SOCIETAT CATALANA D'ESTADÍSTICA

MODA ESTADÍSTICA

Ús i aplicació de l'estadística

Exposició de treballs de divulgació en l'àmbit de l'ESTADÍSTICA, organitzada per la Societat Catalana d'Estadística amb l'objectiu d'acostar la feina dels estadístics al públic general.

L'exposició s'inaugurarà a la 25ª Setmana de la Ciència 2020.

http://soce.iec.cat/
@socestadistica #SoCE

SOCIETAT CATALANA D'ESTADÍSTICA

CONTINGUT

- Estadística aplicada a la Neurogenètica: El mètode d'Aleatorització Mendeliana
- Anàlisi i visualització de dades per a estudiar l'evolució de l'epidèmia de Covid-19
- Dades composicionals (CoDa): La importància dels valors relatius
- El paper de l'estadística en el mètode científic: aplicació en l'epidemiologia del càncer
- Les proves de diagnòstic in-vitro i l'estadística
- Factors associats amb l'ús dels recursos sanitaris de pacients amb insuficiència cardíaca crònica
- Composició corporal i salut en VIH+ / SIDA
- L'estadística als assajos clínics de COVID-19
- El món dels llibres ple de dades i estadística
- Estudi de l'herbivoria en quatre espècies de plantes: models amb excés de zeros

SOCIETAT CATALANA D'ESTADÍSTICA

Estadística aplicada a la Neurogenètica: El mètode d'Aleatorització Mendeliana

INTRODUCCIO

- Nombrosos factors de risc augmenten la probabilitat a desenvolupar la **malaltia d'Alzheimer** (MA)[1] [***Figura 1***].
- La **longitud dels telòmers** (LT) és un biomarcador d'envelliment biològic i de malalties relacionades amb l'edat com la MA[2] [***Figura 2***].

Figura 1. Factors de risc de la MA no modificables (edat, sexe, genètica) i modificables (tabac, depressió, tabac, contaminació, poc exercici, dieta).

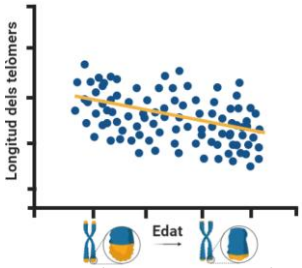


Figura 2. Relació de LT amb l'edat. Els telòmers s'escurcen a una velocitat similar en tots els teixits de el cos.

- Els **estudis observacionals no poden determinar** si la **relació** entre LT i la MA és **causal**, **o** si la LT és un marcador d'un **procés patològic subjacent**[3].
- Tanmateix, la **LT i la MA es veuen afectats per factors de risc comuns** relacionats amb l'estil de vida, que **poden induir la confusió i la causalitat inversa** en estudis observacionals[4].

OBJETIU

- L'objectiu és investigar el possible **paper causal de la LT** en la progressió de la MA des de la seva fase preclínica (sense símptomes), estudiada a través del **rendiment cognitiu**, la **vulnerabilitat cerebral** i **biomarcadors de la MA en el líquid cefalorraquidi (LCR)**.

MÈTODES

- **2,473 subjectes** de l'estudi **ALFA (Alzheimer i Families)**[5] van ser seleccionats per a aquest estudi. ALFA inclou una cohort de subjectes **cognitivament sans** amb un risc alt de desenvolupar MA.
- El mètode d'**Aleatorització Mendeliana (AM)**[6] utilitza variants genètiques associades a una exposició determinada, com a **variables instrumentals** per examinar l'**efecte causal** entre **l'exposició** i **la malatia o condicions d'interés**.
- Utilitzant 7 variants genètiques associades a la LT[7], obtindrem les estimacions de la **predisposició genètica a telòmers més curts** y calcularem els **efectes causals imparcials** d'aquesta associació [***Figura 3***].

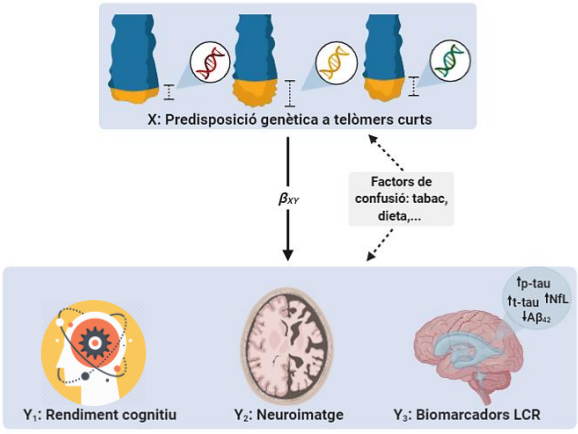


Figura 3. Model d'Aleatorització Mendeliana. L'efecte causal (β_{XY}) entre l'exposición (X) i cada condició d'interés (Y_i) es calcula mitjançant el mètode de variància inversa ponderada (IVW).

RESULTATS

- AM revela una **associació causal significativa** entre les variants genètiques que prediuen LT curta i un **baix rendiment cognitiu** en el domini de la **funció executiva**. No s'observen altres relacions causals significatives.

		Aleatorització Mendeliana – Mètode IVW		
Condició d'interés	**Tamany Mostral (N)**	**Efecte causal (β_{XY})**	**Desviació Estàndar (SE)**	**Significació Estadística (p-valor)**
Rendiment cogn. Memòria	2,233	-3.08	4.59	0.50
Rendiment cogn. Funció Executiva	2,233	-9.12	4.64	**< 0.05**
Rendiment cogn. Global MA Preclínic	2,233	-4.44	2.74	0.10
Neuroimatge	453	-10.09	55.33	0.85
Biomarcador t-tau	304	-0.26	32.22	0.99
Biomarcador p-tau	304	0.45	6.27	0.94
Biomarcador Amyloide	304	-3.59	6.00	0.55
Biomarcador NfL	304	6.15	11.55	0.59

Taula 1. Resultats de l'anàlisi de Aleatorització Mendeliana (AM) mitjançant el mètode de variància inversa ponderada (*IVW*).

CONCLUSIONS

- Es pot concloure un **potencial efecte causal del rol de la LT** en el **rendiment cognitiu** dels subjectes de l'estudi ALFA.
- **L'estadística ens ha permès modelitzar** el càlcul d'aquests effectes, i **valorar els resultats significatius**.

Referències: 1. Livingston, Lancet 2020; 2. Calado, N Eng J Med 2009; 3. Hägg, Transl Psychiatry 2017; 4 Lin, Mutat Res 2012; 5. Molinuevo, Alzheimer's Dement Transl Res Clin Interv. 2016; 6. Zhan, JAMA Neurol 2015; 7. Codd, Nat Genet 2013.

