Models per a dades de recompte amb mesures repetides i errors de mesura

LLORENÇ BADIELLA BUSQUETS

Tesi Doctoral

Models per a dades de recompte amb mesures repetides i errors de mesura

LLORENÇ BADIELLA BUSQUETS

Director: PERE PUIG CASADO

Departament de Matemàtiques

Doctorat en Matemàtiques

2023

Universitat Autònoma
de Barcelona

I si un p-valor és molt petit, és insignificant?

Reflexions Estadístiques, Llorenç Badiella, 2022

Agraïments

Fa molts i molts anys, a la vila de Montblanc hi havia un drac ferotge d'alè pudent, la gent viva atemorida. Així comença una de les versions que tenim per casa de la llegenda de sant Jordi, un conte que he llegit infinitat de cops a la Berta abans d'anar a dormir. Diuen que escriure una tesi és un procés intens, exigent i solitari com si es tractés d'una lluita cruenta contra el drac ferotge de la llegenda, però també pot esdevenir una experiència agraïda, enriquidora i reconfortant.

Vaig iniciar els estudis de doctorat deu fer vint-i-cinc anys, així que en el meu cas, més que una batalla ha estat un combat amb múltiples assalts. Finalment, una estranya confluència de factors ha propiciat ara l'espai i l'energia suficient per a dedicar-hi els darrers esforços per culminar el projecte. A part de diferents elements exògens com la pandèmia de la COVID, la deriva incerta de la recerca a la Universitat, les múltiples crisis econòmiques, polítiques i climàtiques recents i la marxa de Messi del Barça, hi ha altres factors que han influït de forma més directa i, per tant, m'agradaria agrair explícitament:

Al Pere Puig com a director de la tesi, pel seu suport i dedicació al present treball, com a referent estadístic i com a guia vital, amb les seves particulars, constructives i sempre encertades contribucions.

Al Joan del Castillo, qui tant en l'etapa de professor, com en l'etapa de director del Servei d'Estadística, com ara en l'actualitat ha estat engrescant-me i orientant-me, sempre present al moment adequat amb un consell sota la màniga.

A la resta de coautors dels treballs que formen part de la tesi. Per un costat, voldria donar les gràcies al Martí Casals i al Carlos Lago, el treball sobre les targetes vermelles

ha estat especialment entretingut i agraït. Espero poder donar continuïtat a aquesta línia juntament amb vosaltres. Per altra banda, voldria agrair a l'Elena Santamariña, a la Katherine Pérez, a la María José López i al Josep Ferrando que em donessin l'oportunitat de sumar-me al projecte sobre els camins escolars i poder gaudir així de l'eficiència del seu grup de treball, espero tenir la possibilitat de continuar col·laborant-hi.

Als companys i amics del Servei d'Estadística (Anabel Blasco, Ester Boixadera, Oliver Valero, Ana Vázquez, Javier Martín) i a tots els qui en algun moment hi han format part (especialment al Joan Valls i a l'Anna Espinal), per haver-me ajudat a créixer professionalment. Gràcies a vosaltres, treballar al Servei d'Estadística és una aventura apassionant i un veritable luxe.

També a la multitud d'usuaris i col·laboradors que té el Servei d'Estadística, els reptes que ens plantegen dia rere dia requereixen aprendre constantment nous aspectes relacionats amb l'Estadística i amb les branques de coneixement implicades. En particular al Sergi Simon, al David Touzón, al Rafel Sala i a la Isabel Serra, emprenedors i amics, per la confiança dipositada i per haver-me engrescat en multitud de projectes apassionants.

De manera especial vull donar les gràcies al Joaquim Bruna per haver estat el principal instigador del Servei d'Estadística i per la seva insistència tant directa com indirecta en relació amb el desenvolupament de la tesi. Periòdicament teníem converses de passadís com aquesta que reprodueixo: *Badiella, com portes la tesi? Qui és el teu director? Doncs ja parlaré amb el Pere*.

Als amics i companys del Departament de Matemàtiques, especialment al Ramon Antoine i al Francesc Perera, autors de majestuosos treballs en C*-àlgebres i semigrups de Cuntz (Antoine et al., 2018, 2022) que han estat una veritable font d'inspiració, juntament amb els enriquidors debats a l'hora de dinar.

També vull agrair el suport de l'afició: als meus pares, Roser i Josep Maria, al meu germà Gil, a la Cata i a la Mari, pel seu entusiasme i ànims constants i també als amics de l'ànima, fans incondicionals.

I finalment els meus agraïments per a les veritables princeses d'aquest conte, la Maite i la Berta. La seva fe xamànica m'ha donat el coratge i la fortalesa per a fer realitat aquest somni.

Moltes gràcies a tots, va per tots vostès.

Índex

CAPÍTOL 1

Objectiu i resum

La distribució de Poisson, representa un punt de referència per a modelar dades de recompte, ja sigui en el cas d'observacions independents, amb mesures repetides o en presència de factors aleatoris. Però a la pràctica, en l'anàlisi d'aquest tipus de dades en dissenys experimentals més o menys complexos apareixen adversitats.

Per un costat, la distribució presenta la restricció que les dades ajustades han de ser equidisperses i sovint cal considerar distribucions més sofisticades. Per altra part, la naturalesa de les eines de modelització provoca dificultats per a comparar propostes alternatives, per a quantificar la bondat de l'ajust o per a validar les suposicions del model.

L'objectiu general d'aquesta tesi doctoral consisteix en descriure les estratègies principals per a l'anàlisi de dades de recompte amb mesures repetides i errors de mesura incidint en les seves limitacions operatives i, complementàriament, introduir noves propostes alternatives.

El capítol 2 està dedicat a presentar i revisar les principals propostes de modelització que s'utilitzen a la pràctica estadística amb dades independents o en presència de dades correlacionades: models lineals, models lineals generalitzats, models mixtos i models lineals generalitzats mixtos. Per a cada cas s'exposa la corresponent formulació, les situacions que permeten analitzar i detalls per al seu ajust, validació i aplicació de tasques

inferencials. Pel que fa als models lineals mixtos i als models lineals generalitzats mixtos s'emfatitzen dues visions de modelització contraposades: el model condicional i el model marginal.

En aquest mateix capítol també s'inclouen seccions específiques per als models amb dades de recompte, exposant les particularitats que en fan un cas d'especial interès.

El capítol 3 presenta un cas pràctic on s'usen models lineals generalitzats mixtos condicionals per a analitzar el recompte de sinistres en diferents cruïlles de la ciutat de Barcelona sota certa intervenció preventiva.

En el capítol 4 s'exposa un altre cas d'estudi on s'utilitzen models lineals generalitzats marginals per a analitzar l'impacte de les targetes vermelles en el nombre de gols marcats en diferents partits de futbol.

Al capítol 5 es presenta una nova proposta original per a la modelització de dades de recompte d'experiments amb subrèpliques, un cas especial de mesures repetides on les observacions són replicades sense modificar es condicions experimentals, només amb l'objectiu de controlar els errors de mesura.

El darrer capítol inclou valoracions i conclusions en relació amb cadascun dels articles, incloent algunes línes de treball futures i una valoració conjunta de tot el treball.

CAPÍTOL 2

Models estadístics

Segons va escriure Fisher (1922), els problemes estadístics són de dues classes:

- Descobrir quines quantitats són necessàries per a la descripció adequada d'una població, fet que inclou la consideració d'expressions matemàtiques per representar les distribucions de freqüència.
- Determinar quanta informació i de quin tipus respecte a tals valors poblacionals es reflecteixen en una o diverses mostres aleatòries.

Aquestes dues classes de situacions fan referència a la modelització estadística i a la inferència estadística basada en aquests models.

En general, els models estadístics són eines emprades per explicar o predir esdeveniments on intervé l'atzar a més d'altres causes conegudes i desconegudes. Els models estadístics permeten analitzar aquestes relacions a partir de l'estudi detallat de les dades d'una mostra d'observacions dels fenòmens d'interès.

Els models tenen en compte les particularitats de la variable objectiu, del disseny experimental associat amb l'obtenció de dades, dels efectes rellevants que poden influir en el resultat i de les fonts d'error potencials. Els models estadístics, a més incorporen certes suposicions associades a la incertesa del model.

L'estudi d'aquests models sovint es duu a terme a partir de la funció de versemblança de les dades, quantificant la densitat conjunta de tota la mostra d'observacions sota el model considerat. La funció de versemblança obre la via a l'ús de tècniques com el mètode de la raó de versemblances per a comparar propostes de modelització alternatives o el mètode de màxima versemblança per a trobar les estimacions òptimes dels coeficients del model. En cert sentit, aquest mètode permet trobar els coeficients per als quals la probabilitat d'haver obtingut les dades observades és màxima. Les tècniques d'inferència permeten prendre decisions per millorar la proposta, validar-ne les suposicions i quantificar-ne els efectes d'interès.

En la mesura que el model sigui més específic i detallista al mateix temps que parsimoniós, les anàlisis seran, per tant, més precises i vàlides. Aquesta és la tasca principal de la modelització estadística.

2.1 El model lineal

Els models lineals permeten avaluar relacions lineals entre una variable resposta d'interès quantitativa i un conjunt de variables explicatives i representen una generalització dels models de regressió i de l'anàlisi de la variància.

Tenint en compte que es disposa d'una mostra d'n unitats experimentals (individus) per als quals s'ha mesurat una variable resposta anomenada Y i un conjunt de variables explicatives $X_1, X_2, ..., X_p$, el model lineal per analitzar-ne les relacions es pot formular com:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \varepsilon_i$$

essent ε_i l'error del model per a l'individu i amb la propietat que $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, σ^2 és doncs la variabilitat residual, β_0 el terme independent i β_k cadascun dels coeficients associats a les p variables explicatives. Implícitament, el model assumeix que les observacions que constitueixen la variable resposta són mútuament independents, les variables explicatives són mesurades sense error i que el valor esperat de la variable resposta es

pot expressar mitjançant una combinació lineal de variables explicatives. Aquestes suposicions impliquen que la distribució de la variable resposta condicionada a les variables explicatives és normal.

És habitual simplificar la formulació anterior emprant una notació més compacta en forma matricial:

$$Y = X\beta + \varepsilon$$
$$\varepsilon \sim \mathrm{N}(0, I_n\sigma^2)$$
$$Y \sim \mathrm{N}(X\beta, I_n\sigma^2)$$

essent ara Y el vector amb les variables resposta, la matriu X, de dimensió $n \times p+1$ i anomenada matriu de disseny està formada per una columna d'uns i les p variables explicatives, els coeficients del model es representen amb el vector β de dimensió $p+1$ incloent ara el terme independent, mentre que els n termes d'error apareixen al vector ε. I_n és la matriu identitat de dimensió n.

El logaritme de la funció de versemblança per al model $Y \sim \mathrm{N}(X\beta, I_n\sigma^2)$ és:

$$\log(L(\beta; Y)) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sigma^{-2}(Y - X\beta)'(Y - X\beta)$$

El mètode de màxima versemblança per a estimar el vector β coincideix per al model lineal amb el mètode de mínims quadrats ordinaris i dona lloc a:

$$\widehat{\beta} \;=\; (X'X)^{-1}X'Y$$

El model es valida revisant que els efectes han estat especificats adequadament i que l'error del model segueix una distribució normal amb variabilitat constant.

Aquests models són emprats de forma molt generalitzada per a tractar experiments de naturalesa relativament senzilla. Quan els experiments són més sofisticats o bé quan els errors del model no són normals, caldrà dotar el model de major flexibilitat.

2.2 El model mixt

Un model lineal mixt és una generalització del model lineal, contemplant la possibilitat que existeixin observacions correlacionades o bé que posseeixin certa variabilitat hete-

rogènia. Des d'una altra perspectiva, es fa servir el terme model mixt quan el model conté simultàniament efectes fixos i efectes aleatoris.

Els models mixtos són l'eina habitual per a l'anàlisi de dades quantitatives en experiments longitudinals, dissenys creuats, multinivell, jeràrquics o amb dades espacials. Donat que aquests dissenys s'utilitzen en multitud d'àmbits, la nomenclatura emprada al voltant de la tècnica és força rica, tot i que de vegades ambigua:

Unitat experimental: Unitat objecte de les intervencions de l'investigador i de les mesures d'interès. S'assumeix que les unitats experimentals són seleccionades a l'atzar d'una determinada població. En un mateix experiment podrien existir diferents nivells d'unitats experimentals, sovint niuades, donant lloc a unitats i subunitats experimentals.

Factor fix: Una variable explicativa categòrica els nivells de la qual compleixen algun dels següents requisits:

- Els nivells experimentals són tots els possibles nivells observables.
- Els nivells experimentals han estat predeterminats per l'investigador.

En un assaig clínic serien factors fixos la variable Sexe o la Dosi Administrada.

Factor aleatori: Un factor, els nivells del qual són una mostra de possibles nivells. Un factor aleatori apareix per exemple, quan en un mateix experiment hi ha diferents nivells de selecció d'unitats experimentals. Les unitats superiors es corresponen amb els nivells del factor aleatori. Sota aquest plantejament, la variable resposta s'observa més d'una vegada per a cada nivell del factor aleatori. Per exemple:

- En un assaig clínic multicèntric, l'hospital és un factor aleatori i els pacients són subunitats.
- En l'àmbit veterinari, els corrals o el progenitor són factors aleatoris, els animals són subunitats.
- En estudis sociològics, la família o l'àrea d'estudi són en general considerats factors aleatoris, els individus serien subunitats.
- En ecologia, la parcel·la de terreny o el transsecte solen considerar-se factors aleatoris, els arbres dins la parcel·la o la posició dins del transsecte serien les subunitats.

- En estudis clínics longitudinals o en sondejos tipus panel, els participants constitueixen un factor aleatori, essent avaluats en diferents visites o onades.

Rèpliques: Terme utilitzat generalment en l'àmbit del disseny d'experiments per a designar les diferents mesures sota condicions experimentals idèntiques en unitats experimentals diferents. Per exemple, es parla de dissenys amb una sèrie de factors (que serien fixos) i un nombre de rèpliques, indicant que cada combinació de condicions experimentals s'ha avaluat en un nombre d'unitats diferents. Les diferents rèpliques són, per tant, independents. La variabilitat entre rèpliques rep el nom de variabilitat experimental o biològica (en biologia).

Mesures repetides: Terme que dessigna diferents mesures que es duen a terme en la mateixa unitat experimental, ja sigui en diferents subunitats, en períodes diferents, sota diferents condicions experimentals, en diferents ubicacions, al llarg del temps o inclús en components diferenciades. En alguns àmbits (sobretot en biologia) el terme pseudorèpliques s'empra per a identificar aquest tipus de medicions. Donat que les mesures repetides provenen del mateix individu, les mesures repetides estaran correlacionades. Sovint es diferencien les mesures repetides en funció de l'existència o no d'una jerarquia d'unitats i subunitats. En aquest sentit, es parla de mesures repetides jeràrquiques quan existeix una estructura de factors aleatoris que defineixen nivells experimentals, mentre que s'utilitza el terme de mesures repetides no jeràrquiques quan es disposa de dades multivariants per unitat experimental, per exemple, considerant diferents components avaluades en un mateix subjecte. Cal dir que a vegades no és immediat decidir si les mesures repetides tenen o no aquesta estructura jeràrquica.

Subrèpliques: Nomenclatura per identificar aquelles mesures repetides que consisteixen en la simple repetició de la medició en la mateixa unitat experimental sense variar cap altra condició. En aquest cas, les dades de diferents subrèpliques per a la mateixa medició tampoc són independents i es tractaria d'un cas particular de mesures repetides jeràrquiques. La variabilitat entre subrèpliques es coneix com a variabilitat residual, de l'error de mesura, instrumental o tècnica.

2.2.1 Efectes aleatoris

L'estratègia per a modelar dades quantitatives quan el disseny contempla una estructura jeràrquica d'unitats experimentals consisteix a habilitar la presència d'efectes aleatoris al model lineal, incorporant una o més variables categòriques els coeficients de les quals seran aleatoris. Per simplicitat, sovint s'assumeix que aquests efectes aleatoris segueixen una distribució normal.

La notació a nivell d'observacions per a un únic factor aleatori seria:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + ... + \beta_p X_{pij} + a_i + \varepsilon_{ij}$$

on i és l'identificador del corresponent nivell del factor aleatori (unitat experimental), j l'indicador de l'observació dins de cada nivell (subunitat), $a_i \sim \mathrm{N}(0, \sigma_A^2)$ és l'efecte aleatori associat al nivell i, σ_A^2 correspon a la variabilitat dels efectes aleatoris en qüestió, mentre que $\varepsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ és l'error residual per a cadascuna de les observacions i finalment σ^2 és la variabilitat residual.

És fàcil comprovar que aquesta proposta de modelització indueix correlacions entre les observacions del mateix nivell del factor aleatori:

$$\mathrm{Cor}(Y_{ij}, Y_{ij'}) = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2}$$

En general, com que poden existir diversos factors aleatoris, niats o creuats, és preferible emprar una notació matricial que proporciona una representació més compacta i que també permet diferenciar entre la part fixa i l'aleatòria:

$$Y \;\; = \;\; X\beta + Zu + \varepsilon$$

essent ara Z la matriu de disseny per als factors aleatoris i u el vector amb els efectes aleatoris associats a cada columna de Z. De fet, Z tindrà una dimensió de $n \times g$ on g és el nombre de columnes considerades a Z.

La variabilitat d'aquest vector u es denota habitualment per la matriu G, de forma que

$$u \sim \mathrm{N}(0, G)$$

Amb aquesta formulació és possible estudiar la distribució del vector Y. Suposant que es coneguessin els efectes aleatoris, és a dir, condicionant pel vector u s'obté:

$$Y|u \sim \mathrm{N}(X\beta + Zu, I_n\sigma^2)$$

Així doncs, la distribució del vector Y condicionada a conèixer els efectes aleatoris és un model lineal estàndard, on es tractarien les observacions com a dades independents.

Tenint en compte aquesta distribució condicionada i incorporant la informació sobre la normalitat dels efectes aleatoris, es deriva la distribució marginal d'Y:

$$Y \sim \mathrm{N}(X\beta, ZGZ' + I_n\sigma^2)$$

on la matriu de variàncies i covariàncies d'Y rep la notació genèrica de V i reflecteix les correlacions induïdes per la proposta de modelització a través de la inclusió de factors aleatoris.

2.2.2 Estructures de covariàncies

De vegades les mesures repetides no són de naturalesa jeràrquica i malgrat tot, la variabilitat d'Y encara té certa estructura, com per exemple en l'anàlisi de variables resposta multivariants (en un mateix individu es mesuren una sèrie de paràmetres diferents). També podria donar-se que fins i tot existint factors aleatoris, les corresponents mesures repetides mostressin certa estructura especial, com succeiria en el cas de mesures repetides al llarg del temps realitzades a intervals no equiespaiats.

D'aquesta manera, s'obté un model més general que l'anterior, on ara la distribució marginal es pot escriure com:

$$Y \sim \mathrm{N}(X\beta, V)$$

on V és la matriu de variàncies i covariàncies de la resposta, una matriu amb la propietat de ser semidefinida positiva, però que podria tenir estructures molt peculiars. Sovint és una matriu configurada per blocs, essent cada bloc una unitat experimental. Les propostes d'estructura més habituals inclouen les estructures anomenades de simetria composta (o intercanviable), autorregressiva d'ordre 1, Toeplitz i sense-estructura (o simètrica). Sota aquest plantejament, també és possible dotar d'heterogeneïtat la variància de l'error ε en funció dels criteris que calgui considerar.

L'estructura de V se sol dissenyar per a cadascun dels nivells que donen lloc a les mesures repetides, ja que s'assumeix que els diferents nivells ja són intrínsecament independents entre si.

2.2.3 Els models lineals mixtos marginals i condicionals

La distinció entre les visions anteriors dona lloc a dues propostes de modelització diferents:

Model mixt condicional (d'efectes aleatoris):

- S'assumeix que les correlacions de les dades venen induïdes per la presència d'un o més factors aleatoris.
- Condicionant als efectes aleatoris es tracta d'un model lineal.
- S'assumeix que els errors i els efectes aleatoris segueixen distribucions normals.
- A la pràctica, aquest tipus de visió se sol emprar per a dades amb estructura jeràrquica, multinivell o amb subrèpliques i també en models per a dades longitudinals.

Model mixt marginal (per a estructures de covariàncies):

- El model s'explicita directament a partir de la distribució marginal d'Y, on la seva matriu de variàncies i covariàncies té una estructura particular configurada per blocs i decidida per l'analista en funció dels criteris experimentals que consideri.
- Aquest tipus de plantejament sol ser aplicable a dissenys amb mesures repetides no jeràrquiques, per exemple quan es disposa de mesures multivariants en els mateixos individus, que defineixen els blocs de la matriu.

Cal apuntar que de fet, ambdues propostes es refereixen en realitat al mateix model, si bé la proposta del model condicional és aparentment més restrictiva. El model condicional porta implícita certa estructura de covariàncies marginal que apareix en considerar factors aleatoris. En canvi, el model marginal tindria més flexibilitat donat que assumeix directament una estructura particular per a aquesta matriu, podent capturar inclús components negatives o heterogeneïtat. En un treball no publicat de Badiella and Brewer (2015), es comprova que donada una proposta de model marginal, mitjançant la parametrització adient existeix una proposta de modelització amb factors aleatoris que dona lloc a la mateixa inferència. Aquesta relació implica que l'equivalència entre propostes és total.

