In vitro veritas 2015;16:35–39

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

Distribucions de freqüències i distribucions de probabilitats

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

Sílvia Miró Cañís [a], Xavier Fuentes Arderiu [b]

[a] *UDIAT Centre Diagnòstic, Corporació Sanitària Parc Taulí, Sabadell*
[b] *Consultoria en Ciències de Laboratori Clínic, Barcelona*

[1]*Membres de la Secció d'Estadística i Metrologia durant la preparació d'aquest document: A. Blanco Font, B. Candás Estébanez, X. Fuentes Arderiu (coordinador), M. Martínez Casademont, M. Mosquera Parrado, J.M. Queraltó Compañó, R. Rigo Bonnin, H. Valbuena Parralejo*

1. Introducció

Una *variable* és un símbol (per exemple, *X*) que representa una propietat. Com és natural, cada valor possible (individual) de la propietat és un valor possible de la variable que la representa. Les variables poden ser controlades i aleatòries. Una *variable controlada* (també anomenada *variable dependent*) pot tenir només els valors que se li permeti tenir, mentre que una *variable aleatòria* (també anomenada *variable independent* o *variable estocàstica*) és una variable capaç d'adoptar qualsevol valor d'un conjunt particular de valors i que té associada una *distribució de probabilitat*.

Les variables aleatòries poden ser contínues o discretes. Una *variable contínua* és una variable aleatòria que pot adoptar tots els valors compresos en un interval finit o infinit; en les ciències de laboratori clínic aquestes variables prenen qualsevol valor dins d'un interval finit de valors numèrics reals (els valors compatibles amb la vida). D'altra banda, una *variable discreta* és una variable aleatòria que només pot adoptar valors aïllats; aquests valors poden ser nombres naturals, nombres ordinals, valors ordinals no numèrics, valors nominals, inclosos els nombres sense significat numèric o ordinal, les categories, els codis, etc (1).

El conjunt de valors d'una variable constitueix una *població* [*estadística*] i un subconjunt representatiu d'una població estadística és una *mostra* [*estadística*]. Les mostres estadístiques es poden dividir en diverses classes segons el nombre, *n*, d'elements que la formen: si $n < 10$ la mostra es considera *molt petita*, si $n < 30$ la mostra es considera *petita*, si $n > 30$ la mostra es considera *gran* i si $n > 100$ la mostra es considera *molt gran* (2). D'altra banda, la freqüència relativa amb què apareix un esdeveniment tendeix a estabilitzar-se cap a un valor fix, i degut a aquest fet, conegut com a *llei dels grans nombres*, es pot definir la probabilitat d'un esdeveniment (probabilitat d'obtenir un cert valor mesurat, per exemple) com el nombre cap el qual tendeix la freqüència relativa en repetir l'experiment o l'observació moltes vegades. És a dir, es poden fer equivalents la freqüència i la probabilitat.

Des d'un punt de vista experimental, els valors d'una propietat (d'una variable) tenen una relació empírica amb la freqüència amb què es produeixen; aquesta relació empírica s'anomena *distribució de freqüències* (3) i sol representar-se textualment per una *taula de freqüències*, o gràficament mitjançant un *histograma de freqüències* (3) (també anomenat *diagrama de freqüències*), un *polígon de freqüències* o una *corba de freqüències*. Un exemple d'això és la mesura de la concentració de substància de colesterol en el plasma de pacients que han patit un infart agut de miocardi. Els valors mesurats s'agrupen en classes, cadascuna de les quals té una certa freqüència; això dóna lloc a una distribució de freqüències que es pot representar gràficament com s'ha dit unes línies més amunt (Figura 1).

En els histogrames els valors solen agrupar-se en *classes* (intervals numèrics) més o menys amples; aquestes classes es caracteritzen pel seu *centre de classe* i els seus *límits de classe*. Això dóna lloc a un seguit de rectangles contigus, la base dels quals és un *interval de classe* i l'alçada la *freqüència* de valors pertanyents a cada classe. Si unim els centres de classe de la part superior de cada rectangle, obtindrem un *polígon de freqüències* (Figura 1), i si considerem que el nombre de valors és infinit (o molt gran), un polígon de freqüències es converteix en una *corba de freqüències* (Figura 2) que és la que correspondrà

http://www.acclc.cat/continguts/ivv178.pdf

aproximadament a una funció densitat de probabilitat, de la que en parlarem més endavant.

