Metodologia quantitativa: Bioestadística bàsica

PID_00265797

Manuel Lozano Relaño

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Manuel Lozano Relaño

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats per les professores: Anna Bach, Marina Bosque

Primera edició: setembre 2019

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

Índex

Introducció

La recerca científica tracta d'explicar certs comportaments en la població o l'efecte sobre un paràmetre de salut. Aquests comportaments de la població es defineixen per mitjà d'una sèrie de característiques «mesurables», com, per exemple, el sexe, l'edat, l'índex de massa corporal (IMC), la zona de residència o el consum de nutrients. Unes característiques podrien afectar el paràmetre de salut d'interès i unes altres no. I les que ho fan ho faran en sentit positiu o negatiu, en major o menor mesura.

L'eina per mesurar com, quant i en quin sentit es produeixen aquestes associacions es coneix com a **recerca o metodologia quantitativa**. Aquesta metodologia és el procediment d'identificar i quantificar, d'entre totes les característiques de la població, quines resulten significatives en l'explicació del patró de comportament del paràmetre de salut que ens interessa. D'aquesta manera, la recerca quantitativa es basa en l'anàlisi de la causa i l'efecte. Perquè existeixi metodologia quantitativa es requereix que entre aquestes característiques i el paràmetre estudiat existeixi una relació la naturalesa de la qual sigui representable per algun model numèric. Per a això s'usen magnituds numèriques que poden ser tractades mitjançant eines del camp de l'estadística. En l'àmbit de la salut, això es coneix com a **bioestadística**.

1. Planificació de la recerca

Tota recerca comença amb la idea que certs aspectes de la població afecten un paràmetre de salut que és del nostre interès.

Per exemple, com afecta el tipus de dieta la incidència de certa malaltia, o com afecten les característiques sociodemogràfiques d'una població el correcte compliment d'una dieta saludable. Aquest efecte es mesura en un **estudi**, que pot ser de diferents tipus (retrospectiu, prospectiu, cohorts, casos i controls...). El tipus d'estudi condiciona el **disseny de la recerca**.

En estadística, fonamentalment tractem de valorar si l'efecte observat en un estudi es deu a l'atzar o si existeix un patró determinat per certs factors associats al paràmetre o a la població que també observem i tractem de relacionar amb l'efecte.

Tant els efectes que volem estudiar com els factors que pretenem relacionar amb aquests es mesuren a través del que anomenem **variables estadístiques**.

Per tant, depenent de què volem investigar, amb quins objectius i amb quina hipòtesi, haurem de plantejar correctament la recerca. Això resulta crucial per observar correctament tant l'efecte com aquests factors. Quina és la forma correcta de plantejar-ho? Contestant a cinc preguntes bàsiques:

1.1. Què?

Per començar, hem de tenir clar què desitgem investigar: identificar les **variables resposta o dependents** (l'efecte estudiat) i les **variables predictives o independents** (els possibles factors associats). En aquest cas, la variable resposta seria el nivell de triglicèrids mesurat abans i després del canvi dietètic, i el canvi o no en la dieta seria el factor predictiu.

Podríem valorar, per exemple, altres factors que pensem que puguin influir en la variable resposta, com el sexe, l'edat, l'índex de massa corporal, la dieta o la medicació.

1.2. Com?

Depenent de l'efecte i els factors que hàgim de mesurar, hem de triar les millors eines per fer-ho. En aquest cas podríem optar per entrevistes amb qüestionaris de freqüència de consum alimentari i la bioquímica dels individus. El com és molt important, ja que d'ell depèn el correcte disseny *a priori* de la recerca.

Per exemple, pensem en un altre estudi que pretén valorar l'eficàcia d'una prova diagnòstica respecte a una altra per a determinada patologia o paràmetre clínic. En aquest cas hem de preveure la inclusió en la mostra d'individus el resultat dels quals amb la nova prova diagnòstica va resultar tant positiu com negatiu, perquè hem d'analitzar els falsos positius i els falsos negatius.

1.3. Quan?

El nostre estudi pot ser **retrospectiu** (per exemple, estudiant el comportament dietètic en el passat d'aquesta població i contrastar-lo amb el seu nivell de triglicèrids) o **prospectiu** (per exemple, estudiant l'evolució del nivell de triglicèrids durant un temps i contrastar-la amb la dieta seguida durant aquest període, amb intervenció o sense ella).

Però, a més, el moment de recollida en si és rellevant. Així, per exemple, el registre alimentari a l'estiu variarà respecte al registre a l'hivern.

1.4. On?

El lloc de realització de la recerca defineix certes característiques de la població que pretenem estudiar.

No és el mateix estudiar el nivell de triglicèrids a Espanya que al Senegal, o en persones grans que viuen al seu domicili particular que en centres residencials on es dispensen dietes estandarditzades.

1.5. Per què?

En el fons, ens estem preguntant quina és la nostra **hipòtesi**. Si hem d'estudiar la relació entre el nivell de triglicèrids (variable resposta) i el canvi dietètic (variable predictiva), estem hipotetitzant sobre una possible relació entre tots dos factors en un sentit o un altre. Referent a això, és un error comú buscar sempre una «associació significativa», ja que la falta d'associació també és un resultat vàlid, fins i tot satisfactori.

Penseu en un estudi que tracti de relacionar el consum d'aigua potable i el càncer de còlon.

A continuació es descriuran i abordaran els processos que permeten descriure les nostres dades, així com explorar-les i analitzar-les d'una manera molt bàsica.

És el que es coneix com a **anàlisi exploratòria de dades**, i permet enfocar una anàlisi estadística completa posterior, d'acord amb els objectius del nostre estudi.

Tècniques d'anàlisi estadística avançades

Queden fora d'aquest breu text les tècniques d'anàlisi estadística més avançades, a causa de la gran quantitat de possibilitats i a la seva complexitat. Al final del document es troben diversos recursos que poden ser útils en el desenvolupament de les més comunes.

2. Estadística bàsica Conceptes generals

Les dades recopilades en un estudi epidemiològic s'organitzen en una **base de dades** (casos en files i variables en columnes), però, com extraiem la informació que busquem a partir d'elles?

2.1. Organització i tipus de dades

La figura 1 representa un extracte de la base de dades. S'hi observen les dades obtingudes per a sis de les variables dels nou primers participants (d'un total de cinquanta). L'objectiu de l'estudi podria ser l'anàlisi dels canvis en el nivell de triglicèrids entre la setmana 0 i la 24 havent canviat o no la dieta, considerant l'edat i el gènere.

Figura 1. Extracte d'un conjunt de dades d'un estudi epidemiològic

	A	B	C	D	E	F
1	Age	Gender	Triglycerides_w0	Triglycerides_w24	Change_med	Adverse_effect
cas → 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 quantitativa contínua variable qualitativa nominal

Font: INDESTAP. Departament d'Estadística i Recerca Operativa de la Universitat de València.

Les variables es diuen així perquè canvien entre cas i cas (en cas contrari estaríem davant de constants), i poden ser:

1) Qualitatives: presenten els seus resultats en forma de categories (com en el cas del sexe, homes-dones) i poden ser nominals (classificacions independents, com tipus específics de dieta) o ordinals (amb un ordre, com baix, mitjà, alt). Solen anomenar-se variables categòriques o factors.

2) Quantitatives: presenten els seus resultats en forma de valors numèrics (com en el cas de l'edat, la ingesta de lípids o el nivell de triglicèrids). Les variables quantitatives poden ser contínues, quan poden prendre qualsevol valor numèric entre dos valors establerts, i poden aconseguir un nombre infinit de valors diferents, o discretes, quan els seus resultats procedeixen per exemple d'un recompte, en aquest cas poden agafar un nombre finit o infinit numerable de valors.

És freqüent convertir variables quantitatives en qualitatives agrupant els valors numèrics en diferents categories. Això sol ser útil per diferenciar l'efecte mesurat respecte a aquests grups, com, per exemple, grups d'edat o rangs d'índex de massa corporal.

Fulls de càlcul

Els fulls de càlcul (Excel, Open Office) són ideals per recollir la nostra informació i es poden carregar fàcilment en una aplicació estadística com R-Commander, SPSS, PSPP, Stata, etc.

2.2. Població i mostreig

Hem de prestar especial atenció als conceptes de **població (N)** i **mostra (n)**.

El conjunt dels individus objecte del nostre estudi és el que es denomina **població**. Si pretenem estudiar el nivell de triglicèrids en una ciutat, la població serien tots els habitants de la ciutat.

Així, si l'objectiu anés estudiar el nivell de triglicèrids només en majors de 65 anys d'aquesta ciutat, la població estaria formada per tots els majors de 65 anys que hi resideixen. En la majoria d'ocasions això és inabastable, perquè hauríem de recollir les dades de totes i cadascuna d'aquestes persones. Per això sol estudiar-se solament una part representativa: una mostra. Per a això, se selecciona una mostra de persones d'entre el conjunt de la població objecte de l'estudi. Aquest subconjunt, molt menor i manejable, és sobre el qual estudiarem la característica que ens interessa (en aquest cas, el nivell de triglicèrids i els factors associats).

Així, una mostra serà qualsevol subconjunt d'individus d'una població; a la selecció d'individus d'entre la població se l'anomena **mostreig**.

El mostreig és correcte quan garanteix que les característiques de la població queden reflectides apropiadament a la mostra. Preservar la **representativitat** és l'atribut més important que ha de reunir el mostreig, la qual cosa ens permetrà extrapolar la població els resultats obtinguts en el nostre estudi.

En cas que la mostra sigui poc representativa de la població, direm que està esbiaixada (o que té **biaix**). Aquest biaix sol donar-se quan alguns sectors de la població estan més representats dins de la mostra que uns altres.