Blanca Rodríguez-Fernández[1], Natalia Vilor-Tejedor[1-3] , Marta Crous-Bou[1,4-5]

[1] *BarcelonaBeta Brain Research Center, Barcelona, Espanya.*
[2]*Center for Genomic Regulation, Barcelona.*
[3]*Erasmus University Medical Center, Rotterdam, Països Baixos.*
[4]*Catalan Institute of Oncology - Bellvitge Biomedical Research Institute.* [5]*Harvard TH Chan School of Public Health*

alzheimer's association
barcelonaβeta BRAIN RESEARCH CENTER
CRG Centre for Genomic Regulation
PERIS 2016 2020 Pla estratègic de recerca i innovació en salut
Erasmus MC
Agencia Estatal de Investigación

SOCIETAT CATALANA D'ESTADÍSTICA

Anàlisi i visualització de dades per a estudiar l'evolució de l'epidèmia de Covid-19

INTRODUCCIO

L'anàlisi de dades i la seva correcta visualització i interpretació resulta essencial per explorar i comunicar resultats en la recerca mèdica, especialment en vigilància epidemiològica. En el context actual i, davant la situació generada per l'epidèmia de SARS-CoV-2, resulta d'interès oferir a la comunitat científica, agents polítics i de salut pública i a la ciutadania en general, eines per a la correcta monitorització de la pandèmia. Hem desenvolupat dos eines online, COVID19-Tracker i COVID19-World que analitzen dades de Espanya i del tots els països del món respectivament. Les dades es llegeixen de forma instantània a partir de repositoris de dades oficials.

METODE ESTADISTIC QUE S'APLICA

Les anàlisis implementades empren principalment el model de regressió de Poisson. Aquest model exponencial permet avaluar la tendència de les diferents taxes d'interès (incidència, mortalitat, raons) de forma adequada, emprant polinomis de diferents graus. També s'han implementat eines per al càlcul de prediccions futures així com l'estimació del conegut nombre bàsic de reproducció o el temps d'infecció.

RESULTATS

Les eines desenvolupades, COVID19-Tracker (https://ubidi.shinyapps.io/covid19/) i COVID19-World (https://ubidi.shinyapps.io/covid19world i https://ubidi.shinyapps.io/covid19global) ofereixen una estructura amigable i intuïtiva que, mitjançant menús, permeten obtenir figures i taules personalitzades. Tots els menús estan disponibles en català, castellà i anglès, permetent la comparació entre Comunitats Autònomes, en el cas d'Espanya, i països, en el cas del món.

Fig. 1 Home page of the COVID19-World application, for comprehensive country-specific data visualization for the SARS-CoV-2 epidemic. Available at [https://ubidi.shinyapps.io/covid19world/]

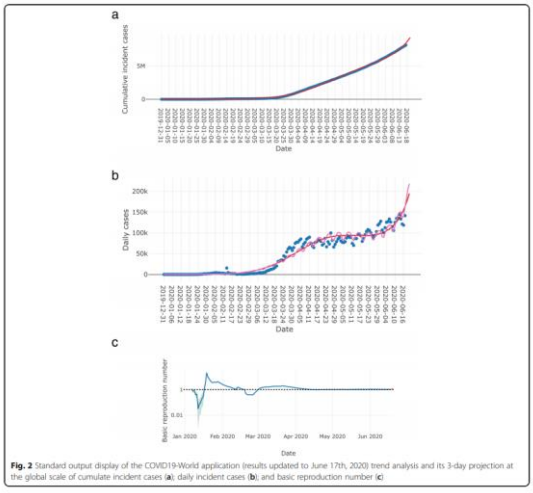


Fig. 2 Standard output display of the COVID19-World application (results updated to June 17th, 2020) trend analysis and its 3-day projection at the global scale of cumulate incident cases (**a**), daily incident cases (**b**), and basic reproduction number (**c**)

CONCLUSIONS

COVID19-Tracker i COVID19-World ofereixen un conjunt d'eines per a realitzar una anàlisi i visualització de dades adequada per al millor coneixement de l'evolució de la pandèmia de COVID-19 a Catalunya, a Espanya i al món. Tot i que les dades no estan exemptes de biaixos i limitacions per diferents motius que poden afectar la seva qualitat i interpretabilitat, és important analitzar-les de forma adequada mitjançant models teòrics que busquin aproximar-se a la realitat del problema i a la natura de les dades. De fet, la freqüència d'un esdeveniment és un mirall de la realitat i l'estadística la disciplina que pot ajudar a veure'n el reflex. Deixem doncs que les dades parlin

Referències:

Tebé C, Valls J, Satorra P, Tobías A. COVID19-world: a shiny application to perform comprehensive country-specific data visualization for SARS-CoV-2 epidemic. MC Med Res Methodol. 2020 Sep 21;20(1):235

Valls J, Tobías A, Satorra P, Tebé C. COVID19-Tracker: a shiny app to analise data on SARS-CoV-2 epidemic in Spain. Gac Sanit. 2020 Apr 27;S0213-9111(20)30085-6

Joan Valls[1], Aurelio Tobias [2] , Pau Satorra [3] i Cristian Tebé [3]

[1] Unitat de Bioestadística, Institut de Recerca Biomèdica de Lleida, Lleida, Catalunya

[2] Institut de Diagnòstic Ambientat I Estudis de l'Aigua, Consell Superior d'Investigacions Científiques, Barcelona, Catalunya

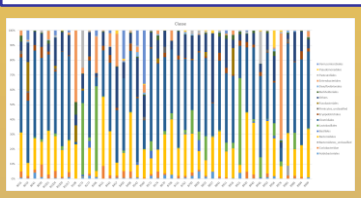
[3] Unitat de Bioestadística, Institut d'Investigació Biomèdica de Bellvitge, Barcelona, Catalunya

SOCIETAT CATALANA D'ESTADÍSTICA

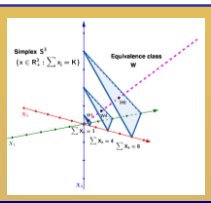
DADES COMPOSICIONALS (CoDa)
La importància dels valors relatius

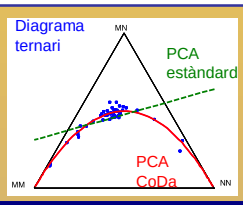
QUÈ SÓN LES DADES COMPOSICIONALS? PERQUÈ SÓN "DIFERENTS" ?

- **CoDa**: són dades multivariants que **representen parts o proporcions** respecte d'un total, per tant, només contenen informació relativa.
- **Alguns exemples**: les dades del microbioma, les dades electorals, les composicions químiques d'aliments, de mostres de terres, d'aire o d'aigües (ppm,...), la distribució del consum familiar, el temps diari dedicat a cadascuna de les activitats, i la despesa dels turistes.
- Tenen un **espai mostral diferent**: les dades composicionals són classes d'equivalència i viuen en el símplex, que amb 3 parts es representa en el diagrama ternari.
- Les **tècniques estadístiques estàndards no són adequades** per aquest tipus de dades atès que poden sorgir problemes d'interpretació i incoherències..



Estem interessats en valors relatius; la suma total no és informativa





DIFICULTATS EN L'ANÀLISI CoDa: CORRELACIONS ESPÚRIES I COHERÈNCIA SUBCOMPOSICIONAL

- Les correlacions calculades amb tècniques estàndards són **correlacions espúries** (no autèntiques) perquè les dades són relatives

•Quan reduïm la dimensió o seleccionem variables, les tècniques estàndards poden mostrar **incoherència subcomposicional**: si treballem amb totes les parts o amb un subconjunt d'aquestes trobem resultats contradictoris.