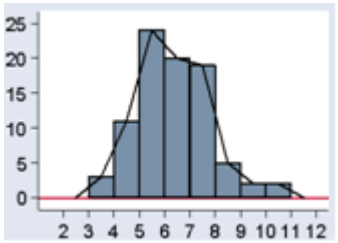

Figura 1. Histograma i polígon de freqüències relatives de la concentració de substància (mmol/L) de colesterol en el plasma de pacients que han patit un infart agut de miocardi.

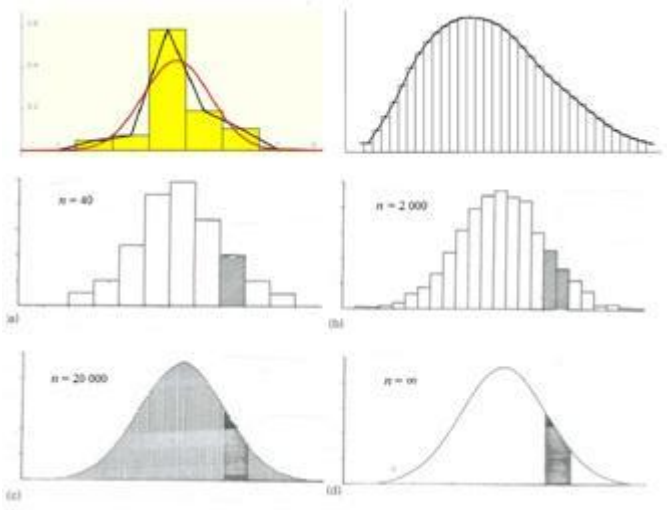

Figura 2. Histograma i corbes de polígon de freqüència.

Les corbes de freqüències poden ser *simètriques* respecte a un eix central que passa per la moda, o esbiaixades a la dreta (*biaix positiu*, com el de la Figura 1) o a l'esquerra (*biaix negatiu*) respecte a l'eix esmentat. Depenent del nombre de modes que tinguin poden ser *unimodals*, *bimodals* o *multimodals*.

2. Distribució de probabilitats

En l'exemple del colesterol exposat a la introducció, fent l'equivalència entre freqüència i probabilitats, es pot passar dels conceptes empírics de distribució freqüències i de corba de freqüències als conceptes teòrics de distribució de probabilitats i de funció de densitat de probabilitat, respectivament.

Una *distribució de probabilitats* (o *llei de probabilitat*) és un conjunt de parells (x, $f(x)$) associat a una variable aleatòria X (3). Si aquesta variable és discreta, $f(x) = P\,[X = x]$ (4). Si X és una variable contínua, $f(x)$ és la *funció de densitat de probabilitat*, que descriu la probabilitat relativa que X assoleixi un valor donat (4, 5). [En molts textos escrits en anglès la funció de densitat de probabilitat l'abrevia com PDF].

Una activitat empiricoteòrica és estudiar quina funció de densitat de probabilitat, que expressa una llei de probabilitats, millor s'aproxima a la corba de freqüències. Això és el cas de les proves estadístiques de normalitat («gaussianitat») que es troben a les aplicacions informàtiques d'estadística.

En els apartats que segueixen revisarem les principals distribucions de probabilitats (lleis de probabilitats) que per una raó o una altra tenen interès en les ciències de laboratori clínic.

3. Distribució de Poisson

La *distribució de Poisson* és la distribució de probabilitats duna variable aleatòria discreta, X, de tal manera que:

$$P[X = x] = \frac{m^x}{x!} e^{-m}$$

on m és l'esperança matemàtica, és a dir la mitjana, que en aquesta distribució és igual a la variància (3). Els valors de x són nombres naturals que tenen lloc amb freqüències petites, raó per la qual a aquestes distribucions també se les anomena *distribucions d'esdeveniments rars* (o *poc freqüents*).

Aquestes distribucions, en les ciències de laboratori clínic, tenen interès especial en citohematologia (6, 7) i en microbiologia (8-10).

4. Distribució de Laplace-Gauss

La *distribució de Laplace-Gauss* o *distribució normal*, $\mathcal{N}(\mu, \sigma)$, és la distribució de probabilitats d'una variable aleatòria contínua la funció de densitat de probabilitats de la qual és (Figura 3):

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

on x és un valor possible d'X, que pot anar de $-\infty$ a $+\infty$, μ és la mitjana i σ^2 la variància (3).