Per a l'ajust i estimació de paràmetres d'aquests models, ja vinguin plantejats com a model condicional o com a model marginal s'utilitza la funció de log-versemblança marginal, obtinguda de forma exacta a partir de la distribució marginal.

$$\log(L(\beta;\alpha;Y)) - \frac{n}{2}\log|2\pi V(\alpha)| - \frac{1}{2}(Y - X\beta)'V^{-1}(\alpha)(Y - X\beta)$$

essent α el conjunt de paràmetres a estimar de la matriu V i β el conjunt de coeficients de la part fixa del model. Per a estimar els paràmetres continguts al vector α, s'utilitza el mètode de la màxima versemblança restringida (REML) que incorporar una correcció contemplant la pèrdua de graus de llibertat atribuïble al fet d'haver estimat els paràmetres β.

A la pràctica estadística és habitual dur a terme els mètodes següents per a la inferència dels paràmetres del model:

- Per als coeficients dels efectes fixos i covariables, el test de Wald, on els graus de llibertat s'estimen amb el mètode de Satterthwaite o el mètode de Kenward-Roger.
- Per a les components de l'estructura de covariàncies en propostes niades, el test de raó de versemblances.
- Per a comparar propostes de modelització alternatives, el criteri Akaike Information Criterion (AIC) o bé el Bayesian Information Criterion (BIC) com a mesures penalitzades de la funció de log-versemblança.

Per a validar les suposicions del model se sol comprovar que:

- El model d'efectes fixos i l'estructura de covariàncies han estat especificats adequadament.
- L'error del model i els efectes aleatoris segueixen distribucions normals amb l'estructura especificada.

És imprescindible notar que independentment de la naturalesa de les correlacions existents en les dades, la proposta de modelització i d'ajust és única. Ja sigui en presència de subrèpliques, factors aleatoris, dades multivariants o tot alhora, el model es podrà ajustar per màxima versemblança sota la mateixa proposta genèrica de modelització i es podrà comparar l'ajust de les diferents alternatives i consideracions.

Malauradament, i com es veurà en les següents seccions, aquesta propietat desapareix en els models per a dades no normals amb mesures repetides.

2.3 El model lineal generalitzat

El model lineal desenvolupat anteriorment contempla que la variable resposta es modelitza a partir d'una combinació lineal d'efectes i que els errors, segueixen una distribució normal. En conseqüència, la distribució de la variable objectiu, és també normal.

Moltes vegades, però, la variable resposta no compleix aquesta darrera propietat. Pot passar que es tracti d'una variable discreta, una taxa, que representi una proporció, etc. Quan la variable resposta no segueix una llei normal, apareixen diferents dificultats:

- Les relacions deixen de ser lineals.
- Els errors no solen ser homogenis i depenen de la magnitud de la variable resposta.
- La inferència emprant models lineals no és adequada, donat que les suposicions en què es basa no són vàlides.

La proposta de modelització emprada habitualment per a ajustar dades tenint en compte la distribució natural de la variable resposta rep el nom de model lineal generalitzat (GLM). Nelder and Wedderburn (1972) i McCullagh and Nelder (1983) van definir els GLM per a distribucions de probabilitat de la família exponencial, permetent tenir en compte la distribució de les dades per a modelar el seu valor esperat, en lloc de fer-ho a través de les observacions en si.

Aquesta família de distribucions inclou entre d'altres, les distribucions binomial, Poisson, binomial negativa, gamma, beta, inversa gaussiana i dins la qual també la distribució normal n'és un cas particular. Per a una variable aleatòria Y de la família de distribucions exponencial, la funció de densitat (o funció de probabilitats, en el cas de variables discretes) es pot escriure com:

$$f_Y(y, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right)$$

on ϕ representa un paràmetre d'escala conegut com a paràmetre de dispersió i θ és el paràmetre natural de la distribució, conegut també com a paràmetre canònic.

Sota aquest plantejament,

$$\mathrm{E}(Y) = b'(\theta)$$
$$\mathrm{Var}(Y) = b''(\theta)a(\phi)$$

La proposta contempla expressar la relació entre la variable resposta i les variables explicatives a partir d'un model per al seu valor esperat fent servir una determinada funció d'enllaç (una funció monòtona i diferenciable). El model expressat matricialment és:

$$\mathrm{E}(Y) = \mu$$
$$g(\mu) = X\beta$$
$$\mathrm{Var}(Y) = V(\mu)\phi$$

on Y és el vector de dades, μ és el vector per al valor esperat de la variable d'interès, $g()$ és la funció d'enllaç (*link*) que especifica la relació entre μ i les variables explicatives X, $V(\mu)$ és la funció variància associada a la distribució de les dades (específica per a cada distribució), mentre que ϕ és una constant associada a la dispersió no recollida per aquesta funció variància.

Per a les principals distribucions de treball, aquestes funcions són:

Taula 2.1: Model lineal generalitzat per a les principals distribucions

	Binomial/n	Poisson	Normal
$\theta(\mu)$	$\log(\mu(1-\mu))$	$\log(\mu)$	μ
$a(\phi)$	$1/n$	1	σ^2
$V(\mu)$	$\mu(1-\mu)$	μ	1
$\mathrm{Var}(Y)$	$\mu(1-\mu)/n$	μ	σ^2
Enllaç canònic	logit	log	identitat

Per a dades normals, l'enllaç canònic és la identitat, donant lloc al model lineal general clàssic, per a dades Poisson tenim el logaritme, donant lloc als models log-lineals i final-

ment, per a dades binàries, l'enllaç canònic és la funció *logit*, donant lloc a la regressió logística.

L'estimació dels paràmetres del vector β del model es duu a terme a partir de la maximització de la funció de versemblança mitjançant mètodes iteratius. La inferència en relació amb els paràmetres del model es pot dur a terme a partir del test de Wald, el test de raó de versemblances o també amb els criteris AIC o BIC. La validació del model es pot realitzar partir de la deviança, una generalització de la suma de quadrats de l'error dels models lineals. La deviança del model es calcula com:

$$D(\widehat{\mu}; y) = 2\left(l(y; y) - l(\widehat{\mu}; y)\right)$$

on $l(y; y)$ és la log-versemblança avaluada en $\mu = y$ (per tant, representa la versemblança màxima) i $l(\widehat{\mu}; y)$ és la log-versemblança sota el model en consideració.

Quan el paràmetre d'escala ϕ és conegut, cal calcular la deviança escalada:

$$D^*(\widehat{\mu}; y) = D(\widehat{\mu}; y)/\phi$$

que aproximadament segueix una distribució χ^2_{n-p}. D'aquesta manera, la deviança proporciona una prova de bondat d'ajust.

Quan el paràmetre d'escala és desconegut i es vol estimar-lo, es pot emprar la següent expressió:

$$\widehat{\phi} = D(\widehat{\mu}; y)/(n - p)$$

Un altre índex amb propietats similars a la deviança és l'estadístic χ^2 de Pearson, que donaria lloc a la clàssica prova de bondat d'ajust χ^2:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

essent i l'indicador d'observació. La versió escalada d'aquest estadístic és χ^2/ϕ.

La classe de models lineals generalitzats permet estendre la teoria i els mètodes del model lineal a variables resposta amb distribució no normal. Tanmateix, aquest model assumeix que les observacions són independents, de manera que no és una proposta adient per a dissenys que contemplin mesures repetides.

2.4 El model lineal generalitzat mixt

L'anàlisi mitjançant models lineals generalitzats mixtos (GLMM) obre noves alternatives de modelització i permet generalitzar el model lineal mixt (flexibilitzant els requisits distribucionals), al mateix temps que permet ampliar el model lineal generalitzat (incorporant efectes aleatoris i dotant-lo de flexibilitat en relació amb l'estructura de covariàncies).

De manera paral·lela als models lineals mixtos, es disposa també de dues visions per a especificar el model, basades en el model marginal o bé el model condicional.

Mentre que anteriorment existia una correspondència directa entre ambdues propostes, en fer servir altres famílies de distribucions juntament amb les respectives funcions d'enllaç, aquesta correspondència desapareix, esdevenint dues propostes diferents. Com a conseqüència, el plantejament emprat tindrà una important repercussió en la interpretació dels paràmetres del model, a banda d'afectar l'ajust i les estimacions realitzades.

2.4.1 El model GLMM condicional

La proposta sota el model GLMM condicional es distingeix per la incorporació d'efectes aleatoris dins del model per als valors esperats.

Aquest model emprant notació matricial és:

$$\begin{aligned} \mathrm{E}(Y|u) &= g^{-1}(X\beta + Zu) \\ u &\sim \mathrm{N}(0, G) \end{aligned}$$

Amb les propietats següents:

- S'assumeix que $Y|u$ segueix una distribució dins de la família exponencial.
- Condicional als efectes aleatoris es tracta d'un model lineal generalitzat les observacions del qual serien independents.
- La distribució dels efectes aleatoris u és normal, tot i que aquesta suposició es pot arribar a flexibilitzar.

- La introducció d'efectes aleatoris al model, indueix certa estructura de covariàncies a la distribució marginal.

L'estimació dels paràmetres en els models lineals generalitzats mixtos condicionals es basa en la maximització de la versemblança integrada (o marginal) construïda a partir de la funció de versemblança ampliada, que contempla les dades observades i els efectes aleatoris no observats.

$$L(\beta, \alpha|Y) = f_y(Y|\beta, \alpha) = \int_u f_{y|u}(Y|u, \beta, \alpha) f_u(u|\alpha) du$$

on $f_y(Y|\beta, \alpha)$ és la funció de densitat conjunta del vector de dades Y, $f_{y|u}(Y|u, \beta, \alpha)$ la funció de densitat condicionada d'Y donats els efectes aleatoris u, i $f_u(u|\alpha)$ la funció de densitat dels efectes aleatoris u. β representa el vector de coeficients del model i α el conjunt de paràmetres associats a l'estructura de covariàncies, en particular la variabilitat dels efectes aleatoris.

En el cas de dades normals, aquesta versemblança no plantejava excessius problemes donat que la distribució dels errors i dels efectes aleatoris és normal, i per tant també ho és la distribució marginal. Per als models amb dades no normals, però, la integració sovint és intractable.

Una solució habitual consisteix a aproximar l'integrant pel mètode de Laplace (Breslow and Clayton, 1993). Una altra alternativa és estimar la integral aplicant el mètode de quadratura de Gauss-Hermite (Pinheiro and Bates, 1995). El primer mètode és una mica més simple i alhora més flexible. No obstant això, cap de les dues alternatives proporciona resultats del tot satisfactoris si el disseny és una mica complex a causa de la dificultat de la integral en qüestió.

Donades aquestes limitacions en el tractament de la versemblança, sol ser més pràctic aproximar el model a través d'una linearització (en sèries de Taylor) de manera que sigui ara viable resoldre la integral (Wolfinger and O'Connell, 1993). El mètode de linearització genera pseudodades per a les quals s'ajusta un model lineal mixt i s'estima els paràmetres necessaris. El procés es repeteix fins a assolir convergència. Per aquest motiu es diu que l'estimació de paràmetres es duu a terme emprant pseudo (o quasi) versemblança (PQL). També cal dir que hi ha alternatives bayesianes per estimar el model, tot i que solen ser computacionalment molt intensives i no especialment flexibles.

Els models ajustats amb els mètodes de Laplace o de la quadratura de Gauss-Hermite, poden comparar-se amb les eines habituals basades en la versemblança: el test de raó de versemblances en casos niats i els índexs AIC o BIC altrament.

Per a models condicionals ajustats amb pseudoversemblança també és possible aplicar una versió adaptada del test de raó de versemblances. Una revisió més detallada d'aquests mètodes es pot trobar a Lee et al. (2018).

2.4.2 El model GLMM marginal

La visió marginal per a l'anàlisi de models lineals generalitzats mixtos pretén modelitzar el valor esperat de la resposta en funció de les covariables, deixant al marge l'estructura de covariàncies, que és tractada de forma separada.

En certa manera, aquesta proposta és una extensió natural del model lineal generalitzat per a dades independents exposat anteriorment, quan es vol tenir en compte la presència de correlació entre observacions dins d'un mateix bloc de dades.

El model marginal es concreta de la manera següent:

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ g(\mu) &= X\beta \\ \mathrm{Var}(Y) &= V(\mu)\phi \\ \mathrm{Corr}(Y_{ij}, Y_{ij'}) &= \alpha_{jj'} \\ \mathrm{Corr}(Y_{ij}, Y_{i'j'}) &= 0 \end{aligned}$$

essent i l'índex per als diferents blocs, i j l'índex per identificar les mesures dins de cada bloc.

Així doncs, és possible dotar d'estructura la matriu de variàncies i covariàncies, podent emprar les estructures presentades anteriorment: simetria composta, autorregressiva d'ordre 1, sense estructura, etc.

Aquest model es pot ajustar utilitzant el mètode PQL basat en la linearització del model i l'ajust de pseudodades mitjançant reiterats models mixtos. Si bé la proposta de modelització i ajust és viable, cal dir que malauradament aquesta proposta no fa explícit

el model de dades subjacent (no es disposa explícitament d'un model que generi dades amb les propietats desitjades) i, per tant, tampoc es disposa d'una funció de versemblança legítima.

Això provoca que no sigui possible disposar de les eines adequades per a la validació i comparació genèrica de models. Únicament serà possible comparar models que hagin estat ajustats amb les mateixes pseudodades fent servir la pseudoversemblança d'aquestes dades particulars. Dit d'una altra manera, només es podran comparar propostes d'estructures de covariàncies marginals entre si.

2.5 Model GLMM marginal vs. model GLMM condicional

Les dues visions per a ajustar els models GLMM porten implícites interpretacions diferenciades.

En el model condicional (en presència de factors aleatoris):

- Els coeficients representen l'efecte de les variables explicatives a cada nivell (unitat experimental).
- La correlació entre observacions va lligada al seu valor esperat donada la naturalesa de la distribució emprada.

En el model marginal:

- La mitjana de la variable resposta es modelitza només a partir de les covariables, l'estructura de covariàncies és tractada a part.
- Els paràmetres (coeficients) representen l'efecte de les variables explicatives a la mitjana poblacional.
- Els contrastos sobre paràmetres avaluen subpoblacions que comparteixen els mateixos valors per a la resta de covariables.
- Els paràmetres tenen la mateixa interpretació que si es tractés d'anàlisis transversals.

En resum, els models marginals presenten i comparen efectes des de la perspectiva poblacional, mentre que els models condicionals ofereixen interpretació individual. Per exemple, en el cas d'un estudi on s'avaluen dues intervencions (una experimental i una altra control) realitzades en els mateixos individus en períodes diferents i mitjançant un disseny experimental creuat, el model marginal vindria a estimar l'efecte mitjà de la intervenció en el conjunt de la població, mentre que el model condicional avaluaria l'efecte per a l'individu mitjà. Donat que els models GLMM incorporen funcions d'enllaç no lineals, que hi poden haver dades perdudes i que possiblement el model contempla altres variables explicatives, aquests dos càlculs no són idèntics.

Des d'un punt de vista pràctic el model condicional és la visió natural per a estudis multinivell amb selecció d'unitats jeràrquiques i estudis longitudinals, mentre que el model marginal és més adient per a modelitzar variables multivariants o mesures repetides amb estructures de covariàncies especials.

Des d'un punt de vista més teòric, diferents autors mostren la seva discrepància sobre l'eina analítica que cal emprar:

- Diggle, Liang and Zeger (1994) recomanen l'ús de models marginals quan l'objectiu de l'estudi és fer inferències respecte a la població d'interès (com és habitual en epidemiologia) i models condicionals quan calgui inferir sobre les respostes individuals.

- Molenberghs and Verbeke (2005) indiquen que idealment s'haurien d'escollir models marginals sempre que hi hagi preguntes d'investigació marginals i també quan calgui quantificar l'associació entre les mesures repetides.

- Fieberg, et al. (2009) indiquen que el mètode escollit hauria de dependre principalment de la pregunta d'interès. Moltes preguntes de l'àmbit de l'ecologia discorren al voltant de la mesura del rendiment de certa població (supervivència, reproducció, etc.) en funció de les característiques generals de l'entorn, per a les quals la interpretació del model marginal pot ser més natural.

- Lindsey and Lambert (1998) enumeren diversos inconvenients de la visió marginal, principalment la manca d'un mecanisme probabilístic de generació de dades. Proporcionen un exemple en què un tractament podria ser superior en mitjana, alhora

que inferior per a cada individu. Conclouen que els models marginals poden ser adequats per a estudis observacionals descriptius, però s'han d'utilitzar amb molta cura en entorns experimentals causals, com ara els assajos clínics.

- Lee and Nelder (2004) argumenten que els models condicionals sempre s'han de preferir als models marginals, ja que els models condicionals permeten obtenir efectes condicionals i marginals. Demostren que és falsa l'afirmació que s'han d'utilitzar models marginals quan es desitgen inferències sobre poblacions.

- Muff et al. (2016) assenyalen que teòricament quan es modelen dades no normals, no existeix una definició inequívoca de model marginal. El model marginal és matemàticament anàleg a ometre covariables amb capacitat predictiva i, en conseqüència, s'introdueix deliberadament un error de mesura a les covariables. Consideren que, en la majoria dels casos, el model condicional és l'opció més eficient per explicar com s'associen les covariables amb una variable resposta no normal. Tot i això, els models marginals poden ser útils atès que la qüestió científica sovint requereix explícitament una formulació d'aquest model.

Així doncs, per a l'anàlisi de dades no normals amb mesures repetides, el model marginal resulta ser una eina versàtil i eficaç per a l'estimació de paràmetres, però no ofereix un model probabilístic per a les dades amb el qual poder avaluar la qualitat de l'ajust o fer prediccions a nivell individual per a observacions futures. El model condicional, tot i ser menys flexible, supera d'entrada aquestes deficiències, tot i que condueix a estimacions de paràmetres que s'han d'interpretar de forma condicionada als efectes aleatoris i no proporciona la mateixa robustesa.

Aquesta preocupació sobre la robustesa del model condicional prové principalment del fet que el model GLMM condicional assumeix que la distribució dels efectes aleatoris és normal. Al voltant d'aquest tema també hi ha certa controvèrsia sobre l'impacte que té la violació d'aquesta hipòtesi i les alternatives per a una modelització més vàlida:

- Verbeke and Lesaffre (1997) varen comprovar que en el model lineal mixt, desviacions d'aquesta suposició de normalitat dels efectes aleatoris tenen molt poc impacte en l'estimació dels paràmetres del model. En canvi, per als models GLMM condi-

cionals, l'especificació incorrecta de la distribució d'efectes aleatoris pot conduir a estimacions esbiaixades dels paràmetres del model, inclosos els efectes fixos.

- McCulloch and Neuhaus (2011) indiquen que, segons comprovacions teòriques i estudis de simulació, la majoria d'aspectes de la inferència estadística són molt robustos en relació amb la hipòtesi de normalitat dels efectes aleatoris en el model GLMM condicional. Especialment robusta és la inferència per als efectes dins les subunitats, essent sovint aquests efectes el motiu per a considerar dissenys amb mesures repetides.

- Litière et al. (2008) conclouen que el biaix induït en els paràmetres dels efectes fixos és generalment petit, en la mesura que la variabilitat de la distribució d'efectes aleatoris subjacent també és petita. Tanmateix, les estimacions d'aquesta variabilitat sempre estan molt esbiaixades. Atès que els components de la variància són l'única eina per estudiar la variabilitat de la distribució real, és difícil avaluar els potencials problemes en l'estimació dels efectes fixos.

Recentment, han aparegut propostes de modelització per a incorporar efectes aleatoris amb distribució diferent de la normal, tot i que la seva implementació en el programari habitual i el seu ús a la pràctica no es troba especialment estès (Molenberghs et al., 2017).

Per exemple:

- El mètode *integrated nested Laplace approximation* (INLA) per a ajustar models jeràrquics amb una visió bayesiana (Rue et al., 2017).

- Els procediments que utilitzen la versemblança jeràrquica, (*hierarchical generalized linear models*, HGLM), també anomenada *h-likelihood* (Lee et al., 2017).

- A través d'una reformulació de la funció de versemblança (Liu and Yu, 2008).

Els models lineals generalitzats mixtos són, en resum, models que permeten contemplar la presència de dades correlacionades i, per tant, de mesures repetides en l'anàlisi de dades amb distribució no normal.

En el cas de dades normals, s'ha fet distinció entre diferents tipus de mesures repetides en funció de la seva naturalesa jeràrquica/no jeràrquica i del seu paper dins del disseny de

l'experiment. Una primera valoració permetia concloure que el model mixt no necessita fer-ne especial distinció donat que la proposta de modelització i d'ajust és única. Per altra part, el model és robust en relació amb la suposició de normalitat dels efectes aleatoris.

En dades no normals, la proposta de modelització no tan sols no és única sinó que a més provoca interpretacions diferents. Per altra banda, no sempre és viable dur a terme comparacions entre models ajustats amb propostes diferents (condicional vs marginal). També hi ha incertesa sobre l'impacte que pot produir la distribució assumida per als efectes aleatoris.

En aquest context, és d'especial interès el cas de les dades de recompte atès que el model de partida sota la distribució de Poisson incorpora a més la limitació que les dades ajustades han de ser equidisperses i sovint cal recórrer a distribucions alternatives. A continuació s'exposen les particularitats del tractament específic d'aquest tipus de dades.

2.6 Dades de recompte

Les dades de recompte són un tipus de particular de dades estadístiques emprades per a descriure processos en què les variables són mesurades amb valors discrets i no-negatius. La distribució més popular per a tractar dades de recompte és la distribució de Poisson, un cas especialment interessant de la família exponencial. La distribució fou introduïda per Siméon Denis Poisson el 1837 (Poisson, 1837). Els primers treballs sobre dades amb aquesta distribució van aparèixer a finals del segle XIX i principis del XX, incloent-hi alguns exemples més o menys quotidians, però també d'altres de natura força exòtica:

- El nombre d'estrelles per unitat d'espai (Newcomb, 1860).
- El recompte de coces de cavall en l'exèrcit prussià (von Bortkewicz, 1898).
- La quantitat de cèl·lules de llevat per a elaborar cervesa Guiness (student, 1907).
- La quantitat de trucades per minut que arriben a una centraleta telefònica (Erlang, 1909).

