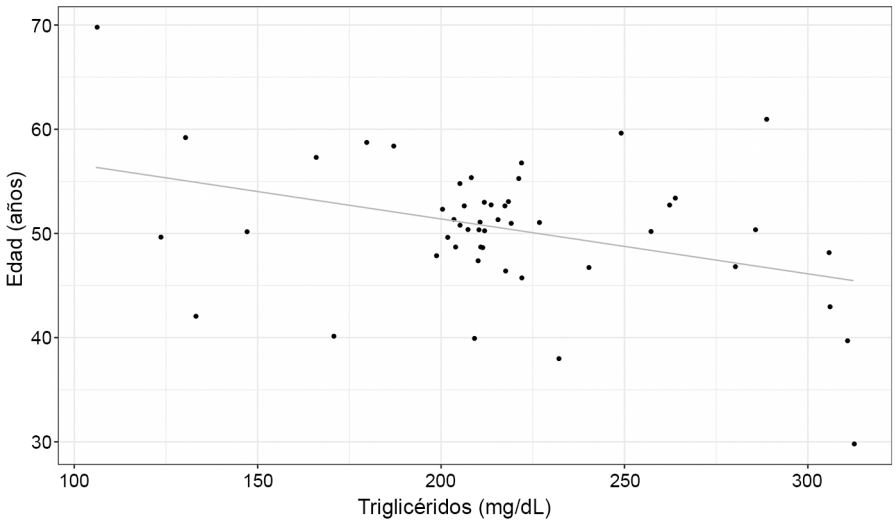
Figura 3. Diagrama de sectores. Cambio en la medicación en función del sexo



2) Variables cuantitativas. La información contenida en este tipo de variables es mayor que en el caso cualitativo y tiene más posibilidades de representación. Se pueden representar los estadísticos descriptivos y observar los valores más habituales, cómo varían, cómo de dispersos son, si se observa algún patrón de distribución, etc. Y todo ello se puede hacer, además, en función de otras variables, tanto cuantitativas como categóricas (agrupando los resultados por grupos). Así, se pueden analizar de un solo vistazo las relaciones entre variables. Veamos algunos de los casos más sencillos:

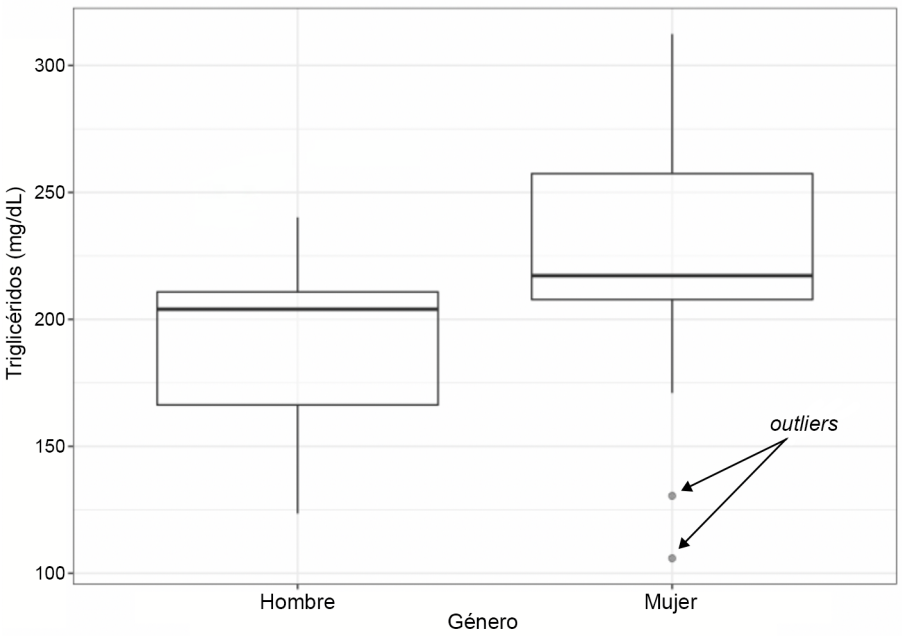
a) Diagramas de puntos (*scatter plot*). Puede representarse una variable cuantitativa en función de otra cuantitativa. En este caso, usando el conjunto de datos de muestra, representamos en la figura 4 el nivel de triglicéridos al inicio del estudio (eje x) en función de la edad (eje y). Añadimos una recta de regresión que marca la relación entre ambas variables. Se aprecia una relación inversa entre ambas (a mayor edad, menor nivel de triglicéridos).

Figura 4. Diagrama de puntos. Triglicéridos en función de la edad, ambas variables cuantitativas



b) Diagramas de caja (*boxplot*). Permite apreciar fácilmente la distribución de los datos, representando en una caja los datos contenidos entre el cuartil superior y el inferior (la línea horizontal dentro de la caja es la mediana). Las líneas verticales, o bigotes, marcan el extremo superior e inferior. Los puntos aislados más alejados son los *outliers*. En la figura 5 se observa que existe una clara diferencia de distribución en el nivel de triglicéridos entre géneros.

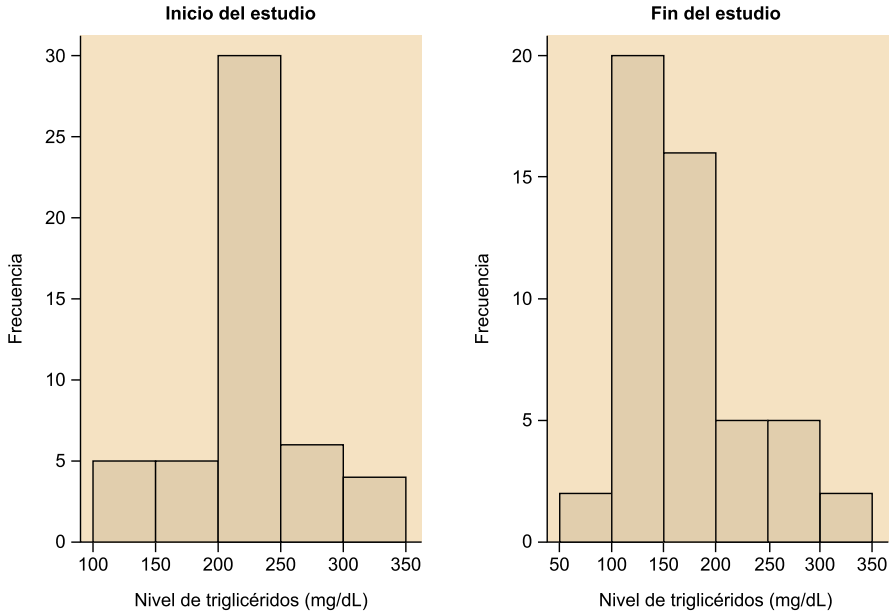
Figura 5. Diagrama de cajas (*boxplots*). Triglicéridos en función del género, cuantitativa frente a categórica



c) Histograma. Es una representación gráfica de una variable en forma de barras, donde la altura de cada barra es proporcional a la frecuencia de los valores representados. Sirve para estimar la distribución de la muestra, respecto a una característica, cuantitativa y continua (como el IMC, la edad, el peso o la con-

centración de triglicéridos...). De esta manera, ofrece una visión sobre posibles tendencias. En la figura 6 se observa que existen tendencias muy distintas en los niveles de triglicéridos antes y después del estudio.