Aquesta funció es caracteritza, entre d'altres propietats, per originar una corba simètrica respecte a un eix central corresponent a la mitjana i que coincideix amb la mediana i la moda. D'altra banda, l'interval $[\mu \pm 1\sigma]$ conté el 68,28 % dels valors possibles, l'interval $[\mu \pm 2\sigma]$, el 95,46 % i l'interval $[\mu \pm 3\sigma]$, el 99,73 %. Per extensió, sempre que la mostra sigui molt gran ($n > 100$), això és aplicable als estadístics $\bar{x}$ i s.

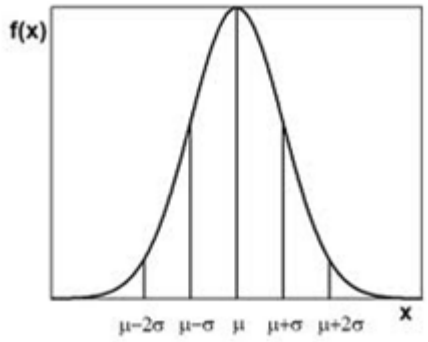

Figura 3. Funció de densitat de probabilitat de la distribució de Laplace-Gauss, on a l'eix de les abscisses es representen els valors de la variable i al de les ordenades la freqüència de cada valor.

D'altra banda, encara que una població d'una variable aleatòria contínua no segueixi la llei de Laplace-Gauss, si es prenen a l'atzar mostres grans ($n > 30$) d'aquesta població, les seves mitjanes sí que la segueixen. Aquest fet es coneix com el *teorema central del límit* o *teorema del límit central.*

Com ja hem vist anteriorment, les *proves de normalitat* o d'ajust a la distribució de Laplace-Gauss s'apliquen a una sèrie de valors que pren una variable (valors obtinguts en una mostra) amb l'objectiu de saber si la distribució d'aquesta variable en la població d'on procedeixen aquests valors s'ajusta o no a una distribució de Laplace-Gauss. S'han descrit diverses proves de normalitat, de les quals les més utilitzades en ciències de

laboratori clínic són les de Shapiro-Wilk i de Anderson-Darling, ambdues incloses en les aplicacions informàtiques d'estadística d'ús habitual (SPSS, Analyse-it, etc.). Les proves de normalitat no confirmen que una corba de freqüències obeeixi la llei de Laplace-Gauss, només indiquen que s'hi aproxima força.

Pot ser que una variable aleatòria contínua, *X*, no quedi representada per una distribució de Laplace-Gauss —fet freqüent en el cas de magnituds biològiques humanes (11)—, però si aquesta variable es transforma matemàticament d'una manera apropiada (12), llavors pot ser que s'ajusti a una distribució de Laplace-Gauss. Les transformacions més emprades són les següents:

- en cas que la corba de freqüències mostri una asimetria cap a la dreta (o positiva), es poden aplicar les transformacions $Y = \log(X + c)$ o $Y = \sqrt{(X + c)}$
- en cas que la corba de freqüències mostri una asimetria cap a l'esquerra (o negativa), es poden aplicar les transformacions $Y = 10^{(X + c)}$ o $Y = (X + c)^2$

on *c* és una constant que acostuma a ser 1. Cal no oblidar que un cop tractats els valors transformats, s'ha de fer la transformació inversa per retornar els valors a la seva escala de mesura original.

En les ciències de laboratori clínic, aquesta distribució de probabilitats és molt emprada en el control intern de la qualitat, en l'avaluació externa de la qualitat i en l'estimació dels intervals de referència biològics, entre d'altres.

Un cas particular de la distribució de Laplace-Gauss, utilitzada per al càlcul de probabilitats, és la *distribució normal estandarditzada* (o *tipificada*). És aquella que té una mitjana $\mu = 0$ i una desviació estàndard $\sigma = 1$ (Figura 4). Una variable *X* ajustada a una distribució de Laplace-Gauss $\mathcal{N}(\mu, \sigma)$ es pot transformar a una variable *Z* ajustada a una distribució normal estandarditzada $\mathcal{N}(0, 1)$ mitjançant la següent transformació:

$$Z = \frac{X - \mu}{\sigma}$$

Cal remarcar que aquesta transformació sovint no es pot dur a terme ja que habitualment no es disposa d'una bona estimació de la desviació estàndard poblacional.

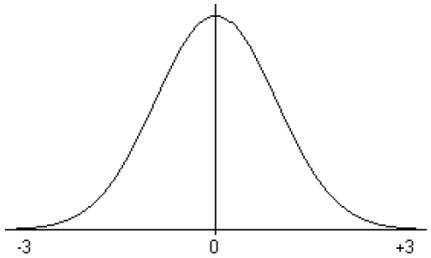

Figura 4. Distribució normal estandarditzada $\mathcal{N}(0, 1)$.