- Col·lisions de partícules alfa emeses per una barra de poloni (Rutherford et al., 1910).
- El nombre de bacteris en una mostra (Fisher, 1922).
- El nombre de polls en caps de presoners hindús a Cannamore (Williams, 1944).
- El nombre de bombes caigudes a Londres durant la Segona Guerra Mundial en diferents parcel·les de terreny (Clarke, 1946).

Posteriorment, amb l'aparició dels models lineals generalitzats el 1983 (McCullagh and Nelder, 1983), l'ús dels models de Poisson per a dades de recompte s'estengué i els exemples esdevingueren més interessants, contemplant al mateix temps dissenys experimentals més complexos: atacs d'epilèpsia al llarg del temps, incidències en l'àmbit de les assegurances de cotxe, recomptes de fauna aviària, etc.

Avui en dia, les dades de recompte són tan populars com qualsevol altre tipus de dades (Sellers et al., 2012). Gràcies a la recollida sistemàtica d'informació es capturen dades arreu i es disposa d'exemples en àmbits molt diversos i en situacions especialment sofisticades, per exemple visites a planes web, canvis de preu d'una acció que cotitza en borsa per interval de temps, incidència de malalties segons àrees geogràfiques, recomptes de cèl·lules malmeses en biodosimetria, etc. Exemples recents inclouen l'avaluació d'intervencions urbanes en la reducció d'accidents de trànsit o l'impacte de les targetes vermelles en el nombre de gols marcats en partits de futbol.

A partir de la formulació general (2.1) i les funcions particulars per a la distribució de Poisson (taula 2.1), el model expressat de forma matricial es pot concretar com:

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i \\ \log(\mu_i) &= X_i\beta + \log(w_i) \\ Y_i &\sim \mathrm{Poi}(\mu_i) \end{aligned}$$

on Poi denota la distribució de Poisson, i és l'índex de l'observació, Y el vector resposta, X és la matriu de disseny per als efectes fixos, β és el vector dels coeficients del model i w és un vector amb els corresponents pesos de compensació (*offset*) que permet modular el valor esperat d'Y_i segons certa mesura de magnitud w_i.

Com s'ha indicat prèviament, una de les principals característiques dels models basats en la distribució de Poisson és que s'assumeix que la distribució és equidispersa, és a dir, per a cada observació el valor esperat i la variància són iguals: $\text{Var}(Y_i) = \text{E}(Y_i) = \mu_i$. Aquest fet constitueix una limitació important ja que en la majoria de conjunts reals de dades de recompte sovint la dispersió és excessiva.

La discrepància entre la variància empírica i l'esperança de les dades ajustades pot produir-se per múltiples mecanismes que caldria validar i revisar amb cura per tal de corregir el model si fos necessari. A continuació s'exposen els fenòmens més rellevants que donen lloc a sobredispersió o bé infradispersió (Xekalaki, 2014):

- **Model mal especificat:**

  Quan els efectes considerats en el model ajustat difereixen del model real es diu que el model està mal especificat. Per exemple, si s'omet una variable explicativa rellevant, o bé si al model hi falta una interacció, un terme quadràtic o si la variable explicativa requereix algun tipus de transformació, aleshores es genera sobredispersió. De fet, les conseqüències d'un model mal especificat solen ser força incertes, d'entrada el model és més ineficient i les observacions deixen de ser independents.

- **Dades correlacionades:**

  Si les mesures presenten algun tipus de clusterització, per exemple perquè el pla de mostreig contempla algun factor aleatori, aleshores, ometre aquesta consideració del model també és una forma de mala especificació que pot donar lloc a un increment de la dispersió.

- **Alteració de zeros:**

  Un altre escenari que dona lloc a sobredispersió (o infradispersió) és l'excés (o la manca) de zeros. Aquesta situació normalment té lloc a causa de limitacions en la recollida de dades, per exemple si s'ha recollit un excés d'observacions amb zeros estructurals o bé si s'han descartat de la recollida observacions amb valors nuls. Els models que admeten aquest fenomen reben el nom de models per a zeroinflació o zerodeflació. El primer cas es pot representar a partir de la mixtura d'una distribució de recompte i una variable discreta.

- **Heterogeneïtat:**

En el cas en què les observacions no siguin homogènies també es detectarà sobredispersió. És a dir, la taxa d'ocurrència dels esdeveniments pot ser diferent entre individus idèntics, com si cada individu tingués la seva pròpia taxa al marge de les seves característiques. Per a la modelització d'aquest sovint s'assumeix que les taxes segueixen determinada distribució (normal, gamma, etc.). Aquests models es coneixen com a composicions o mixtures de models (*mixture models*), essent la composició Poisson-gamma la més popular donat que és equivalent a la distribució binomial negativa.

- **Outliers:**

  També cal tenir en compte que en cas de presència de valors anòmals (*outliers*) la distribució resultant serà sobredispersa.

- **Taxa d'esdeveniments no constant:**

  Les distribucions de recompte també es poden caracteritzar a partir del recompte d'esdeveniments per unitat de temps en processos estocàstics. Per exemple, si la taxa d'ocurrència d'esdeveniments és constant, aleshores el temps entre esdeveniment és una variable aleatòria amb distribució Exponencial, mentre que el recompte d'esdeveniments per unitat de temps segueix una distribució de Poisson. Per altra banda, si a mesura que no apareixen esdeveniments la taxa creix (*increasing failure rate*) el procés serà infradispers, per exemple quan la distribució del temps entre esdeveniments segueix una distribució tal que la funció de densitat és log-còncava. Per contra, si la taxa és decreixent (*decreasing failure rate*) el procés serà sobredispers.

- **Composició de processos:**

  De vegades, els recomptes d'interès provenen de la composició de processos. Per exemple, un primer procés estaria associat a l'ocurrència d'esdeveniment d'interès i l'altre a la severitat de cadascun d'ells. Aquesta combinació genera un increment de la variabilitat i, per tant, sobredispersió.

La majoria de les situacions esmentades tenen solució dins de la família de distribucions exponencial, així que els mètodes per a models lineals generalitzats seran igualment vàlids.

2.6.1 Models de recompte amb sobredispersió / infradispersió

La solució habitual per a ajustar dades de recompte amb sobredispersió o infradispersió és la de considerar un model més general que el model 2.1, dotant-lo de més flexibilitat. En funció del tipus de fenomen responsable de la manca d'equidispersió existeixen propostes de model diverses. A continuació es presenten les més habituals:

- **Model amb distribució binomial negativa:**

  Possiblement l'alternativa més usual per a ajustar dades amb sobredispersió és el model amb distribució binomial negativa. Aquest model consisteix en una mixtura Poisson-gamma i assumeix que la sobredispersió és fruit de l'heterogeneïtat entre els individus de l'estudi. La relació entre la mitjana i la variància és quadràtica, $\text{Var}(Y_i) = \mu_i + \theta\mu_i^2 = \mu_i(1+\theta\mu_i)$ donant lloc a un tipus particular de sobredispersió. També existeixen proves estadístiques específiques per a contrastar aquesta forma de sobredispersió. El model es pot formular així:

$$\begin{aligned} \text{E}(Y_i) &= \mu_i\tau_i \\ \log(\mu_i) &= X_i\beta + \log(w_i) \\ Y_i &\sim \text{Poi}(\mu_i\tau_i) \\ \tau_i &\sim \text{Gamma}(1/\theta, 1/\theta) \end{aligned}$$

  de manera que $\text{E}(\tau_i) = 1$ i $\text{Var}(\tau_i) = \theta$ i, per tant, θ representa un paràmetre associat a la sobredispersió de les dades.

  Aquest model també es pot formular a partir de la distribució binomial negativa (BN):

$$\begin{aligned} \text{E}(Y_i) &= \mu_i \\ \log(\mu_i) &= X_i\beta + \log(w_i) \\ Y_i &\sim \text{BN}(1/(1+\theta\mu_i), 1/\theta) \end{aligned}$$

- **Model mixtura Poisson-normal**

  Una altra alternativa amb propietats similars al cas anterior és la mixtura Poisson-normal. Aquest model s'ajusta incorporant un factor aleatori amb tants nivells com

observacions. També assumeix que la sobredispersió és fruit d'heterogeneïtat entre individus.

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i e^{e_i} \\ \log(\mu_i) &= X_i\beta + \log(w_i) \\ Y_i &\sim \mathrm{Poi}(\mu_i e^{e_i}) \\ e_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

- **Model amb distribució Conway-Maxwell-Poisson (COM-Poisson):**
  La disitribució COM-Poisson constitueix un model vàlid tant per a contemplar sobredispersió com infradispersió. Aquesta distribució va ser proposada inicialment per (Conway and Maxwell, 1962), però la seva implementació pràctica s'atribueix a Shmueli et al. (2005). Posteriorment, (Guikema and Coffelt, 2008) suggeriren una reparametrització més interpretable en el context dels models lineals generalitzats. La seva funció de densitat és:

$$\mathrm{P}(Y_i = y_i) = \frac{1}{S(\mu_i, \nu)} \left(\frac{\mu_i^{y_i}}{y_i!}\right)^{\nu}$$

  on $S(\mu_i, \nu)$ és una constant de regularització i ν és un paràmetre associat amb la sobredispersió. Sota aquest model:

$$\begin{aligned} \mathrm{E}(Y_i) &\approx \mu_i + \frac{1}{2\nu} - \frac{1}{2} \\ \mathrm{Var}(Y_i) &\approx \frac{\mu_i}{\nu} \end{aligned}$$

  El model es concreta establint que:

$$\log(\mu_i) = X_i\beta + \log(w_i)$$

- **Generalized Poisson Model (GPM):**
  El model GPM fou introduït per Consul (1989) i també permet contemplar sobredispersió o bé infradispersió. Es defineix així:

$$\mathrm{P}(Y_i = y_i) = \left(\frac{\mu_i}{1+\phi\mu_i}\right)^{y_i} \frac{(1+\phi y_i)^{y_i-1}}{y_i!} \exp\left(\frac{-\mu_i(1+\phi y_i)}{1+\phi\mu_i}\right)$$

  on ϕ és un paràmetre associat amb la sobredispersió. En aquest cas, $\mathrm{E}(Y_i) = \mu_i$ i el model es concretaria establint la relació amb les variables explicatives com en el cas anterior.

Els models exposats es poden ajustar per màxima versemblança i es poden comparar utilitzant el test de raó de versemblances si estan niats o mitjançant mesures relatives com ara els índexs AIC o BIC.

2.6.2 Dades de recompte amb mesures repetides

El model lineal generalitzat mixt amb efectes aleatoris i distribució de Poisson (GLMM-Poisson), vindria a ser una ampliació del model GLM-Poisson per a recomptes habilitant la presència de dades correlacionades, ja sigui a través de factors aleatoris (model condicional) o a través de certa estructura de covariàncies (model marginal).

En presència de sobredispersió o infradispersió, també seria possible considerar un model GLMM condicional amb altres distribucions de recompte, per exemple amb la distribució binomial negativa o bé la distribució COM-Poisson. Aquestes propostes donarien lloc als models que es podrien anomenar GLMM-binomial negativa, o GLMM-COM-Poisson.

Per altra part, el model GLMM-Poisson marginal contempla la flexibilitat necessària per a adaptar la manca d'equidispersió i l'estructura de covariàncies marginal desitjada, de forma que des d'un punt de vista pràctic no és necessari considerar les versions marginals per a altres distribucions de recompte.

En aquest sentit, la tria del mètode marginal o condicional es veu condicionada per diferents aspectes:

- En funció de la interpretació dels efectes que es desitgi: poblacional o individual.
- Segons la naturalesa de les mesures repetides:
  - Mesures repetides jeràrquiques.
  - Dades multivariants.
  - Errors de mesura.
- Segons la presència de sobredispersió o infradispersió.
- En funció de la distribució dels efectes aleatoris que calgui considerar (normalitat dels efectes aleatoris).

Però el principal motiu per a escollir una metodologia o una altre és el fet de disposar o no d'un model de dades adient. És a dir, en cas que es pugui plantejar un model de dades vàlid que permeti tenir en compte les particularitats de les dades de recompte i la naturalesa de l'experiment, aleshores seria preferible utilitzar el model condicional, altrament caldrà emprar models marginals. En bona mesura, la necessitat d'haver de recórrer a models marginals és un símptoma de la mancança de propostes de models de dades adients.

A continuació es presenten tres exemples d'aplicació de la metodologia d'anàlisi de dades de recompte amb mesures repetides considerant dissenys complexos amb mesures repetides i per tant en presència de dades correlacionades. En el primer cas, s'analitza el recompte de sinistres al llarg del temps en diferents cruïlles de la ciutat de Barcelona sota certa intervenció preventiva mitjançant un model condicional amb diferents distribucions de recompte. En el segon exemple s'utilitzen models lineals generalitzats marginals amb distribució de Poisson per a analitzar l'impacte de les targetes vermelles en el nombre de gols marcats en diferents partits de futbol. En tercer lloc es presenta una proposta original per a la modelització de dades de recompte en experiments amb subrèpliques, un cas especial de mesures repetides on les observacions són replicades amb l'objectiu de controlar l'error de mesura.

CAPÍTOL 3

Effectiveness of a Road Traffic Injury Prevention Intervention in Reducing Pedestrian Injuries, Barcelona, Spain, 2002–2019

3.1 Article

American Journal of Public Health

Effectiveness of a Road Traffic Injury Prevention Intervention in Reducing Pedestrian Injuries, Barcelona, Spain, 2002–2019

Katherine Pérez, MPH, PhD [a,b,c], Elena Santamariña MPH, PhD [a], Josep Ferrando MD, MPH, PhD [a], María José López MPH, PhD [a,b,c], Llorenç Badiella MSC [d]

[a] Agència de Salut Pública de Barcelona (ASPB), Barcelona, Spain

[b] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain

[c] Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain

[d] Departament de Matemàtiques - Universitat Autònoma de Barcelona - Cerdanyola del Vallès, Barcelona, Spain

Resum

En aquest estudi s'analitzà l'efectivitat de la intervenció Rutes Segures a l'Escola (SRTS) dut a terme a Barcelona entre el 2009 i el 2016, per reduir el nombre de col·lisions i ferits de trànsit a les proximitats dels centres escolars. El disseny contemplava una avaluació pre-post, quasiexperimental amb un grup de comparació aparellat.

L'anàlisi basada en models lineals generalitzats condicionals amb distribució de Poisson va permetre detectar que la intervenció provoca reduccions significatives en el nombre total de col·lisions i lesionats, especialment en joves i vianants a les escoles que implementen el programa en comparació de les escoles que no ho fan.

Per tant, l'SRTS va contribuir de manera significativa a millorar la seguretat viària entre els nens i adolescents en un entorn urbà. Aquest tipus d'intervenció en seguretat viària urbana pot ajudar a assolir els objectius de desenvolupament sostenible per a l'Agenda 2030.

Effectiveness of a Road Traffic Injury Prevention Intervention in Reducing Pedestrian Injuries, Barcelona, Spain, 2002–2019

Katherine Pérez, PhD, MPH, Elena Santamariña-Rubio, PhD, MPH, Josep Ferrando, MD, PhD, MPH, Maria José López, PhD, MPH, and Llorenç Badiella, MSC

This study aimed to evaluate the effectiveness of the Safe Routes to School (SRTS) intervention in Barcelona, Spain, at reducing the number of road traffic collisions and injuries in the school environment. It was a pre–post, quasi-experimental evaluation with a matched comparison group. Road traffic injuries were significantly reduced in the intervention schools—especially among school-age pedestrians—but not in the comparison schools. The SRTS program significantly improved road safety among children. (*Am J Public Health*. Published online ahead of print February 23, 2023:e1–e5. https://doi.org/10.2105/AJPH.2022.307216)

AJPH Published online ahead of print February 23, 2023

Many cities have promoted Safe Routes to School (SRTS) programs to make it easier for children to walk or cycle to school safely. Most studies have found that implementation of these programs increases active travel to school[1–3] and decreases road traffic injuries,[4–10] although there is controversy because of methodological limitations.[11]

INTERVENTION AND IMPLEMENTATION

Barcelona's SRTS program, called *Camí escolar, espai amic* (Safe route to school, friendly space), began with the aim of increasing children's and adolescents' personal autonomy, responsibility, and quality of life on their way to school or while walking around the neighborhood. The program promotes road safety education in schools through an educational program conducted within the school and the community, and through changes in the environment around the school.[12] After initial piloting, full deployment of SRTS began in 2006. Available data allowed us to evaluate a real-life policy with important public health implications. (For more details, see the Appendix, available as a supplement to the online version of this article at http://www.ajph.org).

PLACE, TIME, AND PERSONS

This is a pre–post, quasi-experimental evaluation study, with a matched comparison group. The intervention group was schools with the SRTS program, and the comparison group comprised schools without the SRTS program. The study area was defined as a buffer around the schools with a radius of about 200 meters (656 feet).

The inclusion and exclusion criteria were as follows: of the 152 schools with SRTS, we selected schools whose SRTS implementation year was after 2005 and whose inauguration year (street works and program implemented) was prior to 2016. All of the selected intervention schools had a pre- and postintervention period of at least four years per period; therefore, injury data include the years 2002 to 2019.

Inclusion and exclusion criteria for traffic collisions were as follows: we included traffic collisions with casualties occurring in the study area (buffer with a 200-m radius around the schools) from Monday to Friday from 7:00 AM to 9:30 AM, 12:00 PM to 3:00 PM, and 4:00 PM to 6:00 PM, from September 15 to June 23. Collisions occurring during Christmas and Easter holiday periods were excluded. (For more details, see the Appendix).

Outcome variables included the number of road traffic collisions involving

AJPH Published online ahead of print February 23, 2023

casualties (total, children [defined as aged 0–16 years], and pedestrian children) and number of people injured (total, children, and pedestrian children).

Exposure variables included population in the area, available family income, and data on motor and active mobility and structure streets (for more detail, see Appendix).

To compare the results in the postintervention period versus the preintervention period, for each outcome measure, we fitted a generalized linear mixed model with Poisson distribution using the logarithm as a link function between expected values and explanatory variables. The explanatory variables included in the model were the group, the period (pre- vs postintervention), the interaction between group and period, and the year. To obtain a more precise fit, the model was also adjusted by the exposure variables.

PURPOSE

This study aimed to evaluate the effectiveness of the SRTS program carried out in Barcelona between 2006 and 2016 in reducing the number of road traffic collisions and injuries in the school environment.

EVALUATION AND ADVERSE EFFECTS

The study included 64 schools with SRTS programs implemented between 2006 and 2016, and 63 comparison schools, reaching 49 092 students in 2018 (intervention and comparison schools). A total of 15.0% of the schools were preschools (students aged 0–3 years), 55.1% were primary schools (students aged 4–11 years), and 29.9% were secondary schools (students aged 12–18 years). The proportion of public schools was higher in intervention than in comparison schools (75% and 60.3%, respectively), but there were no significant differences in the mean number of students per school: 367.8 (95% confidence interval [CI] = 306.7, 428.9) and 405.6 (95% CI = 320.2, 491.0), respectively.

The environmental characteristics of the intervention and comparison schools were similar. Differences were only found in the mean neighborhood income in 2017 and in the concentration of injured pedestrians in the school neighborhood in 2018. Available family income in the intervention school neighborhoods was significatively higher than in the comparison school neighborhoods (relative index = 112.1 and 99.8, respectively). The number of injured pedestrians per 100 meters of street was significatively lower for intervention schools (7.8) than for comparison schools (10.1).

In the intervention schools overall (aggregated), the total number of people injured was 2994 (annual mean = 272.2) in the preintervention period and 2284 (annual mean = 228.4) in the postintervention period. In the comparison schools, this number was 4061 (annual mean = 369.2) and 3196 (annual mean = 319.6), respectively (Table 1).

Per school, in the preintervention period, the annual mean number of injury road traffic collisions involving children and pedestrian children was significantly higher in the comparison schools than in the intervention schools. There were no differences in the annual school mean number of collisions involving children and pedestrian children (Table 1). In the postintervention period, the pattern was the same, although in general the annual school means were lower than in the preintervention period in both the intervention and comparison schools.

When we compared the results of the pre- and postintervention periods, the final adjusted models showed a significant reduction in the risk of collisions and people injured in the intervention schools, with a reduction of 11.7% in the number of injury collisions, 41.1% in the number of injury collisions involving children, and 43.3% in the number of injury collisions involving children pedestrians. For people injured, there was a reduction of 9.1% in the total injured, 36.6% in the number of children injured, and 39.9% in the number of children pedestrians injured (Table 2).

Among the comparison schools, there were no significant changes in outcomes between the pre- and postintervention periods (Table 2).

The significant difference in percentage change in the post- versus the preintervention period between intervention and comparison schools (significance of the interaction between intervention group and period) showed that the reduction in the intervention schools in the number of injury collisions involving children and pedestrian children could be attributable to the implementation of the SRTS program (Table 2).

SUSTAINABILITY

The SRTS program is currently beginning a new phase, with a greater focus on increasing safety in front of the school (*protegim les escoles*: we protect the schools).