Figura 6. Histograma. Frecuencias del nivel de triglicéridos antes y después del estudio



4. Estadística inferencial

Hasta este punto se han descrito técnicas de organización, presentación y resumen de datos que proporcionan información sobre las variables estudiadas. Este tipo de exploración nos da una idea del comportamiento de las variables y las posibles relaciones entre ellas, generalmente de dos en dos. Sin embargo, uno de los objetivos básicos en la investigación es extraer conclusiones basadas en la evidencia científica.

Para afirmar que un efecto responde a cierto patrón de comportamiento no basta con la estadística descriptiva, sino que debemos emplear test estadísticos basados en el contraste de hipótesis, que ya hemos introducido y que evalúan matemáticamente la probabilidad de que el efecto observado se deba simplemente al azar.

Para medir el grado de asociación entre variables, los test estadísticos realizan el contraste de hipótesis. Estos test tratan de contrastar así la variable de interés en nuestro estudio (aquella cuyo comportamiento pretendemos explicar respecto a las demás y que se conoce como **variable respuesta**, o dependiente) respecto al resto de las variables (aquellas que también medimos en el estudio y que pretendemos asociar al comportamiento de la primera, y que se conocen como **variables predictivas**, o independientes).

Para realizar este contraste, un test estadístico calcula:

- Un **coeficiente estadístico** (específico de cada prueba), asociado a la variable predictiva, que mide el grado de asociación entre ambas variables. También indicará el sentido de esta asociación: negativo o positivo, dependiendo de si el coeficiente es negativo o positivo, respectivamente. A título orientativo, cuanto más se aleje del 0, mayor será la asociación. Se entiende como el número de unidades que cambia la variable respuesta cuando aumenta la variable predictiva en una unidad.
- Un **p-valor**, que indicará si la asociación expresada por el coeficiente es significativa o no, bajo cierta probabilidad. Es decir, la probabilidad de que el cambio en la variable respuesta no se deba al azar sino al efecto de la variable predictiva. Esta probabilidad se considera que debe ser al menos del 95 % (o 0,95, expresada sobre 1), por lo que el valor de referencia, o nivel de significatividad, es del 5 % (o 0,05). Por ello, se considera que el test indica «significatividad estadística» cuando p-valor es 0,05.

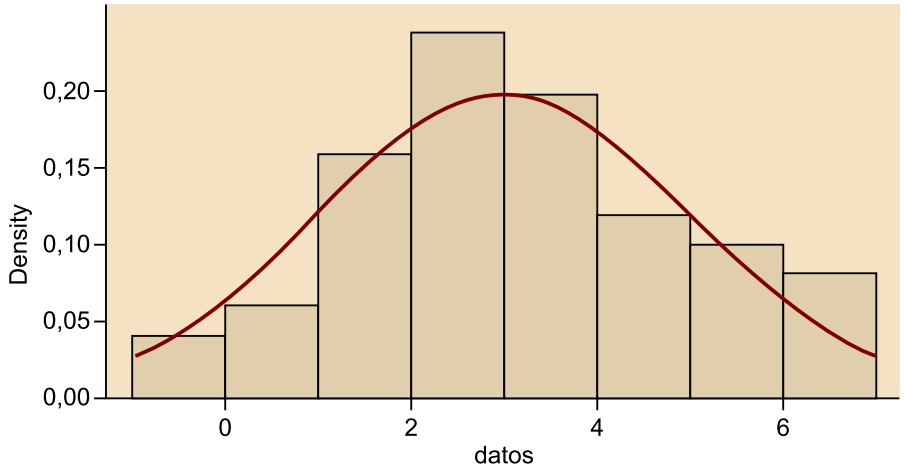
4.1. Tipos de test estadísticos y condiciones de aplicabilidad

Hay dos grandes tipos de test estadísticos: los paramétricos y los no paramétricos, y no pueden utilizarse unos u otros indistintamente.

a) Las pruebas paramétricas se basan en la media, la desviación estándar y demás medidas de dispersión (parámetros) y requieren el cumplimiento de unas condiciones de aplicación más estrictas. Principalmente que los datos sigan una distribución normal y que las varianzas de los grupos o variables contrastadas sean homogéneas (que su dispersión sea similar, lo que se denomina homocedasticidad).

En estadística se denomina **distribución normal**, o distribución de Gauss, a una de las distribuciones de los valores de una variable continua que con más frecuencia aparece en estadística. Al representar gráficamente los valores obtenidos de nuestra variable para cada individuo, estos valores se distribuirán en torno al estadístico descriptivo (en general, la media). Si estos datos siguen una distribución normal, la mayoría de los valores de nuestra variable se concentrarán alrededor del valor de la media, y disminuirán progresivamente los valores menores y mayores de la media a medida que se alejan de esta, con lo que se obtiene una forma acampanada y simétrica. Habrá más o menos valores que se alejen de la media en función de si la desviación estándar es mayor o menor, respectivamente. A esta curva se la conoce como **campana de Gauss**. En la figura 7, se observa un ejemplo basado en unos datos generados para que se distribuyan según una distribución normal. La campana será más o menos picuda o plana (lo que llamamos curtosis) en función de su menor o mayor desviación estándar, respectivamente; se acepta como normal cuando tiene el 95 % de sus observaciones dentro del intervalo media 1,96 veces la desviación estándar.

Figura 7. Histograma + curva de distribución normal de unos datos aleatorios (n = 50, media = 3, desviación = 2)



Por tanto, estas condiciones solo pueden cumplirlas las variables cuantitativas. Cuando las variables cuantitativas no cumplen estos requisitos o se trata de variables categóricas, deben usarse las pruebas no paramétricas.

En caso de que los test indiquen que no existe normalidad, podremos igualmente utilizar test paramétricos si el tamaño de la muestra es grande.

Esto se debe al **teorema central del límite**, que indica que, si tenemos un grupo grande de observaciones, su distribución «se aproxima» al comportamiento de una distribución normal.

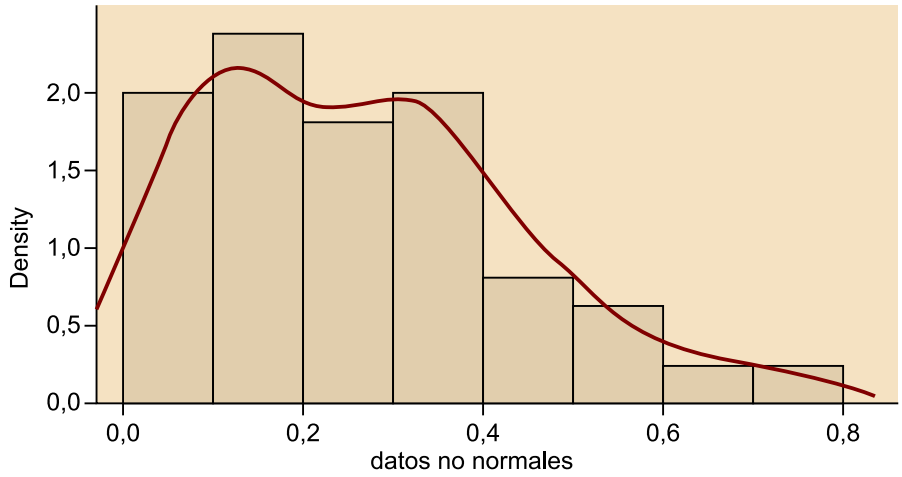
Otra alternativa es probar a transformar la variable respuesta en forma logarítmica, o como raíz cuadrada, lo cual a veces aproxima su distribución a la normal y supera las pruebas de normalidad. Cuando aun así no exista normalidad, la muestra sea pequeña y/o sean datos categóricos, deberemos usar pruebas no paramétricas.