Però, com se selecciona la mostra i quina mida és l'adequada? Perquè sigui representativa de la població, el procés de selecció de la mostra s'ha de basar en el principi d'**aleatorització**, és a dir, triant a l'atzar, d'entre la població, els individus que formaran part de la mostra. Hi ha diferents tècniques de mostreig. Vegem-ne algunes de les més rellevants utilitzades en salut pública:

1) Mostreig aleatori simple. Aquesta tècnica assegura que tots els individus de la població tenen la mateixa probabilitat de ser inclosos i que els individus se seleccionin de manera independent els uns dels altres.

2) Mostreig aleatori estratificat. Si la nostra població objectiu és especialment heterogènia, podent diferenciar-se en diversos grups (el que es coneix com a estrats) que constitueixen categories importants per a la recerca (per exemple, per sexes), l'elecció de la mostra no ha de fer-se de la mateixa manera en tots els estrats, ja que uns podrien acabar més representats que d'altres. Per determinar els estrats se sol recórrer a variables espacials (comunitats, municipis, entre altres) o d'altres que ens interessin en funció de l'objectiu de l'estudi (la capacitat econòmica, l'ètnia, etc.). Aquest mostreig pretén assegurar la representació de cada grup a la mostra. Una vegada definits els estrats, es realitza un mostreig aleatori simple dins de cadascun d'ells.

3) Mostreig per conveniència. Aquest tipus de mostreig no és aleatori i es basa en triar la mostra segons la facilitat d'accés a ella, la disponibilitat dels participants per formar part de l'estudi, o qualsevol requisit estricte de la població d'estudi que restringeixi la mida de la població. Quan s'utilitza aquesta tècnica, es poden observar hàbits, opinions, i punts de vista de manera més fàcil, per la qual cosa és molt utilitzada en salut pública, encara que ha de justificar-se suficientment el seu ús.

Tot estudi porta implícit en la fase de disseny la determinació de la mida de mostra necessària. És molt important que la mostra sigui representativa de la població, i la mida necessària depèn de l'error que estiguem disposats a assumir en els nostres resultats. Vegem a continuació a què ens referim.

2.3. Contrastos d'hipòtesis

Existeix una relació entre la mida de la mostra i l'**error estadístic**. De la nostra població d'estudi ens plantegem una sèrie de preguntes. Aquestes preguntes es tradueixen en **hipòtesis** estadístiques. Es plantegen dues hipòtesis i s'ha de triar entre l'anomenada **hipòtesi nul·la (H_0)** i la **hipòtesi alternativa (H_a)**. Això és el que es coneix com a contrast d'hipòtesis.

Se sol considerar la hipòtesi nul·la com aquella que determinarà que no hi ha relació entre l'efecte mesurat i la variable estudiada, i l'alternativa estipula que sí que n'hi ha, de manera «estadísticament significativa». Veurem què vol dir això últim.

En aquest procés es poden cometre dos errors possibles:

1) Error de tipus I: rebutjar H_0 quan és certa. L'investigador arriba a la concussió que hi ha diferències entre les hipòtesis (o variables d'estudi) quan en realitat no n'hi ha. Per tant l'error tipus I dóna falsos positius.

2) Error de tipus II: acceptar H_0 quan és falsa. L'investigador és incapaç de trobar diferències quan en realitat existeixen. Per tant equival a un fals negatiu.

Les seves probabilitats, freqüentment denominades riscos, venen donades per:

1) α = Probabilitat de rebutjar H_0 quan és certa (error de tipus I). L'investigador arriba a la conclusió que hi ha diferències entre les hipòtesis (o variables d'estudi) quan en realitat no n'hi ha. Per tant l'error tipus I dóna falsos positius.

2) β = Probabilitat d'acceptar H_0 quan no és certa (error de tipus II). L'investigador és incapaç de trobar diferències quan en realitat existeixen. Per tant equival a un fals negatiu.

Normalment s'estableix per consens el valor del risc α en 0,05 (5 %) i es denomina **nivell de significació**, mentre que 1 – β es coneix com a **potència** de la prova, que ha de ser almenys de 0,8 (80 %). Com més incrementem *n* (mida de la mostra), més decreixeran α i β i augmentarà la potència de la prova, amb la qual cosa disminuiran els errors.

El grau de risc de cometre aquests errors el podem quantificar amb el nivell crític *p* (o **p-valor**).

El p-valor és la probabilitat d'obtenir una discrepància major o igual a l'observada en la mostra quan H_0 és certa, i pot considerar-se com l'error de tipus I que cometem amb les dades observades a la mostra.

Dit d'una altra manera, el **p-valor** és la probabilitat (mesurada en escala de 0 a 1) que, en repetir l'experiment en idèntiques condicions en una altra mostra aleatòria de la mateixa població, trobem resultats discrepants.

Per tant, un p-valor = 0,05 s'entén com un 5 % de probabilitat que en repetir l'experiment obtinguem resultats diferents. α (nivell de significació) serà el límit acceptable que imposem a aquesta probabilitat, que sol ser 0,05 (5 %) per consens entre la comunitat científica. Així:

si p-valor < α → Rebutgem H_0

si p-valor ≥ α → No rebutgem H_0

2.4. Estadística descriptiva i inferencial

L'estadística té com a objectiu últim extreure informació de les dades recollides. I ho aborda de dues maneres:

1) Estadística descriptiva: s'encarrega de descriure, analitzar i representar un grup de dades utilitzant mètodes numèrics i gràfics que resumeixen i presenten informació continguda en ells. Solament descriu la mostra.

2) Estadística inferencial: a partir del càlcul de probabilitats i de dades mostrals, s'encarrega del càlcul d'estimacions i prediccions de la relació entre variables. Treu conclusions sobre el possible patró de comportament en una mostra.

3. Estadística descriptiva

Les nostres variables es poden descriure de manera tant numèrica com gràfica. Quan descrivim una sola variable parlem d'**anàlisi univariable**, quan descrivim una variable en funció d'una altra ens referim a **anàlisi bivariant**, i quan en l'anàlisi hi ha implicades més variables (per exemple, una regressió lineal amb una variable resposta i diverses predictives), parlem d'**anàlisi multivariant o múltiple**. Tant la descripció de variables com les anàlisis estadístiques es realitzen fàcilment amb l'ajuda de diferents programes estadístics.

3.1. Descripció numèrica de les dades

La descripció numèrica de les dades es realitza per mitjà del que anomenem **estadístics descriptius** Qualsevol programari estadístic les proporciona fàcilment.

Quines són? Depèn de si la variable és categòrica o quantitativa. Quan treballem amb variables categòriques solen utilitzar-se freqüències per a la seva descripció:

1) Freqüència absoluta: nombre de vegades que apareix en l'estudi un determinat valor o categoria.

2) Freqüència relativa: quocient entre la freqüència absoluta i la mida de la mostra, n.

3) Percentatge: freqüència relativa multiplicada per 100.

En el cas de les variables quantitatives ens referim a les mesures de tendència central, dispersió i posició, que comparen qualsevol individu de la mostra en relació amb el valor central i permeten comparar resultats mitjans obtinguts per dos o més grups d'individus:

1) Mitjana: és la suma de tots els valors de la mostra de la variable dividida entre n.

2) Mediana: és el valor de la variable que deixa per sota la meitat de les dades, una vegada que aquestes estan ordenades de menor a major. Equival al percentil 50 de la distribució (p50).

3) Moda: és la dada més repetida, el valor de la variable amb major freqüència absoluta.

4) Rang: és la diferència entre el valor mínim i el valor màxim de la variable.

5) Màxim: és el valor màxim de la variable.

6) Mínim: és el valor mínim de la variable.

7) Variància: és una mesura estadística que mesura la dispersió dels valors respecte de la seva mitjana.

8) Desviació típica o estàndard: és l'arrel quadrada de la variància i té la mateixa funció.

9) Percentil p: és el valor que deixa per sota el p% d'observacions de la mostra. Es correspon amb els quartils: percentil 25 (quartil 1), percentil 50 (quartil 2), percentil 75 (quartil 3). No obstant això, se solen utilitzar molts altres percentiles per comparar mesuraments individuals respecte a la resta de la població.

10) Interval de confiança (IC): és un interval que ens indica entre quins valors es troba el paràmetre que estem contrastant al nostre nivell de confiança. Per a una confiança habitual del 95 % (= 0,05), de cada 100 experiments, en el 95 % l'IC comprendrà el veritable valor. L'IC és equivalent al p-valor per fer contrastos d'hipòtesis. Rebutgem H_0 quan l'IC no contingui el valor del paràmetre.

Tant la variància com la desviació típica són sensibles a l'existència de valors extrems o ***outliers*** (valors atípics molt distants de la resta de les dades. No s'han de confondre amb el **valor màxim** i el **valor mínim**). Aquests valors poden semblar prescindibles perquè distorsionen els estadístics descriptius, però sovint són les dades més interessants, especialment en recerca sanitària, per la qual cosa hem de valorar com abordar-les de la manera que millor convingui al nostre estudi.

3.2. Descripció gràfica de les dades

Una vegada descrites les diferents variables de forma numèrica mitjançant els estadístics descriptius, podem inspeccionar el seu comportament, especialment respecte a la resta de les variables, per mitjà de **tabulacions** i de la **representació gràfica**.

Aquest procés serà també diferent en el cas de variables categòriques o quantitatives.