PUBLIC HEALTH SIGNIFICANCE

The SRTS program, carried out in Barcelona between 2006 and 2016, showed a significant reduction in injuries in the intervention schools, which

TABLE 1— Injury Traffic Collisions and People Injured in Areas Surrounding Schools With an SRTS Program (200-Meter Buffer) and in Areas Surrounding Comparison Schools, by Intervention Period: Barcelona, 2002–2019

	Intervention Schools (n = 64)				Comparison Group Schools (n = 63)				Per School *P*[a] (Intervention/ Comparison)
	All Schools Total	All Schools Annual Mean (95% CI)	Per School Range	Per School Annual Mean (95% CI)	All Schools Total	All Schools Annual Mean (95% CI)	Per School Range	Per School Annual Mean (95% CI)	
No. of road traffic collisions with injuries									
Preintervention	2994	272.2 (180.3, 364.1)	0–28	6.0 (5.6, 6.5)	4061	369.2 (249.1, 489.2)	0–37	8.2 (7.5, 8.9)	.001
Postintervention	2284	228.4 (119.9, 336.9)	0–28	5.7 (5.2, 6.2)	3196	319.6 (168.8, 470.4)	0–50	8.2 (7.3, 9.0)	.002
No. of collisions involving any injured person aged 0–16 y									
Preintervention	240	21.8 (13.2, 30.5)	0–4	0.5 (0.4, 0.5)	262	23.8 (14.5, 33.1)	0–4	0.6 (0.5, 0.6)	.033
Postintervention	120	12 (6.6, 17.4)	0–4	0.3 (0.2, 0.4)	169	16.9 (8.9, 24.9)	0–6	0.4 (0.4, 0.5)	.022
No. of collisions involving any injured pedestrians aged 0–16 y									
Preintervention	135	12.3 (6.8, 17.7)	0–3	0.3 (0.2, 0.3)	124	11.3 (6.7, 15.8)	0–3	0.3 (0.2, 0.3)	.76
Postintervention	66	6.6 (2.8, 10.4)	0–3	0.2 (0.1, 0.2)	97	9.7 (4.7, 14.7)	0–4	0.2 (0.2, 0.3)	.053
No. of people injured									
Preintervention	3478	316.2 (207.2, 425.2)	0–34	7 (6.4, 7.6)	4774	434 (292.4, 575.6)	0–47	9.6 (8.8, 10.5)	.001
Postintervention	2715	271.5 (141.7, 401.3)	0–33	6.8 (6.2, 7.4)	3720	372 (199.8, 544.2)	0–58	9.5 (8.5, 10.5)	.005
No. of injured persons aged 0–16 y									
Preintervention	251	22.8 (13.6, 32.1)	0–4	0.5 (0.4, 0.6)	288	26.2 (15.5, 36.8)	0–6	0.6 (0.5, 0.7)	.02
Postintervention	131	13.1 (7.1, 19.1)	0–4	0.3 (0.3, 0.4)	177	17.7 (9.2, 26.2)	0–6	0.5 (0.4, 0.5)	.024
No. of injured pedestrians aged 0–16 y									
Preintervention	136	12.4 (6.8, 17.9)	0–3	0.3 (0.2, 0.3)	131	11.9 (7.0, 16.8)	0–4	0.3 (0.2, 0.3)	.74
Postintervention	70	7 (3.0, 11.0)	0–3	0.2 (0.1, 0.2)	98	9.8 (4.7, 14.9)	0–4	0.3 (0.2, 0.3)	.06

Note. SRTS = Safe Routes to School.

[a]Significance of the nonparametric Wilcoxon rank-sum test (Mann-Whitney).

AJPH Published online ahead of print February 23, 2023

AJPH Published online ahead of print February 23, 2023

TABLE 2— Mean Number of Adjusted Annual Injury Collisions and Injured People, Adjusted Relative Risk, and Pre–Post Percentage Change in Surrounding Areas of Intervention and Comparison Schools: Barcelona, 2002–2019

	Intervention Schools (n = 64)				Comparison Group Schools (n = 63)				
	Adjusted Annual Mean Per School	SE	RR (95% CI)	Post/Pre % Change (95% CI)	Adjusted Annual Mean Per School	SE	RR (95% CI)	Post/Pre % Change (95% CI)	*P*[a]
No. of road traffic collisions with injuries									.14
Preintervention	4.72	0.34	1 (Ref)	1 (Ref)	5.03	0.36	1 (Ref)	1 (Ref)	
Postintervention	4.16	0.30	0.88 (0.80, 0.97)	−11.7 (−19.9, −2.7)	4.83	0.35	0.96 (0.88, 1.05)	−4.1 (−12.3, 5.0)	
No. of collisions involving injured persons aged 0–16 y									.019
Preintervention	0.43	0.04	1 (Ref)	1 (Ref)	0.44	0.04	1 (Ref)	1 (Ref)	
Postintervention	0.25	0.03	0.59 (0.47, 0.73)	−41.1 (−52.6, −27.0)	0.37	0.03	0.84 (0.68, 1.03)	−16.5 (−32.3, 3.1)	
No. of collisions involving injured pedestrians aged 0–16 y									.003
Preintervention	0.20	0.03	1 (Ref)	1 (Ref)	0.19	0.02	1 (Ref)	1 (Ref)	
Postintervention	0.12	0.02	0.57 (0.42, 0.77)	−43.3 (−58.3, −22.7)	0.20	0.03	1.05 (0.79, 1.38)	4.5 (−21.2, 38.5)	
No. of people injured									.87
Preintervention	5.44	0.39	1 (Ref)	1 (Ref)	6.14	0.44	1 (Ref)	1 (Ref)	
Postintervention	4.94	0.36	0.91 (0.82, 1.00)	−9.1 (−17.7, 0.4)	5.63	0.41	0.92 (0.84, 1.01)	−8.2 (−16.5, 0.9)	
No. of injured persons aged 0–16 y									.1
Preintervention	0.43	0.04	1 (Ref)	1 (Ref)	0.47	0.04	1 (Ref)	1 (Ref)	
Postintervention	0.28	0.03	0.63 (0.50, 0.81)	−36.6 (−50.4, −18.8)	0.39	0.04	0.82 (0.66, 1.03)	−17.8 (−34.3, 2.9)	
No. of injured pedestrians aged 0–16 y									.011
Preintervention	0.20	0.03	1 (Ref)	1 (Ref)	0.20	0.03	1 (Ref)	1 (Ref)	
Postintervention	0.12	0.02	0.60 (0.44, 0.82)	−39.9 (−55.9, −18.0)	0.20	0.03	1.00 (0.76, 1.34)	0.5 (−24.5, 33.6)	

Note. CI = confidence interval; RR = relative risk. Surrounding area defined as within a 200-meter buffer. Adjusted models include the explanatory variables, intervention group, period, the interaction between both terms and year, and the exposure variables, number of students of the school (log), number of inhabitants in the neighborhood (log), kilometers traveled by motor vehicles in the study area (log), and a quadratic term for the latter effect.

[a]Significance of the interaction between period and type of school (intervention, comparison), which allows us to assess whether the differences between intervention and comparison schools can be attributed to the intervention. It shows whether the difference pre–post in the intervention group is significantly different from the difference pre–post in the comparison group.

was not observed in the comparison schools. There was a notable decrease in the number of injured pedestrians, especially school-age pedestrians, which is the target population of the SRTS.

These results are relevant for two reasons. On the one hand, injuries were significantly reduced in the intervention schools but not in the comparison group, in the context of increasing road traffic injury rates in the city (although with decreasing severity). On the other hand, our results provide evidence of the effectiveness of the SRTS program in improving road safety and reducing road crashes and injuries, particularly among children, when there is controversy in the scientific literature.[9,11] Our study aimed to overcome the limitations reported in previous studies by using a quasi-experimental study, which controlled for major confounding factors through the study design and statistical analysis.

This study evaluates the health impacts of a policy developed outside the health sector. It provides evidence on how an infrastructure intervention contributes to health benefits, implementing health in all policies and reducing social inequities. AJPH

ABOUT THE AUTHORS

Katherine Pérez, Elena Santamariña-Rubio, Josep Ferrando, and Maria José López are with Agència de Salut Pública de Barcelona (ASPB), Barcelona, Spain. Katherine Pérez and Maria José López are also with CIBER Epidemiología y Salud Pública (CIBERESP), Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain. Llorenç Badiella is with Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain.

CORRESPONDENCE

Correspondence should be sent to Katherine Pérez, Agència de Salut Pública de Barcelona (ASPB), Pl. Lesseps, 1, 08023 Barcelona, Spain (e-mail: cperez@aspb.cat). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

PUBLICATION INFORMATION

Full Citation: Pérez K, Santamariña-Rubio E, Ferrando J, L Maria José, Badiella L. Effectiveness of a road traffic injury prevention intervention in reducing pedestrian injuries, Barcelona, Spain, 2002–2019. *Am J Public Health*. Published online ahead of print February 23, 2023:e1–e5.

Acceptance Date: December 18, 2022.

DOI: https://doi.org/10.2105/AJPH.2022.307216

CONTRIBUTORS

K. Pérez and E. Santamariña designed and conceptualized the study. K. Pérez and J. Ferrando did the literature review. E. Santamariña and L. Badiella analyzed the data. All authors contributed to results and discussion. K. Pérez drafted the initial article, and all authors contributed to subsequent edits of the revised article.

ACKNOWLEDGMENTS

Llorenç Badiella received funding from the Spanish Ministry of Science, Innovation and Universities (grant RTI2018-096072-B-I00).

We thank Miquel Ruscalleda and Carme Ruiz (Road Safety and Mobility Program, Barcelona City Council); Pilar Lleonart, Gretel Vila, and Encarna Isanda (Municipal Institute of Education of Barcelona, IMEB); and Isaac Aparicio and Josep Clotet (Municipal Institute of Informatics, IMI).

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

HUMAN PARTICIPANT PROTECTION

The present article did not require institutional review board approval because we do not report human participant data.

REFERENCES

1. Chriqui JF, Taber DR, Slater SJ, Turner L, Lowrey KMG, Chaloupka FJ. The impact of state safe routes to school-related laws on active travel to school policies and practices in US elementary schools. *Health Place.* 2012;18(1):8–15. https://doi.org/10.1016/j.healthplace.2011.08.006
2. Hoelscher DM, Ganzar LA, Salvo D, et al. Effects of large-scale municipal safe routes to school infrastructure on student active travel and physical activity: design, methods, and baseline data of the Safe Travel Environment Evaluation in Texas Schools (STREETS) natural experiment. *Int J Environ Res Public Health.* 2022;19(3):1810. https://doi.org/10.3390/ijerph19031810
3. Stewart O, Moudon AV, Claybrooke C. Multistate evaluation of safe routes to school programs. *Am J Health Promot.* 2014;28(3 suppl):S89–S96. https://doi.org/10.4278/ajhp.130430-QUAN-210
4. DiMaggio C, Li G. Effectiveness of a safe routes to school program in preventing school-aged pedestrian injury. *Pediatrics.* 2013;131(2):290–296. https://doi.org/10.1542/peds.2012-2182
5. DiMaggio C, Brady J, Li G. Association of the Safe Routes to School program with school-age pedestrian and bicyclist injury risk in Texas. *Inj Epidemiol.* 2015;2(1):15. https://doi.org/10.1186/s40621-015-0038-3
6. DiMaggio C, Frangos S, Li G. National Safe Routes to School program and risk of school-age pedestrian and bicyclist injury. *Ann Epidemiol.* 2016;26(6):412–417. https://doi.org/10.1016/j.annepidem.2016.04.002
7. Hagel BE, Macpherson A, Howard A, et al. The built environment and active transportation safety in children and youth: a study protocol. *BMC Public Health.* 2019;19(1):728. https://doi.org/10.1186/s12889-019-7024-6
8. Yu CY. How differences in roadways affect school travel safety. *J Am Plann Assoc.* 2015;81(3):203–220. https://doi.org/10.1080/01944363.2015.1080599
9. Lizarazo CG, Hall T, Tarko A. Impact of the Safe Routes to School Program: comparative analysis of infrastructure and noninfrastructure measures in Indiana. *J Transp Eng A Syst.* 2021;147(1). https://doi.org/10.1061/JTEPBS.0000480
10. Muennig PA, Epstein M, Li G, DiMaggio C. The cost-effectiveness of New York City's Safe Routes to School program. *Am J Public Health.* 2014;104(7):1294–1299. https://doi.org/10.2105/AJPH.2014.301868
11. Kang B. Identifying street design elements associated with vehicle-to-pedestrian collision reduction at intersections in New York City. *Accid Anal Prev.* 2019;122:308–317. https://doi.org/10.1016/j.aap.2018.10.019
12. Saurí E, Sintes E, Truñó M. Avaluació Del Programa Camí Escolar, Espai Amic. 2017. Available at: https://institutinfancia.cat/es/mediateca/informe-davaluacio-programa-cami-escolar-espai-amic. Accessed August 7, 2022.

AJPH Published online ahead of print February 23, 2023

Title: Effectiveness of a road traffic injury prevention intervention in reducing pedestrian injuries, Barcelona, 2002-2019

Katherine Pérez, MPH, PhD a,b,c, Elena Santamariña MPH, PhD, Josep Ferrando MD, MPH, PhD, Maria José López MPH, PhD a,b,c, Llorenç Badiella d

a Agència de Salut Pública de Barcelona (ASPB), Barcelona, Spain;

b CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain;

c Institut d'Investigació Biomèdica Sant Pau (IIB SANT PAU), Barcelona, Spain

d Departament de Matemàtiques - Universitat Autònoma de Barcelona - Cerdanyola del Vallès, Barcelona, Spain

SUPPLEMENTAL APPENDIX

METHODS

Safe Routes to School Program in Barcelona

Although there have been previous pilot experiences of SRTS programs, The "Ca*mí escolar, espai amic*" ("Safe Route to School, friendly space") program began with the aim of increasing children´s and adolescents personal autonomy, responsibility and quality of life on their way to school or while walking around the neighborhood. The program promotes road safety education in schools through an educational program conducted within the school and the community, and changes in the environment around the school. It is led by the Municipal Institute of Education of Barcelona (IMEB) and carried out in collaboration with the Department of Safety & Mobility of Barcelona. In each education center it includes four phases: Phase 0: *We start to walk*, for the definition of the project; Phase I: *We explore the path*, to carry out the diagnosis. Phase II. *We go out into the neighborhood*, to create the network of friendly spaces and celebrate the work done. Phase III: *We keep the path alive*, to evaluate and guarantee the sustainability of the project in the school and the educational community [1].

Study design and methodology

This was a pre-post quasi-experimental evaluation study, with a matched comparison group. The study population consisted of people who moved around Barcelona between 2002 and 2019 in Barcelona city.

The intervention group consisted of schools in which the SRTS program was conducted, and the comparison group comprised schools without the SRTS program.

The study area was defined as a buffer around the schools within a radius of about 200 meters.

The inclusion and exclusion criteria were as follows: of the 152 schools in Barcelona in which the SRTS program was implemented, we selected schools whose SRTS implementation year started after 2005 and whose inauguration (street works and program implemented) year was prior to 2016. Therefore, all the selected intervened schools had a pre-intervention and post-intervention period of at least 4 years per period.

For each intervention school, was included a close non-intervention school as a comparison school, preferably in the same neighborhood (although this was not always possible), with the same level of education (nursery, infant, primary, and secondary- or high- schools). We excluded comparison schools overlapping with the study area of an intervention school. Finally, 64 intervention schools fulfilled the criteria, with a matched comparison school meeting the inclusion criteria for 63 of them.

The distribution of intervention and comparison schools is showed in Figure S1.

Information Sources:

- Barcelona Local Police Register of Road Traffic Accidents and Victims, which provided information on traffic collisions in the city from 2002 to 2019. Data on road traffic collisions and people injured included geocodes and allowed identification of collisions occurring in the intervention and comparison study zones.

- The IMEB provided information on the intervention schools. For each calendar year, the IMEB defined the schools that had started the activities or interventions related to the program in that year (start year), as well as those that had carried out most of the agreed actions on the public road and their SRTS program had already been inaugurated (inauguration year).

- The Municipal Institute of Informatics (IMI) provided a map with the geolocation of all the schools in the city, which allowed identification of the location of the intervention and comparison schools, and their study areas (polygon around the schools within a radius of about 200 meters).

- The Department of Safety & Mobility of Barcelona, provided a map of all the street sections in the city with information on the location of points for measuring traffic congestion in the city and kilometers of the street. This allowed estimation of the annual average daily traffic (ADT) and the

kilometers traveled by motor vehicles (ADT/km of street) for all the interventions and comparison study areas in 2018.

Study period

The study period comprised 2002 to 2019. Three periods were defined, specifics for each school: The pre-intervention period, from 2002 to the year before the program started; The implementation period, from the program (works) start year to the inauguration year; The post-intervention period, from the year following the inauguration year until 2019. Not all schools started the program in the same year, nor did they inaugurate in the same year. Therefore, not all schools had the same pre- and post- period, nor did the periods have the same number of years, although they all had a minimum of 4 pre- and 4 post- years. The two periods analyzed were the pre- and post-intervention periods.

Inclusion and exclusion criteria for traffic collisions

Traffic collisions with casualties occurring in the study area (buffer with a 200 meters radius around the schools): from Monday to Friday, from 7:00 to 9:30, 12:00-15:00 and 16:00 to 18:00, and from September 15 to June 23, were included.

Collisions occurring on Christmas and Easter holiday periods were excluded.

Study variables

Dependent or outcome variables:

- Number of road traffic collisions with casualties
- Number of collisions involving 0-16 years-old injured
- Number of collisions involving 0-16 years-old pedestrians injured

- Number of people injured
- Number of 0-16 years-old injured
- Number of 0-16 years-old pedestrians injured

Exposure variables:

- Number of students at the school in 2018
- Total population and population aged 0-16 years old resident in the school neighborhood for each study year (2002-2019)

- Family Available Income in the school neighborhood in 2017
- Number of injured people and number of injured pedestrians per 100 meter of street in the school neighborhood in 2018
- Number of injured people and number of injured pedestrians on weekdays per 10 million km traveled by motor vehicles in the school neighborhood in 2018
- Kilometer of street and kilometer traveled by motor vehicles in the study area around the school in 2018
- Kilometers traveled by motor vehicles in Barcelona in 2003-2017
- Number of total trips and on foot trips, made by people in Barcelona 2003-2018

The other variables included were as follows:

- Group: Intervention schools; Comparison schools
- Period: pre-intervention; post-intervention
- Year: 2002 to 2019

Main characteristics pre-post intervention of the surrounding area for intervention and comparison groups is showed in Table S1.

Statistical analysis

First, a descriptive analysis of the characteristics of the school's surrounding area (neighborhood or study area around the school) for both the intervention and comparison schools, was carried out using the annual mean and its 95% confidence interval (95%CI) and the median and interquartile range. A descriptive analysis of the 6 dependent variables was also performed in each period (pre- and post-intervention) for both the intervention and comparison schools, for each school and for the total number of schools (aggregated), using the annual minimum, maximum, mean with its 95%CI and median with its interquartile range. The results between the intervention and comparison schools, both in the pre- and post-intervention periods, were compared using the non-parametric Wilcoxon rank-sum test (Mann-Whitney).

To compare the results in the post-intervention period versus the pre-period and to assess the effectiveness of the SRTS program in increasing road safety, for each outcome measure a GLMM with Poisson distribution using the logarithm as a link function between expected values and explanatory variables were fitted. The model can be formulated as:

$$\log\big(E(Y_i)\big) = X_i\beta + Z_i\gamma$$

$$Y_i \sim \mathrm{Poi}(\mu_i)$$

where *i* is the observation index, *Y* is the response variable, *X* is the design matrix for fixed effects, is the vector of model coefficients, *Z* is the design matrix for the random effects and is a vector of normally distributed random coefficients.

The explanatory variables included in the model were the group, period, the interaction between group and period, and the year. To obtain a more precise fit, the model was also adjusted by the exposure variables. Among all the available exposure variables, those finally included in the model were: the number of students of the school (log), number of inhabitants in the neighborhood (log), kilometers traveled by motor vehicles in the study area (log), and a quadratic term for the latter effect (the relationship between the volume of traffic and the rate of accidents is not linear). These adjustment variables represented all the sources of variation we wanted to include and offered a better fit compared with: number of students, inhabitants in the neighborhood and kilometers traveled by motor vehicles.

To take into account the correlation between measurements and obtain a valid inference, various random factors were also evaluated: school, matched pair and year, and also random slopes for year at the school level.

For each outcome measure, the final model was selected based on the smallest Akaike Information Criteria, removing non-relevant random components.

Finally, in case the model showed overdispersion, the model was fitted again considering a GLMM with negative binomial distribution.

All models were validated by revising the absence of pattern in the residual plots against the predicted values.

Annual average estimates, relative risks (RR) and their 95% confidence intervals (95%CI) were obtained from the model least square means (LSMEANS, also called empirical marginal means), i.e. point estimates of different levels of interest evaluated as the average of other explanatory variables or random effects. The inverse transformation of the link function was applied to provide the results in the response scale.

To quantify the impact of the intervention, from the RRs, the number of traffic collisions, the number of collisions involving any injured 0-16 year-old, the number of collisions involving any injured 0-16 years-old pedestrians, the number of injured people, the number of 0-16 years-old injured and the number of 0-16 years-old pedestrians injured, in the post-intervention period versus the pre-intervention period were obtained, as the % change = -(1-RR)). Prevented cases were estimated from the difference in the % change between the intervention and the comparison schools, multiplied by the collisions in the first period of the intervention group.

Figure S1. Schools with the Safe Routes to School program in Barcelona and comparison schools included in the study and its 200-meter buffer.