Existen test estadísticos que contrastan las hipótesis de normalidad (H_0) frente a la no normalidad (H_a) a disposición en cualquier software estadístico. Algunos de los más conocidos son el **test de Kolmogorov-Smirnov** y el **test de Shapiro-Wilk**. Para evaluar la homogeneidad de varianzas suele usarse el **test de Levene**. Valores de $p < 0.05$ en estos tests rechazan la hipótesis de normalidad o de homogeneidad de varianzas.

b) Las pruebas no paramétricas solo tienen en cuenta el orden que ocupa cada observación en el conjunto ordenado de observaciones. Esto permite realizar pruebas de contraste de hipótesis que no asumen ninguna distribución específica (no es necesario que los datos sigan una distribución normal). Por lo tanto, se emplean cuando las variables cuantitativas no cumplen los principios de normalidad y homogeneidad de varianzas, o bien cuando se trata de variables categóricas.

Por ejemplo, el mismo conjunto de datos de la figura 7 podría seguir una distribución no normal, tal como se ve en la figura 8, donde se aprecia una forma irregular y una cola hacia la derecha.

Figura 8. Histograma + curva de distribución de unos datos ideales aleatorios



Como ya hemos comentado, nuestro estudio pretende observar y explicar el comportamiento de un parámetro de salud (o varios), medido a través de una variable. Por ejemplo, y siguiendo con los datos presentados al inicio del módulo, el nivel de triglicéridos a las 24 semanas de un tratamiento.

Por otra parte, el objetivo de nuestro estudio se basará en observar otro/s factor/es en los participantes que pensamos que puede/n afectar al cambio en el nivel de triglicéridos, o no, concepto en el que se basará nuestro contraste de hipótesis. Estos factores son los que llamamos **variables predictivas**. En nuestro caso puede ser el cambio de dieta o el cambio de medicación, o la edad, o todos ellos.

Tanto la variable respuesta como las predictivas pueden ser categóricas o cuantitativas. Y el tipo de test estadístico que utilicemos debe elegirse en función de este aspecto. Veamos a continuación algunas de las pruebas estadísticas más usuales que se utilizan de forma exploratoria en salud pública.

4.2. Estadística inferencial bivariante

Retomemos los datos presentados al inicio del módulo (figura 1). De ahora en adelante supongamos que nuestra variable respuesta es el nivel de triglicéridos de la muestra (medido al inicio del estudio y en la semana 24), y el resto de las variables se considerarán predictivas (edad, género, cambio en la dieta, cambio en la medicación, etc.). Nuestro objetivo es determinar si estas variables predictivas afectan a la variable respuesta y cómo lo hacen. El contraste de hipótesis será por tanto evaluar la H_0, definida como un cambio observado sobre la variable respuesta debido al azar, frente a la H_a, que determinará un cambio observado sobre la variable respuesta debido a la variable predictiva, con cierta probabilidad (recordemos que se suele considerar un mínimo del 95 % para afirmar que la relación es significativa).

En primer lugar, suelen realizarse pruebas seleccionando las variables de dos en dos (pruebas por pares), las cuales contrastan el parámetro de la variable respuesta (concentración de triglicéridos en este caso) frente a cada una de las variables predictivas por separado (concentración frente a género, concentración frente a edad, concentración frente a cambio en la dieta, etc.). Estos test tratan de evaluar el efecto que cada variable predictiva ejerce o no por sí misma, aislada, como si no hubiese presente ningún otro parámetro que pudiese afectar a la variable respuesta.

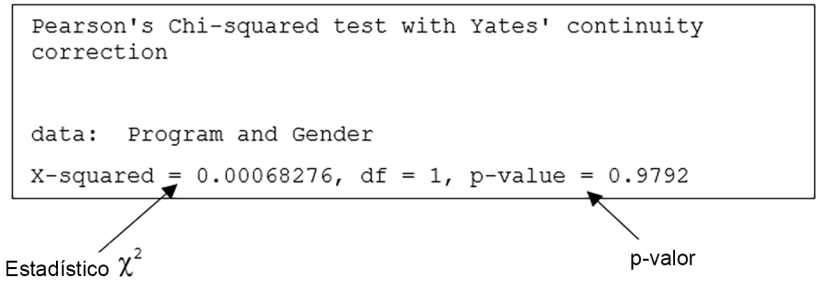
En las figuras 4 y 5, hemos podido observar ciertas diferencias aparentes en los valores del nivel de triglicéridos (variable cuantitativa) en función de la edad (variable cuantitativa), o el género (variable categórica). Sin embargo, la observación muestral es algo subjetivo. Debemos aplicar los test estadísticos apropiados que validen estos resultados poblacionalmente, es decir, que podríamos extrapolar los resultados de nuestra muestra al resto de la población que representa. Para poder aplicar test estadísticos, necesitamos un software específico. A este software se exportan de manera sencilla los datos que han sido recopilados y organizados en un formato tipo Excel o similar. Hay disponibles programas gratuitos, como R-Commander o PSPP, y con licencia, como SPSS o STATA.

4.2.1. Inferencia entre variables categóricas

Al ser categóricas, solo podrán ser aplicadas pruebas no paramétricas. Cada variable tendrá dos o más categorías, y el contraste de hipótesis se basa en comparar las frecuencias en las distintas combinaciones de una categoría de una variable con las categorías de la otra. El test utilizado con más frecuencia es 1) Test de *Pearson o* $\chi 2$(Chi cuadrado) de Pearson. Se trata de un test no paramétrico que se utiliza para determinar si existe una asociación significativa entre dos variables categóricas. La H_0 es la independencia de las dos variables, es decir, la igualdad de proporciones entre los grupos. El estadístico de contraste $\chi 2$ se calcula a partir de las frecuencias observadas y las frecuencias esperadas. Se le asocia un p-valor en función del cual tomaremos nuestra decisión. Es imprescindible que cada grupo de la muestra tenga al menos cinco individuos. Si alguno de los grupos de la muestra tiene menos de 5 individuos, entonces el test más aconsejado es el test Exacto o de Fisher

> Por ejemplo, supongamos que tenemos una muestra de hombres y mujeres a los que se aconseja un cambio de dieta. Pasados unos meses, se recoge para cada individuo si se ha producido un cambio de dieta (sí o no). Tendremos cuatro grupos de individuos: **1)** mujeres con cambio de dieta; **2)** mujeres sin cambio de dieta; **3)** hombres con cambio de dieta; **4)** hombres sin cambio de dieta. Si queremos determinar si la dieta es seguida de forma diferente por hombres o mujeres, el resultado del test sería el siguiente:

Figura 9



En este caso vemos que no habría diferencias entre sexos, pues p-valor ≥ 0,05.

