In vitro veritas 2015;16:64–67
ISSN: 1697-5421

Continguts: www.acclc.cat/ivv_docs.php?any= 2015

In vitro veritas

Pàgina web de la revista: www.acclc.cat/ivv.php

Document docent

Aplicació de la correlació i la regressió en les ciències de laboratori clínic

Associació Catalana de Ciències de Laboratori Clínic
Secció d'Estadística i Metrologia [1]

Sílvia Miró Cañís [a], Natàlia Claver Belver [a], Beatriz Candás Estébanez [b]

[a] *UDIAT Centre Diagnòstic, Corporació Sanitària Parc Taulí, Sabadell*
[b] *Laboratori Clínic, Hospital Universitari de Bellvitge, L'Hospitalet de Llobregat*

[1] *Membres de la Secció d'Estadística i Metrologia durant la preparació d'aquest document: B. Candás Estébanez, X. Fuentes Arderiu (coordinador), M. Martínez Casademont, S. Miró Cañís, J.M. Queraltó Compañó, H. Valbuena Parralejo*

1. Introducció

Dins de l'àmbit de les ciències de laboratori clínic és freqüent estudiar el tipus d'associació que pot existir entre dues magnituds biològiques, entenent associació com la relació que existeix entre dues o més variables. Els estudis estadístics de regressió i correlació permeten descriure i definir aquesta associació. Malauradament, aquests termes es confonen provocant un mal ús de les eines estadístiques que descriuen.

L'objectiu d'aquest document docent és aclarir els conceptes de correlació i regressió, facilitar la seva interpretació i reflectir quina és la seva aplicació pràctica dins de l'àmbit de les ciències del laboratori clínic.

2. Correlació i regressió

En general, per establir l'associació entre dues o més variables s'utilitzen les proves estadístiques de correlació o de regressió.

La correlació identifica l'associació entre dues variables quantitatives independents. El seu estudi proporciona informació sobre la fortalesa d'aquesta associació o de la seva inexistència. S'acostuma a utilitzar en el context d'un estudi observacional retrospectiu entre dues variables independents aleatòries.

La regressió s'utilitza per establir quina és la l'equació matemàtica que millor defineix la relació entre dues o més variables (una variable dependent i una o més variables independents) mitjançant un model matemàtic i habitualment és emprada en el context d'un estudi experimental prospectiu (Taula 1).

	Correlació	Regressió
Objectiu de l'estudi	Fortalesa de l'associació lineal entre variables	Model matemàtic de relació entre variables
Àmbit d'aplicació	Retrospectiu	Prospectiu
Variables implicades	Dues variables independents	Una variable dependent i una o més independents
Eines matemàtiques	Coeficient de correlació	Funció matemàtica

Taula 1. Diferències principals entre correlació i regressió.

2.1. Correlació

La correlació lineal es defineix com el grau en què dues variables independents es troben associades linealment (1).

La correlació mesura la intensitat d'aquesta associació entre les variables mitjançant un nombre: el *coeficient de correlació*, simbolitzat per *r*. L'estadístic *r* permet quantificar el grau l'associació entre dues variables però en cap cas indica causalitat d'aquesta associació. És a dir, permet explicar si els canvis en una variable també es produeixen en l'altra variable i conèixer si es correlacionen directament o inversament (correlació positiva o negativa respectivament), o no es correlacionen, però en cap cas permet dir si els canvis en una variable són producte dels canvis en l'altra variable. Destaquem que en els estudis de correlació la hipòtesi nul·la és la inexistència d'associació entre les dues variables considerades.

http://www.acclc.cat/continguts/ivv185.pdf

La manera més simple de fer una primera aproximació a un estudi de correlació és fer un cop d'ull al *gràfic* o *diagrama de dispersió* corresponent (2), com la de l'exemple de la Figura 1. Quan existeix una bona correlació, els punts es disposen de forma estreta al voltant d'una línia del sistema cartesià que té el seu origen a prop de l'extrem inferior esquerre i va cap el superior dret (correlació positiva), o a la inversa (correlació negativa). Quan la disposició dels punts és aleatòria, la correlació és nul·la,

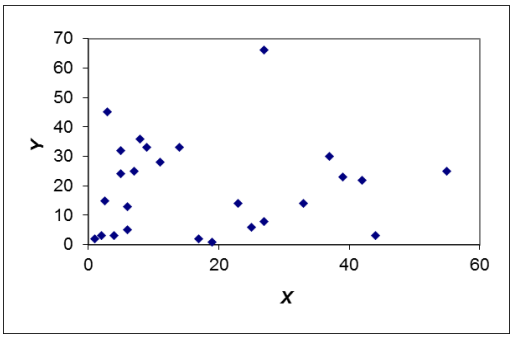

Figura 1. Exemple de gràfic de dispersió.