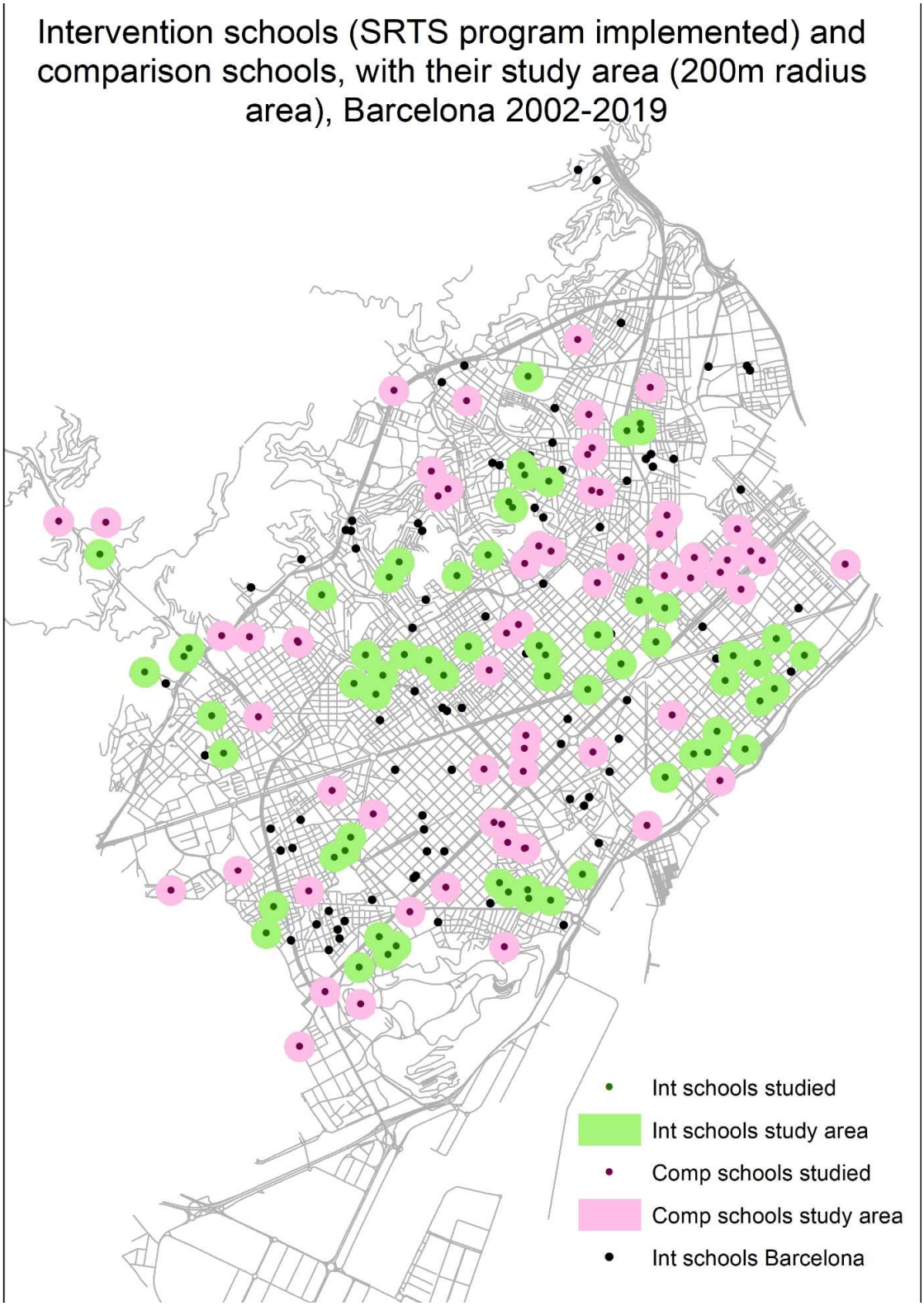

Table S1. Characteristics of the surrounding area (neighborhood or study area defined around the school) of the schools with an SRT program and comparison schools included in the study. Barcelona, 2018.

	Intevention Schools (n= 64)				**Comparison Group Schools** (n= 63)				
Characteristics of the schools' surrounding area	**Mean**	**95%CI**	**Median**	**Interquartile Range**	**Mean**	**95%CI**	**Median**	**Interquartile Range**	**p** [8]
Family Available Income[1]	112.1	[102.4-121.8]	99.9	[82.9-144.4]	99.8	[89.6-110.0]	81.7	[69.3-114.2]	**0.010**
Concentration of people injured [2]	77.9	[63.8-92.1]	60.5	[51.4-89.0]	91.7	[74.2-109.2]	67.9	[51.9-98.9]	0.238
Ratio of injured people by motor vehicle mobility [3]	33.3	[30.0-36.7]	30.2	[24.7-43.0]	35.2	[31.8-38.6]	34.6	[26.2-43.3]	0.296
Concentration of pedestrian injured [4]	7.8	[6.6-9.0]	6.4	[4.0-10.9]	10.1	[8.5-11.7]	8.2	[5.9-12.8}	**0.019**
Ratio of injured pedestrians by motor vehicle mobility [5]	3.9	[3.2-4.5]	3.6	[2.3-4.5]	4.6	[3.8-5.4]	4.3	[2.7-5.6]	0.103
Street kms in the stydy area around the school [6]	3.5	[3.2-3.7]	3.4	[2.9-4.0]	3.5	[3.2-3.8]	3.4	[2.8-4.1]	0.960
Motor vehicle mobility in the study area around the school[7] (millions)	7.2	[5.1-9.3]	3.9	[2.5-8.8]	9.7	[6.9-12.6]	5.5	[3.2-10.3]	0.168

1: Family Available Income in the school neighbourhood in 2017
2: Number of people injured per 100 m of street in the school neighbourhood in 2018
3: Number of people injured on weekdays per 10 million km traveled by motor vehicles in the school neighborhood in 2018
4: Number of injured pedestrians per 100 m of street in the school neighbourhood in 2018
5: Number of injured pedestrians on weekdays per 10 million km traveled by motor vehicles in the school neighborhood in 2018
6: Kilometers of street in a 200-meter radius around the school in 2018
7: Kilometers traveled on weekdays by motor vehicles in a 200-meter radius around the school in 2018
8: Significance of the nonparametric Wilcoxon rank-sum test (Mann-Whitney) comparing intervention and comparison schools

References

1. Saurí E, Sintes E, Truñó M. *Avaluació Del Programa Camí Escolar, Espai Amic*.; 2017. Accessed August 7, 2022. https://institutinfancia.cat/es/mediateca/informe-davaluacio-programa-cami-escolar-espai-amic/

CAPÍTOL 4

Influence of Red and Yellow cards on team performance in elite soccer

4.1 Article

Annals of Operations Research

Influence of Red and Yellow cards on team performance in elite soccer

Llorenç Badiella[1,2], Pedro Puig[2,3], Carlos Lago-Peñas[4], Martí Casals[5,6,7]

[1] Servei Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

[2] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

[3] Centre de Recerca Matemàtica, Cerdanyola del Vallès, Barcelona, Spain

[4] Faculty of Education and Sports Sciences, Universidad de Vigo, Pontevedra, Spain

[5] Sport and Physical Activity Studies Centre (CEEAF), Faculty of Medicine, University

of Vic-Central University of Catalonia (UVic-UCC), Barcelona, Spain
[6] Sport Performance Analysis Research Group, University of Vic-Central University of Catalonia (UVic-UCC), Barcelona, Spain
[7] National Institute of Physical Education of Catalonia (INEFC), University of Barcelona, Barcelona, Spain

Resum

En aquest estudi s'analitzaren els efectes de les targetes vermelles i grogues en la taxa de gols marcats en el futbol d'elit. La mostra de dades estava constituïda per 1826 partits de les cinc principals lligues europees disputats durant la temporada 2017/18. Els esdeveniments ocorreguts en els diferents partits es van estructurar en intervals de 5 minuts. El nombre de gols marcats per interval es van analitzar emprant un model lineal mixt generalitzat amb distribució de Poisson, considerant la presència de dades correlacionades, tenint en compte la informació contextual de l'interval i del partit com a variables explicatives. La informació a nivell partit com ara la fortalesa de cada equip, si juga a casa o la posició a la classificació es va considerar implícitament al model mitjançant una nova variable construïda a partir de les quotes ofertes per diferents cases d'apostes. Com a criteris contextuals el model també incorporà variables com el marcador en curs, temps de joc, targetes vermelles, grogues, etc.

Globalment, es va detectar que després d'una expulsió, la taxa de gols de cada equip canvia significativament, perjudicant l'equip penalitzat i afavorint el seu adversari. Quan el jugador expulsat pertany a l'equip visitant, l'impacte d'una targeta vermella es manté més o menys al llarg del temps. L'efecte de la targeta vermella, en canvi, tendeix a esvair-se amb el temps quan l'afectat és de l'equip és més fort. La taxa de gols marcats també es veu influenciada pel nombre de jugadors amonestats amb targeta groga, essent una mica més baixa per l'equip que té més jugadors amonestats si a més a més va per davant en el marcador.

L'anàlisi presentada permet estimar la taxa de gols acumulada esperada al llarg del temps per a diversos escenaris de targetes vermelles. En particular, si s'expulsa un jugador visitant quan resten 30 min de joc, l'impacte esperat a favor de l'equip local és de 0,39 gols. Si, en canvi, l'expulsat és de l'equip local, l'impacte és de 0,50 gols a favor de l'e-

quip visitant. Els entrenadors i analistes podrien utilitzar aquesta informació per establir objectius particulars i preparar-se per a escenaris de superioritat o inferioritat numèrica.

Annals of Operations Research
https://doi.org/10.1007/s10479-022-04733-0

ORIGINAL RESEARCH

Influence of Red and Yellow cards on team performance in elite soccer

Llorenç Badiella[1,2] · Pedro Puig[2,3] · Carlos Lago-Peñas[4] · Martí Casals[5,6,7]

Accepted: 12 April 2022

Abstract

The aim of the current study is to analyze the effects of red and yellow cards on the scoring rate in elite soccer. The sample was composed of 1826 matches in the top five European leagues. All events were structured in 5-min intervals and were analyzed by means of a Generalized Linear Mixed Model with Poisson distribution, considering the presence of correlated data, where the dependent variable is represented by scoring rate. Team strength and home advantage were considered implicitly by means of a transformation of the betting odds for each game. The model also took into account the goal difference and time evolution. Overall, we found that after a sending off, each team's scoring rate changes significantly, damaging the penalised team and favouring its opponent. When the player who is sent off belongs to the Away team, the impact of a red card is more or less maintained over time intervals. The red card effect, on the other hand, tends to fade over time when the affected team is stronger. The relative difference in scoring rates is also affected by the goal difference and the difference in booked players, being slightly lower for the team going ahead if it has more booked players. Our approach allows estimating the expected cumulative soring rate through time for various red card scenarios. Particularly if a red card is given with 30 min of remaining time, the expected impact is 0.39 goals if the guilty player is on the visiting team and 0.50 if he plays for the home team. Coaches and analysts could use this information

✉ Llorenç Badiella
badiella@mat.uab.cat

1 Servei Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

2 Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Barcelona, Spain

3 Centre de Recerca Matemàtica, Cerdanyola del Vallès, Barcelona, Spain

4 Faculty of Education and Sports Sciences, Universidad de Vigo, Pontevedra, Spain

5 Sport and Physical Activity Studies Centre (CEEAF), Faculty of Medicine, University of Vic-Central University of Catalonia (UVic-UCC), Barcelona, Spain

6 Sport Performance Analysis Research Group, University of Vic-Central University of Catalonia (UVic-UCC), Barcelona, Spain

7 National Institute of Physical Education of Catalonia (INEFC), University of Barcelona, Barcelona, Spain

Published online: 16 May 2022

to establish objectives for players and teams in training and matches and to be prepared for these very different scenarios of numerical superiority or inferiority.

Keywords Soccer · Performance analysis · Red and yellow cards · Generalized linear mixed model · Poisson distribution

1 Introduction

The analysis of strategies plays a major role in soccer. Managers use scouting, video footage, and soccer data feeds to collect information about tactics and player performance. Performance analysis in soccer requires objective recording and examination of behavioral events involving one or more players during training or competition. The primary goal of performance analysis is to provide information to coaches and players about player and/or team performance in order to plan subsequent training sessions to improve performance or to support preparations for the next match (McGarry et al. 2013).

However, the validity of the data generated by most studies, especially regarding the prescriptive function of tactical analysis, may be questionable due to the lack of assessment of the role of the opposing team in their analysis (McGarry et al. 2013). Other critical issues related to the conceptual and methodological shortcomings of contemporary research that require attention in future research may include the development of a theoretical framework, research on critical behaviors, consideration of situational and game contexts, and the inclusion of spatial and temporal dynamics (Bornn et al. 2018).

At the highest level of competition, team performances are generally affected by the smallest of details that can imply considerable advantages in the quest for success. It is relatively common for referees to decide to send off soccer players at all levels and in all competitions and it is therefore likely that such events could change the outcome of matches. Red cards are a significant event that can influence the outcome of a game. Previous research demonstrates that advantaged teams exhibit better team performance after a player has been sent off and that teams modify their tactics and style of play in response to this new match scenario (Bar-Eli et al. 2006; Carling and Bloomfield 2010; Lago-Peñas et al. 2016). A sending off represents a disadvantage and can trigger negative momentum leading to decreased performance. Previous literature suggests it may reduce self-confidence, increase demoralization and reduce the team-cohesion effect (Bar-Eli and Tenenbaum 1989; Bar-Eli et al. 1996).

Team performances have been analyzed previously by different authors, using data from different countries or competitions and based on substantially different quantitative approaches guided by the type of data available (Anders and Rotthoff 2011; Bar-Eli et al. 2006; Carmichael and Thomas 2005; Chowdhury 2015; Gómez-Déniz et al. 2019; Greenberg 2015; Mattera 2021; Mechtel et al. 2011; Titman et al. 2015). Regarding the effects and occurrence of red cards, most of the literature finds, as might be expected, that red cards produce a decrease in the performance of the penalized team while causing an increase in that of the opposing team, either in terms of the probability of winning the game or in terms of expected goals, points achieved or scoring rate; although the magnitude of the effect can vary considerably. One study performed on 743 Bundesliga matches suggested that red cards weaken the affected team in terms of goals conceded and final score following the sending off (Bar-Eli et al. 2006). The study subsequently highlighted that the scoring or eventually winning the match chances of a team that has had a player sent off are substantially reduced.

Carling and Bloomfield (2010) examined the effects of an early dismissal (after 5-min play) on work-rate in a professional soccer match. Their results suggest that playing with ten men leads to a greater total distance being covered than normal (particularly in moderate-intensity activities) revealing shorter recovery times between high-intensity activities. Furthermore, this study suggests that in 11 versus 11 situations, players might not always use their full physical potential, because they are able to increase their overall work-rate when reduced to 10 players. The study also suggests that a team with 10 players should change its strategy and tactical set-up in order to minimize the effects of the higher levels of fatigue. Lago-Peñas et al. (2016) found that playing 11 versus 10 increases the time spent in possession, number of total passes, short passes, total touches and the percentage of successful passes compared with playing 11 versus 11. Advantaged teams also spent less time defending. The punished team performed worse in all variables after the dismissal. Sapp et al. (2019) also found similar rates of aggressive play in the top English soccer leagues, possibly due to a standardized refereeing style.

While everyone understands the circumstantial effect of red cards (the immediate loss of a player), statistical analysis helps us better understand the extent to which teams are affected when they receive one (Chowdhury 2015). The main methodological criticism of previous research is the use of aggregate data at match level (e.g. in Anders and Rotthoff (2011), Carmichael and Thomas (2005) and Chowdhury (2015) that only provides a partial view of the problem without considering the temporal coherence of the events. Other authors used a similar approach including information from the remaining time at the issuing of red cards (e.g. in Albanese et al. 2020; Červený et al. 2018; Mechtel et al. 2011).

In a similar manner, some other studies have compared the occurrence of post-card goals versus the pre-card situation (e.g. in Bar-Eli et al. 2006; Caliendo and Radic 2006; Ridder et al. 1994) or analyzed time to goal. These approaches do not suffer the aforesaid bias but the series analyzed are relatively small: only games with players sent off are considered and no other event data is included.

In much of the literature home/away effects are evaluated. Adjustments for team strength are also common, using the previous year's rankings (Anders and Rotthoff 2011; Mechtel et al. 2011), rankings provided by third parties (Červený et al. 2018; Albanese et al. 2020) or betting odds (Červený et al. 2018; Titman et al. 2015). Only in a small number of studies have authors examined the effect of yellow cards in combination with red cards (Anders and Rotthoff 2011; Titman et al. 2015).

Considering all previous limitations, the objective of this study is to propose a more valid statistical approach to analyze the consequences of red and yellow cards. Our proposal is based on the analysis of a large series of games using aggregated data per short time intervals from a dataset of event data. The response variable is the number of goals scored in an interval, whereas the explanatory variables are a series of variables that characterize the match, the team and the game situation at the beginning of the interval. Hence there is temporal coherence between the response variable and explanatory variables (including the goal difference, number of players that have received yellow or red cards, remaining time and time elapsed since a superiority scenario); injury time can be included as a special interval and Team strength is considered using information from betting odds. In this way, our method will allow us to estimate the scoring rate over time for specific red card scenarios and team characteristics, and in consequence will provide a framework to quantify the average impact of having a player sent off.

2 Data processing

The original database was taken from (Pappalardo et al. 2019), version 3. It consists of the 1826 games played in the 2017/2018 season in the five biggest European soccer leagues (French Ligue 1, Spanish La Liga, English Premier League, Italian Serie A and German Bundesliga). The dataset consists of spatio-temporal match events, from which we extracted goals, red and yellow cards and substitutions for each team and its opposing team. The sample was classified into 5-min periods. The response variable is the number of goals scored in a given interval for a particular team. In order to characterize the particularities of the match and the situation at each interval, we considered the following explanatory variables:

1. Home advantage (*HomeAdv*) and competition (*Competition*). Home advantage is a pervasive phenomenon in sport and soccer. The variable *HomeAdv* identifies whether a team is playing as a Home (H) or Away (A) team, whereas the variable *Competition* identifies each soccer league: French Ligue 1, Spanish La Liga, English Premier League, Italian Serie A and German Bundesliga.
2. Team strength and winning odds (*WinOdds*). To adjust the model more accurately, information has been included on team strength by employing a new variable constructed on the basis of betting odds. As gambling on professional soccer is a highly liquid market, it can be assumed that betting odds include the full combination of information about team strength and other game-related effects. To a certain extent, betting odds are much more accurate than any other measure of team strength as it can be assumed that these take into account, when relevant, such effects as (de)motivation prior to the game, form, stadium, rivalry between clubs, home advantage, competition, etc. Hence they are more sensitive than any other method for quantifying team strength, such as the Elo ratings or points at the end of the season. The Football-Data.co.uk (2020) database of European soccer leagues results was used to obtain the betting odds from different bookmakers (Bet and Win, Bet365 and William Hill) for each game. These values were transformed into probabilities by eliminating the bookmaker overround in accordance with the standard methodology described by Wunderlich and Memmert (2018), assuming a proportional margin for all outcomes of a game:
   Let $odds_{i,j}$ be the betting odds offered by bookmaker i ($i = 1 \ldots B$) of outcomes $j = 1$ (home win), 2 (draw) and 3 (away win). Then the margin m_i for the ith bookmaker can be calculated as:

$$m_i = \sum_{j=1}^{3} \frac{1}{odds_{i,j}} \tag{1}$$

   and the probabilities $p_{i,j}$ and p_j for outcome j are:

$$p_{i,j} = \frac{1}{m_i odds_{i,j}} \tag{2}$$

$$p_j = \sum_{i=1}^{B} \frac{p_{i,j}}{B} \tag{3}$$

   The resulting probabilities p_j add up to 1. The probability of a draw is distributed equally to both teams, obtaining probabilities of a win/draw for each team and game. Thus, $p_W = p_1 + p_2/2$ for a team playing at home and $p_W = 1 - (p_1 + p_2/2)$ for a team playing away. These probabilities are transformed using a logit function to avoid floor/ceiling effects, and linearize the relation with the goal rate, ultimately obtaining a

quantitative measure of the chances of a win or loss. Hereinafter this variable is called *WinOdds*, which is formally the log-odds of the probability of victory (contemplating part of the probabilities of a draw): $WinOdds = \log(p_W/(1 - p_W))$. In fact, for any game between any two teams A and B, $WinOdds_A + WinOdds_B = 0$. This means that the models only require the specific term of the team in question, without the need to include information about the opponent.

3. Goal difference (*GoalDif*). The goal difference at the beginning of the interval has been included as a quantitative variable. The purpose of including this variable is to adjust the scoring rate by the current goal difference, and to evaluate the extent to which the effect of a red or yellow card depends on this variable.
4. Red cards (*RedDif*) and yellow cards (*YellowDif*). For each interval of time, we obtained the number of players to have received red or yellow cards. Thus, the superiority scenario at the beginning of the interval (0: 11 vs. 11, +1: 11 vs. 10, −1: 10 vs. 11, etc.) is considered using a categorical variable named *RedDif*. In the case of yellow cards, a quantitative variable measuring the difference in number of booked players between opposing teams has been included. A relevant consideration regarding this last variable is that the number of booked players has been corrected when one of the players on a yellow card is substituted or sent off.
5. Current clock time (*ClockTime*) and superiority scenario time (*ScenarioTime*). In order to take into account game evolution, the current clock time at the beginning of each interval has been considered. The elapsed time from the beginning of the red card scenario has also been included, i.e. the elapsed time from the last shown red card in either team. If no red cards are shown *ScenarioTime* is set to 0. Both variables are regarded as continuous variables and are recorded in minutes. However, for comparability and computational reasons, we standardized them to take values in the interval [0, 1] by dividing by the total match duration. Note that in practice, *Clocktime* provides the same information as the remaining time.
6. Interval length (*IntervalLength*) and injury time (*InjuryTime*). The database has included specific intervals for each of the periods of time added on for stoppages. The inclusion of this phase is more than relevant as this is when some of the most intense soccer gets played, generating a higher rate of goals and yellow and/or red cards, and it tends to be the tightest games that produce the longest injury times. The dataset for each team and match consists of 20 intervals, 18 for normal time and 2 for stoppage time. Although the intervals are generally five minutes long, injury times are of variable duration, from just a few seconds to periods sometimes greater than 10 min. In order to adequately weight the number of goals scored in these intervals, a weighting variable was created (*IntervalLength*), dividing the interval duration in minutes by 5. Finally, injury time has been taken into account by means of a categorical variable named *InjuryTime* with different values for regular intervals (*RT*) or injury times at the end of the first or second half (*I1* and *I2*).