4.2.2. Inferencia entre una variable cuantitativa y una variable categórica

En este caso se trata de describir la distribución de la variable cuantitativa en cada categoría de la variable categórica. Hemos de comparar los valores de la variable respuesta en los grupos o categorías de la variable predictiva, o lo que es lo mismo, comparar las respuestas medias de cada grupo.

Las categorías de la variable categórica forman grupos que hay que comparar. Hay que considerar si los grupos (o subcategorías) son independientes entre sí o no. Por ejemplo, la muestra puede analizarse en función de la variable sexo, donde encontramos las categorías hombre-mujer, que son independientes porque dividen la muestra en casos distintos. Lo mismo ocurre con otros factores con más de dos categorías, como el IMC, que puede dividir la muestra en casos con <18,5 kg/m^2, 18,5-25 kg/m^2, 25-30 kg/m^2, >30 kg/m^2. En estos casos hablamos de **muestras independientes**. Sin embargo, la variable categórica puede hacer referencia a un antes y un después en la medida de un mismo parámetro bioquímico (la variable respuesta cuantitativa) medido en cada paciente antes y después de una intervención dietética. Se trata en este caso de observaciones relacionadas, dependientes o emparejadas (son términos habituales para el mismo concepto). Suele referirse en este caso a **muestras dependientes o emparejadas**.

Tanto en un caso como en el otro debemos determinar si la diferencia entre las medias observadas en las distintas categorías es estadísticamente significativa, es decir, si podemos determinar que el resultado de *Y* (la variable cuantitativa) cambia en función de *X* (el factor categórico). Para ello, debemos realizar un **test de contraste de medias**.

En el caso de muestras independientes, la H_0 del test indicará que no existen diferencias entre los grupos, que hay ausencia de efecto (por ejemplo, que no hay diferencias entre sexos), mientras que la H_a indicará que sí las hay, bajo cierta probabilidad de equivocarnos. En el caso de muestras dependientes la H_0 indicará una diferencia nula entre las distintas observaciones del mismo caso (por ejemplo, no hay diferencia en el nivel de triglicéridos antes y después del

tratamiento nutricional), mientras que la H_a determinará que el tratamiento sí produjo un efecto sobre el nivel de triglicéridos, también bajo cierta probabilidad de equivocarnos.

Debemos destacar que las condiciones de aplicabilidad de las pruebas paramétricas deben satisfacerse en cada uno de los grupos en los que se subdivide la muestra por categorías. Por ello, si necesitamos analizar una variable cuantitativa de una muestra no muy grande en función de una variable cualitativa con varias subcategorías, la muestra se subdividirá consecuentemente. De esta manera, las condiciones de normalidad y homocedasticidad en el total de la muestra pueden perderse cuando subdividimos los datos en categorías, y en ese caso deberemos recurrir a test no paramétricos si no queda en cada categoría un número lo suficientemente grande de casos (al menos treinta).

Los test más habituales en ciencias de la salud son:

1) Cuando la variable cualitativa tiene dos categorías

a) Paramétricos: prueba de la **t-Student para muestras independientes** o prueba de la **t-Student para muestras emparejadas**.

> La **distribución t de Student** es parecida a la distribución normal y la sustituye cuando no se conoce la desviación estándar poblacional (que suele indicarse como σ, sigma) y hay que recurrir a calcular la desviación estándar en la propia muestra (*s*).

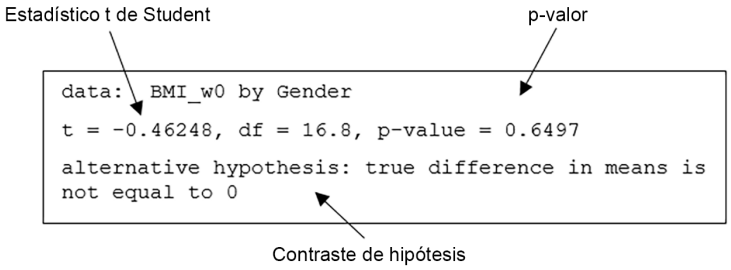
Como casi nunca se dispone de σ, el uso de la t de Student es muy usual. Hay una *t* distinta para cada tamaño de muestra (a esto se refieren los llamados «grados de libertad»). Recordemos que *n* es el tamaño de nuestra muestra.

El valor de *t* se obtiene al dividir la diferencia entre las dos medias (el «efecto») por el error estándar de la diferencia de medias (o «error»). Así obtenemos el estadístico de contraste *t*, que se asocia a un p-valor. Como en la mayoría de los test estadísticos, lo que se evalúa es la diferencia observada entre medias dividida por un término de error que representa la variabilidad biológica aleatoria. Si la diferencia entre las medias es mucho mayor que la variabilidad biológica, *t* será grande y el p-valor, pequeño. Si la diferencia observada es pequeña en relación con el error estándar, entonces *t* será pequeña y el p-valor, grande, lo que indica que no hay diferencias significativas. Así, la H_0 será la igualdad de medias entre categorías y la H_a será la diferencia bajo cierta probabilidad de equivocarnos.

Para aplicar la prueba de la t-Student, como prueba paramétrica que es, la muestra debe cumplir (y cada categoría en la que se subdivida la muestra) las condiciones de aplicabilidad: normalidad y homocedasticidad.

En el caso de existir normalidad, pero no homocedasticidad, puede recurrirse al test de Welch (test t para dos medias independientes con varianzas heterogéneas).

Por ejemplo, veamos el resultado de un test de t-Student que contrasta el IMC de los sujetos del estudio antes del inicio de este (variable respuesta cuantitativa) respecto del sexo (variable predictiva categórica):



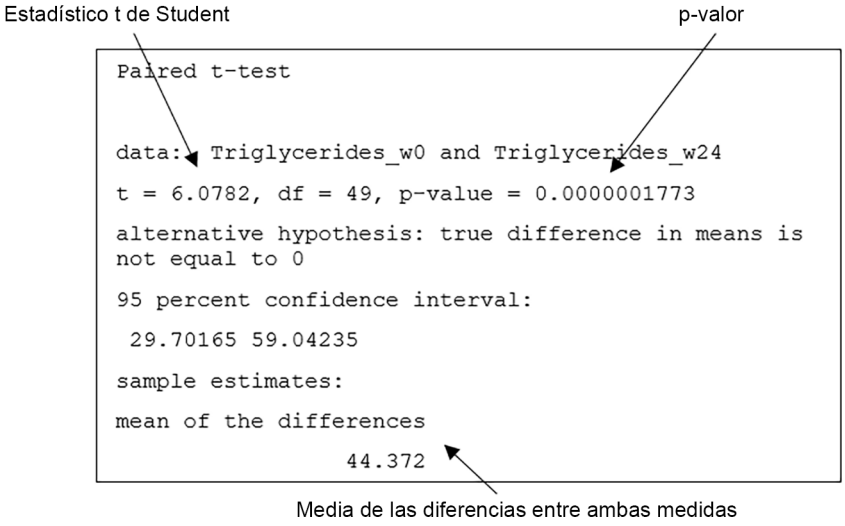
Vemos que el p-valor es mayor que 0,05, lo que indica que no existen diferencias entre ambos grupos. Sin embargo, si realizamos el mismo test usando el nivel de triglicéridos:

```
data:  Triglycerides_w0 by Gender
t = -2.9843, df = 27.261, p-value = 0.005935
alternative hypothesis: true difference in means
is not equal to 0
```

podemos ver que sí existen diferencias significativas entre hombres y mujeres, pues el p-valor es menor que 0,05.