$\mathbf{x} = (x_1, \ldots, x_D) \quad \sum_{i=1}^{D} x_i = \kappa \quad cov(x_i, x_1) + \cdots + cov(x_i, x_D) = 0$

sample	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.2	0.3	0.3
3	0.3	0.3	0.1	0.3

Correlacions espúries de Pearson

corr	x_1	x_2	x_3	x_4
x_1	1.000	0.866	0.000	-0.866
x_2	0.866	1.000	-0.500	-0.500
x_3	0.000	-0.500	1.000	-0.500
x_4	-0.866	-0.500	-0.500	1.000

Exemple: dos científics A i B registren la composició de mostres de sòl: A registra composicions (animals, vegetals, minerals, aigua); B registra composicions (animals, vegetals, minerals) després d'assecar la mostra.

sample A	x_1	x_2	x_3	x_4
1	0.1	0.2	0.1	0.6
2	0.2	0.1	0.2	0.5
3	0.3	0.3	0.1	0.3

sample B	x_1^*	x_2^*	x_3^*
1	0.25	0.50	0.25
2	0.40	0.20	0.40
3	0.43	0.43	0.14

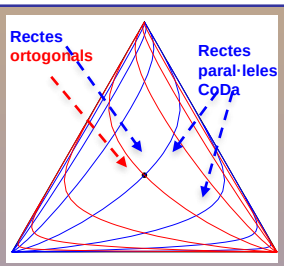
Correlacions incoherents de Pearson

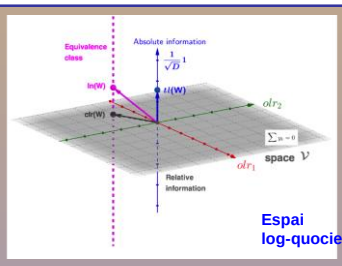
corr A	x_1	x_2	x_3	x_4
x_1	1.00	0.50	0.00	-0.98
x_2		1.00	-0.87	-0.65
x_3			1.00	0.19
x_4				1.00

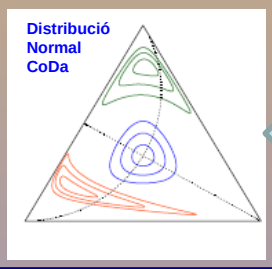
corr B	x_1^*	x_2^*	x_3^*
x_1^*	1.00	-0.57	-0.05
x_2^*		1.00	-0.79
x_3^*			1.00

GEOMETRIA D'AITCHISON: COORDENADES LOG-QUOCIENT

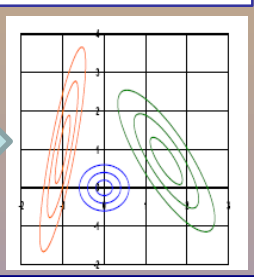
- **Nova geometria**: operacions vectorials, producte escalar, norma, i distància, elements necessaris en els mètodes de l'anàlisi multivariant, són definits d'acord amb la geometria de l'espai mostral.
- **Coordenades log-quocient:** el simplex és isomorf a l'espai log-quocient. Les CoDa es poden expressar en coordenades log-quocient (alr, clr, ilr). Les anàlisis estadístiques es poden realitzar en coordenades log-quocient ortonormals (**olr**)







OLR



EINES CoDa a l'abast desenvolupades pel CoDa-research group

- Les dades CoDa són diferents: necessiten tècniques estadístiques multivariants específiques per a la geometria del seu espai mostral.
- La representació de les dades CoDa en coordenades logquocient permet l'ús de les tècniques multivariants.

•Programari CoDa: CoDaPack i paquets de R (zCompositions, coda.base).
•Lloc Internet: *http://www.compositionaldata.com: notícies, material, activitats i cursos de formació.*

Referències: **Aitchison J (1982)** *The statistical analysis of compositional data*. J R Stat Soc Ser B 44(2):139–177
Barceló-Vidal C, Martín-Fernández JA (2016) *The mathematics of compositional analysis*. Aust J Stat 45:57–71
Mateu-Figueras, G. Pawlowsky-Glahn, V & Egozcue, JJ & (2013). *The normal distribution in some constrained sample spaces*. SORT (Statistics and Operations Research Transactions). 37(1). 29-56.

Marc Comas-Cufí, Pepus Daunis-i-Estadella, Glòria Mateu-Figueras,
Josep A. Martín-Fernández, Santiago Thió-Henestrosa, Marina Vives-Mestres
CoDa-research group: http://www.compositionaldata.com
Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona

CoDa Association

Universitat de Girona
Departament d'Informàtica, Matemàtica Aplicada i Estadística

SOCIETAT CATALANA D'ESTADÍSTICA

El paper de l'estadística en el mètode científic: aplicació en l'epidemiologia del càncer

INTRODUCCIÓ

El **rol** del **professional de l'estadística** en **l'àmbit de la salut**, com els que treballen dins del "Programa d'Epidemiologia i Recerca del Càncer" de l'ICO l'Hospitalet, per exemple, és dinàmic, interactiu, i sobretot molt enriquidor.

La seva **implicació** en els estudis duts a terme dins el Programa és **global**, amb una **participació activa** des del primer moment on es dissenya l'estudi fins a la redacció dels resultats i ajuda en la seva publicació.

A més, la **interacció amb altres investigadors** de l'ICO ofereix la possibilitat de participar en estudis ja començats, **assessorant** i aportant les eines necessàries per a dur-los a terme.

D'altra banda, la **formació continuada** és de vital importància pel bon desenvolupament de les seves tasques i totalment recomanable per a mantenir en un nivell adequat la **qualitat** de les tasques dutes a terme.

MÈTODE ESTADISTÍC QUE S'APLICA

Amb el **mètode científic**, basat en l'observació, es formulen hipòtesis i dissenyen estudis experimentals per extreure conclusions i elaborar teories. Tot el procés ha de garantir l'objectivitat, reproductibilitat i el principi de falsabilitat. L'aplicació de les tècniques estadístiques adequades garanteix el procediment.

Els **mètodes estadístics** més utilitzats en l'àmbit de la recerca mèdica són:

- Disseny:
  - Càlcul de la mida mostral de l'estudi
  - Formularis de recollida de dades
  - Generació i depuració de grans bases de dades
- Anàlisi descriptiva:
  - Taules de contingència
  - Mesures d'associació
  - Tests (paramètrics i no paramètrics)
  - Revisions sistemàtiques
- Models de regressió
  - Lineal
  - Logística
  - Supervivència (regressió de Cox)
- Altres més específics:
  - Models de cost-efectivitat
  - Calibratge de qüestionaris
  - Simulació

Els **softwares** disponibles habitualment en l'àmbit sanitari són:

- Gestors de bases de dades:
- Anàlisi:
- Entorns de programació:

EXEMPLE D'UN CAS D'ÈXIT

Figure 1: Anatomical sites included from each country. The number of cancer samples included at each anatomical site varies among country.