We therefore obtained a database made up of 1826 games. For each game we have information about the two teams and 20 time intervals, with a total of 73,040 observations.

Table 1 shows the processed data corresponding to the English Premier League game between Chelsea and Burnley on 12/08/2017 that ended 3-2 to the visiting team. In this particular game, the home team had two players sent off after 13 and 81 minutes. The first was a direct red card and the latter was for a second bookable offense. It can be observed that the data on *WinOdds*, goal difference, red and yellow cards for one team are the same measure for the other team with the opposite sign.

Table 1 Game details per time interval for the Premier League game between Chelsea and Burnley in the 2017/18 season

Home Adv	WinOdds	Period	Injury time	Clock time	Interval duration	Goal	Goal Dif	Yellow Dif	Red Dif	Superiority Time
H	1.74	1H	RT	0.0	5.0	0	0	0	0	0.0
H	1.74	1H	RT	5.0	5.0	0	0	1	0	0.0
H	1.74	1H	RT	10.0	5.0	0	0	1	0	0.0
H	1.74	1H	RT	15.0	5.0	0	0	1	−1	2.0
H	1.74	1H	RT	20.0	5.0	0	0	2	−1	7.0
H	1.74	1H	RT	25.0	5.0	0	−1	2	−1	12.0
H	1.74	1H	RT	30.0	5.0	0	−1	2	−1	17.0
H	1.74	1H	RT	35.0	5.0	0	−1	2	−1	22.0
H	1.74	1H	RT	40.0	5.0	0	−2	2	−1	27.0
H	1.74	1H	I1	45.0	3.3	0	−3	2	−1	32.0
H	1.74	2H	RT	48.3	5.0	0	−3	2	−1	37.0
H	1.74	2H	RT	53.3	5.0	0	−3	2	−1	42.0
H	1.74	2H	RT	58.3	5.0	0	−3	2	−1	47.0
H	1.74	2H	RT	63.3	5.0	0	−3	1	−1	52.0
H	1.74	2H	RT	68.3	5.0	1	−3	0	−1	57.0
H	1.74	2H	RT	73.3	5.0	0	−2	0	−1	62.0
H	1.74	2H	RT	78.3	5.0	0	−2	0	−1	67.0
H	1.74	2H	RT	83.3	5.0	0	−2	0	−1	72.0
H	1.74	2H	RT	88.3	5.0	1	−2	−1	−2	2.7
H	1.74	2H	I2	93.3	4.5	0	−1	0	−2	7.7
A	−1.74	1H	RT	0.0	5.0	0	0	0	0	0.0
A	−1.74	1H	RT	5.0	5.0	0	0	−1	0	0.0
A	−1.74	1H	RT	10.0	5.0	0	0	−1	0	0.0
A	−1.74	1H	RT	15.0	5.0	0	0	−1	1	2.0
A	−1.74	1H	RT	20.0	5.0	1	0	−2	1	7.0
A	−1.74	1H	RT	25.0	5.0	0	1	−2	1	12.0
A	−1.74	1H	RT	30.0	5.0	0	1	−2	1	17.0
A	−1.74	1H	RT	35.0	5.0	1	1	−2	1	22.0
A	−1.74	1H	RT	40.0	5.0	1	2	−2	1	27.0
A	−1.74	1H	I1	45.0	3.3	0	3	−2	1	32.0
A	−1.74	2H	RT	48.3	5.0	0	3	−2	1	37.0
A	−1.74	2H	RT	53.3	5.0	0	3	−2	1	42.0
A	−1.74	2H	RT	58.3	5.0	0	3	−2	1	47.0
A	−1.74	2H	RT	63.3	5.0	0	3	−1	1	52.0
A	−1.74	2H	RT	68.3	5.0	0	3	0	1	57.0
A	−1.74	2H	RT	73.3	5.0	0	2	0	1	62.0
A	−1.74	2H	RT	78.3	5.0	0	2	0	1	67.0
A	−1.74	2H	RT	83.3	5.0	0	2	0	1	72.0
A	−1.74	2H	RT	88.3	5.0	0	2	1	2	2.7
A	−1.74	2H	I2	93.3	4.5	0	1	0	2	7.7

Table 2 Betting odds and corrected probabilities for the Premier League game between Chelsea and Burnley in the 2017/18 season

	Odds			Corrected probabilities		
Bookmaker	Home	Draw	Away	Home (%)	Draw (%)	Away (%)
BW	1.22	6.5	12.5	77.80	14.60	7.60
B365	1.25	6.5	15	78.40	15.10	6.50
WH	1.25	5.5	13	75.60	17.20	7.30

As for the *WinOdds* variable, the odds on a win, draw or loss for the home team offered by different bookmakers are shown in Table 2.

The estimated margin is 5.35%, 2.05% and 5.87% for each bookmaker respectively. The probabilities of a win/draw for the home team given by each bookmaker are 85.1%, 85.9% and 84.1%. The average p_W is 85% and the *WinOdds* for that team is 1.74. So, for the visiting team, $p_W = 15\%$ and $WinOdds = -1.74$.

3 Methodological approach

To study the influence of red cards on the performance of the different teams, the number of goals scored was modeled in 5 min intervals, taking into account contextual and other variables that could be predictive of the rate of goals scored, including the variables described before, quadratic effects and interactions.

The most natural model for count data is a generalized linear model (GLM) with Poisson distribution using the logarithm as a link function between expected values and explanatory variables (McCullagh and Nelder 2019). This model assumes that events occur randomly at a constant rate over the observed time period conditionally on the explanatory variables. It can be formulated as follows:

$$\begin{aligned} \log(\mathrm{E}(Y_i)) &= X_i\beta + \log(w_i) \\ Y_i &\sim \mathrm{Poi}(\mu_i) \end{aligned} \tag{4}$$

where i is the observation index, Y is the response variable, X is the design matrix for fixed effects, β is the vector of model coefficients and w is an offset variable that if considered, enables modulation of the expectation in relation to its magnitude (in our case, the interval length is included as an offset). This model can be fitted by restricted maximum likelihood. It assumes that observations are independent and that the distribution is equidispersed, i.e. $\mathrm{V}[Y_i] = \mathrm{E}[Y_i] = \mu_i$.

In the present case, given the repeated measures that appear in the data, observations of the same individuals (teams, match, etc.) may be correlated. In order to adapt the former model to the presence of correlated data it is common to include random effects (also called random intercepts) associated to experimental units that are sampled repeatedly. In longitudinal data, it is also common to evaluate whether there are random time trends associated to those units (also called random slopes). These models can be formulated as:

$$\begin{aligned} \log(\mathrm{E}[Y_i|\gamma]) &= X_i\beta + Z_i\gamma + \log(w_i) \\ Y_i|\gamma &\sim \mathrm{Poi}(\mu_i) \\ \mathrm{V}[Y_i|\gamma] &= R \end{aligned} \tag{5}$$

where Z is the design matrix for the random effects, γ is a vector of normally distributed random coefficients (or random effects) and R is a matrix that contains the variance functions of the model (determined by the distribution of the data) and when necessary, other terms related to the correlation structure of the data. The matrix R allows to take into account other correlation structures that random effects cannot account for, such as negative correlations between groups of units. When there are no effects in R, these models can be fitted by maximum likelihood and are referred as conditional GLMMs (Lee et al. 2018). Alternatively they can also be fitted using a pseudo-likelihood approach (Molenberghs and Verbeke 2005). When part of the correlation structure is also specified through the matrix R, the models are referred as marginal GLMMs and can only be fitted by the pseudo-likelihood method.

On the other hand, the Poisson distribution has the property that the expectation is equal to the variance. The equidispersion assumption can be evaluated allowing for the presence of a free dispersion parameter, estimated using Pearson's χ^2 statistic, so that $\mathrm{V}[Y_i] = \mu_i\phi$:

$$\hat{\phi} = \sum_i \frac{(y_i - \mu_i)^2}{\mu_i} \tag{6}$$

In order to select the most appropriate model, different steps were applied. First, in order to simplify the fixed part of the model (i.e. explanatory variables), a hierarchical backward stepwise selection procedure based on the smallest Akaike Information Criteria (AIC) index was performed. The selection was hierarchical in the sense that the main effects were not removed whenever a higher order term involving this variable was present. Secondly, the best variance and covariance structure was chosen from a set of proposals considering AIC (when available) and pseudo-AIC indexes. Conditional GLMMs can be compared using the AIC criterion, however marginal GLMMs can only be compared with other models when they share the same fixed and random effects (i.e. if they only differ in the R matrix) by means of a pseudo-AIC. Finally, in order to quantify the explained variability of the initial and final models a pseudo-R^2 for GLMMs (Nakagawa and Holger 2013) and RMSE were also computed.

The model estimates presented in the results section are least square means (LSMEANS, also called empirical marginal means) estimations, i.e. point estimates of different levels of interest evaluated at the average of other explanatory variables or random effects. The inverse transformation of the link function was applied in order to provide the results in terms of scoring rates. In graphical representations, 95% confidence intervals are also shown.

After the initial exploration, we decided to exclude from the analysis such intervals where more than one player had been sent off from either team, since these scenarios are very infrequent and scoring rates cannot be estimated with enough precision. This affected 196 observations out of 73,040.

The model was validated by revising the lack of pattern in the residual plots against the predicted values. Moreover, we validated our proposal based on 5-min intervals using shorter lengths of 2 and 3 min. The results obtained from these models are essentially the same, but offer slightly more sensitivity related to a larger sample size and require much greater computational effort when evaluating random components. Since, by dividing games in 5-minutes intervals the count of goals scored per interval is very rarely 2 or more, We also used a GLMM with Bernoulli distribution and a logit link to validate the model (with response values greater than 1 replaced with 1). Given the low event rate, models fitted with a logit or a log link produce similar results.

Table 3 Descriptive summary for goals, yellow, red cards and substitutions per match and match location for the different competitions in the dataset

Venue	Competition	Goals	1st Yellow cards	2nd Yellow cards	Direct Red cards	Total Red cards	Substituted players
Home	England	1.53 (1.34)	1.51 (1.27)	0.02 (0.15)	0.03 (0.16)	0.05 (0.23)	2.73 (0.57)
	France	1.53 (1.35)	1.76 (1.21)	0.04 (0.19)	0.07 (0.26)	0.11 (0.31)	2.84 (0.43)
	Germany	1.60 (1.28)	1.56 (1.28)	0.03 (0.17)	0.04 (0.20)	0.07 (0.26)	2.89 (0.34)
	Italy	1.46 (1.31)	1.86 (1.27)	0.04 (0.21)	0.05 (0.24)	0.10 (0.32)	2.92 (0.27)
	Spain	1.55 (1.38)	2.27 (1.48)	0.06 (0.24)	0.05 (0.21)	0.11 (0.32)	2.89 (0.34)
Away	England	1.15 (1.18)	1.59 (1.28)	0.03 (0.16)	0.03 (0.18)	0.06 (0.23)	2.75 (0.52)
	France	1.19 (1.13)	2.01 (1.26)	0.05 (0.21)	0.07 (0.25)	0.12 (0.32)	2.83 (0.41)
	Germany	1.19 (1.14)	1.82 (1.22)	0.04 (0.20)	0.03 (0.16)	0.07 (0.25)	2.90 (0.34)
	Italy	1.22 (1.19)	2.11 (1.31)	0.07 (0.26)	0.07 (0.26)	0.14 (0.37)	2.93 (0.28)
	Spain	1.15 (1.19)	2.63 (1.53)	0.05 (0.22)	0.03 (0.17)	0.08 (0.28)	2.89 (0.33)

Values represent mean and standard deviation

4 Results

Table 3 presents a descriptive summary of the number of goals, red and yellow cards and substitutions per game for the different competitions in the countries in the sample and depending on whether they are home or away teams. Although the data used are based on 5-min intervals, the diagrams show overall rates.

Figure 1 shows the scoring, card and substitution raw rates per interval of time depending on whether the teams are at home or away. The average number of players sent off per game is 0.18. Of these, 51.5% are direct red cards, whereas 48.5% correspond to second bookings. The average time when sendings off occur is at minute 67.4. In particular, second bookings occur on average at minute 74.4 and direct red cards at minute 60.8. An average of 3.84 players receive a yellow card per game, approximately 20% of whom are substituted before the game ends, while 2.3% receive a second booking and are sent off.

4.1 Model fitting

The following models were initially evaluated: a model considering all explanatory variables, quadratic terms for all quantitative variables, and all interactions involving *RedDif* or *YellowDif* (M1) and a simplified version of M1 after a hierarchical backward stepwise selection procedure (M2). Subsequently a set of models was fitted to evaluate M2 with different variance components: random intercepts for Team, Match or Match nested in Team (M3, M4, M5); random intercepts and time slopes for Team, Match or Match nested in Team (M6, M7, M8); including the covariance between opposing teams for the same interval (M9) and the final model (M10) that considers all the random components found to be relevant in the previous models.

Table 4 shows the RMSE, AIC and pseudo-AIC fit indexes for the different evaluated models. M2 is the model with the smallest AIC considering independent observations. The variance components fitted in models M4, M5, M7 and M8 are not relevant, essentially because the presence of *WinOdds* in the model eliminates any variability that may exist

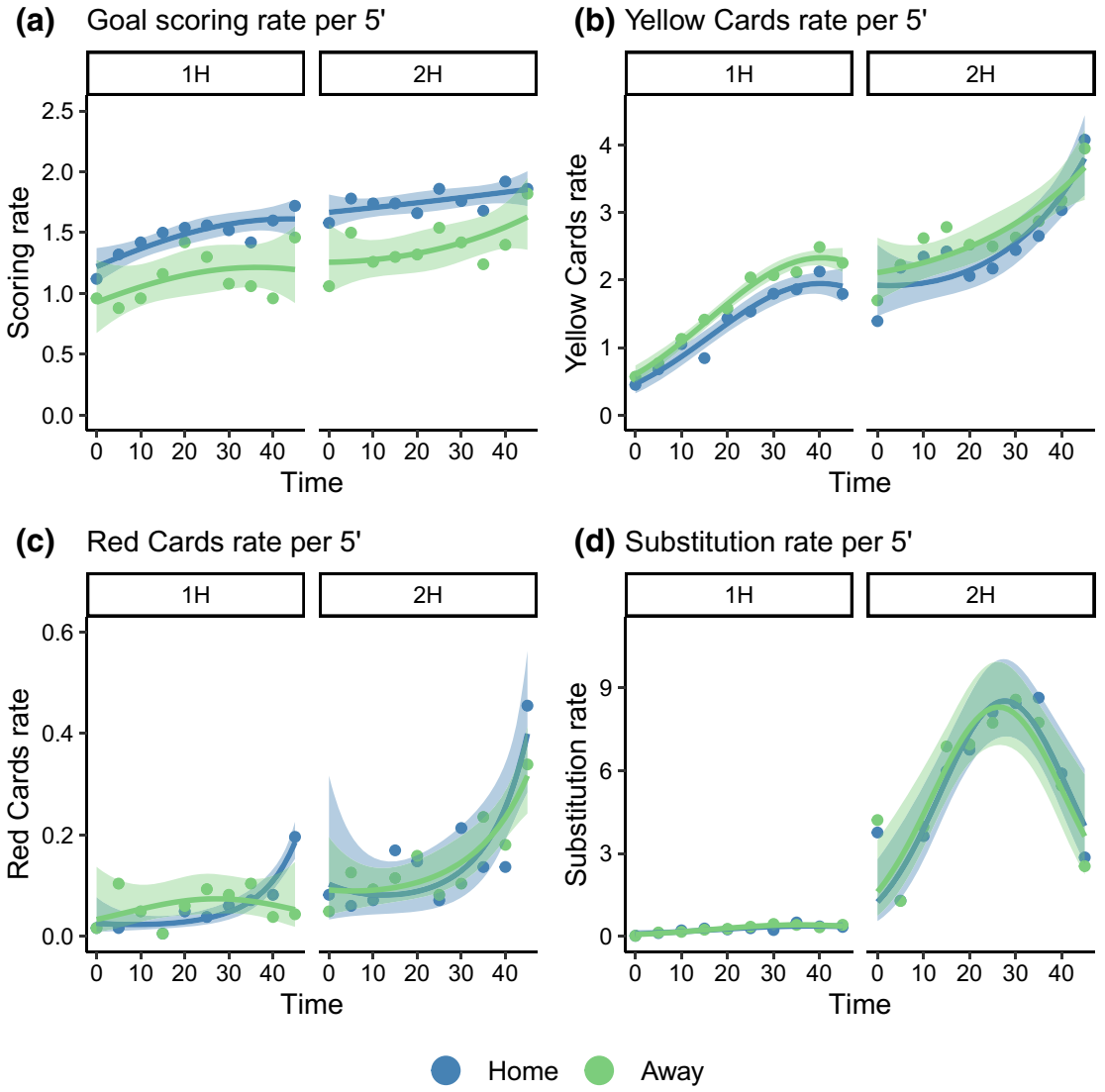

Fig. 1 **a** Goals, **b** red cards, **c** yellow cards and **d** substitution rates for Home and Away teams over time and match period. Rates are adjusted to reflect the total number of events if the rate was constant for the full duration of the match. Smooth fits using the loess method are shown as well as 95% confidence intervals (Cleveland, 1979)

Table 4 Fit indexes for different model proposals

Model	Variance components	AIC	Pseudo-AIC	RMSE
M2:	None (independent observations)	35475.1	411240.7[a]	0.06479
M3:	Team (random intercepts)	35470.8	410817.2[b]	0.06471
M9:	Interval|Match (random intercepts)	NA	411204.4[a]	0.06479
M10:	Team (random intercepts) and Interval|Match (random intercepts)	NA	410750.5[b]	0.06470

[a]Pseudo-AIC with different letters cannot be compared
[b]Models not presented have irrelevant variance components and do not improve the model fit

Table 5 Results of the Poisson generalized mixed model, for the analysis of the 5-minute scoring rate

Variable	M1			M10		
	Estimate	Std. error	*P* value	Estimate	Std. error	*P* value
Intercept	−2.958	0.055	< 0.001	−2.958	0.043	< 0.001
HomeAdv—Home	−0.025	0.031	0.420			
Competition—E	−0.045	0.045	0.318			
Competition—F	0.006	0.044	0.886			
Competition—G	0.061	0.047	0.196			
Competition—I	−0.039	0.045	0.381			
Extratime—I1	0.092	0.113	0.416			
Extratime—I2	0.124	0.079	0.119			
WinOdds	0.421	0.020	< 0.001	0.397	0.018	< 0.001
$WinOdds^2$	0.008	0.012	0.513			
GoalDif	−0.038	0.014	0.006	−0.039	0.013	0.003
$GoalDif^2$	0.008	0.005	0.103	0.009	0.005	0.067
RedDif -1	−1.821	1.177	0.122	−1.496	1.114	0.179
RedDif +1	1.498	0.499	0.003	1.623	0.479	0.001
YellowDif	0.054	0.109	0.623	0.010	0.016	0.532
$YellowDif^2$	−0.007	0.008	0.350			
ClockTime	0.710	0.213	0.001	0.601	0.193	0.002
$ClockTime^2$	−0.354	0.221	0.109	−0.227	0.193	0.239
YellowDif*RedDif -1	−0.207	0.126	0.101			
YellowDif*RedDif +1	−0.216	0.109	0.048			
WinOdds*RedDif -1	−0.001	0.138	0.995	0.089	0.125	0.476
WinOdds*RedDif +1	−0.309	0.088	0.001	−0.267	0.086	0.002
$WinOdds^2$ *RedDif -1	0.096	0.085	0.256			
$WinOdds^2$ *RedDif +1	0.051	0.062	0.416			
GoalDif*RedDif -1	0.034	0.076	0.656			
GoalDif*RedDif +1	0.010	0.056	0.866			
$GoalDif^2$ *RedDif -1	0.018	0.027	0.505			
$GoalDif^2$ *RedDif +1	0.018	0.020	0.386			
ClockTime*RedDif -1	1.836	3.689	0.619	1.125	3.547	0.751
ClockTime*RedDif +1	−4.021	1.764	0.023	−4.382	1.658	0.008
$ClockTime^2$ *RedDif -1	−0.922	2.762	0.739	−0.469	2.665	0.860
$ClockTime^2$ *RedDif +1	3.107	1.417	0.028	3.554	1.340	0.008
ScenarioTime*RedDif -1	5.017	2.106	0.017	5.194	2.056	0.012
ScenarioTime*RedDif +1	0.102	1.115	0.927			
$ScenarioTime^2$ *RedDif -1	−8.072	3.142	0.010	−8.062	3.071	0.009
$ScenarioTime^2$ *RedDif +1	0.114	1.418	0.936			
WinOdds*YellowDif	−0.003	0.019	0.882			
$WinOdds^2$ *YellowDif	0.003	0.012	0.790			
GoalDif*YellowDif	−0.001	0.011	0.939	−0.001	0.010	0.879

Table 5 continued

Variable	M1			M10		
	Estimate	Std. error	*P* value	Estimate	Std. error	*P* value
GoalDif^2 *YellowDif	−0.007	0.004	0.060	−0.008	0.004	0.034
ClockTime*YellowDif	0.120	0.565	0.832			
ClockTime^2 *YellowDif	0.152	0.477	0.750			
ScenarioTime*YellowDif	−0.138	0.464	0.767			
ScenarioTime^2 *YellowDif	−0.209	0.430	0.626			
Covariance team				0.011	0.005	
Covariance interval|Match				−0.025	0.005	
Dispersion	0.969			0.991		
Degrees of freedom	72800				72728	
RMSE		0.06475		0.06470		
$R^2_{GLMM(m)}$	6.16%			6.02%		

Model M1 includes the explanatory variables, quadratic terms and potential interactions of interest, without variance components. Model M10 consists of those variables remaining in the model after a hierarchical backward stepwise selection procedure including the variance components for Team and Interval|Match. For the categorical variable *RedDif*, the reference category is 0 (11 vs. 11)

between games whereas time variables capture the potential autocorrelation between consecutive observations of the same game. Model M3 considers a random effect associated to the Team that turns out to be relevant. This component might reflect playing styles and other team-related characteristics that *WinOdds* does not capture. However adding random slopes (M6) does not improve the fit. Model M9 includes the covariance between opposing teams for the same interval. This component is also relevant and has a negative sign, reflecting a negative correlation between both teams at the interval level. The final model M10 considers simultaneously both covariance components included in models M3 and M9. Adding a free dispersion parameter does not improve the model, since the empirical dispersion in model M10 is 0.991 (close to 1).