En el caso de muestras emparejadas no se estudia la variabilidad entre individuos (interindividual), sino dentro de un mismo individuo (intraindividual). Por eso no se habla de observaciones independientes, sino de datos apareados, relacionados o emparejados. El tratamiento estadístico es distinto porque la variabilidad aleatoria intraindividual es menor que la interindividual. Aquí la H_0 indicará que las diferencias en los niveles del parámetro antes y después del estudio son nulas (se deben a la aleatoriedad), mientras que la H_a indicará que son diferentes.

Veamos por ejemplo el caso de considerar el nivel de triglicéridos de los pacientes antes y después del estudio:



Vemos que indica que, efectivamente, han cambiado los niveles de forma estadísticamente significativa. Observad que el test también indica la media muestral de las diferencias. Es decir, que existe una disminución media de 44,372 mg/dL en el nivel de triglicéridos.

b) No paramétricos: test U de Mann-Whitney para muestras independientes o **test de Wilcoxon para muestras dependientes o emparejadas**.

La U de Mann-Whitney es la alternativa no paramétrica que sustituye a la t-Student para comparar medias de dos grupos o muestras independientes. Como requiere ordenar los valores antes de hacer el test (algo que el software informático hace por sí solo), no compara realmente las dos medias sino las dos medianas. Hay softwares que denominan este test como test de Wilcoxon para muestras independientes. Se basa en calcular el estadístico *U* comparando cada individuo de un grupo con cada individuo de otro para contabilizar el número de veces que alguien de un grupo presenta un valor superior a alguien de otro. Por eso los individuos de cada grupo deben estar ordenados de mayor a menor. Las H_0 y H_a serán equivalentes a la t-Student, y a este estadístico se le asocia un p-valor que se interpreta de igual manera.

El **test de Wilcoxon** para muestras emparejadas es el sustituto no paramétrico de la t-Student para muestras emparejadas. Se calculan en este caso las diferencias entre cada par de observaciones relacionadas y se ordenan estas diferencias (valor absoluto) de menor a mayor. Después se contabiliza el número de veces que la diferencia entre observaciones ha sido positiva y el número de veces en el que ha sido negativa.

Es decir, el número de veces que se ha incrementado o disminuido el valor del parámetro medido en cada individuo antes y después de aplicar el factor categórico (por ejemplo, el valor de triglicéridos antes y después de aplicar el cambio de dieta). Sabiendo cuántas veces ocurre un cambio positivo y cuántas

uno negativo, se calcula el estadístico y el p-valor. En este caso la H_a indica cambio significativo y lo hace en el sentido de que existe cierta probabilidad de que haya más cambios debidos al factor categórico (por ejemplo, la dieta) que al azar.

2) Cuando la **variable cualitativa tiene tres o más categorías**. Cuando hay más de dos grupos no es correcto utilizar la t-Student, pues ello supondría hacer varios test por parejas, con lo que se acumularía la tasa de error con cada prueba (error). En el caso de que pretendamos estudiar cómo se ve afectado el parámetro cuantitativo (por ejemplo, el nivel de triglicéridos) en función de un factor categórico con más de dos categorías (por ejemplo, más de dos tipos de dietas o tratamientos farmacológicos), hablamos de la comparación de medias en $k > 2$ muestras independientes.

a) Prueba paramétrica: ANOVA (análisis de la varianza). La H_0 del contraste es la ausencia de diferencias, es decir, la igualdad entre las medias de los k grupos (por ejemplo, la igualdad en los niveles de triglicéridos en los individuos tratados con tres tratamientos distintos). Sin embargo, la H_a indicará la diferencia entre grupos. Se utiliza en ese caso el estadístico de contraste F (de Fisher), que se calcula a partir de las varianzas como el cociente de la varianza intergrupos dividido entre la varianza intragrupo (de ahí que se conozca como análisis de la varianza, aunque realmente compare medias):

$$F = \frac{\text{efecto en los datos debido a la pertenencia a los grupos}}{\text{dispersión de los datos debida al azar (efecto aleatorio)}}$$

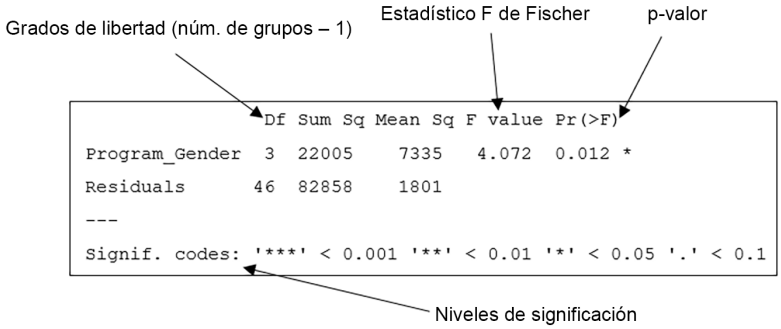
A partir de este estadístico (y los grados de libertad) se calcula un p-valor que se interpreta de modo parecido: ¿cuál sería la probabilidad de que las medias de las muestras difiriesen tanto o más que lo observado? Es decir, en caso de obtener un p-valor =0,01, diríamos que existe un 1 % de probabilidad de que las diferencias entre las tres medias se deban al azar; si fuese un p-valor de 0,05, sería una probabilidad del 5 %; si fuese un p-valor de 0,20, sería una probabilidad del 20 %. En todo caso se suele aceptar un valor máximo de 0,1 (idealmente 0,05, como hemos visto con anterioridad).

Los **grados de libertad** se refieren al número de grupos menos uno ($k - 1$). Este grupo restado es el que el test utiliza de referencia para comparar con el resto.

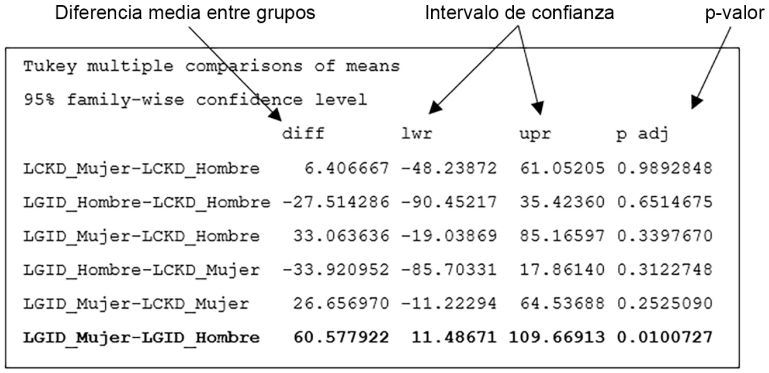
Cabe resaltar que el test ANOVA tan solo indica que existen o no diferencias entre grupos (heterogeneidad u homogeneidad de medias), pero no identifica entre qué grupos, ya que puede que haya diferencias significativas entre los grupos con los tratamientos 1 y 3, pero no entre los grupos con los tratamientos 1 y 2, y/o los tratamientos 2 y 3. Esto se determina posteriormente con las **pruebas *post hoc***.