5. Distribució *F*

La *distribució F* (també anomenada *distribució F d'Snedecor*) és una distribució de probabilitats d'una variable aleatòria contínua la funció de densitat de probabilitats de la qual és (3):

$$f(F;\nu_1,\nu_2) = \frac{\Gamma((\nu_1+\nu_2)/2)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)}(\nu_1)^{\nu_1/2}(\nu_2)^{\nu_2/2}\frac{F^{(\nu_1/2)-1}}{(\nu_1 F+\nu_2)^{(\nu_1+\nu_2)/2}}$$

on $F \geq 0$, i ν_1 i $\nu_2 = 1, 2, 3, \ldots$ (paràmetres equivalents a l'estadístic *grau de llibertat* [2] en el cas d'una mostra) i Γ = funció gamma = $\Gamma(m)$:

$$\Gamma(m) = \int_0^\infty x^{m-1}e^{-x}dx$$

on *x* és un valor d'una variable aleatòria contínua *X* i *m* és un paràmetre que determina la forma de la distribució; si *m* és un nombre enter, llavors $\Gamma(m) = (m - 1)!$.

Gràficament es caracteritza per ser asimètrica i prendre només valors positius (Figura 5).

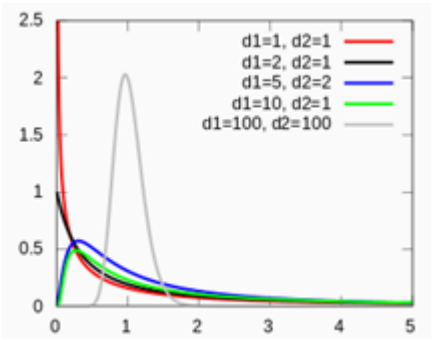

Figura 5. Funció de densitat de probabilitats de la distribució *F*. L'eix de les abscisses representa els valors possibles d'*F* i el de les ordenades la freqüència de cada valor. Els símbols d1 i d2 corresponen als dos graus de llibertat ν_1 i ν_2. S'observa que al augmentar els graus de llibertat la distribució és més propera a la distribució de Laplace-Gauss.

Aquesta distribució de probabilitats és la base de les proves paramètriques de comparació de variàncies i per a l'anàlisi de la variància, com es veurà en uns altres documents docents.

6. Distribució *t*

La *distribució t* o *distribució d'Student* (també anomenada *distribució d'Student-Fisher*) és una distribució de probabilitats d'una variable aleatòria contínua la funció de densitat de probabilitats és (3):

$$f(t,\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{n\pi}}(1+\frac{t^2}{\nu})^{-\frac{1}{2}(\nu+1)}$$

on $-\infty < t < +\infty$, $\nu = 1, 2, 3, \ldots$ (paràmetre equivalent a l'estadístic grau de llibertat en el cas d'una mostra) i Γ = funció gamma = $\Gamma(m)$:

$$\Gamma(m) = \int_0^\infty x^{m-1}e^{-x}dx$$

on *x* és un valor d'una variable aleatòria contínua *X* i *m* és un paràmetre que determina la forma de la distribució; si *m* és un nombre enter, llavors $\Gamma(m) = (m - 1)!$.

La funció densitat de probabilitats de la distribució *t* canvia segons el valor del paràmetre ν, aproximant-se a una distribució de Laplace-Gauss a mesura que ν s'apropa a infinit (Figura 6). En el cas d'una mostra, el paràmetre ν correspon als graus de

[2] En general, els graus de llibertat corresponen a la grandària de la mostra menys el nombre de paràmetres que s'han d'estimar a partir d'aquesta mostra.

llibertat, és a dir, a la grandària de la mostra menys el nombre de paràmetres que s'han d'estimar a partir d'aquesta mostra.

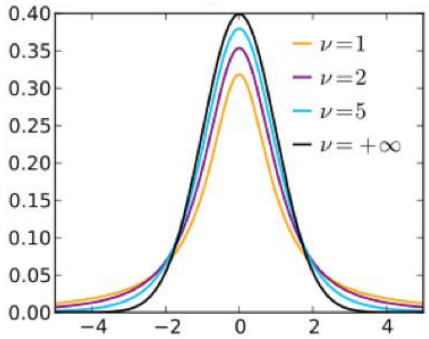

Figura 6. Funció de densitat de probabilitats de la distribució *t*. L'eix de les abscisses representa els valors possibles de *t* i el de les ordenades la freqüència de cada valor. S'observa que al augmentar els graus de llibertat, ν, la distribució és més propera a la distribució de Laplace-Gauss.