2.1.1. El coeficient de correlació

En un model de dues variables (bivariant) el coeficient de correlació mesura la intensitat amb la què les dues variables es troben associades linealment, té uns valors continguts dintre del interval [-1,+1] i és adimensional. No es troba afectat, per tant, per canvis en les unitats de les variables. Els valors aberrants d'alguna de les variables poden alterar considerablement el seu valor.

Si existeix una associació lineal i amb pendent positiva, el coeficient de correlació s'aproxima a +1. Si la correlació és positiva, sempre que una variable tingui un increment (o decrement) en el seu valor, l'altra variable tindrà la mateixa tendència. Si per el contrari, la correlació és negativa, els canvis en una variable es manifestaran de forma oposada en l'altra variable (si una augmenta, l'altra disminueix o viceversa). Per últim, si el coeficient de correlació és 0 significa que no existeix associació lineal entre les variables (3). A la Figura 2 es poden veure exemples de gràfics de dispersió on hi ha diferents valors del coeficient de correlació:

1. Coeficient correlació = 1, les variables estan perfectament associades linealment i els canvis en una variable reflectiran canvis del mateix tipus en l'altra variable.
2. Coeficient correlació = 0, les variables observades no estan associades linealment. Els canvis en una variable no es reflectiran linealment en l'altra, encara que puguin estar associades de forma no lineal.
3. Coeficient correlació > 0, la relació entre les dos variables és directa i augments en una variable (o disminucions) reflectiran augments (o disminucions) en l'altra.
4. Coeficient correlació < 0, l'associació entre les variables és inversa i augments en una variable reflectiran disminucions en l'altra.

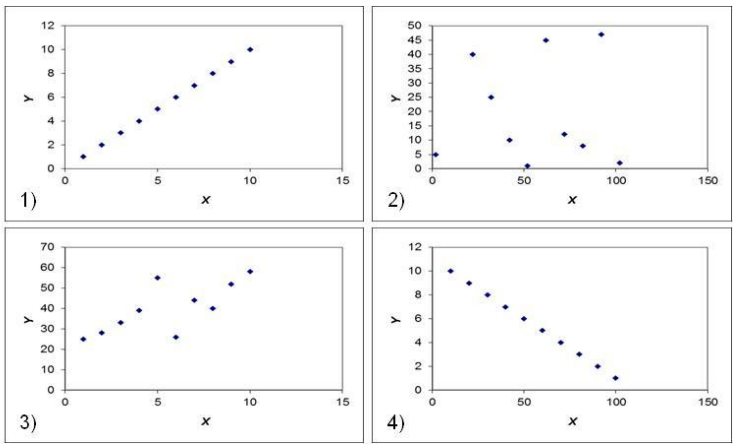

Figura 2. Exemples en gràfics de dispersió de valors correlacionats.

Les característiques a tenir en compte per a la correcta interpretació del coeficient de correlació són les següents:

1. El conjunt de dades, és a dir, la mostra poblacional, ha de ser representativa de la població que pretenem estudiar.
2. La mostra poblacional no ha d'incloure valors aberrants, ja que el coeficient correlació és excessivament sensible a la seva presència.
3. Las variables han d'estar associades linealment. Les relacions d'altre tipus (exponencials, etc.) generalment proporcionaran coeficients de correlació no significatius.

La distribució de freqüències (4) a la que s'ajusten els valors de les variables determinarà el coeficient de correlació a emprar. En el cas que les variables que volem estudiar s'ajustin a una distribució de freqüències de Laplace-Gauss, emprarem el *coeficient de correlació de Pearson*, i en el cas contrari, emprarem alternatives no paramètriques com el *coeficient de correlació de Spearman*, simbolitzat per ρ (5).