Las pruebas *post hoc* son todas las posibles comparaciones de medias por pares. Se realizan $k(k-1)/2$ contrastes, siendo k el número de grupos que vamos a contrastar. Por tanto, si hay tres grupos que comparar ($k = 3$), se pueden hacer tres comparaciones por pares (el primero con el segundo y con el tercero, y el segundo con el tercero). Todo software estadístico realiza estos contrastes automáticamente aplicando test de comparaciones múltiples por parejas. El más usual es el **test de Tukey**, en el caso de que se cumpla el supuesto de igualdad de varianzas (aplicando el test de Levene para comprobarlo), o bien el **test Games-Howell**, cuando no se cumple esta condición.

Así pues, usando los datos del ejemplo, podemos valorar si existen diferencias en el valor de triglicéridos respecto al tipo de dieta que sigue cada sexo (variable *Program_Gender*, la cual consta de cuatro grupos posibles). Primero realizaríamos un ANOVA para determinar si existen o no diferencias entre grupos:



El test nos indica que hay diferencias estadísticamente significativas entre los grupos (p-valor <0,05) pero no entre cuáles de ellos. Para ello, realizamos el test de Tukey en este caso:



Puede observarse que la diferencia en el nivel de triglicéridos se detecta tan solo entre hombres y mujeres que siguieron la dieta de bajo índice glucémico (LGID).

b) Prueba no paramétrica: test de **Kruskal-Wallis**. Análogamente a lo que ocurre con ANOVA, no deberíamos usar Mann-Whitney en caso de tener más de dos grupos/muestras debido a que implicaría sucesivos test por parejas, lo que incrementaría el error. Kruskal-Wallis permite comparar las medianas de un conjunto de k muestras independientes y se utiliza cuando no se cumple la condición de normalidad, aunque la potencia estadística que ofrece es menor que en el caso del ANOVA.

Funciona parecido a Mann-Whitney (de hecho, si se aplica con un factor categórico con solo dos categorías obtendremos un resultado idéntico): se ordenan los datos de menor a mayor, se asignan rangos (número de orden) a cada valor y luego se suman los rangos asignados a cada grupo. Si unos grupos acumulan más valores de rangos inferiores (valores más bajos), al sumarlos dará un resultado menor que aquellos grupos cuyas observaciones se situaron en rangos de orden más altos (valores más elevados). Así, cuanto más se diferencien unos grupos de otros al sumar los rangos de sus valores, mayor será la probabilidad de diferencia significativa entre grupos. Si la H_0 fuese cierta, los rangos medios de cada grupo serían similares al rango medio total. Mediante este procedimiento se calcula el estadístico^2, al cual también se asocia un p-valor. El test de Kruskal-Wallis debe entenderse de igual manera que el ANOVA, pues también se limita a indicar que existen diferencias entre grupos, pero sin especificar entre qué grupos.

En el caso no paramétrico las pruebas ***post hoc*** que deben usarse se basan en comparaciones múltiples por pares usando el **test U de Mann-Whitney**. También lo hará cualquier software estadístico de manera automática.

4.2.3. Relación entre dos variables cuantitativas

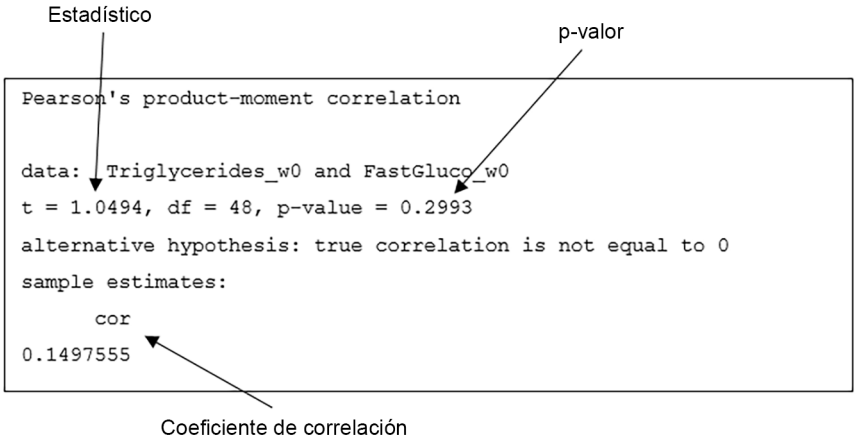
Estudiar la relación entre dos variables cuantitativas es describir cómo cambia la distribución de valores de una variable en función de los valores de la otra. Por ejemplo, ¿cuando aumentan los niveles de triglicéridos, aumentan los niveles de glucosa y viceversa? Una de las dos se considera la variable respuesta (determinada por los objetivos de nuestro estudio) y la otra, la variable predictiva. Si considerásemos que nuestro estudio pretende estudiar los factores que afectan a los niveles de triglicéridos, la pregunta se concretaría en: ¿los niveles de triglicéridos cambian al cambiar los niveles de glucosa?, ¿la relación es directa o inversa?, ¿existe una relación estadísticamente significativa?

Nos centraremos en el estudio de la relación lineal entre dos variables. En el caso de las relaciones no lineales existen técnicas específicas más complejas que quedan fuera del alcance del presente módulo.

Existen dos técnicas muy comunes en epidemiología: la correlación lineal y la regresión lineal.

1) Correlación. El coeficiente de correlación es una medida adimensional (sin unidades) del grado de asociación lineal existente entre dos variables. Suele notarse con la letra *r* y sus posibles valores oscilan entre –1 y 1. Valores próximos a –1 indican fuerte asociación lineal inversa (mayores valores de una variable se asocian a menores valores de la otra) y próximos a 1, fuerte asociación lineal directa (mayores valores de una variable se asocian a mayores valores de la otra). Cuanto más se acerca a 0, menor asociación lineal indica. Para determinar si cierto nivel de correlación es o no significativo, suele usarse

el **test de correlación de Pearson**, a partir del cual podemos obtener un p-valor interpretable. Así, siguiendo con el ejemplo anterior, si relacionamos los valores de triglicéridos y glucosa, obtenemos el siguiente resultado:



Como puede observarse, el coeficiente de correlación positivo y próximo a 0, así como el p-valor, corroboran que no existe significatividad estadística para afirmar que ambas variables se encuentren directamente asociadas. Si las variables de estudio no siguen una distribución normal, una opción es utilizar un test no paramétrico, calculando el coeficiente de correlación de Spearman (rho) cuya interpretación es la misma que el coeficiente de correlación de Pearson y sus valores oscilan entre –1 y 1.