1) Variables categòriques. Solen tabular-se mitjançant **taules de freqüències** o de contingència, en les quals es distribueixen les freqüències de la variable resposta, categòrica en aquest cas, en relació amb un altre factor, també cate-

gòric, respecte al qual volem valorar com es comporta. Per exemple, a partir de les dades de la taula 1, podem descriure el canvi de medicació en els cinquanta participants a l'estudi en funció del gènere:

Taula 1. Canvi de medicació en funció del gènere

	Gender		
Change_med	**Home**	**Dona**	
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

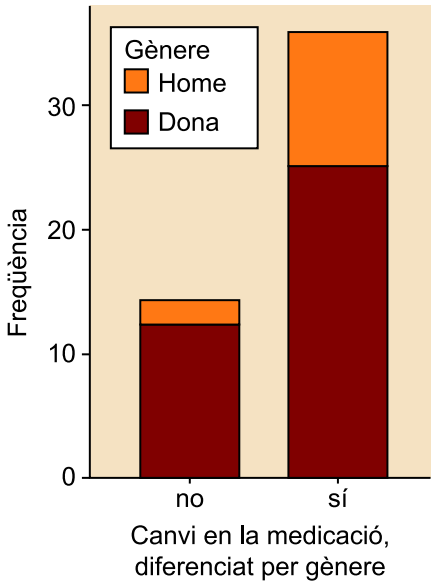
Com pot apreciar-se, sovint s'associa una prova estadística que contrasta la hipòtesi nul·la (no hi ha diferències entre les categories), enfront de la hipòtesi alternativa (si n'hi ha). Explicarem aquest concepte més endavant.

Per representar gràficament aquestes diferències entre categories, disposem de diverses opcions, les més comunes de les quals són:

a) Diagrames de barres (*barplot*). Les gràfiques de barres són una manera de representar freqüències; les freqüències estan associades amb categories. Una gràfica de barres es presenta de dues maneres: horitzontal o vertical.

La **gràfica de barres** serveix per comparar i tenir una representació gràfica de la diferència de freqüències o d'intensitat de la característica numèrica d'interès.

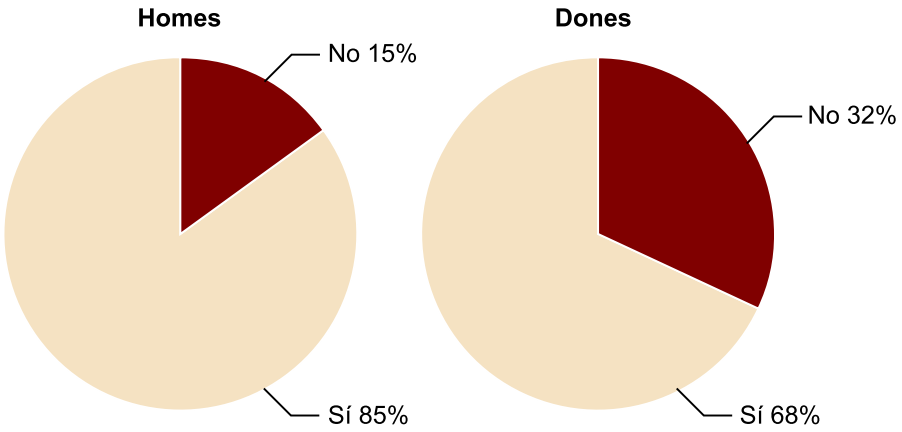
Figura 2. Diagrama de barres categòric



Participants amb canvi en la dieta en funció del sexe. Canvi en la medicació, diferenciat per gènere.

b) Diagrames de sectors (*pie*). Es tracta d'un gràfic que consisteix en un cercle dividit en sectors d'amplitud proporcional a la freqüència de cada valor. Poden afegir-se etiquetes amb les freqüències de cada categoria o els percentatges de cadascuna d'elles.

Figura 3. Diagrama de sectors

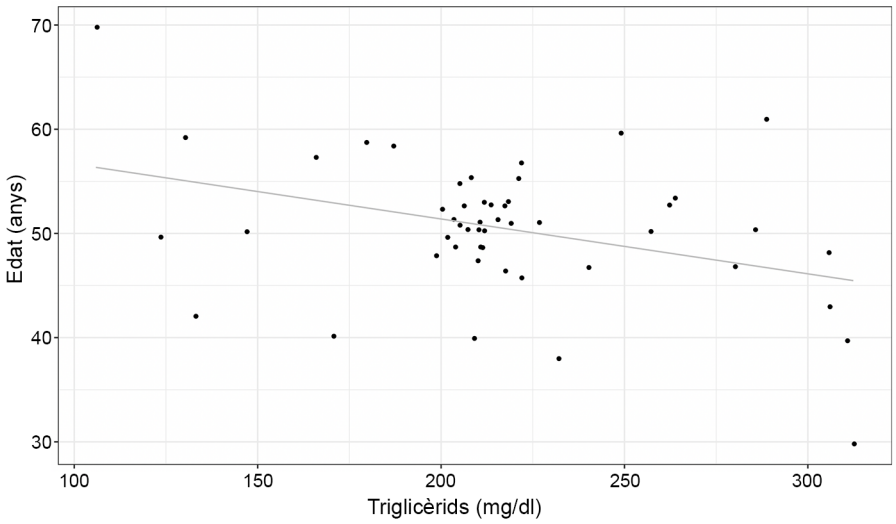


Canvi en la medicació en funció del sexe.

2) Variables quantitatives. La informació continguda en aquest tipus de variables és més gran que en el cas qualitatiu i té més possibilitats de representació. Es poden representar els estadístics descriptius i observar els valors més habituals, com varien, com de dispersos són, si s'observa algun patró de distribució, etc. I tot això es pot fer, a més, en funció d'altres variables, tant quantitatives com a categòriques (agrupant els resultats per grups). Així, es poden analitzar d'un sol cop d'ull les relacions entre variables. Vegem alguns dels casos més senzills:

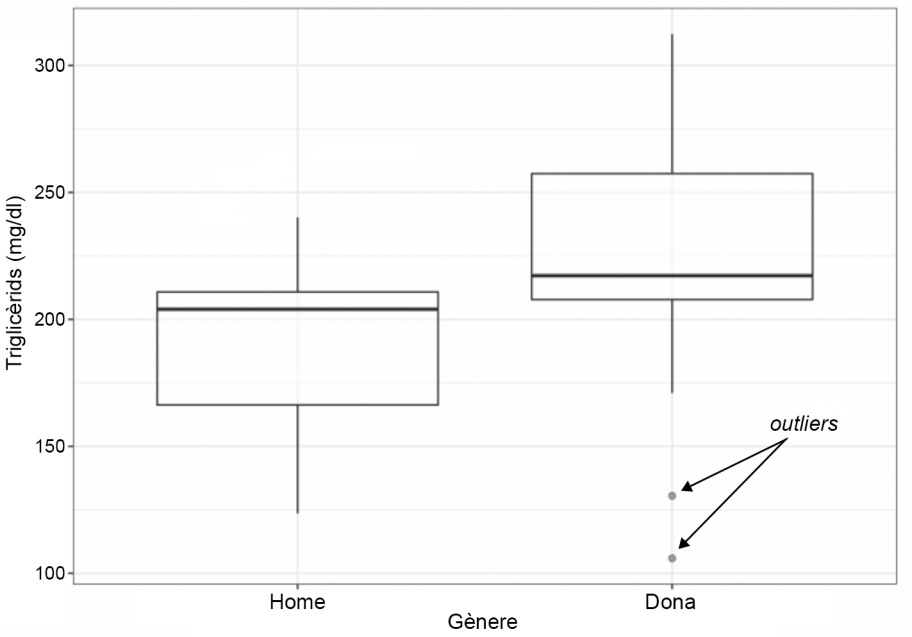
a) Diagrames de punts (*scatter plot*). Pot representar-se una variable quantitativa en funció d'una altra quantitativa. En aquest cas, usant el conjunt de dades de mostra, representem a la figura 4 el nivell de triglicèrids a l'inici de l'estudi (eix x) en funció de l'edat (eix y). Afegim una recta de regressió que marca la relació entre ambdues variables. S'aprecia una relació inversa entre ambdues (a més edat, menys nivell de triglicèrids).

Figura 4. Diagrama de punts. Triglicèrids en funció de l'edat, ambdues variables quantitatives



b) Diagrames de caixa (*boxplot*). Permet apreciar fàcilment la distribució de les dades, representant en una caixa les dades contingudes entre el quartil superior i l'inferior (la línia horitzontal dins de la caixa és la mediana). Les línies verticals, o bigotis, marquen l'extrem superior i inferior. Els punts aïllats més allunyats són els *outliers*. A la figura 5 s'observa que existeix una clara diferència de distribució en el nivell de triglicèrids entre gèneres.

Figura 5. Diagrama de caixes (*boxplots*). Triglicèrids en funció del gènere, quantitativa enfront de categòrica

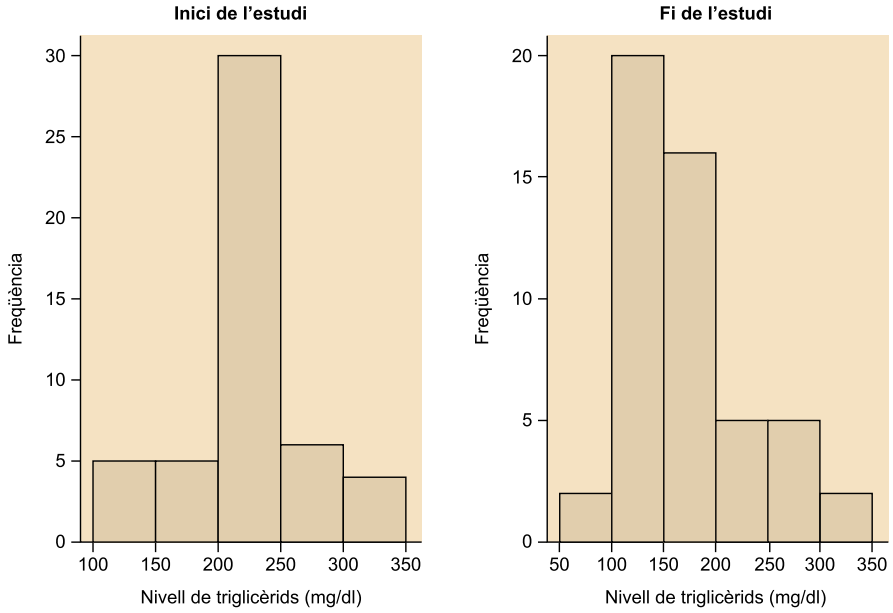


. Triglicèrids enfront de gènere.

c) Histograma. És una representació gràfica d'una variable en forma de barres, on l'alçària de cada barra és proporcional a la freqüència dels valors representats. Serveix per estimar la distribució de la mostra, respecte a una característica, quantitativa i contínua (com l'IMC, l'edat, el pes o la concentració de

triglicèrids...). D'aquesta manera, ofereix una visió sobre possibles tendències. A la figura 6 s'observa que existeixen tendències molt diferents en els nivells de triglicèrids abans i després de l'estudi.