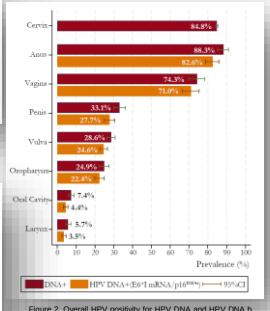


Figure 2: Overall HPV positivity for HPV DNA and HPV DNA þ (E6/I mRNA or p16INK4a) by anatomical site.
95% CI = 95% confidence interval (one-sided, 97.5%); CI calculated when appropriate; HPV = human papillomavirus. For more details, please see the methodological section from Castellsagué et al. 2016; de Sanjosé et al. 2010; de Sanjosé et al. 2013; Alemany et al. 2014, Alemany et al. 2015, and Alemany et al. 2016 (5,16–20).

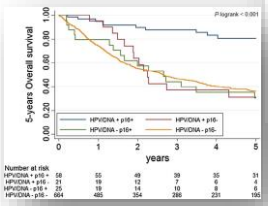


Figure 3. 5 years Overall Survival by HPV-DNA detection and p16INK4a high expression; Legend: Pairwise analyses showed that only patients double positive for HPV-DNA/p16INK4a had a statistically better OS compared to any other combination of those biomarkers. For more details, please see the methodological section from Mena et al. 2018

Quina relació hi ha entre els càncers anogenitals i de cap i coll amb la infecció pel Virus del Papil·loma Humà (VPH)?

a) Hipòtesi:
   a) La prevalença del VPH varia segons la localització del càncer.
   b) La supervivència dels càncer d'orofaringe varia segons la presència o no del VPH.

b) Disseny:
   a) Recollida de mostres i dades clíniques.
   b) Anàlisis centralitzat al laboratori amb estrictes controls de la contaminació.

c) Observació:
   a) Quantificació de la prevalença del VPH segons localització del càncer.
   b) Identificació dels factors determinants de la infecció per VPH.
   c) Estimació de la funció de supervivència i comparació de les corbes.

d) Conclusions:
   a) Hi ha un gradient en les contribucions del VPH: més alta en cèrvix i més baixa en laringe.
   b) La supervivència dels càncers d'orofaringe VPH-relacionats és superior que la dels VPH-no relacionats.

Mètodes estadístics

- Disseny de l'estudi Transversal / Cohort retrospectiva.
- Càlcul de la mida mostral necessària per a observar diferències entre localitzacions tumorals.
- Creació d'un Pla d'Anàlisi Estadístic (SAP)
- Creació de la base de dades.
- Comparació dels resultats obtinguts amb el laboratori de referència.
- Càlcul de les prevalences (determinació de la fórmula, definint quins són el numerador i el denominador en cada cas).
- Aplicació dels tests estadístics Chi-quadrat, F-Fisher, ANOVA, ...
- Ajustament de models de regressió logística incondicional (Odds ratios).
- Estimació de les corbes de Kaplan-Meier, aplicació de test Log-rank i ajustament de models de regressió de Cox (Hazard ratios).
- Redacció i interpretació dels resultats.
- Generació de gràfiques que ajudin a la interpretació dels resultats.

CONCLUSIONS

L'evidència científica ha d'anar sempre acompanyada de dades que la corroborin i la presa de decisions (en aquest cas en l'àmbit de la salut) s'ha de recolzar en ella.

Els estudis duts a terme al PREC han ajudat a determinar el paper etiològic del VPH en els càncers anogenitals i de cap i coll.

L'acompanyament del professional de l'estadística en tot el procés de recerca és essencial i garanteix els principis del mètode científic.

Sara Tous i Belmonte[1,2]

[1] Unitat d'Infeccions i Càncer - Molecular, Programa d'Epidemiologia en Recerca del Càncer. Institut Català d'Oncologia | Institut d'Investigació Biomèdica de Bellvitge, [2] Centro de Investigación Biomédica en Red: Epidemiología y Salud Publica (CIBERESP CB06/02/0073), Madrid, Spain

stous@iconcologia.net

SOCIETAT CATALANA D'ESTADÍSTICA

LES PROVES DE DIAGNÒSTIC *IN-VITRO* I L'ESTADÍSTICA

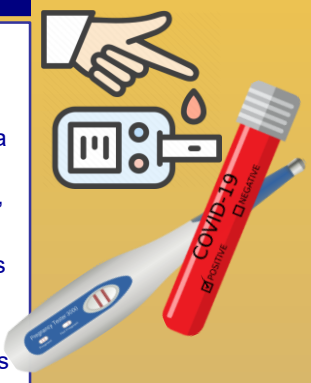
QUÈ SÓN LES PROVES DE DIAGNÒSTIC IN-VITRO?

Les proves de diagnòstic in-vitro són productes sanitaris utilitzats per a la determinació d'una malaltia o d'un cert estat de salut, i són generalment coneguts per l'acrònim "IVD", de l'anglès "*In-Vitro Diagnostics*".

La referencia llatina '*in-vitro*' significa literalment "dins del vidre", però en aquest context es refereix a fer un cert experiment a un tub d'assaig, és a dir, fora de l'organisme viu. En aquest cas el tipus d'experiment és per a l'estudi de mostres procedents del cos humà, incloses les donacions de sang, orina o teixits, només o principalment amb la finalitat de proporcionar informació.

Per exemple, són IVDs les proves d'embaràs d'ús domèstic, les proves d'hepatitis o de VIH, les tires analítiques d'orina o els sistemes de control de sucre en sang per a diabètics.

Els IVD es poden classificar segons la tecnologia que fan servir, alguns dels més coneguts són els "ELISA", els "Western Blot" o les "PCR", però hi ha altres tècniques automatitzades que es troben als grans laboratoris d'anàlisi.



MÈTODES ESTADISTICS NECESSÀRIS PER AL DESENVOLUPAMENT D'UNA PROVA DE DIAGNÒSTIC

GENERACIÓ DE CONCEPTE	Recerca bàsica	Generació, selecció i cribratge de matèries primeres i biomaterials necessaris. Estudis preliminars.	**Disseny d'experiments Proves d'hipòtesis**
	Viabilitat tècnica		
DESENVOLUPAMENT	Disseny	Optimització de paràmetres de l'assaig. Estudis per l'avaluació del disseny: especificacions que compleixin els requeriments, i productes que compleixin les especificacions Proves amb mostres clíniques.	**Límits de Detecció i Quantificació Sensibilitat i Especificitat Avaluació de la linealitat Comparació de mètodes (Gràfics de Bland-Altman, Regressió de Deming o de Passing-Bablok, ...)**
	Verificació + Validació		
	Avaluacions Clíniques		
PRODUCCIÓ	Fabricació	Transferència a les instal·lacions de producció. Avaluació de les capacitats garantint la producció del producte.	**Capacitat de Qualitat Mostreig d'acceptació Control estadístic de procés**
	Control de Qualitat		
LLANÇAMENT	Aprovació regulatòria	Compliment de les normatives i requisits previs al mercat. Avaluació del client.	**Rol d'expert especialista Suport en la presentació a agències regulatòries Integració amb dades del món real**
	Suport tècnic		

VOLEU SABER-NE MÉS?

COM ANALITZEM LES DADES?

Amb software d'anàlisi estadística:

R, Analyse-it, jmp, Minitab®

NORMATIVES O GUIES

Són documents informatius o de rigorós compliment que defineixen com i en quines condiciones cal fer els estudis o proves.