2) Regresión lineal. Antes de intentar una regresión lineal suele haberse comprobado previamente con una correlación que la relación entre ambas variables es efectivamente lineal. Si es así, resulta muy útil encontrar la mejor recta en una gráfica que permita predecir valores en una de las variables (variable respuesta) a partir de la otra (variable predictiva). Esta es una de las mayores cualidades de la regresión lineal.

En el caso bivariante, una regresión lineal se expresa como una relación entre dos variables cuantitativas, en forma de **modelo estadístico bivariante**, de la siguiente manera:

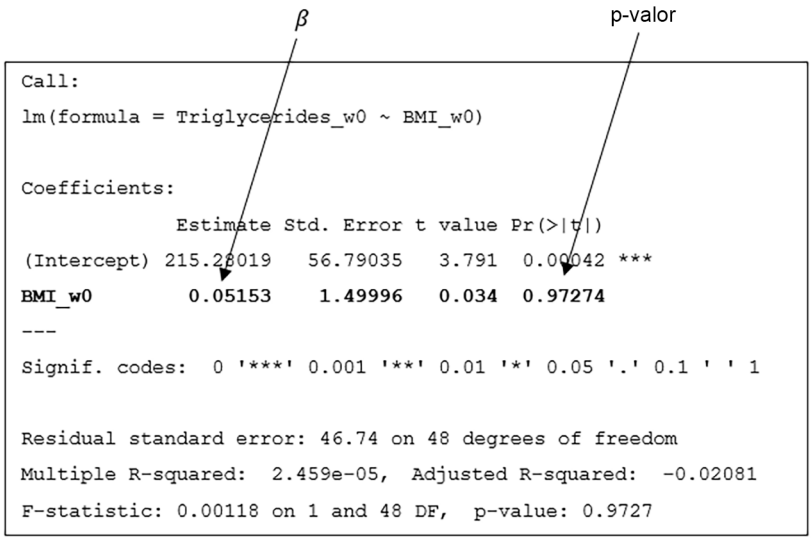
$$Y = \alpha + \beta X$$

donde Y representa la variable respuesta, α la ordenada en el origen (punto donde la recta corta el eje vertical, o de ordenadas, o de las y), β es la pendiente de la recta y X es la variable independiente o predictiva. Es importante destacar que β se entiende como el mayor o menor efecto (pendiente) que X produce sobre Y. Concretamente su valor numérico debe interpretarse como el cambio de unidades producido en Y al aumentar X en una unidad. Además, su signo + o – nos indicará si al aumentar el valor de X aumenta el de Y (β positivo) o disminuye (β negativo).

Por ejemplo, supongamos que queremos valorar el efecto del IMC sobre el nivel de triglicéridos de los individuos de la muestra al inicio del estudio. La expresión anterior puede entenderse de esta manera:

$$\text{Triglicéridos} = \alpha \pm \beta \times \text{IMC}$$

Lo que nos interesa es saber el valor de β y si el cambio que indica β es significativo o no mediante un test estadístico. El resultado del test, algo más complejo, pero de fácil interpretación, será el siguiente:



En este caso, la línea del resultado que nos interesa es la marcada en negrita, que se corresponde con la variable respuesta. Como puede deducirse, el valor de β es muy pequeño y el p-valor indica que no existe un efecto significativo de *X* sobre *Y* (del IMC sobre el nivel de triglicéridos), ya que es ≥ 0,05.

5. Exploración de las relaciones entre variables

En la mayoría de las ocasiones, nuestro estudio no contará con una única variable predictiva cuyo comportamiento queramos observar para determinar cómo afecta a nuestra variable respuesta. Normalmente medimos varias de ellas, que pueden o no estar relacionadas.

> Siguiendo con el ejemplo propuesto, supongamos que queremos estudiar cómo afecta la combinación de todas las variables predictivas de la base de datos al nivel de triglicéridos (que será nuestra variable respuesta). ¿Afectará de igual manera al nivel de triglicéridos que se haya seguido una dieta u otra, teniendo en cuenta el sexo, la edad, el nivel previo de glucosa de cada paciente, el de colesterol, el tipo de medicación que utiliza...?

Como puede intuirse, hay combinaciones inasumibles desde el punto de vista bivariante. Debemos, por tanto, observar el comportamiento de todas las variables de manera conjunta, ejerciendo su efecto sobre la variable respuesta todas las predictivas a la vez. Así pues, el objetivo es observar un **patrón de comportamiento**. A esto nos solemos referir con «explicar la variabilidad de los datos».

Hemos introducido algunos de los test básicos más comunes en el análisis exploratorio de datos. Existe un gran número de pruebas que pueden utilizarse dependiendo del objetivo de nuestro estudio y de la naturaleza categórica o cuantitativa de la variable respuesta y de las variables predictivas. En la tabla 2 se describen algunas de ellas:

Tabla 2. Tipos básicos de test de elección para la realización de inferencias estadísticas

Distribución	Variable predictiva	Variable respuesta	Prueba estadística de elección
Pruebas paramétricas	Categórica	Cuantitativa	Prueba t-Student para una muestra única
			Prueba t-Student para dos muestras
			1º Análisis de la varianza de tres o más muestras (ANOVA) 2º Pruebas de rango *post hoc* y comparaciones múltiples (Tukey si las varianzas son iguales; Games-Howell en caso contrario)
	Cuantitativa	Cuantitativa	Regresión lineal
	Cuantitativa	Categórica dicotómica	Regresión logística
	Cuantitativa	Categórica politómica	Regresión logística multinomial

Distribución	Variable predictiva	Variable respuesta	Prueba estadística de elección
Pruebas no paramétricas	Categórica	Cuantitativa	Prueba de U de Mann-Whitney para dos muestras
			Kruskal-Wallis para más de dos muestras
	Categórica	Categórica	Tablas de contingencia con prueba de χ^2 de Pearson
	Cuantitativa	Cuantitativa	Coeficiente de correlación de Spearman

A continuación, nos detendremos en una técnica muy común en salud pública ya vista: la regresión lineal.

5.1. Modelos de regresión lineal múltiple

Los fenómenos que afectan a la salud suelen tener múltiples causas, por lo que solemos estudiar cualquier parámetro considerando distintas variables simultáneamente.

Supongamos que queremos explicar el comportamiento que ha seguido el nivel de triglicéridos al final del estudio en función del efecto de varios factores observados, como la edad, el sexo, el IMC, el cambio o no en la medicación, el tipo de dieta seguida y el nivel de glucosa resultante. Diremos que todas juntas, unas en presencia de las otras, afectan a la variable respuesta en un grado (más o menos) y en un sentido u otro (incrementando el valor de la variable respuesta al incrementar el suyo, o disminuyéndolo). **El efecto de cada una de las variables puede ser estadísticamente significativo o no**.