Figura 6. Histograma



Freqüències del nivell de triglicèrids abans i després de l'estudi.

4. Estadística inferencial

Fins a aquest punt s'han descrit tècniques d'organització, presentació i resum de dades que proporcionen informació sobre les variables estudiades. Aquest tipus d'exploració ens dona una idea del comportament de les variables i les possibles relacions entre elles, generalment de dues en dues. No obstant això, un dels objectius bàsics en la recerca és extreure conclusions basades en l'evidència científica.

Per afirmar que un efecte respon a cert patró de comportament no n'hi ha prou amb l'estadística descriptiva, sinó que hem d'emprar tests estadístics basats en el contrast d'hipòtesis, que ja hem introduït i que avaluen matemàticament la probabilitat que l'efecte observat es degui simplement a l'atzar.

Per mesurar el grau d'associació entre variables, els tests estadístics realitzen el contrast d'hipòtesis. Aquests tests tracten de contrastar així la variable d'interès en el nostre estudi (aquella el comportament de la qual pretenem explicar respecte a les altres i que es coneix com a **variable resposta**, o dependent) respecte a la resta de les variables (aquelles que també mesurem a l'estudi i que pretenem associar al comportament de la primera, i que es coneixen com a **variables predictives**, o independents).

Per realitzar aquest contrast, un test estadístic calcula:

- Un **coeficient estadístic** (específic de cada prova), associat a la variable predictiva, que mesura el grau d'associació entre ambdues variables. També indicarà el sentit d'aquesta associació: negatiu o positiu, depenent de si el coeficient és negatiu o positiu, respectivament. A títol orientatiu, com més s'allunyi del 0, major serà l'associació. S'entén com el nombre d'unitats que canvia la variable resposta quan augmenta la variable predictiva en una unitat.
- Un **p-valor**, que indicarà si l'associació expressada pel coeficient és significativa o no, sota certa probabilitat. És a dir, la probabilitat que el canvi en la variable resposta no es degui a l'atzar sinó a aquest efecte de la variable predictiva. Aquesta probabilitat es considera que ha de ser almenys del 95 % (o 0,95, expressada sobre 1), per la qual cosa el valor de referència, o nivell de significativitat, és del 5 % (o 0,05). Per això, es considera que el test indica «significativitat estadística» quan p-valor és 0,05.

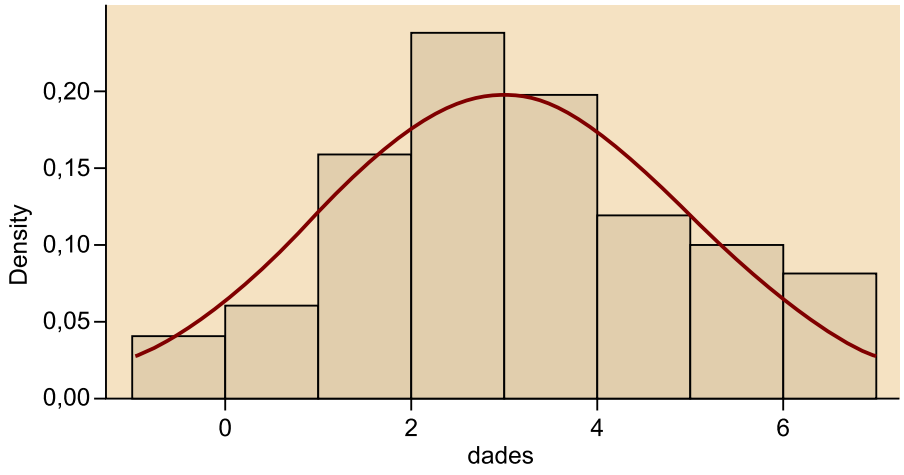
4.1. Tipus de tests estadístics i condicions d'aplicabilitat

Hi ha dos grans tipus de tests estadístics: els paramètrics i els no paramètrics, i no poden utilitzar-se uns o uns altres indistintament.

a) Les proves paramètriques es basen en la mitjana, la desviació estàndard i altres mesures de dispersió (paràmetres) i requereixen el compliment d'unes condicions d'aplicació més estrictes. Principalment que les dades segueixin una distribució normal i que les variàncies dels grups o variables contrastades siguin homogènies (que la seva dispersió sigui similar, la qual cosa es denomina homocedasticitat).

En estadística es denomina **distribució normal**, o distribució de Gauss, a una de les distribucions dels valors d'una variable contínua que amb més freqüència apareix en estadística. En representar gràficament els valors obtinguts de la nostra variable per a cada individu, aquests valors es distribuiran entorn de l'estadístic descriptiu (en general, la mitjana). Si aquestes dades segueixen una distribució normal, la majoria dels valors de la nostra variable es concentraran al voltant del valor de la mitjana, i disminuiran progressivament els valors menors i majors de la mitjana a mesura que s'allunyen d'aquesta, amb el que s'obté una forma acampanada i simètrica. Hi haurà més o menys valors que s'allunyin de la mitjana en funció de si la desviació estàndard és major o menor, respectivament. A aquesta corba se la coneix com a **campana de Gauss**. A la figura 7, s'observa un exemple basat en unes dades generades perquè es distribueixin segons una distribució normal. La campana serà més o menys punxeguda o plana (el que anomenem curtosi) en funció de la seva menor o major desviació estàndard, respectivament; s'accepta com a normal quan té el 95 % de les seves observacions dins de l'interval mitjana 1,96 vegades la desviació estàndard.

Figura 7. Histograma + corba de distribució normal d'unes dades aleatòries (n = 50, mitjana = 3, desviació = 2)



Freqüències del nivell de triglicèrids abans i després de l'estudi.

Per tant, aquestes condicions solament poden complir-les les variables quantitatives. Quan les variables quantitatives no compleixen aquests requisits o es tracta de variables categòriques, han d'usar-se les proves no paramètriques.

En cas que els tests indiquin que no existeix normalitat, podrem igualment utilitzar tests paramètrics si la mida de la mostra és gran.

Això es deu al **teorema central del límit**, que indica que, si tenim un grup gran d'observacions, la seva distribució «s'aproxima» al comportament d'una distribució normal.

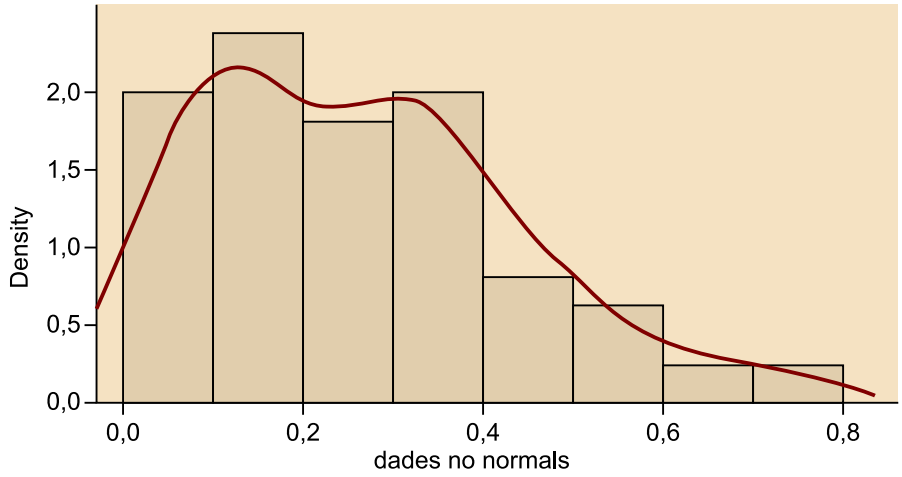
Una altra alternativa és provar de transformar la variable resposta en forma logarítmica, o com a arrel quadrada, la qual cosa de vegades aproxima la seva distribució a la normal i supera les proves de normalitat. Quan així i tot no existeixi normalitat, la mostra sigui petita i/o siguin dades categòriques, haurem d'usar proves no paramètriques.

Existeixen tests estadístics que contrasten les hipòtesis de normalitat (H_0) enfront de la no normalitat (H_a) a disposició a qualsevol programari estadístic. Alguns dels més coneguts són el **test de Kolmogorov-Smirnov** i el **test de Shapiro-Wilk**. Per avaluar l'homogeneïtat de variàncies sol usar-se el **test de Levene**. Valors de $p < 0.05$ en aquests tests rebutgen la hipòtesi de normalitat o d'homogeneïtat de variàncies.

b) Les proves no paramètriques solament tenen en compte l'ordre que ocupa cada observació en el conjunt ordenat d'observacions. Això permet realitzar proves de contrast d'hipòtesi que no assumeixen cap distribució específica (no és necessari que les dades segueixin una distribució normal). Per tant, s'empren quan les variables quantitatives no compleixen els principis de normalitat i homogeneïtat de variàncies, o bé quan es tracta de variables categòriques.