El *coeficient de correlació de Pearson* es pot definir com un coeficient que permet estudiar la relació lineal entre dues variables *X* i *Y*, els valors de les quals s'ajusten a una distribució de freqüències de Laplace-Gauss i es calcula matemàticament a partir del quocient entre la covariància S_{xy} i el producte de variàncies de les variables $S_x S_y$, mitjançant l'expressió següent (6):

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

La variància (S_x i S_y) és la mesura de dispersió estadística de cada una de les variables *X* o *Y* per separat. La covariància (S_{xy}) és la mesura de la variabilitat conjunta entre les dues variables *X* i *Y*.

$$S_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) \qquad S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Tant la variància com el coeficient de correlació estudien l'associació entre les variables, no obstant la primera depèn de la unitat de mesura, mentre que el segon, al estar normalitzat pel producte de variàncies, és independent de la unitat de mesura.

El *coeficient de correlació de Spearman* es pot definir com una mesura del grau de correlació entre dues variables (7) sense necessitat de que aquestes s'ajustin a una distribució de Laplace-Gauss. És a dir, és una prova no paramètrica basada en el coeficient de correlació entre els parells de nombres d'ordre

corresponents a cada dada dins del seu conjunt, tal i com veurem tot seguit.

L'expressió matemàtica del coeficient de correlació d'Spearman és:

$$\rho = 1 - \frac{6\sum d^2}{n(n-1)}$$

on d^2 és el quadrat de les diferències dels ordinals corresponents a *x* i *y*, i *n* és el nombre de parells de dades.

Per calcular ρ el primer que s'ha de fer és ordenar els valors de les variables *X* i *Y*, en ordre ascendent o descendent, donant el corresponent nombre d'ordre. En cas que alguna variable tingui valors repetits, farem la mitjana del nombre d'ordre i l'assignarem a ambdós valors (8).

Exemple: tenim dues variables *X* i *Y* que prenen els valors mostrats a la Taula 2. Assignem un nombre d'ordre als valors d'aquestes variables de manera creixent, tenint en compte que si la variable pren dues vegades el mateix valor, fem la mitjana dels seus nombres d'ordre (és el cas del valor 8 que és donat dues vegades en la variable *X*, com que correspondria al nombre d'ordre 3 i 4, fem la mitjana i a ambdós valors els assignem el nombre d'ordre 3,5).

Calculem la diferència entre els nombres d'ordre dels valors d' *X* i *Y*.

X	*Y*	Nombre d'ordre d'x_i	Nombre d'ordre d'y_i	Diferències
8	3	3,5	1	*2,5*
17	11	7	5,5	*1,5*
22	18	8	7	*1*
5	23	1	8	*-7*
16	10	6	4	*2*
8	11	3,5	5,5	*-2*
10	7	5	2	*3*
7	*10*	*2*	*3*	*-1*

Taula 2. Càlcul de les diferències per tal d'obtenir ρ.

Apliquem l'expressió matemàtica següent:

$$\rho = 1 - \frac{6\sum d^2}{n(n-1)}$$

ρ = 1- [6 [2,5^2+1,5^2+1^2+(-7)^2+2^2+(-2)^2+3^2+(-1)^2]/8 (64-1)]= 0,09

El coeficient de correlació ρ té uns valors continguts dintre del interval [-1,+1] i s'interpreta igual que el coeficient de correlació *r*. En canvi, el coeficient de correlació ρ, a diferència del coeficient de correlació *r*, és menys influenciable per la presència de valors aberrants, no requereix que les variables s'ajustin a cap distribució de freqüències concreta, ni que la relació entre elles sigui lineal, però requereix un nombre de dades més gran.

2.2. Regressió

La regressió expressa l'associació entre una variable dependent *Y* i una o més variables independents *X* a partir d'una funció matemàtica $Y = f(X)$ a la qual s'ajusten els valors de les variables. L'objectiu de l'estudi de regressió serà descriure l'associació entre ambdues variables, preveure el comportament de la variable dependent a partir dels canvis en les independents i valorar la contribució de cada variable independent sobre la variable dependent.

La funció matemàtica pot representar qualsevol tipus d'associació a partir d'un model matemàtic (lineal, parabòlic, exponencial) entre les dues variables (9). A més, segons el nombre de variables independents implicades tenim regressions simples o bivariants (una variable independent) o múltiples o multivariants (més d'una variable independent). La distribució de freqüències a la que s'ajustin les dades decidirà l'estudi de regressió a realitzar. Així quan les dades s'ajustin a una distribució de Laplace-Gauss es realitzarà un estudi de regressió paramètrica, i en cas contrari, no paramètric.

En aquest document docent ens centrarem en les regressions lineals simples paramètriques i no paramètriques.