Table 5 presents the results of models M1 and the selected final model M10 obtained after fitting a Poisson Generalized Mixed Model with the variance components described earlier.

The final model detects a highly relevant effect of the *WinOdds* criterion. There is no effect associated to competition or playing as a home team as this information is in fact already reflected by the aforesaid variable. The model also depicts large differences associated to playing with a player less or having an extra player and several interactions between this variable and other terms: with *WinOdds*, *ClockTime*, *ScenarioTime* and their respective quadratic terms. As for Yellow cards, a slight interaction with the quadratic term for goal difference is detected. In average terms, the model reveals that the scoring rate is approximately 0.065 goals per 5-min interval. The difference between teams when *WinOdds* is 0.38 versus −0.38 is around 37%, these values represent the difference in favor of the home side, since the average *WinOdds* for Home teams is 0.38. This ratio is maintained throughout the game. However, the scoring rate does have an upward tendency, right at the start of the game the average scoring rate is 0.054 and reaches rates of 0.079. The relative performance between teams also depends on the goal difference and the difference in booked players. The winning team's performance drops if it has more booked players, see Fig. 2. With equal numbers of

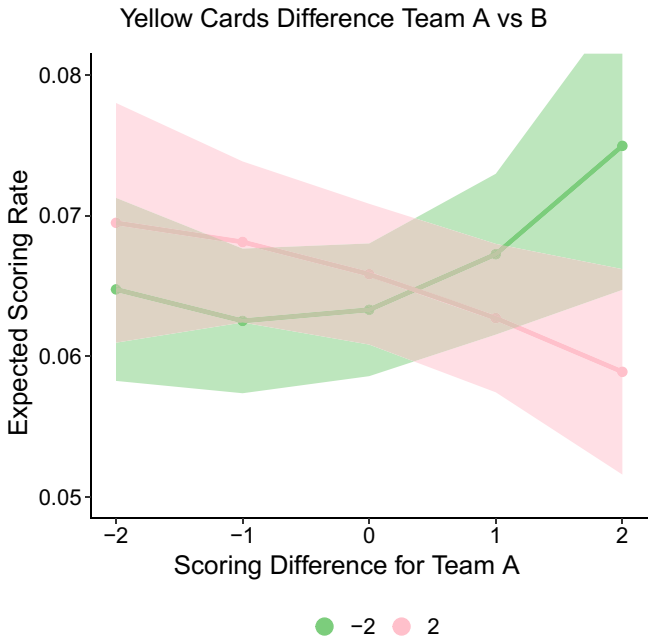

Fig. 2 Scoring rate for different number of players warned by Scoring difference

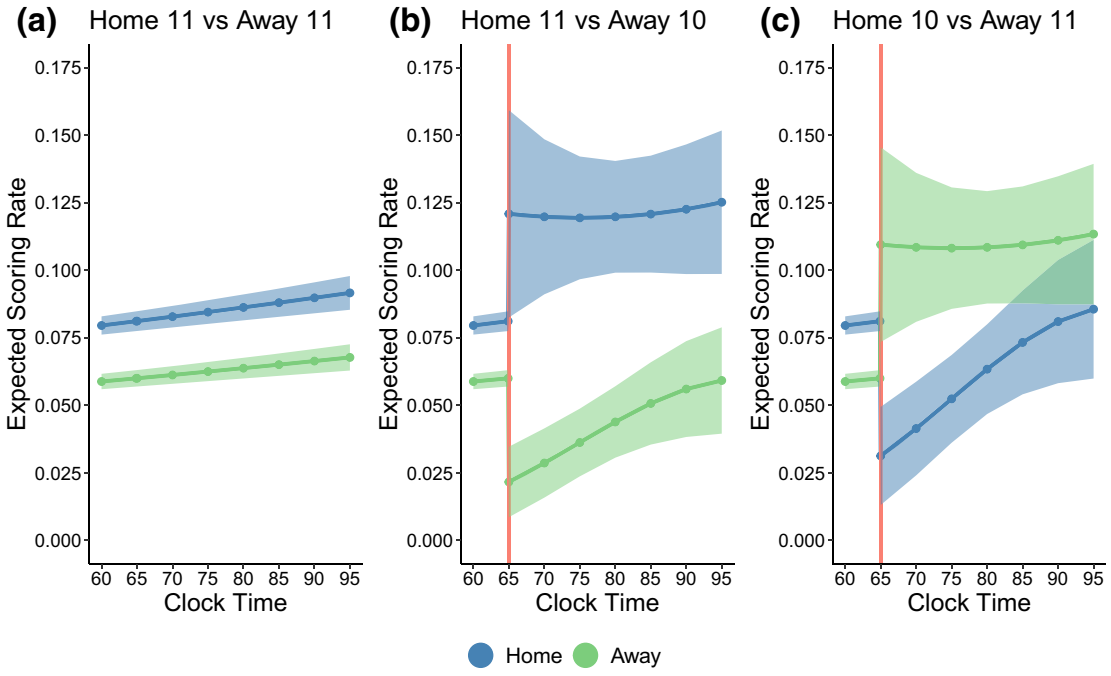

Fig. 3 Scoring rate for different Red Card Scenarios for Home and Away teams through time **a** 11 versus 11, **b** 11 versus 10 and **c** 10 versus 11, respectively

booked players (not shown in the plot), the scoring rate is slightly lower (around 10%) when the team is one or two goals ahead.

The effect over time of sending off a player is presented graphically. To this end, Fig. 3 presents the expected scoring rates and their 95% confidence intervals at different times for the following situations: no players sent off and a player sent off at minute 65 for teams with either *WinOdds* = 0.38 or *WinOdds* = −0.38. The idea behind these particular estimates is again to reproduce the situation where a Home or an Away player is sent off, and taking into account that the average time for a player to be sent off is approximately at minute 67.5.

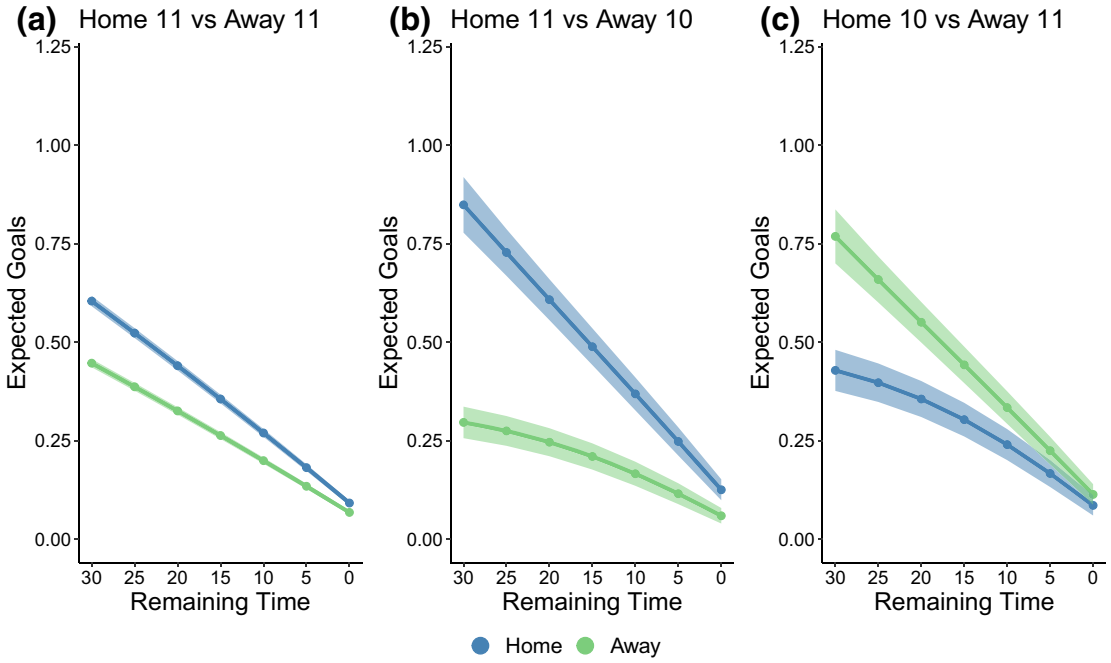

Fig. 4 Cummulative scoring rates for different Red Card scenarios for Home and away teams through time **a** 11 versus 11, **b** 11 versus 10, **c** 10 versus 11, respectively

After a sending off, scoring rates undergo an abrupt change. To exemplify the situation, at minute 65 the scoring rates per interval are on average 0.081 and 0.060 for the Home and Away teams. If a red card is shown to the visiting team, the scoring rates become 0.121 and 0.022, whereas if the affected team is playing at home, the rates are 0.031 and 0.109 respectively. The relative difference in performance is more or less maintained when the player sent off belongs to the Away team. However, when the affected team is stronger, the relative difference tends to decrease over time.

The accumulated scoring rate presented in Fig. 4 allows us to quantify the consequences of a sending off. For instance, when the remaining time is 30 minutes, the difference in accumulated expected goals is 0.16 in favor of the Home team. If an Away player is sent off, the difference becomes 0.55, whereas if the player dismissed plays for the Home team, the difference would be -0.34. In this sense, the expected impact of a sending off at 30 min to finish, is approximately 0.39 or 0.50 goals if the guilty player is on the visiting team or on the home team.

5 Discussion

This study analyses the effect of red and yellow cards on performance, using the scoring rate from the perspective of cohesion over time. The analysis is based on Poisson distribution models, which are more appropriate for count data than the linear regression models used in other studies, such as (Carmichael and Thomas 2005) and (Mechtel et al. 2011). The analysis also takes into account the correlated nature of the data incorporating the pertinent variance components (random effects) that provide a more valid inference framework.

Events that happen in injury time are also included and given the appropriate weight. This consideration allows for more concise modeling of the scoring rate as the game progresses. The model also incorporates as an adjustment variable all variables associated to

each game and team including in particular the home/away and team strength effects based on a transformation of betting odds. This latter criterion gives rise to a more efficient model that can be used to more clearly evaluate the study goals. The analysis also takes into account two time-related variables, the current clock time and the time elapsed from the red card scenario. This consideration enables differentiation of the effect of the remaining time from the effect of playing in numerical inferiority over time. Unlike the findings of (Caliendo and Radic 2006), but in agreement with (Lago-Peñas et al. 2016) and (Bar-Eli et al. 2006), this analysis does detect a relevant change in scoring rates associated to red cards. As suggested by Červený et al. (2018) our model detects an interaction with playing time, in particular the red card effect diminishes over time when the player sent off belongs to the strongest team.

The study by Mechtel et al. (2011) detected certain asymmetry depending on whether a team is playing at home or away, whereby "sending-offs against home teams have a negative impact on their performance. However, for guest teams, the impact depends on the time remaining after the sending-off and can be positive if the sending-off occurs late in the game." Our analysis does not detect such a pattern.

The number of players that have been cautioned with a yellow card has a slight effect on the scoring rate, in agreement with (Anders and Rotthoff 2011) and (Titman et al. 2015). In our case, the effect appears through an interaction with the goal difference. In particular, teams with a greater number of booked players have a lower scoring rate when they are winning. Since the first yellow card is a precursor to a second booking, the effect of yellow cards could be indirect. It would be interesting to conduct a more detailed analysis of the relationship between yellow and red cards, and how bookings affect a player's performance and how recommendable it is for booked players to be substituted.

As for playing time, a certain upward tendency has been detected as a game progresses, with the scoring rate increasing by around 2% per interval. Other variables measured at game level, such as competition, team strength, home advantage, attendance, etc. that are commonly included in similar studies (e.g. Chowdhury 2015; Mechtel et al. 2011) have been analyzed implicitly within the *WinOdds* variable, a transformation of the odds of a win, and thus cannot be discussed individually. Indeed, *WinOdds* constitutes a very relevant variable, and shows an interaction with the red card situation. It should be noted that there are some limitations to our study. While it is comprehensive in that it considers the top five European leagues, it would be interesting to compare our results to others for knockout competitions or weaker leagues (2nd divisions). On the other hand, the use of the *WinOdds* variable based on betting odds has the advantage of including all variables related to teams and the particular match; but it does not distinguish between the individual effects of each component. The current study found that a sending off is a significant event that has a dramatic influence on the outcome of a match and particularly produces a decrease in the scoring rate for the penalized team and an increase for the opposing team, and if taking place with 30 minutes of remaining time, it translates to more or less 0.5 goals. In some cases, red cards are received as a punishment for preventing an obvious scoring occasion, and it is therefore interesting to evaluate the extent to which it is better to concede a goal or receive a red card. It should be noted that in this situation, in addition to the sending off, the penalized team will also be punished with a penalty or a free kick.

For instance:

- At the 2010 World Cup in South Africa, in the final minute of overtime in the quarter final between Ghana and Uruguay and with the score at 1–1, Luis Suárez stopped a clear goal with his hand, leading to a red card and penalty. Ghanaian captain Asamoah Gyan took the penalty and the ball hit the crossbar. The game went to a penalty shootout and

Uruguay won. In this situation, Luis Suárez clearly made the right decision, but it would not have been such a clever move had he done so in the first minute of overtime: there would have been a penalty (the conversion rate of which is almost 80%) and his team would have played the remaining 30 min a man down.

- In the Spanish Liga game between Real Madrid and FC Barcelona in December 2017, with Barça leading 1-0 with 30 min to go, Real's Dani Carvajal stopped a goal with his hand: red card and penalty. Leo Messi's converted penalty as good as clinched the win in a game that ended 3-0.
- In the 2020 Spanish Super Cup Final between Real Madrid and Atlético Madrid, with the teams level at 1-1 with minutes to go before the end of overtime, Real's Federico Valverde fouled Álvaro Morata outside the penalty area when the latter was through on goalkeeper Thibaut Courtois. This was a clear goalscoring opportunity and the offender was sent off. The scoring rate in one-on-one situations is approximately 40% while for well-positioned direct free kicks it is between 10 and 20%. As the game was coming to an end, and the foul was outside the penalty area, Valverde would appear to have made the right decision. Real Madrid won the game in a penalty shootout.

Nowadays, thanks to Video Assistant Referee (VAR) technology, the possibility of a referee not spotting an offense is negligible. Hence, just at the end of the game, it will make sense to prevent a clear goalscoring opportunity and be punished with a red card plus the consequent penalty.

Coaches and players should be very cautious and try to avoid situations where players might receive a red card. Otherwise, teams need to be prepared for these scenarios of numerical inferiority.

Acknowledgements We thank the anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript.

Funding Open Access Funding provided by Universitat Autònoma de Barcelona. This work was partially funded by the grants RTI2018-096072-B-I00 and PID2019-104830RB-I00/AEI DOI:10.13039/501100011033 from the Spanish Ministry of Science, Innovation and Universities, and the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D (CEX2020-001084-M) from the Spanish Research Agency.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Albanese, A., De Meyere, A., Vanruymbeke, W., & Baert, S. (2020). Player dismissal and full-time results in the UEFA champions league and Europa league. *International Journal of Sport Finance, 15*(1), 27–38.

Anders, A., & Rotthoff, K. W. (2011). Yellow cards: Do they matter? *Journal of Quantitative Analysis in Sports, 7,* 1–12.
Bar-Eli, M., Sachs, S., Tenenbaum, G., Pie, J. S., & Falk, B. (1996). Crisis-related observations in competition: A case study in basketball. *Scandinavian Journal of Medicine and Science in Sports, 6,* 313–321.
Bar-Eli, M., & Tenenbaum, G. (1989). A theory of individual psychological crisis in competitive sport. *Applied Psychology, 38,* 107–120.
Bar-Eli, M., Tenenbaum, G., & Geister, S. (2006). Consequences of players' dismissal in professional soccer: A crisis-related analysis of group-size effects. *Journal of Sports Sciences, 24,* 1083–1094.
Bornn, L., Cervone, D., & Fernandez, J. (2018). Soccer analytics: Unravelling the complexity of the "beautiful game." *Significance,15,* 26–29.
Caliendo, M., & Radic, D. (2006). *Ten do it better, do they?: An empirical analysis of an old football myth*. IZA Discussion Paper 2158. Institute for the Study of Labor: Bonn.
Carling, C., & Bloomfield, J. (2010). The effect of an early dismissal on player workrate in a professional soccer match. *Journal of Science Medicine in Sport, 2010*(13), 126–28.
Carmichael, F., & Thomas, D. (2005). Home-field effect and team performance evidence from English premiership football. *Journal of sports Economics, 6,* 264–281.
Červený, J., van Ours, J. C., & van Tuijl, M. A. (2018). Effects of a red card on goal-scoring in World Cup football matches. *Empirical Economics, 55,* 883–903.
Chowdhury, A. (2015). Can ten do it better? Impact of red card in the English Premier League. No. 2015-01. Marquette University, Center for Global and Economic Studies and Department of Economics.
Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836.
Football-Data.co.uk. (2020). Available online: http://www.football-data.co.uk. Accessed on July 2020.
Gómez-Déniz, E., Cárdenes, N. D., & Sánchez Pérez, J. M. (2019). A probabilistic model for explaining the points achieved by a team in football competition. Forecasting and regression with applications to the Spanish competition. *SORT-Statistics and Operations Research Transactions, 43,* 95–112.
Greenberg, A. (2015). The red card cliché. *Significance, 12,* 30–33.
Lago-Peñas, C., Gómez-Ruano, M. A., Owen, A. L., & Sampaio, J. (2016). The effects of a player dismissal on competitive technical match performance. *International Journal of Performance Analysis in Sport, 16,* 792–800.
Lee, Y., Nelder, J. A., & Pawitan, Y. (2018). *Generalized Linear Models with Random effects, 2nd Edition*. Chapman & Hall/CRC.
Mattera, R. (2021). Forecasting binary outcomes in soccer. *Annals of Operations Research*, 1–20.
McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. London: Routledge.
McGarry, T., O'Donoghue, P., & de Eira Sampaio, A. J. (2013). *Routledge handbook of sports performance analysis*. London: Routledge.
Mechtel, M., Bäker, A., Brändle, T., & Vetter, K. (2011). Red cards: not such bad news for penalized guest teams. *Journal of Sports Economics, 12,* 621–646.
Molenberghs, G., & Verbeke, G. (2005). *Models for discrete longitudinal data*. Berlin: Springer.
Nakagawa, S., & Holger, S. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142.
Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data, 6,* 1–15.
Ridder, G., Cramer, J. S., & Hopstaken, P. (1994). Down to ten: Estimating the effect of a red card in soccer. *Journal of the American Statistical Association, 89,* 1124–1127.
Sapp, R. M., Spangenburg, E. E., & Hagberg, J. M. (2019). Markers of aggressive play are similar among the top four divisions of English soccer over 17 seasons. *Science and Medicine in Football, 3,* 125–130.
Titman, A., Costain, D., Ridall, P., & Gregory, K. (2015). Joint modelling of goal and bookings in association football. *Journal of the Royal Statistical Society: Series A, 178*(3), 659–683.
Wunderlich, F., & Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. *PloS One, 13,* e0198668.

CAPÍTOL 5

Ultra log-concavity of discrete order statistics

5.1 Article

Statistics and Probability Letters

Ultra log-concavity of discrete order statistics

Llorenç Badiella[a,b], Joan del Castillo[b], Pedro Puig[b]

[a] Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola, Spain

[b] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola, Spain

Resum

En aquest treball es demostra que els estadístics d'ordre de distribucions discretes preserven les propietats de log-concavitat i ultra log-concavitat. Per aquest propòsit s'empra una expressió recursiva dels estadístics d'ordre i el concepte de seqüències sincronitzades. Aquest resultat permet concloure que els estadístics d'ordre de la distribució de Poisson són infradispersos.

En conseqüència, l'ús d'estadístics d'ordre (com ara el màxim, el mínim o la mediana) per a resumir dades de recompte amb subrèpliques, consisteix en una estratègia raonable per eliminar el soroll associat als errors de mesura i simplificar el disseny experimental, doncs els estadístics d'ordre donen lloc a una reducció de la dispersió original tot mantenint la natura discreta de les dades.

Nota: A continuació de l'article s'adjunten materials suplementaris. Aquests materials no s'han arribat a publicar per restriccions d'espai de la revista.

Statistics and Probability Letters 201 (2023) 109900

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Ultra log-concavity of discrete order statistics

Llorenç Badiella [a,b,*], Joan del Castillo [b], Pedro Puig [b,c]

[a] Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola, Spain
[b] Departament de Matemàtiques, Universitat Autònoma de Barcelona, Cerdanyola, Spain
[c] Centre de Recerca Matemàtica, Cerdanyola, Spain

ARTICLE INFO

Article history:
Received 8 November 2022
Received in revised form 12 June 2023
Accepted 23 June 2023
Available online 26 June 2023

Keywords:
Poisson distribution
Underdispersion

ABSTRACT

In this work we show that discrete order statistics preserve log-concavity and ultra log-concavity. We use a recursive expression for discrete order statistics and the concept of synchronized sequences. This finding allows to conclude that Poisson order statistics are underdispersed.