Per exemple, el mateix conjunt de dades de la figura 7 podria seguir una distribució no normal, tal com es veu en la figura 8, on s'aprecia una forma irregular i una cua cap a la dreta.

Figura 8. Histograma + corba de distribució d'unes dades ideals aleatòries



Com ja hem comentat, el nostre estudi pretén observar i explicar el comportament d'un paràmetre de salut (o varis), mesurat a través d'una variable. Per exemple, i seguint amb les dades presentades a l'inici del mòdul, el nivell de triglicèrids a les 24 setmanes d'un tractament.

D'altra banda, l'objectiu del nostre estudi es basarà a observar un altre/s factor/és en els participants que pensem que pot/n afectar al canvi en el nivell de triglicèrids, o no, concepte en el qual es basarà el nostre contrast d'hipòtesis. Aquests factors són els que anomenem **variables predictives**. En el nostre cas pot ser el canvi de dieta o el canvi de medicació, o l'edat, o tots ells.

Tant la variable resposta com les predictives poden ser categòriques o quantitatives. I el tipus de test estadístic que utilitzem ha de triar-se en funció d'aquest aspecte. Vegem a continuació algunes de les proves estadístiques més usuals que s'utilitzen de forma exploratòria en salut pública.

4.2. Estadística inferencial bivariant

Reprenguem les dades presentades a l'inici del mòdul (figura 1). D'ara endavant suposem que nostra variable resposta és el nivell de triglicèrids de la mostra (mesurat a l'inici de l'estudi i a la setmana 24), i la resta de les variables es consideraran predictives (edat, gènere, canvi en la dieta, canvi en la medicació, etc.). El nostre objectiu és determinar si aquestes variables predictives afecten la variable resposta i com ho fan. El contrast d'hipòtesis serà per tant avaluar la H_0, definida com un canvi observat sobre la variable resposta a causa de l'atzar, enfront de la H_a, que determinarà un canvi observat sobre la variable resposta a causa de la variable predictiva, amb certa probabilitat (recordem que se sol considerar un mínim del 95 % per afirmar que la relació és significativa).

En primer lloc, solen realitzar-se proves seleccionant les variables de dues en dues (proves per parells), les quals contrasten el paràmetre de la variable resposta (concentració de triglicèrids en aquest cas) enfront de cadascuna de les variables predictives per separat (concentració enfront de gènere, concentració enfront d'edat, concentració enfront de canvi en la dieta, etc.). Aquests tests tracten d'avaluar l'efecte que cada variable predictiva exerceix o no per si mateixa, aïllada, com si no hi hagués present cap altre paràmetre que pogués afectar la variable resposta.

A les figures 4 i 5, hem pogut observar certes diferències aparents en els valors del nivell de triglicèrids (variable quantitativa) en funció de l'edat (variable quantitativa), o el gènere (variable categòrica). No obstant això, l'observació mostral és una cosa subjectiva. Hem d'aplicar els tests estadístics apropiats que validin aquests resultats poblacionalment, és a dir, que podríem extrapolar els resultats de la nostra mostra a la resta de la població que representa. Per poder aplicar tests estadístics, necessitem un programari específic. A aquest

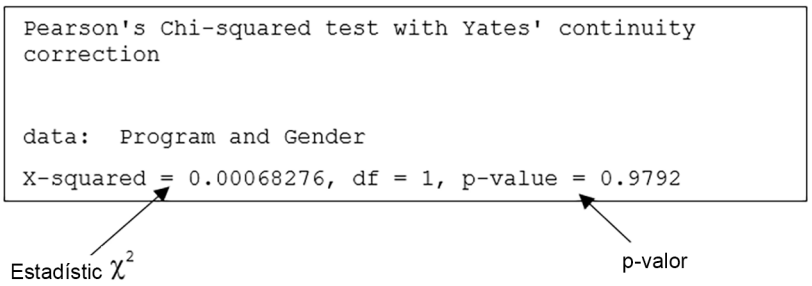
programari s'exporten de manera senzilla les dades que han estat recopilades i organitzades en un format tipus Excel o similar. Hi ha disponibles programes gratuïts, com R-Commander o PSPP, i amb llicència, com SPSS o STATA.

4.2.1. Inferència entre variables categòriques

En ser categòriques, solament podran ser aplicades proves no paramètriques. Cada variable tindrà dues o més categories, i el contrast d'hipòtesis es basa en comparar les freqüències en les diferents combinacions d'una categoria d'una variable amb les categories de l'altra. El test utilitzat amb més freqüència és 1) Test de *Pearson o* $\chi 2$(Chi quadrat) de Pearson. Es tracta d'un test no paramètric que s'utilitza per determinar si existeix una associació significativa entre dues variables categòriques. La H_0 és la independència de les dues variables, és a dir, la igualtat de proporcions entre els grups. L'estadístic de contrast $\chi 2$ es calcula a partir de les freqüències observades i les freqüències esperades. Se li associa un p-valor en funció del qual prendrem la nostra decisió. És imprescindible que cada grup de la mostra tingui almenys cinc individus. Si algun dels grups de la mostra té menys de cinc individus, llavors el test més aconsellat és el test Exacte o de Fisher.

Per exemple, suposem que tenim una mostra d'homes i dones als quals s'aconsella un canvi de dieta. Passats uns mesos, es recull per a cada individu si s'ha produït un canvi de dieta (sí o no). Tindrem quatre grups d'individus: **1)** dones amb canvi de dieta; **2)** dones sense canvi de dieta; **3)** homes amb canvi de dieta; **4)** homes sense canvi de dieta. Si volem determinar si la dieta és seguida de forma diferent per homes o dones, el resultat del test seria el següent:

Figura 9



En aquest cas veiem que no hi hauria diferències entre sexes, doncs p-valor $\geq 0{,}05$.

4.2.2. Inferència entre una variable quantitativa i una variable categòrica

En aquest cas es tracta de descriure la distribució de la variable quantitativa en cada categoria de la variable categòrica. Hem de comparar els valors de la variable resposta en els grups o categories de la variable predictiva, o el que és el mateix, comparar les respostes mitjanes de cada grup.

Les categories de la variable categòrica formen grups que cal comparar. Cal considerar si els grups (o subcategories) són independents entre si o no. Per exemple, la mostra pot analitzar-se en funció de la variable sexe, on trobem les categories home-dona, que són independents perquè divideixen la mostra en casos diferents. El mateix ocorre amb altres factors amb més de dues categories, com l'IMC, que pot dividir la mostra en casos amb < 18,5 kg/m^2, 18,5-25 kg/m^2, 25-30 kg/m^2, > 30 kg/m^2. En aquests casos parlem de **mostres independents**. No obstant això, la variable categòrica pot fer referència a un abans i un després en la mesura d'un mateix paràmetre bioquímic (la variable resposta quantitativa) mesurat en cada pacient abans i després d'una intervenció dietètica. Es tracta en aquest cas d'observacions relacionades, dependents o aparellades (són termes habituals per al mateix concepte). Sol referir-se en aquest cas a **mostres dependents o aparellades**.

Tant en un cas com en l'altre hem de determinar si la diferència entre les mitjanes observades en les diferents categories és estadísticament significativa, és a dir, si podem determinar que el resultat de Y (la variable quantitativa) canvia en funció de X (el factor categòric). Per a això, hem de realitzar un **test de contrast de mitjanes**.

En el cas de mostres independents, la H_0 del test indicarà que no existeixen diferències entre els grups, que hi ha absència d'efecte (per exemple, que no hi ha diferències entre sexes), mentre que la H_a indicarà que sí n'hi ha, amb una certa probabilitat d'equivocar-nos. En el cas de mostres dependents la H_0 indicarà una diferència nul·la entre les diferents observacions del mateix cas (per exemple, no hi ha diferència en el nivell de triglicèrids abans i després del tractament nutricional), mentre que la H_a determinarà que el tractament sí que va produir un efecte sobre el nivell de triglicèrids, també amb una certa probabilitat d'equivocar-nos.

Hem de destacar que les condicions d'aplicabilitat de les proves paramètriques han de satisfer-se en cadascun dels grups en els quals se subdivideix la mostra per categories. Per això, si necessitem analitzar una variable quantitativa d'una mostra no gaire gran en funció d'una variable qualitativa amb diverses subcategories, la mostra se subdividirà conseqüentment. D'aquesta manera, les condicions de normalitat i homocedasticitat en el total de la mostra poden perdre's quan subdividim les dades en categories, i en aquest cas haurem de recórrer a tests no paramètrics si no queda en cada categoria un nombre prou gran de casos (almenys trenta).

Els tests més habituals en ciències de la salut són:

1) Quan la variable qualitativa té dues categories

a) Paramètrics: prova de la **t-Student per a mostres independents** o prova de la **t-Student per a mostres aparellades**.

La **distribució t de Student** és semblada a la distribució normal i la substitueix quan no es coneix la desviació estàndard poblacional (que sol indicar-se com σ, sigma) i cal recórrer a calcular la desviació estàndard en la pròpia mostra (*s*).

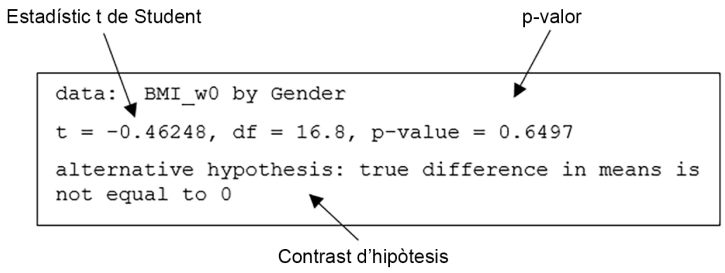
Com gairebé mai es disposa de σ, l'ús de la t de Student és molt usual. Hi ha una *t* diferent per a cada mida de mostra (a això es refereixen els anomenats «graus de llibertat»). Recordem que *n* és la mida de la nostra mostra.