CLSI, ICH, ISO

AGÈNCIES REGULATÒRIES

Són organismes o autoritats públiques que asseguren que els productes disponibles als consumidors compleixen les normatives i són segurs.

FDA U.S. FOOD & DRUG ADMINISTRATION, CE, Health Canada

Susana Pérez Álvarez[1,2]

[1] *Biokit Research&Development SLU,*
[2]*Dpt. Estadística i Investigació Operativa, Universitat Politècnica de Catalunya – Barcelona Tech*

Biokit
A Werfen Company

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH
Departament d'Estadística i Investigació Operativa

SOCIETAT CATALANA D'ESTADÍSTICA

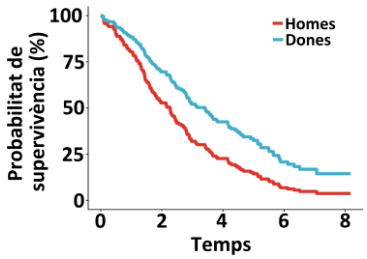
Factors associats amb l'ús dels recursos sanitaris de pacients amb insuficiència cardíaca crònica

INTRODUCCIÓ

Conèixer el temps fins a cert esdeveniment d'interès és fonamental per a una bona gestió en salut. Així, trobem aplicacions en el càlcul de la supervivència de pacients amb cert tipus de càncer, o el temps fins a una recurrència d'esdeveniment cardiovascular, i molts d'altres. Pots pensar-ne algun?

Per altra banda, la correcta gestió del sistema sanitari requereix de la justificació dels recursos que s'hi destinen. Així, inicialment cal saber quina és la despesa associada a diferents estats de salut.

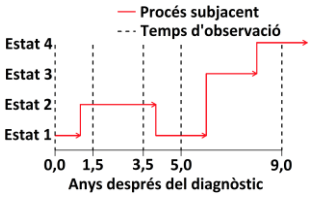
En l'estudi que presentem, es van analitzar dades poblacionals del Departament de Salut de la Generalitat de Catalunya per obtenir probabilitats de transició entre estats de salut de pacients amb insuficiència cardíaca crònica i la despesa sanitària associada.

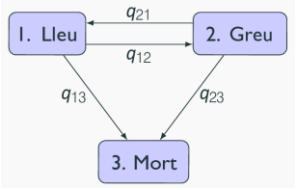


MÈTODE ESTADÍSTIC QUE S'APLICA

El conjunt de dades en estudi corresponia a 23.343 residents a Catalunya, de 50 anys o més d'edat, amb un primer diagnòstic d'insuficiència cardíaca crònica durant el 2011, i amb seguiment fins a març del 2018. Les dades de despesa sanitària estaven basades en l'ús de recursos en el sistema públic de salut.

Es van considerar estats de salut en base als **Grups de Morbiditat Ajustada** (GMA), que estratifiquen la població en quatre grups segons el seu risc de mort [1].





Per ajustar les probabilitats de transició es va fer servir un **model multi-estat** [2]. Aquest tipus de models permeten ajustar probabilitats de transició en situacions on les observacions es produeixen en determinats moments del temps. És a dir, no es disposa del moment exacte en què s'ha produït la transició o canvi d'estat.

Es va ajustar un model amb tres estats de salut: lleu (GMA 2-4), greu (GMA 1) i mort (estat absorbent).

Amb les transicions ajustades amb el model multi-estat, es va modelitzar la despesa sanitària associada a les transicions entre els estats de salut fent servir **models conjunts de dades longitudinals i temps fins a l'esdeveniment** [3]. La modelització conjunta de la despesa sanitària i la mortalitat permet eliminar els biaixos que es produirien modelitzant-les per separat, pel fet de no tractar adequadament que un pacient mort no consumeix recursos sanitaris.

Submodel longitudinal

$$y_i(t) = m_i(t) + \epsilon_i(t)$$

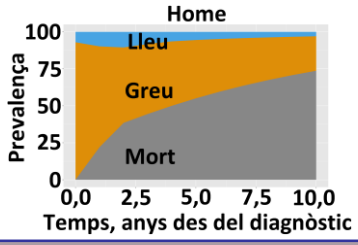
on $m_i(t)$ és funció de l'edat, el sexe i les probabilitats de transició obtingudes amb el model multi-estat anterior.

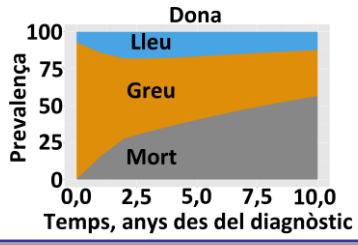
Submodel de supervivència

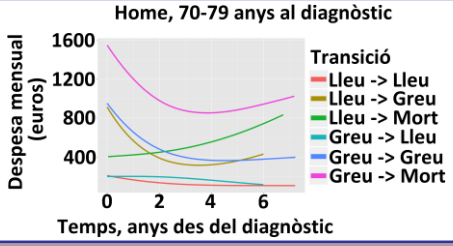
$$h_i(t) = h_0(t) \times \exp\left(\gamma_1 Edat + \gamma_2 Sexe + \cdots + \alpha\, m_i(t)\right)$$

El risc de mort es va modelitzar en termes de la funció de risc, segons el sexe, l'edat, diverses comorbiditats (diabetis, demència, etc.) i el valor de despesa sanitària que ens aporta el submodel longitudinal.

RESULTATS







CONCLUSIONS

- L'edat i la demència són els factors de risc més associats amb el risc de mort.
- La durada del temps en estat lleu s'associa positivament a la transició a l'estat greu. És a dir, com més temps en estat lleu, s'incrementa el risc d'empitjorar a estat greu.
- En canvi, el temps en estat greu s'associa negativament amb el risc de mort. Quelcom que es pot interpretar com "si no mors inicialment, et fas més fort"
- El model conjunt facilita estimacions de despesa sanitària associada a la morbiditat dels pacients, que suposa una valuosa informació per a la gestió del sistema sanitari i la presa de decisions.

Referències: **[1] Vela E, Clèries M, Vella VA, Adroher C, García-Altés A.** Análisis poblacional del gasto en servicios sanitarios en Cataluña (España): qué y quién consume más recursos? Gaceta Sanitaria. 2019;33:24–31; **[2] Jackson C.** Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011;38:1–28; **[3] Rizopoulos D.** Joint Models For Longitudinal and Time-to-Event Data with Applications in R. CRC Press, Biostatistics Series: Boca Raton, FL, 2012.

Montse Rué[1], Emili Vela[2], Montse Clèries[2], David Monterde[2], Carles Forné[1]

[1] Departament de Ciències Mèdiques Bàsiques, Universitat de Lleida, [2] Departament de Salut, Generalitat de Catalunya

Universitat de Lleida

Generalitat de Catalunya
Departament de Salut

SOCIETAT CATALANA D'ESTADÍSTICA

Composició corporal i salut en VIH+ / SIDA

INTRODUCCIO

Perquè una persona estigui saludable ha de tenir –entre altres coses- una bona composició corporal, és a dir, bon equilibri entre

- la quantitat d'os,
- greix i
- massa magra.