Aquesta distribució de probabilitats serveix per a l'estimació intervalar (estimació d'intervals de confiança) de les mitjanes i és la base de les proves paramètriques de comparació de mitjanes i de comparació de dues variàncies de valors aparellats, com es veurà en uns altres documents docents.

7. Distribució χ^2

La *distribució* χ^2 o *distribució chi-quadrat* és una distribució de probabilitats d'una variable aleatòria contínua la funció de densitat de probabilitats de la qual és (3):

$$f(\chi^2,\nu)=\frac{(\chi^2)^{(\frac{\nu}{2})-1}}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}e^{(\frac{-\chi^2}{2})}$$

on $\chi^2 \geq 0$, i ν_1 = 1, 2, 3, ... (paràmetre equivalent a l'estadístic grau de llibertat en el cas d'una mostra) i Γ = funció gamma = $\Gamma(m)$:

$$\Gamma(m)=\int_0^\infty x^{m-1}e^{-x}dx$$

on *x* és un valor d'una variable aleatòria contínua *X* i *m* és un paràmetre que determina la forma de la distribució; si *m* és un nombre enter, llavors $\Gamma(m) = (m-1)!$.

Gràficament la distribució χ^2 es caracteritza per ser asimètrica i amb valors ≥ 0 (Figura 7).

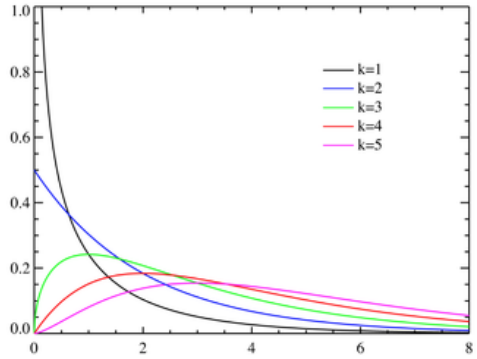

Figura 7. Funció de densitat de probabilitats de la distribució χ^2. L'eix de les abscisses representa els valors possibles de χ^2 i el de les ordenades la freqüència de cada valor. El símbol k correspon als graus de llibertat ν.

Aquesta distribució de probabilitats serveix per a l'estimació intervalar (estimar intervals de confiança) de les variàncies i és la base de les proves paramètriques de comparació de proporcions, amb valors independents o aparellats, proves d'independència entre dos valors nominals, o entre dos valors ordinals binaris, com es veurà en uns altres documents docents.

8. Distribució uniforme contínua

La *distribució uniforme contínua* o *distribució rectangular contínua* és una distribució de probabilitats d'una variable aleatòria contínua (Figura 8) la funció de densitat de probabilitats és constant dins un interval finit [*a*, *b*] i zero fora d'aquest interval (3). En aquesta distribució, qualsevol valor *x* de la variable aleatòria contínua *X* té la mateixa probabilitat d'ocórrer (és equiprobable). En aquesta distribució de probabilitats la desviació estàndard és igual a l'amplitud de la distribució [*a*, *b*] dividida per √12.

En ciències de laboratori clínic, aquesta distribució de probabilitats serveix per estimar la incertesa estàndard de mesura corresponent a la imprecisió interdiària (precisió en condicions intermèdies) d'un dispositiu volumètric (proveta, bureta, pipeta graduada, etc.), d'una balança, de les variacions de temperatura (cita) o de l'arrodoniment d'un número (13, 14).

9. Distribució triangular rectangular contínua

La *distribució triangular rectangular contínua* és una distribució de probabilitats d'una variable aleatòria contínua (Figura 8) en la que desviació estàndard és igual a l'amplitud de la distribució [*a*, *b*] dividida per √18.

En ciències de laboratori clínic, aquesta distribució de probabilitats serveix per estimar la incertesa estàndard de mesura corresponent a les magnituds influents (15) o a la inestabilitat (16).

10. Distribució triangular isòsceles contínua

La *distribució triangular isòsceles contínua* és una distribució de probabilitats d'una variable aleatòria contínua (Figura 8) en la que la desviació estàndard és igual a l'amplitud de la distribució [*a*, *b*] dividida per √24.