El valor de *t* s'obté en dividir la diferència entre les dues mitjanes («l'efecte») per l'error estàndard de la diferència de mitjanes (o «error»). Així obtenim l'estadístic de contrast *t*, que s'associa a un p-valor. Com en la majoria dels test estadístics, el que s'avalua és la diferència observada entre mitjanes dividida per un terme d'error que representa la variabilitat biològica aleatòria. Si la diferència entre les mitjanes és molt més gran que la variabilitat biològica, *t* serà gran i el p-valor, petit. Si la diferència observada és petita en relació amb l'error estàndard, llavors *t* serà petita i el p-valor, gran, la qual cosa indica que no hi ha diferències significatives. Així, la H_0 serà la igualtat de mitjanes entre categories i la H_a serà la diferència amb una certa probabilitat d'equivocar-nos.

Per aplicar la prova de la t-Student, com a prova paramètrica que és, la mostra ha de complir (i cada categoria en la qual se subdivideixi la mostra) les condicions d'aplicabilitat: normalitat i homocedasticitat.

En el cas d'existir normalitat, però no homocedasticitat, pot recórrer-se al test de Welch (test t per a dues mitjanes independents amb variàncies heterogènies).

Per exemple, vegem el resultat d'un test de t-Student que contrasta l'IMC dels subjectes de l'estudi abans de l'inici d'aquest (variable resposta quantitativa) respecte del sexe (variable predictiva categòrica):



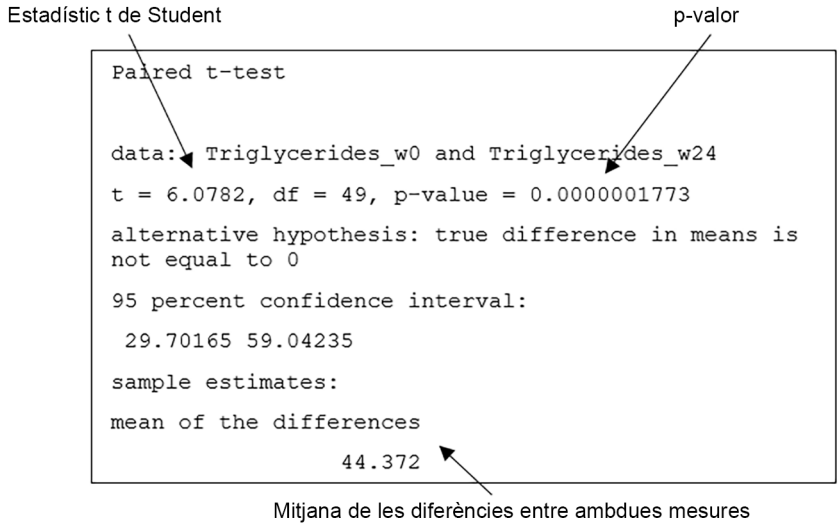
Veiem que el p-valor és més gran que 0,05, la qual cosa indica que no existeixen diferències entre tots dos grups. No obstant això, si realitzem el mateix test usant el nivell de triglicèrids:

```
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
```

podem veure que sí que existeixen diferències significatives entre homes i dones, doncs el p-valor és menor que 0,05.

En el cas de mostres aparellades no s'estudia la variabilitat entre individus (interindividual), sinó dins d'un mateix individu (intraindividual). Per això no es parla d'observacions independents, sinó de dades apariades, relacionades o aparellades. El tractament estadístic és diferent perquè la variabilitat aleatòria intraindividual és menor que la interindividual. Aquí la H_0 indicarà que les diferències en els nivells del paràmetre abans i després de l'estudi són nul·les (es deuen a l'aleatorietat), mentre que la H_a indicarà que són diferents.

Vegem per exemple el cas de considerar el nivell de triglicèrids dels pacients abans i després de l'estudi:



Veiem que indica que, efectivament, han canviat els nivells de forma estadísticament significativa. Observeu que el test també indica la mitjana mostral de les diferències. És a dir, que existeix una disminució mitjana de 44,372 mg/dl en el nivell de triglicèrids.

b) No paramètrics: test U de Mann-Whitney per a mostres independents o **test de Wilcoxon per a mostres dependents o aparellades**.

La U de Mann-Whitney és l'alternativa no paramètrica que substitueix la t-Student per comparar mitjanes de dos grups o mostres independents. Com que requereix ordenar els valors abans de fer el test (alguna cosa que el programari informàtic fa per si sol), no compara realment les dues mitjanes sinó les dues medianes. Hi ha programaris que denominen aquest test com a test de Wilcoxon per a mostres independents. Es basa en calcular l'estadístic *U* com-

parant cada individu d'un grup amb cada individu d'un altre per comptabilitzar el nombre de vegades que algú d'un grup presenta un valor superior a algú d'un altre. Per això els individus de cada grup han d'estar ordenats de major a menor. Les H_0 i H_a seran equivalents a la t-Student, i a aquest estadístic se li associa un p-valor que s'interpreta d'igual manera.

El **test de Wilcoxon** per a mostres aparellades és el substitut no paramètric de la t-Student per a mostres aparellades. Es calculen en aquest cas les diferències entre cada parell d'observacions relacionades i s'ordenen aquestes diferències (valor absolut) de menor a major. Després es comptabilitza el nombre de vegades que la diferència entre observacions ha estat positiva i el nombre de vegades en el qual ha estat negativa.

És a dir, el nombre de vegades que s'ha incrementat o disminuït el valor del paràmetre mesurat en cada individu abans i després d'aplicar el factor categòric (per exemple, el valor de triglicèrids abans i després d'aplicar el canvi de dieta). Sabent quantes vegades ocorre un canvi positiu i quantes un de negatiu, es calcula l'estadístic i el p-valor. En aquest cas la H_a indica canvi significatiu i ho fa en el sentit que existeix certa probabilitat que hi hagi més canvis deguts al factor categòric (per exemple, la dieta) que a l'atzar.

2) Quan la variable qualitativa té tres o més categories. Quan hi ha més de dos grups no és correcte utilitzar la t-Student, doncs això suposaria fer diversos tests per parelles, amb el que s'acumularia la taxa d'error amb cada prova (error). En el cas que pretenguem estudiar com es veu afectat el paràmetre quantitatiu (per exemple, el nivell de triglicèrids) en funció d'un factor categòric amb més de dues categories (per exemple, més de dos tipus de dietes o tractaments farmacològics), parlem de la comparació de mitjanes en $k > 2$ mostres independents.

a) Prova paramètrica: ANOVA (anàlisi de la variància). La H_0 del contrast és l'absència de diferències, és a dir, la igualtat entre les mitjanes dels k grups (per exemple, la igualtat en els nivells de triglicèrids en els individus tractats amb tres tractaments diferents). No obstant això, la H_a indicarà la diferència entre grups. S'utilitza en aquest cas l'estadístic de contrast F (de Fisher), que es calcula a partir de les variàncies com el quocient de la variància intergrups dividit entre la variància intragrup (per aquest motiu es coneix com a anàlisi de la variància, encara que realment compari mitjanes):

$$F = \frac{efecte\ en\ les\ dades\ a\ causa\ de\ la\ pertinença\ als\ grups}{dispersió\ de\ les\ dades\ deguda\ a\ l'atzar\ (efecte\ aleatori)}$$

A partir d'aquest estadístic (i els graus de llibertat) es calcula un p-valor que s'interpreta de manera semblant: quina seria la probabilitat que les mitjanes de les mostres diferissin tant o més que l'observat? És a dir, en cas d'obtenir un

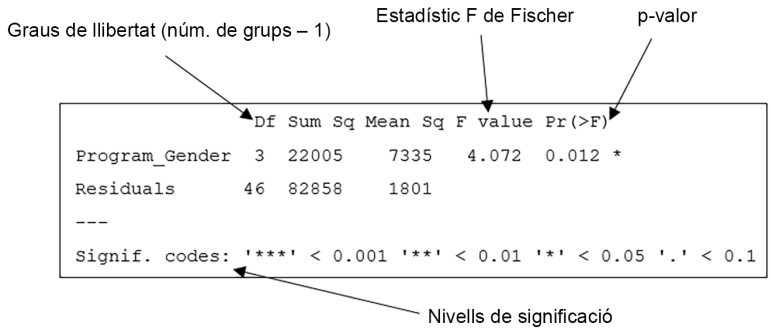
p-valor = 0,01, diríem que existeix un 1 % de probabilitat que les diferències entre les tres mitjanes es deguin a l'atzar; si fos un p-valor de 0,05, seria una probabilitat del 5 %; si fos un p-valor de 0,20, seria una probabilitat del 20 %. En tot cas se sol acceptar un valor màxim de 0,1 (idealment 0,05, com hem vist amb anterioritat).

Els **graus de llibertat** es refereixen al nombre de grups menys un (k – 1). Aquest grup restat és el que el test utilitza de referència per comparar amb la resta.