De fet, alteracions en algun dels components en sí mateix indiquen la presència de dolències com la osteoporosis (relacionat amb fractures òssies), lipodistròfia o lipoatròfia, o sarcopènia -baixa quantitat/qualitat de massa muscular-. Les alteracions poden estar relacionats entre elles o amb desequilibris metabòlics (diabetis, hipertensió arterial, dislipèmies, ...)

En aquest cas es va estudiar la composició corporal en pacients amb VIH+/SIDA ja que són una població més susceptible de patir aquestes malalties degut a:

- La inflamació crònica que provoca la infecció del VIH+/SIDA
- La toxicitat de prendre medicació antiretroviral per controlar la replicació del virus.

OBJECTIU: conèixer de manera objectiva la composició corporal dels pacients amb VIH/SIDA.

LES DADES

La densitometria DEXA, mitjançant l'absorciometria per raigs X d'energia dual està basada en la diferent atenuació dels diferents teixits: os, massa grassa i massa magra enfront dels fotons amb dos nivells d'energia.
Permet quantificar i diferenciar els diferents teixits basant-se en la seva densitat i contingut en minerals.
Amb un test DEXA obtenim mesures dels diferents teixits a diferents regions del cos.

Es va mirar la última DEXA de cada pacient de la nostra unitat, és a dir, és un estudi OBSERVACIONAL i TRANSVERSAL

- 1480 escàners DEXA realitzades entre 2000 a 2016
- 87 característiques o variables (DEXA i demogràfics)

Es va avaluar i millorar la qualitat de les dades:

- Descripció: exploració de les dades
- Validació: aplicar regles de control
- Verificació de les dades: comprovar que les dades són correctes

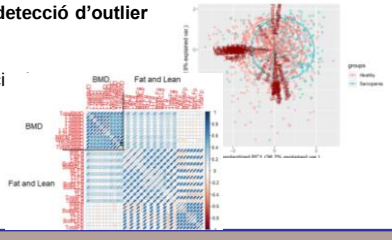
Detecció de valors extrems, valors perduts i revisió de les dades per corregir i completar els registres tant com sigui possible.

MÈTODES ESTADÍSTICS APLICATS

Vam recollir un volum de dades considerables.
Les variables que es van recollir estan bastant relacionades entre si, per exemple, si es mesuren les regions de la dreta del cos, el valor obtingut serà molt semblant a l'obtingut a la part esquerra del cos ->**Seleccionar les variables que tinguin informació "bona" (1)**
Per altra banda voliem **veure com es relacionaven les diferents variables**, p.ex.: si a més greix en una part també podem esperar que hi hagi més quantitat d'os o de massa magra (2).
Amb la informació agrupant per la presència d'una dolència (osteoporosis, lipodistròfia, lipoatròfia, sarcopenia, si/no) ens interessava classificar els pacients fent servir la informació d'altres teixits (3).

Es van aplicar mètodes d'aprenentatge automàtic (machine learning en angles), que poden ser supervisats i no supervisats per 1, 2; i 3.

Reducció de la dimensió i detecció d'outlier (punts 1 i 2)
Biplots
Anàlisis de components princi
Per veure la concordança: (punts 1 i 2)
Scatterplots
coeficients de correlació.
Classificació: (punt 3)
Boscos aleatoris



RESULTATS I CONCLUSIONS

Independentment de la regió del cos, el mateix teixit té un comportament similar
Mitjançant aquestes tècniques d'aprenentatge automàtic hem pogut veure una relació entre la massa magra i les distribucions d'os i greix.

Estem treballant en l'aplicació d'altres mètodes a aquestes dades.
Les aproximacions d'aprenentatge automàtic són molt útils per descriure les característiques de pacients que presenten una malaltia i per resumir grans volums d'informació.

Referències:

- Perez-Alvarez N, Vegas E, Estany C, Bonjoch A, Negredo E. Machine learning methods for assessing and predicting low muscle quantity and/or quality in HIV infected individuals. XVII Conferencia Española y VII Encuentro Iberoamericano de Biometría - CEB-EIB 2019. Valencia, 18 - 21 de junio de 2019.
- Royo Solé, D. "Supervised methods to classify body composition in HIV-infected patients". (2019).
- Perez-Alvarez N, Vegas E, Estany C, Bonjoch A, Negredo E. Machine learning methods for the prediction of abnormal fat and/or lean mass distribution in HIV infected individuals. XXIXth International Biometric Conference (IBC2018). 8-13 July 2018. Barcelona, Spain.
- Bonjoch A, Estany C, Pérez-Alvárez A, Rosales J, Echeverria P, Clotet B, Negredo E. Fat indexes that predict bone composition in HIV infected persons. European Workshop on Healthy Living with HIV. Barcelona, 7 - 8 setembre de 2018.

Nuria Perez-Alvarez[1,2], Dani Royo, Esteban Vegas[3], Carla Estany[1], Anna Bonjoch[1], Eugenia Negredo[1]

[1] Fight against AIDS Foundation, [2] Department of Statistics and Operations Research, Technical University of Catalonia-Barcelona Tech. , [3] Department of Statistics, University of Barcelona

FUNDACIÓ LLUITA CONTRA LA SIDA I LES MALALTIES INFECCIOSES · UNIVERSITAT DE BARCELONA · UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH · GRBIO

SOCIETAT CATALANA D'ESTADÍSTICA

L'estadística als assajos clínics de COVID-19

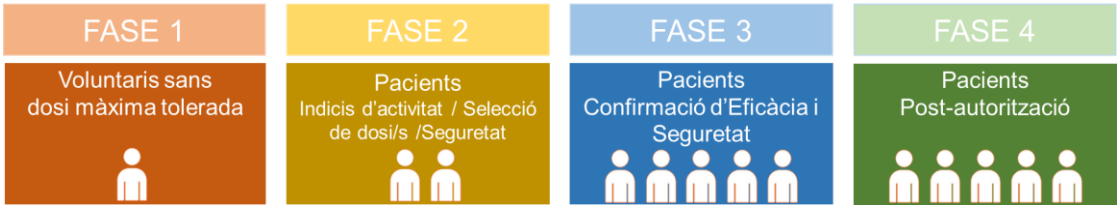
INTRODUCCIÓ

La síndrome respiratòria aguda greu coronavirus 2 (SARS-CoV-2) és l'agent causant de l'actual pandèmia de la malaltia del coronavirus 2019 (COVID-19). La pandèmia ha tingut un efecte arreu del món, sent Catalunya un dels territoris amb més afectació.

Per fer front al problema de salut públic que ha suposat la COVID-19, diferents organitzacions han invertit en la cerca de nous fàrmacs i vacunes que puguin fer front a la pandèmia. Per demostrar la seva efectivitat s'utilitzen els **assajos clínics**, i en ells l'**estadística** hi té un paper rellevant.

ESTRUCTURA DELS ASSAJOS CLÍNICS

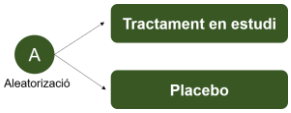
Els assatjos clínics s'acostumen a dividir en 4 fases:



Des de la fase 1 (estudiar si el fàrmac és segur en població sana per poder prosseguir l'avaluació en malalts, població més fràgil) fins a l'estudi de fase 3 (demostrar que el fàrmac és efectiu i segura) milers de pacients són inclosos als assajos clínics. L'objectiu de l'estadística és assegurar un bon disseny de l'estudi, un anàlisi òptim i una interpretació dels resultats l'estudi adient de manera que es puguin treure conclusions sobre els objectius de l'estudi. L'estadístic te una visió de projecte que aplica a cada un dels assajos.