En ciències de laboratori clínic, aquesta distribució de probabilitats serveix per estimar la incertesa estàndard de mesura corresponent al calibratge del dispositiu volumètric emprat (13, 14).

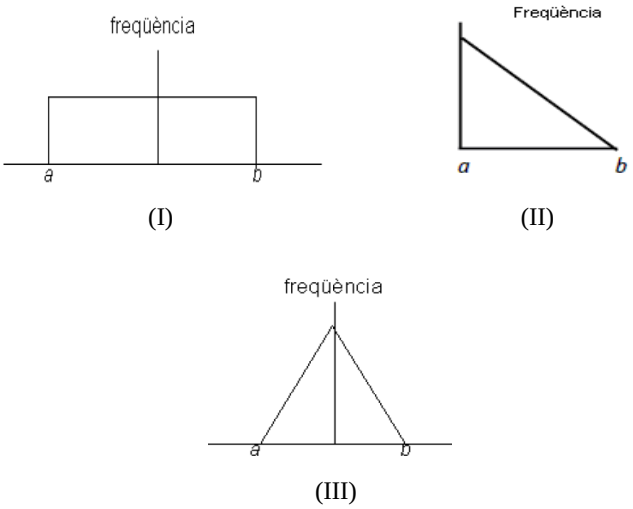

Figura 8. Distribucions (I) uniforme contínua, (II) triangular rectangular contínua i (III) triangular isòsceles contínua.

11. Bibliografia

(1) Fuentes Arderiu X. Estadística i ciències de laboratori clínic. *In vitro veritas* 2014;13:25-30. <http://www.acclc.cat/continguts/ivv148.pdf>. (Consultat: 2014-11-25).

(2) Holický M. Introduction to probability and statistics for engineers. Berlin Heidelberg: Springer-Verlag; 2013:30.

(3) International Organization for Standardization. Statistics — Vocabulary and symbols— Part 1: General statistical terms and terms used in probability. ISO 3534-1:2006. Geneva: ISO; 2006.

(4) Majó Roca J, dir. Diccionari de matemàtiques i estadística. Barcelona: Enciclopèdia Catalana, Universitat Politècnica de Catalunya; 2002.

(5) Wikipedia. Probability density function. <http://en.wikipedia.org/wiki/Probability_density_function>. (Consultat: 2014-11-25).

(6) Weisbrot IM. Poisson Distribution. A: Barnett RN. Clinical laboratory statistics. Boston: Little, Brown, and Co.; 1979; 30-3.

(7) Fuentes-Arderiu X, García-Panyella M, Dot-Bach D. Between-examiner reproductibility in manual differential leukocyte counting. Accred Qual Assur 2007;12:643-5.

(8) International Accreditation New Zealand. Uncertainty of measurement, precision and limits of detection in chemical and microbiological testing laboratories. Auckland: International Accreditation New Zealand; 2004. <http://www.ianz.govt.nz/resources/documents-2/technical-guides>. (Consultat: 2014-11-25).

(9) Niemelä SI (Advisory Commission for Metrology, Chemistry Section, Expert Group for Microbiology). Uncertainty of quantitative determinations derived by cultivation of microorganisms. Helsinki: Centre for Metrology and Accreditation; 2003. <http://www.mikes.fi/mikes/Oppaat/J4_2003.pdf>. (Consultat: 2014-11-25).

(10) Fuentes-Arderiu X. Uncertainty of measurement in clinical microbiology. eJIFCC 2002;13:<http://www.ifcc.org/ejifcc/vol13no4/130401006.htm>. (Consultat: 2014-11-25).

(11) Reed AH, Henry RJ, Mason WB. Influence of statistical method used on the resulting estimate of normal range. Clin Chem 1971;17:275-84.

(12) Osborne, Jason (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation. 2002;8: <http://pareonline.net/getvn.asp?v=8&n=6>. (Consultat: 2014-11-25).

(13) Perruchet C, Priel M. Estimación de la incertidumbre. Medidas y ensayos. Madrid: AENOR; 2000: 77-8.

(14) Associació Catalana de Ciències de Laboratori Clínic. Guia per estimar la incertesa de mesura. *In vitro veritas* 2001;2: <http://www.acclc.cat/continguts/ivv033.pdf>. (Consultat: 2014-11-25).

(15) Fuentes-Arderiu X. Influence quatities and uncertainty of measurement. Clin Chem 2001;47:1327-8.

(16) Fuentes-Arderiu X. Clinical sample stability and measurement uncertainty. Clin Chem Lab Med 2014;52:e37-8.