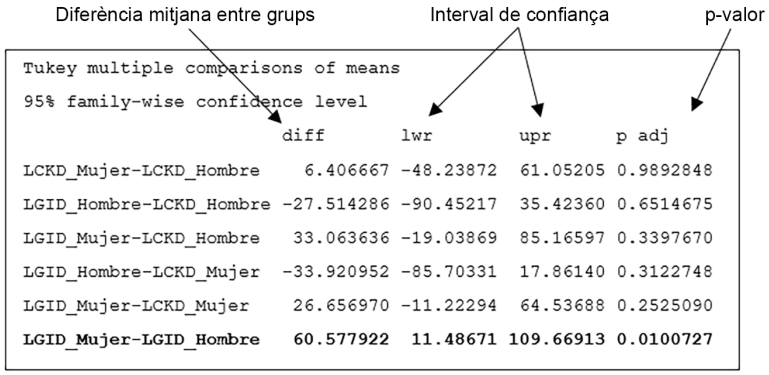
S'ha de ressaltar que el test ANOVA només indica que existeixen o no diferències entre grups (heterogeneïtat o homogeneïtat de mitjanes), però no identifica entre quins grups, ja que pot ser que hi hagi diferències significatives entre els grups amb els tractaments 1 i 3, però no entre els grups amb els tractaments 1 i 2, i/o els tractaments 2 i 3. Això es determina posteriorment amb les proves *posthoc*.

Les **proves *posthoc*** són totes les possibles comparacions de mitjanes per parells. Es realitzen $k(k – 1) / 2$ contrastos, essent k el nombre de grups que contrastarem. Per tant, si hi ha tres grups per comparar ($k = 3$), es poden fer tres comparacions per parells (el primer amb el segon i amb el tercer, i el segon amb el tercer). Tot programari estadístic realitza aquests contrastos automàticament aplicant tests de comparacions múltiples per parelles. El més usual és el **test de Tukey**, en el cas que es compleixi el supòsit d'igualtat de variàncies (aplicant el test de Levene per comprovar-lo), o bé el **test Games-Howell**, quan no es compleix aquesta condició.

Així doncs, usant les dades de l'exemple, podem valorar si existeixen diferències en el valor de triglicèrids respecte al tipus de dieta que segueix cada sexe (variable *Program_Gender*, la qual consta de quatre grups possibles). Primer realitzaríem un ANOVA per determinar si existeixen o no diferències entre grups:



El test ens indica que hi ha diferències estadísticament significatives entre els grups (p-valor < 0,05) però no entre quins d'ells. Per a això, realitzem el test de Tukey en aquest cas:



Pot observar-se que la diferència en el nivell de triglicèrids es detecta només entre homes i dones que van seguir la dieta de baix índex glucèmic (LGID).

b) Prova no paramètrica: test de **Kruskal-Wallis**. Anàlogament al que ocorre amb ANOVA, no hauríem d'usar Mann-Whitney en cas de tenir més de dos grups/mostres perquè implicaria successius tests per parelles, la qual cosa incrementaria l'error. Kruskal-Wallis permet comparar les medianes d'un conjunt de *k* mostres independents i s'utilitza quan no es compleix la condició de normalitat, encara que la potència estadística que ofereix és menor que en el cas de l'ANOVA.

El funcionament s'assembla al de Mann-Whitney (de fet, si s'aplica amb un factor categòric amb només dues categories obtindrem un resultat idèntic): s'ordenen les dades de menor a major, s'assignen rangs (número d'ordre) a cada valor i després se sumen els rangs assignats a cada grup. Si uns grups acumulen més valors de rangs inferiors (valors més baixos), en sumar-los donarà un resultat menor que aquells grups les observacions dels quals es van situar en rangs d'ordre més alts (valors més elevats). Així, com més es diferenciïn uns grups d'uns altres en sumar els rangs dels seus valors, major serà la probabilitat de diferència significativa entre grups. Si la H_0 fos certa, els rangs mitjans de cada grup serien similars al rang mitjà total. Mitjançant aquest procediment es calcula l'estadístic[2], al qual també s'associa un p-valor. El test de Kruskal-Wallis ha d'entendre's d'igual manera que l'ANOVA, ja que també es limita a indicar que existeixen diferències entre grups, però sense especificar entre quins grups.

En el cas no paramètric les proves **post*hoc*** que han d'usar-se es basen en comparacions múltiples per parells usant el **test U de Mann-Whitney**. També ho farà qualsevol programari estadístic de manera automàtica.

4.2.3. Relació entre dues variables quantitatives

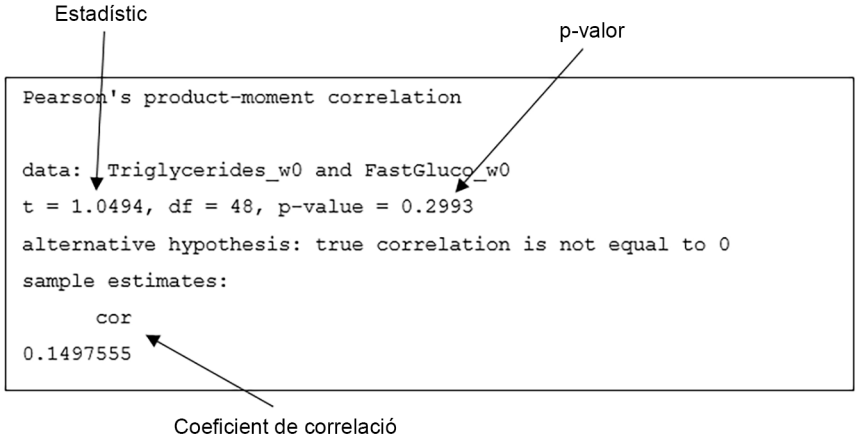
Estudiar la relació entre dues variables quantitatives és descriure com canvia la distribució de valors d'una variable en funció dels valors de l'altra. Per exemple, quan augmenten els nivells de triglicèrids, augmenten els nivells de glucosa i viceversa? Una de les dues es considera la variable resposta (determinada pels objectius del nostre estudi) i l'altra, la variable predictiva. Si conside-

réssim que el nostre estudi pretén estudiar els factors que afecten els nivells de triglicèrids, la pregunta es concretaria en: els nivells de triglicèrids canvien en canviar els nivells de glucosa?, la relació és directa o inversa?, existeix una relació estadísticament significativa?

Ens centrarem en l'estudi de la relació lineal entre dues variables. En el cas de les relacions no lineals existeixen tècniques específiques més complexes que queden fora de l'abast del present mòdul.

Existeixen dues tècniques molt comunes en epidemiologia: la correlació lineal i la regressió lineal.

1) Correlació. El coeficient de correlació és una mesura adimensional (sense unitats) del grau d'associació lineal existent entre dues variables. Sol anotar-se amb la lletra *r* i els seus possibles valors oscil·len entre –1 i 1. Valors propers a –1 indiquen forta associació lineal inversa (valors més grans d'una variable s'associen a valors més petits de l'altra) i propers a 1, forta associació lineal directa (valors més grans d'una variable s'associen a valors més grans de l'altra). Com més s'apropa a 0, menor associació lineal indica. Per determinar si cert nivell de correlació és o no significatiu, sol usar-se el **test de correlació de Pearson**, a partir del qual podem obtenir un p-valor interpretable. Així, seguint amb l'exemple anterior, si relacionem els valors de triglicèrids i glucosa, obtenim el resultat següent:



Com pot observar-se, el coeficient de correlació positiu i proper a 0, així com el p-valor, corroboren que no existeix significativitat estadística per afirmar que ambdues variables es trobin directament associades. Si les variables d'estudi no segueixen una distribució normal, una opció és utilitzar un test no paramètric, calculant el coeficient de correlació de Spearman (rho) la interpretació de la qual és la mateixa que el coeficient de correlació de Pearson i els seus valors oscil·len entre –1 i 1.

2) Regressió lineal. Abans d'intentar una regressió lineal sol haver-se comprovat prèviament amb una correlació que la relació entre ambdues variables és efectivament lineal. Si és així, resulta molt útil trobar la millor recta en una gràfica que permeti predir valors en una de les variables (variable resposta) a partir de l'altra (variable predictiva). Aquesta és una de les majors qualitats de la regressió lineal.

En el cas bivariant, una regressió lineal s'expressa com una relació entre dues variables quantitatives, en forma de **model estadístic bivariant**, de la manera següent:

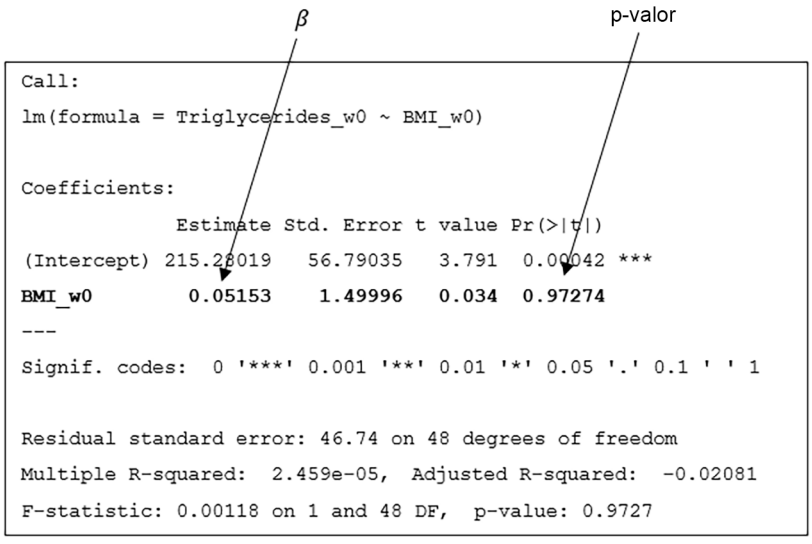
$$Y = \alpha + \beta X$$

on Y representa la variable resposta, α l'ordenada en l'origen (punt on la recta talla l'eix vertical, o d'ordenades o de les y), β és el pendent de la recta i X és la variable independent o predictiva. És important destacar que β s'entén com el major o menor efecte (pendent) que X produeix sobre Y. Concretament el seu valor numèric ha d'interpretar-se com el canvi d'unitats produït en Y en augmentar X en una unitat. A més, el seu signe + o – ens indicarà si en augmentar el valor de X augmenta el de Y (β positiu) o disminueix (β negatiu).