ALEATORITZACIÓ

Per evitar biaixos, un dels principis que segueixen els assajos clínics confirmatoris és l'aleatorització. Els i les pacients reben el tractament o el placebo de manera aleatòria.



VARIABLE PRINCIPAL

Per evitar conclusions errònies, en particular reclamar un resultat positiu en una variable després de haver provat moltes variables, es defineix una/es **variables principals** que permetran concloure si l'objectiu principal de l'estudi s'ha assolit. L'exigència variarà si hi ha una variable o més, quan hi ha més d'una, s'ha de ser més exigent ja que provar per exemple dos variables incrementa la possibilitat que una sigui positiva respecte si només tinguéssim una variable principal.

Algunes de les variables utilitzades als estudis de COVID-19 han sigut: *% de persones infectades per COVID-19, % de mortalitat als 28 dies de la infecció, temps d'estança hospitalària, progressió de la malaltia, entre d'altres.*

ANÀLISIS INTERMEDIS

Existeixen diversos mètodes en el disseny de l'estudi, com per exemple els anàlisis intermedis que permeten fer adaptacions en el disseny d'un estudi un cop està en marxa i s'han obtingut dades parcials. Aquests anàlisis poden tenir com a finalitat poder parar l'estudi prematurament perquè: a) hi ha prou evidència de la seva eficàcia i seguretat, b) el fàrmac no és eficaç i exposar més pacients no canviarà el resultat, c) el fàrmac no té un bon perfil de seguretat; entre d'altres possibles adaptacions.

S'han de planificar i descriure en el protocol de l'estudi tenint en compte que '*xafardejar les dades té un preu*', és a dir, a més anàlisis intermedis més exigents es serà al final de l'estudi per mantenir el nivell de confiança global de l'estudi. És important escollir el mètode adient al nostre estudi i quan s'han de fer en el temps i quin número d'anàlisis intermedis són adequats. En l'avaluació de les vacunes per el COVID-19, els assajos han implementat anàlisis intermedis per poder prendre decisions abans degut a l'emergència sanitària.

CONCLUSIONS

- L'estadística té un paper rellevant en els equips responsables del desenvolupament clínic de fàrmacs.
- Els assajos clínics de COVID-19 tenen un repte específic degut a la pandèmia global, on es requereix una aproximació estadística complexa per poder donar una resposta tant aviat com sigui possible sobre el funcionament o no del fàrmac en termes d'eficàcia i seguretat.

Guillermo Villacampa[1] i Rosa Lamarca [2]

[1] Vall d'Hebron Institute of Oncology (VHIO), [2] AstraZeneca Barcelona

SOCIETAT CATALANA D'ESTADÍSTICA

El món dels llibres ple de dades i estadística

INTRODUCCIÓ

Entre tantes lletres, en les editorials disposem de moltes dades.

Hi ha les de referència dels llibres, que les anomenem metadades ja que són dades qualitatives. I les dades numèriques que es generen desde que es pren la decisió d'imprimir-lo, passant per les dades logístiques i de ventes.

Gràcies a les análisis, podem entendre quina tipologia de llibres es ven més, quan hem de reimprimir, i on hem de servir més ràpid els llibres.

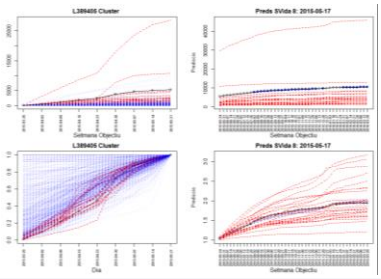
Però l'estadística no només intervé en la logística d'entrega, també s'utilitza desde la creació dels llibres per entendre les tendències i també el comportament dels lectors per saber per exemple tipologies de campanyes de marketing que funcionen i per quins mitjans segons el públic objectiu que volem impactar.

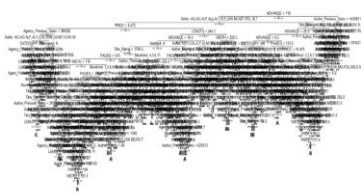
MÈTODE ESTADISTIC QUE S'APLICA

Abans de parlar de models, hem d'entendre que donat el gran volum de dades amb les que es treballa, hi ha processos de **validació** de dades molt complexos ja que la base conté tota la informació de logística, ventes i tipologies de llibres. El segon pas és seleccionar les dades necessàries per a respondre les preguntes d'interès, després s'han de trobar els models amb que s'analitzen les dades, i interpretar-los.

Hi ha una gran part d'anàlisi uni i bivariant de les dades, que ajuda a respondre les preguntes de negoci, ja s'analitza l'impacte en un moment concret del calendari; com pot ser Sant Jordi, Nadal...

Tot i així, és molt important estudiar les **sèries temporals**, ja que ens ajuden a predir el comportament i permeten anticipar la reimpressió de llibres. Aquests anàlisis permeten analitzar dades d'una mateixa variable recollides repetidament durant llargs períodes de temps, i pels quals intentem trobar **patrons** similars en base a diferents covariables i condicionants.





Per estudiar el conjunt de llibres, utilitzem **arbres de decisió** i altres metodologies de **clusterització**, mètodes que funcionen be amb grans bases de dades. Així aconseguim caracteritzar grups de llibres que s'assemblen entre ells i la informació ajuda a anticipar si caldrà fer reimpressions o no. Com també saber de quin catàleg es disposa, quins interessos tenen els lectors.

Tot i disposar de moltes dades, no sempre tenim les dades que responen la nostra qüestió de recerca disponibles. Així doncs, hem d'utilitzar tècniques de captació com pot ser el **webscrapping** per obtenir informació. Com per exemple els seguidors d'un autor/a o què s'està publicant en premsa

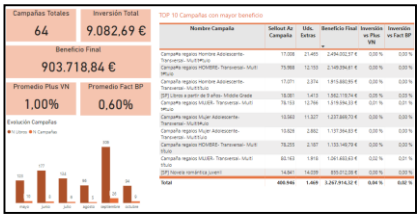
RESULTATS I CONCLUSIONS

Podem concloure que en el sector editorial, l'estadística té un paper molt important en tot el procés d'edició, des de la contractació fins a la producció i la logística.

En els diferents departaments trobem des d'estadística bàsica per caracteritzar, per exemple, el nombre de seguidors que té un autor fins a models més complicats per poder treballar en el departament de logística predint reimpressons i posades als mercat d'alguns títols.

El que sí tenen en comú tots els departaments, és que estan formats per equips multidisciplinars on els estadístics en són una part, i per tal de comunicar els resultats i ajudar a que les decisions es prenguin amb la millor informació possible, els estadístics hem de fer que la interpretació de les dades sigui entenedora per la resta de l'equip.