2.2.1. El model lineal i la estimació de la recta mitjançant el mètode de mínims quadrats

El mètode dels mínims quadrats és el mètode més utilitzat per calcular el pendent i l'ordenada en l'origen de la millor funció matemàtica rectilínia que associa els valors que prenen les variables considerades (10), una independent *X* i una dependent *Y*. Per a cada valor x_i de la variable *X* s'observa un valor y_i per la variable dependent *Y* i s'obtenen parells de dades (x_1, y_1), (x_2,y_2), ... , (x_n,y_n). El conjunt de valors es pot representar gràficament tal i com mostra la Figura 3 i segueix l'equació de la recta:

$$y = bx + a$$

on *b* és el pendent i *a* és l'ordenada en origen de la recta.

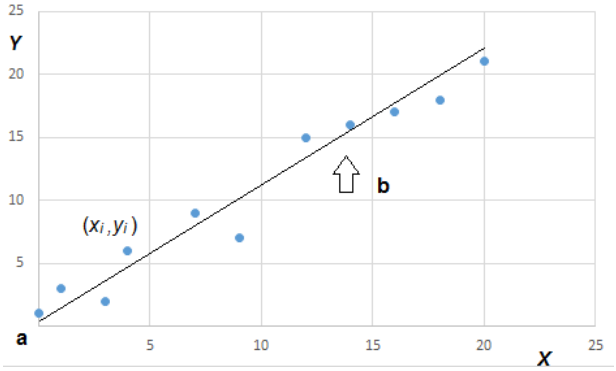

Figura 3. Recta de regressió $y = bx + a$.

A partir de les coordenades dels punts reals (x_i ,y_i) i de la mitjana dels valors de les nostres variables ($\bar{x}$ i $\bar{y}$), mitjançant les equacions que es mostren a continuació, calcularem el pendent *b* i l'ordenada en origen *a* de la recta de regressió lineal paramètrica $y = bx + a$.

$$b = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x} \right)}{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2} \qquad a = \bar{y} - b\bar{x}$$

2.2.2. Coeficient de determinació

El coeficient de determinació, simbolitzat per R^2, es defineix com la variació de la variable dependent *Y* que és explicada per la variable independent *X*, és a dir, indica la proporció de variació comú entre la variable dependent i una o més variables

independents. El coeficient de determinació mesura la bondat del ajust a la recta de regressió.

Els valors de R^2 estan inclosos dins l' interval [0,1], és a dir, el resultat —excepte el zero— sempre és positiu. Si pren el valor de 1, tota la variació de la variable dependent *Y* és explicada per la variació de la variable independent *X* i tots els parells de punts (x_i ,y_i) estan situats sobre la recta. Si, per exemple, pren un valor de 0,70 la variació de la variable *X* explica en un 70% la variació de *Y*, i el 30% restant és explicat per altres variables que no s'han tingut en compte en el nostre model. Si pren un valor de 0 vol dir que les dues variables no estan associades, almenys linealment.

2.2.3. Comparació de sistemes de mesura

La comparació entre els resultats obtinguts per diferents sistemes de mesura és un fet molt freqüent en el laboratori clínic. L'objectiu és esbrinar la relació matemàtica que existeix entre els resultats mesurats pel sistema de mesura objecte de la comparació i el sistema de mesura de referència establert (11).

2.2.3.1 Regressió lineal paramètrica de Deming

Aquesta regressió, també anomenada *regressió ortogonal* o *mètode del component principal*, permet comparar dos sistemes de mesura quan els valors de la variable s'ajusten a una distribució de freqüències de Laplace-Gauss (12) i existeix homogeneïtat de variàncies (homoscedasticitat) en l'interval de valors de les variables estudiades (13).

No obstant, tenint en compte que la majoria de procediments de mesura es comporten de manera heteroscedàstica i que per aplicar aquest mètode cal que el mínim de mostres clíniques en cada interval homoscedàstic sigui 50, en general la regressió ortogonal no és la més recomanada en la comparació de sistemes de mesura (8).

2.2.3.2 *Regressió lineal no paramètrica de Passing Bablok*

La majoria dels resultats de les magnituds biològiques mesurades al laboratori clínic no s'ajusten a una distribució de Laplace-Gauss, raó per la qual la regressió no paramètrica de Passing Bablok és una de les més emprades en les ciències de laboratori clínic (9). La seva principal aplicació és realitzar estudis d'intercanviabilitat o transferibilitat, és a dir, estudis de comparació de valors mesurats obtinguts amb dos sistemes de mesura diferents.