Per exemple, suposem que volem valorar l'efecte de l'IMC sobre el nivell de triglicèrids dels individus de la mostra a l'inici de l'estudi. L'expressió anterior pot entendre's d'aquesta manera:

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

El que ens interessa és saber el valor de β i si el canvi que indica β és significatiu o no mitjançant un test estadístic. El resultat del test, una mica més complex, però de fàcil interpretació, serà el següent:



En aquest cas, la línia del resultat que ens interessa és la marcada en negreta, que es correspon amb la variable resposta. Com pot deduir-se, el valor de β és molt petit i el p-valor indica que no existeix un efecte significatiu de *X* sobre *Y* (de l'IMC sobre el nivell de triglicèrids), ja que és ≥ 0,05.

5. Exploració de les relacions entre variables

En la majoria de les ocasions, el nostre estudi no comptarà amb una única variable predictiva el comportament de la qual vulguem observar per determinar com afecta la nostra variable resposta. Normalment en mesurem diverses, que poden o no estar relacionades.

> Seguint amb l'exemple proposat, suposem que volem estudiar com afecta la combinació de totes les variables predictives de la base de dades el nivell de triglicèrids (que serà nostra variable resposta). Afectarà d'igual manera el nivell de triglicèrids que s'hagi seguit una dieta o una altra, tenint en compte el sexe, l'edat, el nivell previ de glucosa de cada pacient, el de colesterol, el tipus de medicació que utilitza...?

Com pot intuir-se, hi ha combinacions inassumibles des del punt de vista bivariant. Hem d'observar, per tant, el comportament de totes les variables de manera conjunta, exercint el seu efecte sobre la variable resposta totes les predictives alhora. Així doncs, l'objectiu és observar un **patró de comportament**. A això ens solem referir amb «explicar la variabilitat de les dades».

Hem introduït alguns dels tests bàsics més comuns en l'anàlisi exploratòria de dades. Existeix un gran nombre de proves que poden utilitzar-se depenent de l'objectiu del nostre estudi i de la naturalesa categòrica o quantitativa de la variable resposta i de les variables predictives. A la taula 2 se'n descriuen algunes:

Taula 2. Tipus bàsics de test d'elecció per a la realització d'inferències estadístiques

Distribució	Variable predictiva	Variable resposta	Prova estadística d'elecció
Proves paramètriques	Categòrica	Quantitativa	Prova t-Student per a una mostra única
			Prova t-Student per a dues mostres
			1º Anàlisi de la variància de tres o més mostres (ANOVA) 2º Proves de rang *post* hoc i comparacions múltiples (Tukey si les variàncies són iguals; Games-Howell en cas contrari)
	Quantitativa	Quantitativa	Regressió lineal
	Quantitativa	Categòrica dicotòmica	Regressió logística
	Quantitativa	Categòrica politómica	Regressió logística multinomial

Distribució	Variable predictiva	Variable resposta	Prova estadística d'elecció
Proves no paramètriques	Categòrica	Quantitativa	Prova d'O de Mann-Whitney per a dues mostres
			Kruskal-Wallis para més de dues mostres
	Categòrica	Categòrica	Taules de contingència amb prova de χ^2 de Pearson
	Quantitativa	Quantitativa	Coeficient de correlació de Spearman

A continuació, ens detindrem en una tècnica molt comuna en salut pública ja vista: la regressió lineal.

5.1. Models de regressió lineal múltiple

Els fenòmens que afecten la salut solen tenir múltiples causes, per la qual cosa solem estudiar qualsevol paràmetre considerant diferents variables simultàniament.

Suposem que volem explicar el comportament que ha seguit el nivell de triglicèrids al final de l'estudi en funció de l'efecte de diversos factors observats, com l'edat, el sexe, l'IMC, el canvi o no en la medicació, el tipus de dieta seguida i el nivell de glucosa resultant. Direm que totes juntes, unes en presència de les altres, afecten la variable resposta en un grau (més o menys) i en un sentit o un altre (incrementant el valor de la variable resposta en incrementar el seu, o disminuint-lo). **L'efecte de cadascuna de les variables pot ser estadísticament significatiu o no**.

La relació lineal entre la variable resposta i aquelles variables predictives conformarà el **model estadístic múltiple**, que pot entendre's ara de la manera següent:

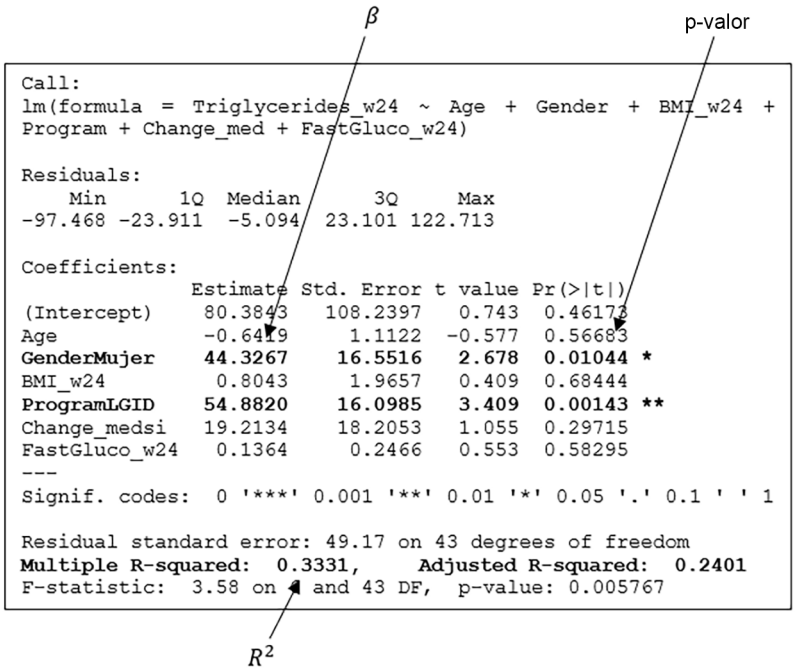
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_n X_n$$

essent n el nombre de variables predictives que vulguem incloure en el model estadístic.

Unes variables predictives tindran un efecte positiu (β positiu) i unes altres, negatiu (β negatiu) sobre la variable resposta Y. Entre totes conformaran el patró de comportament del nivell de triglicèrids a la mostra en funció d'aquestes variables. Com sol dir-se, explicaran la variabilitat de Y en cert grau (normalment mesurat com un percentatge de variabilitat explicat). Aquest percentatge pot ser elevat, en cas que les variables predictives incloses en el model justifiquin per si mateixes el comportament de Y en un alt grau. Per contra, aquest

percentatge pot ser menor, a causa que les variables predictives incloses en el model expliquen en menor mesura el comportament de *Y*, per la qual cosa existiran altres variables que no estiguin incloses en el model (i pot ser que ni en l'estudi) que afectin *Y* de manera rellevant. A aquest percentatge explicatiu se l'anomena **coeficient de determinació R^2**. Per tant, lògicament, R^2 augmentarà conforme augmentem el nombre de variables en el model.

Vegem el resultat de l'exemple anterior:



Com pot observar-se, el model explica el 33,31 % del comportament del nivell de triglicèrids, on el gènere femení i el tipus de dieta de baix índex glucèmic han actuat incrementant (β positiu) el nivell de manera estadísticament significativa (p-valors < 0,05). Dit d'una altra manera, les dones que van seguir aquest tipus de dieta durant l'estudi van incrementar el seu nivell de triglicèrids de manera estadísticament significativa.

Vegem un altre exemple més complex.

A partir d'unes dades obtingudes d'un estudi epidemiològic nutricional real, podem tractar de determinar de quina manera un conjunt de variables sociodemogràfiques i d'estil de vida (variables predictives) afecten el consum de proteïnes (variable resposta) en una mostra de 354 persones més grans de 65 anys. En concret, volem determinar què pot afectar aquest consum de proteïnes entre les variables sexe, edat, IMC, consum de líquids, àrea de residència (metropolitana, rural o mixta), nombre de patologies (comorbilitat) i nombre de fàrmacs (polimedicació).

Si realitzem la regressió lineal corresponent, obtindrem:

```
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 aquest cas es pot observar que el major consum de líquids i viure en un entorn rural millora el consum de proteïnes dels més grans de 65 anys en aquesta mostra. No obstant això, la selecció de variables predictives només explica el 7 % del comportament del consum de proteïnes, per la qual cosa caldria valorar la inclusió de més variables que puguin ser responsables de la resta de la variabilitat d'aquesta variable resposta.

Resum

Hem vist que el resultat satisfactori d'un estudi epidemiològic dependrà del correcte plantejament de la recerca i de l'adequada interpretació de les dades recollides.

Existeixen dues formes d'explorar les nostres dades, la numèrica i la gràfica, i ambdues ens serveixen per intuir patrons de comportament a la mostra que hauran d'avaluar-se mitjançant els tests estadístics adequats.

L'elecció dels tests estadístics dependrà de la naturalesa categòrica o quantitativa de la variable resposta i la/les variable/s predictiva/es, així com del compliment de certes condicions d'aplicabilitat.

La interpretació d'aquests tests ens permet indicar de manera científica que existeixen relacions entre variables de manera significativa, amb certa probabilitat d'equivocar-nos. Aquests resultats s'utilitzen per confirmar o no la hipòtesi inicial que plantegem al principi de l'estudi.

Bibliografia

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

Hernández-Aguado, I. i altres (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. València: Universitat de València.

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

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

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

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