Per això, disposem d'eines de **visualització de dades**, que transformen els càlculs i mètodes estadístics, en visualitzacions que poden entendre la majoria de perfils. Cal triar bé la part de la informació que és rellevant i mostrar-la de manera que la lectura sigui ràpida i àgil; per poder prendre decisions de negoci amb el mínim temps possible.



Referències:

- Pedro A. Castilloa, Antonio M. Mora, Hossam Faris, J.J. Merelo, Pablo García-Sánchez, Antonio J. Fernández-Ares, Paloma De las Cuevas, María I. García-Arenas (2016) Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment
- BurcuYucesoy, XindiWang,JunmingHuang, andAlbert-László Barabási (2018) Success in books:a big data approach to bestsellers

Marta Rams Paradell

SOCIETAT CATALANA D'ESTADÍSTICA

Estudi de l'herbivoria en quatre espècies de plantes: models amb excés de zeros

INTRODUCCIO

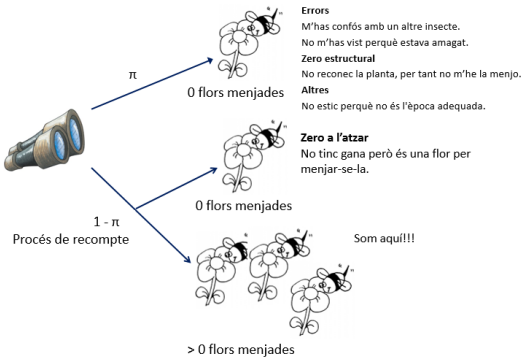
La investigació es centra en estudiar el **nivell de predació** que pateixen **quatre Espècies de plantes** de la família dels *Senecio* situades en la zona del **Montseny**. El terme predació en refereix al dany causat per un insecte que s'alimenta de les plantes. De les quatre varietats considerades, hi ha **dues de natives**, i **dues d'exòtiques** d'origen Sud-africà. Aquestes 4 espècies són d'especial interès ja que són plantes tòxiques que poden ocasionar problemes als animals que se les mengin. La predació pot ajudar a **frenar les invasions d'espècies vegetals**, sempre i quan els insectes de la zona envaïda reconeguin les espècies invasores com a hostes. **Objectiu: Comparar la predació d'aquestes quatre espècies**.

METODOLOGIA ESTADÍSTICA: MODEL ZERO-INFLAT

Des del mes d'abril fins al maig de 2009, es van seleccionar un total de 475 plantes de *Senecio* de les quatre varietats en estudi. Es va comptabilitzar el nombre total de flors produïdes per cada planta i el nombre total de flors danyades.
L'estadística descriptiva va permetre veure que hi havia **MOLTS ZEROS**, és a dir, **flors que no havien estat danyades**.
Per **explicar la predació** de les plantes quan existeix un gran nombre de zeros es fan servir els M**odels zero-inflats** que permeten diferenciar entre diferents fons de zeros (figura de la dreta).
Tenim dos processos barrejats:

- Procés que només genera zeros: són els zeros falsos
- Procés de recompte: general zeros verdaders i recomptes més grans a 0 que es corresponen amb les flors menjades.

Diferenciar el tipus de zero és molt important per saber perquè es mengen unes plantes i les altres no. Permet detectar si les espècies exòtiques esdevenen invasores per falta de predadors.



p: probabilitat de ser un zero fals
1-p: probabilitat de provenir del procés de recompte

RESULTATS

Model predació	ZIP	ZINB
Terme indepedent (β_0)	**-1.67 (0.05)**	**-2.25(0.21)**
Espècie (β_1)		
S. inaequidens	**-1.30 (0.06)**	-0.47(0.24)
S. lividus	**0.72 (0.056)**	**0.99 (0.23)**
S. Pterophorus	**-0.57 (0.07)**	0.12 (0.34)
S. vulgaris	0	0
Model zeros		
Terme indepedent (γ_0)	**1.29 (0.25)**	**0.78 (0.31)**
Espècie (γ_1)		
S. inaequidens	**-3.18 (0.46)**	**-3.74 (1.11)**
S. Lividus	**-2.18 (0.31)**	**-2.63 (0.52)**
S. Pterophorus	0.36 (0.39)	0.82 (0.43)
S. Vulgaris	0	0
Dispersió BN		0.87 (0.14)
% Zeros ($\hat{\pi}$)	78,40%	68,70%

Taula. Coeficients (error estàndard) del model estadístic. En negreta els coeficients estadísticament significatiu. *S. vulgaris* és l'espècie de referència.

Els models zero inflats de Poisson (ZIP) i Binomial Negatiu (ZINB) s'expressen com:

$$log(\mu_i) = \beta_0 + \beta_1 Espècie_i + log(Total\ Capítols)$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \gamma_0 + \gamma_1 Espècie_i$$

on μ_i i π_i són la mitjana de flors menjades i l'excés de zeros en l'espècie *i-éssima*, respectivament. Permeten detectar un 78,4% i un 68,7% d'excés de zeros respectivament.
Les espècies *S. vulgaris* (autòctona) i *S. pterophorus* (exòtica) presenten molts zeros totes dues. Les espècies *S. inaequidens* (exòtica) i *S. Lividus* (autòctona) presenten molt menys zeros que les dues anteriors.
Respecte a la predació, nombre de flors menjades, l'espècie *S. lividus* és la més danyada de totes.

Aquests resultats els extraiem de la taula del costat on valors positius afegeixen zeros (o predació) i valors negatius resten zeros (o predació).

CONCLUSIONS

El factor **Espècie** permet explicar les diferències observades en la predació. Les característiques pròpies de cadascuna de les espècies permeten trobar una explicació raonada a les diferències detectades en el percentatge de flors no menjades, l'excés de zeros. En el cas l'espècie autòctona *S. vulgaris* l'explicació de l'elevat volum de flors intactes es troba en la falta de sincronització entre la floració de la planta i l'època reproductiva de l'insecte. L'explicació de la manca de predació en l'espècie exòtica *S. pterophorus* prové del fet de tractar-se d'una espècie forana present en el Montseny des de fa poc temps i per tant encara no s'ha donat prou temps als insectes depredadors per a que la reconeguin com a aliment.

Referència**:** Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*(7), 949-959.

Anabel Blasco Moreno[1,3], Marta Pérez Casany[2], Pere Puig[3], Maria Morante[4], Eva Castells[4,5]

[1]Servei d'Estadística Aplicada, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain. [2]Departament d'Estadística i Investigació Operativa, Universitat Politècnica de Catalunya, Spain. [3]Departament de Matemàtiques, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain. [4]Departament de Farmacologia, Terapèutica i Toxicologia, Univ. Autònoma de Barcelona, Cerdanyola del Vallès 08193, Spain. [5]CREAF, Cerdanyola del Vallès 08193, Spain.

Servei d'Estadística
Universitat Autònoma de Barcelona

SOCIETAT CATALANA D'ESTADÍSTICA

Agraïments

Agraïm la voluntat i dedicació dels autors dels pòsters que formen part d'aquesta col·lecció, aquests/es socis/es de la nostra societat han volgut compartir un exemple de la seva tasca professional, que és molt diversa i variada i que ocupa moltes àrees de la recerca i la vida quotidiana.

Gràcies pel vostre interès i per visitar aquesta col·lecció virtual.