La relación lineal entre la variable respuesta y aquellas variables predictivas conformará el **modelo estadístico múltiple**, que puede entenderse ahora de la siguiente manera:

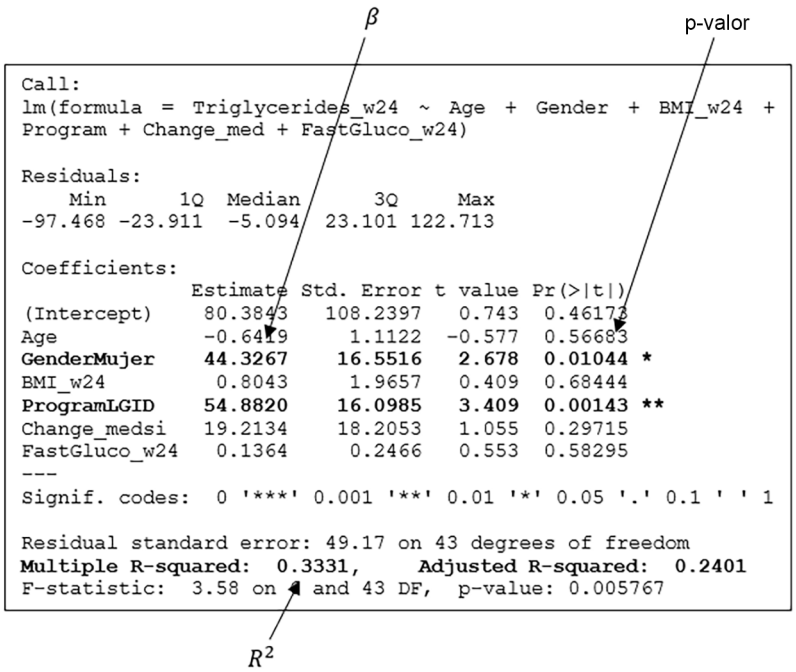
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n$$

siendo n el número de variables predictivas que queramos incluir en el modelo estadístico.

Unas variables predictivas tendrán un efecto positivo (β positivo) y otras, negativo (β negativo) sobre la variable respuesta Y. Entre todas conformarán el patrón de comportamiento del nivel de triglicéridos en la muestra en función de estas variables. Como suele decirse, explicarán la variabilidad de Y en cierto grado (normalmente medido como un porcentaje de variabilidad explicado). Este porcentaje puede ser elevado, en caso de que las variables predictivas incluidas en el modelo justifiquen por sí mismas el comportamiento de Y en

un alto grado. Por el contrario, este porcentaje puede ser menor, debido a que las variables predictivas incluidas en el modelo explican en menor medida el comportamiento de *Y*, por lo que existirán otras variables que no estén incluidas en el modelo (y puede que ni en el estudio) que afecten a *Y* de manera relevante. A este porcentaje explicativo se le llama **coeficiente de determinación R^2**. Por tanto, lógicamente, R^2 aumentará conforme aumentemos el número de variables en el modelo.

Veamos el resultado del ejemplo anterior:



Como puede observarse, el modelo explica el 33,31 % del comportamiento del nivel de triglicéridos, donde el género femenino y el tipo de dieta de bajo índice glucémico han actuado incrementando (β positivo) el nivel de manera estadísticamente significativa (p-valores <0,05). Dicho de otra manera, las mujeres que siguieron este tipo de dieta durante el estudio incrementaron su nivel de triglicéridos de manera estadísticamente significativa.

Veamos otro ejemplo más complejo.

A partir de unos datos obtenidos de un estudio epidemiológico nutricional real, podemos tratar de determinar de qué manera un conjunto de variables sociodemográficas y de estilo de vida (variables predictivas) afectan al consumo de proteínas (variable respuesta) en una muestra de 354 personas mayores de 65 años. En concreto, queremos determinar qué puede afectar a este consumo de proteínas entre las variables sexo, edad, IMC, consumo de líquidos, área de residencia (metropolitana, rural o mixta), número de patologías (comorbilidad) y número de fármacos (polimedicación).

Si realizamos la regresión lineal correspondiente, obtendremos:

```
Call:
lm(formula = PROTEINAS ~ SEXO + EDAD + IMC + TOTALIQUIDOS + PERF_MUN +
    PATTOT + FAR_TOTAL_FARMS)

Residuals:
    Min      1Q  Median      3Q     Max
-57.332 -16.606   0.135  15.710  89.646

Coefficients:
                     Estimate Std. Error t value   Pr(>|t|)
(Intercept)          72.44808   14.84862   4.879 0.00000163 ***
SEXOHombre            1.38487    3.13236   0.442      0.659
EDAD                 -0.14013    0.14448  -0.970      0.333
IMC                   0.31178    0.23753   1.313      0.190
TOTALIQUIDOS          0.28937    0.06647   4.353 0.00001771 ***
PERF_MUNPerfil Mixto  4.46693    3.98306   1.121      0.263
PERF_MUNPerfil Rural 5.69281     3.29577   1.727      0.085
PATTOT                0.65071    0.65318   0.996      0.320
FAR_TOTAL_FARMS      -0.27530    0.53883  -0.511      0.610
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.29 on 345 degrees of freedom
Multiple R-squared:  0.07256,    Adjusted R-squared:  0.05105
F-statistic: 3.374 on 8 and 345 DF,  p-value: 0.0009664
```

En este caso se puede observar que el mayor consumo de líquidos y vivir en un entorno rural mejora el consumo de proteínas de los mayores de 65 años en esta muestra. Sin embargo, la selección de variables predictivas tan solo explica el 7 % del comportamiento del consumo de proteínas, por lo que habría que valorar la inclusión de más variables que puedan ser responsables del resto de la variabilidad de esta variable respuesta.

Resumen

Hemos visto que el resultado satisfactorio de un estudio epidemiológico dependerá del correcto planteamiento de la investigación y de la adecuada interpretación de los datos recogidos.

Existen dos formas de explorar nuestros datos, la numérica y la gráfica, y ambas nos sirven para intuir patrones de comportamiento en la muestra que deberán evaluarse mediante los test estadísticos adecuados.

La elección de los test estadísticos dependerá de la naturaleza categórica o cuantitativa de la variable respuesta y la/s variable/s predictiva/s, así como del cumplimiento de ciertas condiciones de aplicabilidad.

La interpretación de estos test nos permite indicar de manera científica que existen relaciones entre variables de manera significativa, con cierta probabilidad de equivocarnos. Estos resultados se utilizan para confirmar o no la hipótesis inicial que planteamos al principio del estudio.

Bibliografía

Argimón Pallás, J. M.; Jiménez Villa, J. (2004). *Métodos de investigación clínica y epidemiológica*. Madrid: Elsevier.

Hernández-Aguado, I. y otros (2013). *Manual de epidemiología y salud pública para grados en ciencias de la salud*. Madrid: Editorial Panamericana.

INDESTAP (2013). *Aprendiendo de los datos. Un proyecto de innovación docente en estadística aplicada basado en proyectos de investigación*. Grupo de Innovación Docente en Estadística Aplicada. Departamento de Estadística e Investigación Operativa. Valencia: Universitat de València.

Lozano, M. y otros (2017). «Dietary Assesment of Free-Living Elderly Spanish People with Disabilities». *Ecology of Food and Nutrition* (vol. 56, núm. 4, págs. 277-296).

Martínez-González, M. A. y otros (2013). *Bioestadística amigable*. Madrid: Ediciones Díaz de Santos.

Pripp, A. H. (2013). *Statistics in Food Science and Nutrition*. Nueva York: Springer.

Willett, W. (2013). *Nutritional Epidemiology*. Oxford: Oxford University Press.