Els requisits per aplicar Passing Bablok en aquestes comparacions són (14):

1. Les variables a estudi s'han de distribuir de manera contínua i entre elles han de mantenir una associació de tipus lineal.
2. Són necessàries 100 o més mostres clíniques representatives de tot l'interval de mesura del sistema de mesura en estudi.
3. No s'afecta per l'heteroscedasticitat, és a dir, no es veu afectat pel fet que la desviació estàndard corresponent a la imprecisió interdiària depengui de la concentració.

Amb el mètode de Passing Bablok s'obté una recta on les estimacions dels paràmetres *b* (pendent) i *a* (ordenada en l'origen) estan acompanyades del seus intervals de confiança del 95 %. Per tal d'interpretar si els valors mesurats pels dos sistemes de mesura son intercanviables, cal fixar-se amb els dos intervals de confiança del 95 % estimats. Si l'interval de confiança de *b* inclou l'1 i l'interval de confiança de *a* inclou el 0, s'accepta que els dos sistemes de mesura són intercanviables, amb un risc α = 0,05. Si l'interval de confiança de *b* no inclou l'1, entre els dos sistemes hi ha un error proporcional, i no són intercanviables; si l'interval de confiança de *a* no inclou el 0, entre els dos sistemes hi ha un error constant i tampoc són intercanviables.

Exemple: comparació dels sistemes de mesura X i Y d'una magnitud biològica

$$y = 0{,}720\ (\ -0{,}308 - 1{,}748\ ) + 0{,}991\ (\ 0{,}958 - 1{,}024\ )\ x$$

Interpretació: donat que l'interval de confiança del 95 % de l'ordenada en origen inclou el 0 podem dir que no hi ha error constant. Atès que l'interval de confiança del 95 % del pendent inclou l'1 podem dir que tampoc hi ha error proporcional. Per tant, ambdós sistemes de mesura són intercanviables.

3. Bibliografia

(1) Surfstat.Australia. SurfStat *glossary*. <http://surfstat.anu.edu.au/surfstat-home/glossary/glossary.html> (Accés: 2012-11-25).

(2) Yule GU. On the theory of correlation for any number of variables, treated by a new system of notation. Proc R Soc SerA 1907;79:182-193.

(3) Easton BJ, McColl JH. Statistics glossary. <http://www.stats.gla.ac.uk/steps/glossary/paired_data.html#corrcoeff> (Accés: 2015-04-14).

(4) Miró Cañís S, Fuentes Arderiu X. Distribucions de freqüències i distribucions de probabilitats. *In vitro veritas* 2015;16:35-9 <http://www.acclc.cat/continguts/ivv178.pdf> (Accés: 2015-04-14).

(5) Moore David S. The basic practice of statistics. New York: W.H. Freeman; 2000.

(6) Rius Días F, Barón Lopez FJ, Sánchez Font E, Parras Guijosa L. Bioestadística. Métodos y aplicaciones. Universidad de Málaga. <http://www.bioestadistica.uma.es/baron/bioestadistica.pdf> (Accés: 2015-04-14).

(7) Viquipèdia. Coeficient de correlació de Spearman. <https://ca.wikipedia.org/wiki/Coeficient_de_correlació_de_Spearman> (Accés: 2015-04-14).

(8) Vargas Sabadías A. Estadística descriptiva e inferencial. Cuenca: Servicio de Publicaciones de la Universidad de Castilla-La Mancha; 1995.

(9) Fuentes Arderiu X, Castiñeras Lacambra MJ, Queraltó Compañó JM, dir. Bioquímica clínica y patología molecular. Barcelona: Reverté; 1998.

(10) Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method-comparison analysis. Clin Chem 1979;25:432-8.

(11) López Azorín F. La necesidad de mejores evaluaciones metodológicas y nuestra exigencia ante los criterios de aceptabilidad de los resultados. Quím Clín 2003; 22:431-2.

(12) Deming WE. Statistical adjustment of data. New York: Wiley; 1943

(13) González de Aledo Castillo JM, Arbiol Roca A. Estadístics i paràmetres usats en les ciències del laboratori clínic per a variables contínues. *In vitro veritas* 2013;14:42-8. <http://www.acclc.cat/continguts/ivv151.pdf> (Accés: 2015-04-14).

(14) Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part I. J Clin Chem Clin Biochem 1983;21:709-20.