1. Introduction

Order statistics are a subject of interest in statistics and applied probability. Sometimes data come from order statistics of counts, for instance, in the study of avian fauna some researchers use the maximum of a series of replicated observations (Chamberlain et al., 2009; Hartill et al., 2011); in sports such as soccer, the amount of points achieved by a team at the end of the season is modeled according to their final position (Emparanza and Núnez-Antón, 2010); in citometry, the median of individual counts is frequently used to measure the presence of various cell types (Lesko et al., 2013; Nielson et al., 1991), and in discrete process control, it is common to monitorize certain quantiles (Jiang, 2010; Wu et al., 2014).

Discrete and continuous log-concave distributions play an increasingly important role in probability, statistics, optimization theory, econometrics and other areas of applied mathematics. Log-concavity is connected to different branches of mathematics and statistics, including concentration of measure, log-Sobolev inequalities, MCMC algorithms, Laplace approximations, and machine learning (Saumard and Wellner, 2014). The preservation of log-concavity and ultra log-concavity under different operations such as marginalization, convolution, formation of products, and limits in distribution has been object of study by a number of authors (Saumard and Wellner, 2014). The class of ultra log-concave discrete distributions plays a fundamental role in the characterization of the Poisson distribution as a maximum entropy distribution and in the study of the Law of Small Numbers (Harremoës, 2001; Johnson, 2007).

Definition 1. Let $A = (a_k)_{k=-\infty}^{\infty}$ be a sequence of non-negative real numbers. Then

(a) A is said to be log-concave if $a_k^2 \geq a_{k-1}a_{k+1}$ for all k.
(b) A is said to be ultra log-concave if $ka_k^2 \geq (k+1)a_{k-1}a_{k+1}$ for all k.

* Corresponding author at: Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola, Spain.
E-mail address: badiella@mat.uab.cat (L. Badiella).

https://doi.org/10.1016/j.spl.2023.109900

Let X be a discrete distribution with $f_k = P(X = k)$. We identify the sequence $(f_k)_{k=0}^{\infty}$, with the infinite sequence $(f'_k)_{k=-\infty}^{\infty}$, where $f'_k = f_k$ for $k \geq 0$ and $f'_k = 0$ otherwise. For simplicity, hereafter, the sequence (f_k) will be denoted as f_k. In this sense, considering the sequences given by the probability function of different discrete variables, some distributions including Bernoulli, Bernoulli sums, hypergeometric, Poisson, and truncated Poisson, have the property of being ultra log-concave. Other distributions, including the geometric and the negative binomial distributions, are log-concave but not ultra log-concave. The logarithmic distribution and in general, bimodal distributions are not log-concave.

Log-concavity of order statistics has been extensively studied in the continuous case, however, due to the presence of ties, the discrete case presents more difficulties. Recent works by Alimohammadi et al. (2015) and Kim et al. (2018) have focused on studying the strong unimodality of discrete sequences, being this notion equivalent to log-concavity (Keilson and Gerber, 1971). They showed that the probability mass function (p.m.f.) of order statistics preserve the log-concavity of the original discrete distribution.

The aim of the present work is to show that order statistics of an ultra log-concave distribution are also ultra log-concave. Since ultra log-concavity is connected to underdispersion (del Castillo and Pérez-Casany, 2005; Johnson, 2007), it is shown as a corollary that Poisson order statistics are underdispersed. Statisticians should take this property into consideration when modeling this type of data.

2. Discrete order statistics

The distribution of order statistics for discrete distributions does not have a simple formulation due to the presence of ties. Let X be a discrete distribution with $f_k = P(X = k)$, $F_k = \sum_{i=0}^{k} f_i$, and $S_k = 1 - F_k$, denoting the p.m.f., the cumulative distribution function, and the survival function respectively. Given a sample of size n, let $X_{r:n}$ be the rth order statistic. Its p.m.f. can be expressed using the beta integral form (Arnold et al., 2008):

$$P(X_{r:n} = k) = I_{F_k}(r, n+1-r) - I_{F_{k-1}}(r, n+1-r) \tag{1}$$

where $I_x(a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt / \mathrm{B}(a,b)$ is the incomplete beta function and $\mathrm{B}(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the beta function.

Alternatively, we propose an expression for order statistics using a recursive approach which will be used in proving the main theorem. First of all, define $P(X_{r:n} = k) = 0$ whenever $r < 1$ or $r > n$ and $(n-r)\mathrm{B}(r, n-r) = 1$ if $n = r$.

Theorem 1. *Let X be a discrete distribution with $X_{r:n}$, the rth order statistic. Then, for $n > 1$,*

$$P(X_{r:n} = k) = P(X_{r-1:n-1} = k)\, F_k + P(X_{r:n-1} = k)\, S_k + f_k \frac{F_{k-1}^{r-1} S_{k-1}^{n-r}}{(n-r)\mathrm{B}(r, n-r)}$$

Proof. Using expression (1), different properties of the incomplete beta function and taking into account that $F_k = F_{k-1} + f_k$,

$$\begin{aligned}
P(X_{r:n} = k) &= I_{F_k}(r, n+1-r) - I_{F_{k-1}}(r, n+1-r) \\
&= F_k I_{F_k}(r-1, n+1-r) + S_k I_{F_k}(r, n-r) \\
&\quad - F_{k-1} I_{F_{k-1}}(r-1, n+1-r) - S_{k-1} I_{F_{k-1}}(r, n-r) \\
&= F_k \left(I_{F_k}(r-1, n+1-r) - I_{F_{k-1}}(r-1, n+1-r)\right) + \\
&\quad S_k \left(I_{F_k}(r, n-r) - I_{F_{k-1}}(r, n-r)\right) + \\
&\quad f_k I_{F_{k-1}}(r-1, n+1-r) - f_k I_{F_{k-1}}(r, n-r) \\
&= P(X_{r-1:n-1} = k) F_k + P(X_{r:n-1} = k) S_k + \\
&\quad f_k (I_{F_{k-1}}(r-1, n+1-r) - I_{F_{k-1}}(r, n-r))
\end{aligned}$$

Finally, we find that

$$P(X_{r:n} = k) = P(X_{r-1:n-1} = k) F_k + P(X_{r:n-1} = k) S_k + f_k \frac{F_{k-1}^{r-1} S_{k-1}^{n-r}}{(n-r)\mathrm{B}(r, n-r)} \quad \square$$

3. Log-concave and synchronized series

We start by reviewing basic properties and notation for log-concave sequences. Let $A + B$, $A \times B$ denote the sequences with coefficients $(a_k + b_k)$ and $(a_k \times b_k)$ respectively, whereas uA denotes the sequence with coefficients (ua_k), for any constant $u \geq 0$. The convolution of A and B, denoted as $A * B$, is defined to be the sequence with coefficients: $\sum_{i=-\infty}^{\infty} a_{k-i} b_i$. For any sequence $A = (a_k)$, we define the associated offset sequence $A^- = (a_k^-)$ by $a_k^- = a_{k-1}$ for all k. We highlight the following properties of log-concave sequences:

1. Log-concavity of discrete sequences is preserved by products. If A and B are log-concave sequences then the sequence $A \times B$ is also log-concave.
2. The convolution of two log-concave sequences is log-concave.

3. Given a log-concave sequence A, the offset sequence A^- is log-concave.
4. If f_k is a log-concave sequence, so are the sequences F_k and S_k.
5. If f_k is a log-concave sequence, $F_k^i F_{k-1}^j S_k^l S_{k-1}^m$ for any $i, j, l, m \geq 0$ is a log-concave sequence.

Properties (1) and (3) are straightforward using the definition of a log-concave sequence. For a proof of the second property see Theorem 4.7.1 in Prékopa (2013) . To assess property (4), it suffices to consider the convolution between the log-concave sequence f_k with the constant sequence $q_k = 1$, and the convolution between f_k with the constant sequence $q'_k = 0$ respectively (see Theorem 4.7.2. in Prékopa, 2013). An alternative proof of property (4) can also be found in Theorem 2.2 in Alimohammadi et al. (2015). Property (5) is clear using properties (1) and (3).

In a series of recent works focused in the study of topological graph theory and combinatorics, and the Genus distribution of a graph, Gross et al. (2015) introduced new tools to deal with sums and convolutions of log-concave sequences; in particular, the concept of synchronized sequences.

Definition 2. Let A and B be two log-concave series. They are said to be synchronized, denoted as $A \sim B$, if they satisfy

$$a_k b_k \geq a_{k-1} b_{k+1} \quad \text{and} \quad a_k b_k \geq a_{k+1} b_{k-1} \quad \text{for all } k.$$

Proposition 1. *The following properties hold:*

1. *The sequence resulting from a linear combination of synchronized sequences is log-concave, i.e. let A and B be synchronized sequences, and let $u, v > 0$; then $uA + vB$ is log-concave.*
2. *Given a set of n pairwise synchronized sequences denoted as $(A_i)_{i=1}^n$, for any numbers $u_1, v_1, \ldots, u_n, v_n \geq 0$, we have $\sum_{i=1}^n u_i A_i \sim \sum_{i=1}^n v_i A_i$.*
3. *If $A \sim B$ and $C \sim D$, then $(A \times C) \sim (B \times D)$.*

and if f_k is a log-concave sequence,

4. *$F_k \sim F_{k-1}$ and $S_k \sim S_{k-1}$*
5. *$F_k^i \sim F_{k-1}^i$ and $S_k^i \sim S_{k-1}^i$ for $i > 0$.*
6. *$F_k S_k \sim F_k S_{k-1} \sim F_{k-1} S_k \sim F_{k-1} S_{k-1}$.*
7. *$F_k^i F_{k-1}^j \sim F_k^{i'} F_{k-1}^{j'}$ with $i + j = i' + j'$; and $S_k^l S_{k-1}^m \sim S_k^{l'} S_{k-1}^{m'}$ with $l + m = l' + m'$.*
8. *$F_k^i F_{k-1}^j S_k^l S_{k-1}^m \sim F_k^{i'} F_{k-1}^{j'} S_k^{l'} S_{k-1}^{m'}$ with $i + j = i' + j'$ and $l + m = l' + m'$.*

Proof. Property (1) is a particular case of property (2). See Theorem 2.3 from Gross et al. (2015) for a proof of property (2). Property (3) can be assessed using the definition of synchronicity. The first condition for synchronicity in property (4) is obvious. The second condition becomes $F_k F_{k-1} \geq F_{k+1} F_{k-2}$, which is true since F_k is a log-concave sequence. Properties (5) and (6) can be proven using (3) and (4). In order to prove property (7), without loss of generality we assume that $i > i'$. Thus, the property becomes $F_k^{i-i'}(F_k^{i'} F_{k-1}^j) \sim F_{k-1}^{j'-j}(F_k^{i'} F_{k-1}^j)$, which is true by (5) and (3). Finally (8) is a consequence of (7) and (3). □

4. Ultra log-concavity of discrete order statistics

Theorem 2. *Discrete order statistics preserve log-concavity and ultra log-concavity.*

Proof. Consider a discrete random variable X with f_k as its p.m.f. and assume that f_k is a log-concave or ultra log-concave sequence. In order to demonstrate the theorem we will first show that $P(X_{r:n} = k)/f_k$ can be expressed as a linear combination of log-concave synchronized terms:

$$P(X_{r:n} = k)/f_k = \sum_{i=0}^{r-1} \sum_{m=0}^{n-r} c(i, m) F_k^i F_{k-1}^{r-1-i} S_k^m S_{k-1}^{n-r-m} \tag{2}$$

i.e. a linear combination of terms where the sum of the exponents in F_k and F_{k-1} is $r - 1$ and the sum of the exponents in S_k and S_{k-1} is $n - r$ for all $k \geq 1$ and $r \geq 1$.

The proof of this property follows by induction:

- For $r = 1$ $n = 1$ expression (2) holds:

$$P(X_{1:1} = k)/f_k = 1$$

- Suppose that for some $n_0 \in \mathbb{N}$, $n_0 \geq 1$ and $\forall r \in \mathbb{N}$, $r \leq n_0$, the property is true. Using the representation given by Theorem 1 and choosing an arbitrary r_0 ($\leq n_0 + 1$), we can write:

$$P(X_{r_0:n_0+1} = k)/f_k = F_k P(X_{r_0-1:n_0} = k)/f_k$$

$$+S_k P(X_{r_0:n_0}=k)/f_k + \frac{F_{k-1}^{r_0-1} S_{k-1}^{n_0-r_0+1}}{(n_0-r_0)\mathrm{B}(r_0, n_0+1-r_0)}$$

Each one of the three terms can be expressed as a linear combination of terms fulfilling the desired property. Using properties (2) and (7) from Proposition 1 it follows that $P(X_{r:n}=k)/f_k$ is a log-concave sequence. Thus, if f_k is a log-concave sequence so are their order statistics because $P(X_{r:n}=k)$ can be expressed as a product of log-concave sequences and log-concavity is preserved by products. On the other hand, the ultra log-concavity of the p.m.f. sequence for a random variable is equivalent to its log-concavity with respect to the Poisson distribution, given p_k the p.m.f. of the Poisson distribution, f_k is ultra log-concave if and only if f_k/p_k is log-concave. Thus, if f_k is ultra log-concave, $P(X_{r:n}=k)/p_k$ is log-concave, and $P(X_{r:n}=k)$ is ultra log-concave. These results lead us to conclude that discrete order statistics preserve log-concavity and ultra log-concavity. □

Remark 1. The fact that discrete order statistics preserve log-concavity was also proved by Kim et al. (2018) using a totally different approach in terms of strong unimodality.

Corollary 1. *Order statistics of a Poisson distribution are under-dispersed distributions (i.e. variance is strictly smaller than the mean).*

Proof. According to Johnson (2007), given X a discrete ultra log-concave random variable,

$$\mathrm{E}[X(X-1)] \leq (\mathrm{E}[X])^2 \tag{3}$$

Since the Poisson distribution has the maximum entropy property within ultra log-concave distributions (Johnson, 2007) this is the only ultra log-concave distribution fulfilling equality in (3). On the other hand, Poisson order statistics are not Poisson distributed, as can be assessed using formulation (1). □

Remark 2. Corollary 1 can also be proved using Corollary 4 in del Castillo and Pérez-Casany (2005) using the fact that the distribution of Poisson order statistics can be expressed as a weighted Poisson distribution, and the function defining the weights is log-concave.

Funding

This work was partially funded by the grants RTI2018-096072-B-I00 from the Spanish Ministry of Science, Innovation and Universities and CEX2020-001084-M from the Spanish State Research Agency, through the Severo Ochoa and María de Maeztu Program for Centers and Units of Excellence in R&D.

Data availability

No data was used for the research described in the article.

References

Alimohammadi, M., Alamatsaz, M.H., Cramer, E., 2015. Discrete strong unimodality of order statistics. Statist. Probab. Lett. 103, 176–185.
Arnold, B.C., Balakrishnan, N., Nagaraja, H.N., 2008. A First Course in Order Statistics. SIAM.
Chamberlain, D.E., Glue, D.E., Toms, M.P., 2009. Sparrowhawk accipiter nisus presence and winter bird abundance. J. Ornithol. 150 (1), 247–254.
del Castillo, J., Pérez-Casany, M., 2005. Overdispersed and underdispersed Poisson generalizations. J. Statist. Plann. Inference 134 (2), 486–500.
Emparanza, I.D., Núnez-Antón, V., 2010. On the use of simulation methods to compute probabilities: application to the first division spanish soccer league. SORT-Stat. Oper. Res. Trans. 34, 181–200.
Gross, J.L., Mansour, T., Tucker, T.W., Wang, D.G., 2015. Log-concavity of combinations of sequences and applications to genus distributions. SIAM J. Discrete Math. 29 (2), 1002–1029.
Harremoës, P., 2001. Binomial and Poisson distributions as maximum entropy distributions. IEEE Trans. Inform. Theory 47 (5), 2039–2041.
Hartill, B.W., Watson, T.G., Bian, R., 2011. Refining and applying a maximum-count aerial-access survey design to estimate the harvest taken from New Zealand's largest recreational fishery. North Am. J. Fish. Manag. 31 (6), 1197–1210.
Jiang, R., 2010. Discrete competing risk model with application to modeling bus-motor failure data. Reliab. Eng. Syst. Saf. 95 (9), 981–988.
Johnson, O., 2007. Log-concavity and the maximum entropy property of the Poisson distribution. Stochastic Process. Appl. 117 (6), 791–802.
Keilson, J., Gerber, H., 1971. Some results for discrete unimodality. J. Amer. Statist. Assoc. 66 (334), 386–389.
Kim, B., Kim, J., Lee, S., 2018. Strong unimodality of discrete order statistics. Statist. Probab. Lett. 140, 48–52.
Lesko, C.R., Cole, S.R., Zinski, A., Poole, C., Mugavero, M.J., 2013. A systematic review and meta-regression of temporal trends in adult CD4+ cell count at presentation to HIV care, 1992–2011. Clin. Infect. Dis. 57 (7), 1027–1037.
Nielson, L., Smyth, G., Greenfield, P., 1991. Hemacytometer cell count distributions: implications of non-Poisson behavior. Biotechnol. Prog. 7 (6), 560–563.
Prékopa, A., 2013. Stochastic Programming, Vol. 324. Springer Science and Business Media.
Saumard, A., Wellner, J.A., 2014. Log-concavity and strong log-concavity: A review. Stat. Surv. 8 (45).
Wu, H., Gao, L., Zhang, Z., 2014. Analysis of crash data using quantile regression for counts. J. Transp. Eng. 140 (4), 04013025.

Supplementary Material for: Ultra log-concavity of discrete order statistics

S1. Examples of the preservation of log-concavity by discrete order statistics

The following figures (S1-S6) illustrate the preservation of log-concavity by displaying the p.m.f. and the log-concavity rate for various standard discrete distributions and order statistics. The first column in each figure corresponds to the minimum order statistics ranging from $X_{1:1}$ to $X_{1:8}$, the second column corresponds to the maximum order statistics ranging from $X_{1:1}$ to $X_{8:8}$, and the third column is related to the median ranging from $X_{1:1}$ to $X_{6:11}$. The first row of each panel shows the p.m.f. profiles of each order statistics distribution. The black dashed line represents the original distribution, while lighter lines indicate higher order statistics. The logarithm of the log-concavity rate (log-LC) profiles are displayed in the second row of each figure. The log-LC rate for a probability distribution f_k is defined as the sequence $log(f_k^2/(f_{k+1}f_{k-1}))$. For distributions with compact support, in order to avoid values with p.m.f. equal to 0, the log-LC rate is only computed whenever the LC rate is definite. The log LC rate is useful because, according to definition (1a), it is always greater than or equal to zero for log-concave distributions and strictly equal to zero for geometric distributions. Moreover, the log-LC rate is always greater or equal to the sequence $log((k+1)/k)$ for ultra log-concave distributions, but strictly equal for the Poisson distribution, according to definition (1b). The blue line in the second row of each panel represents a reference at 0, while the green line is the logarithm of the sequence $(k+1)/k$. As a result, any distribution with a log-LC rate more than or equal to the blue line is log-concave; if the log-LC rate is greater than or equal to the green line, the distribution is ultra log-concave. The distributions shown are Poisson ($\lambda = 5$), binomial ($p = 0.5, n = 10$), discrete uniform ($a = 0, b = 10$), geometric ($p = 1/6$), negative binomial ($r = 5, p = 0.5$) and a shifted zeta ($s = 2$) starting from

0. To simplify comparisons between distributions, all but the zeta have an average of 5.
The Poisson and the binomial distributions are both ultra log-concave, and so are their respective order statistics (figures S1 and S2). The discrete uniform distribution (figure S3), the geometric (figure S4) and the negative binomial distribution (figure S5) are log-concave, and so are their respective order statistics. For particular median statistics of the discrete uniform distribution, ultra log-concavity is achieved. Minimum order statistics of the geometric distribution are also geometric distributed and all log-LC rate lines for the different minimum order statistics are overlapped at the 0 reference line. Order statistics of a Zeta distribution do not show log-concavity (figure S6).

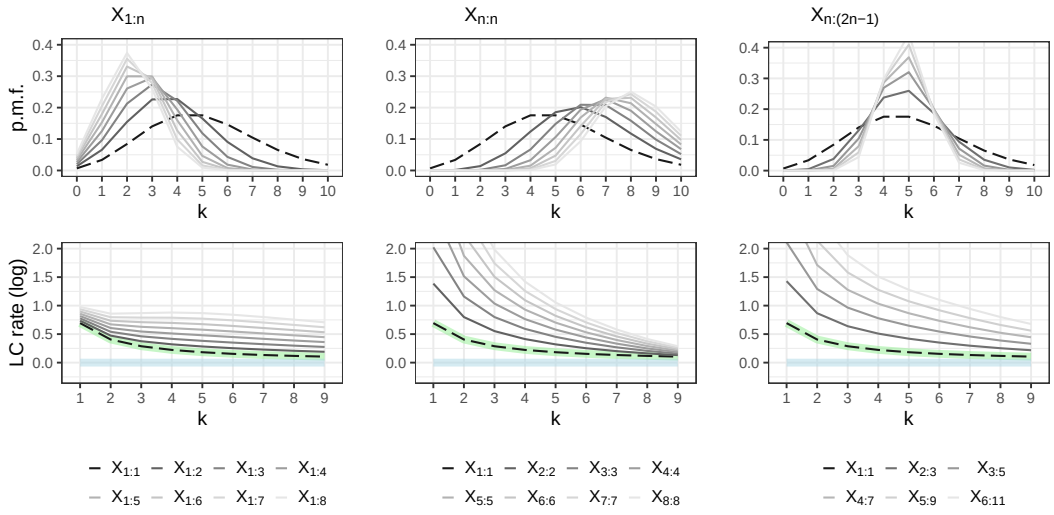

Figure S1: Poisson ($\lambda = 5$)

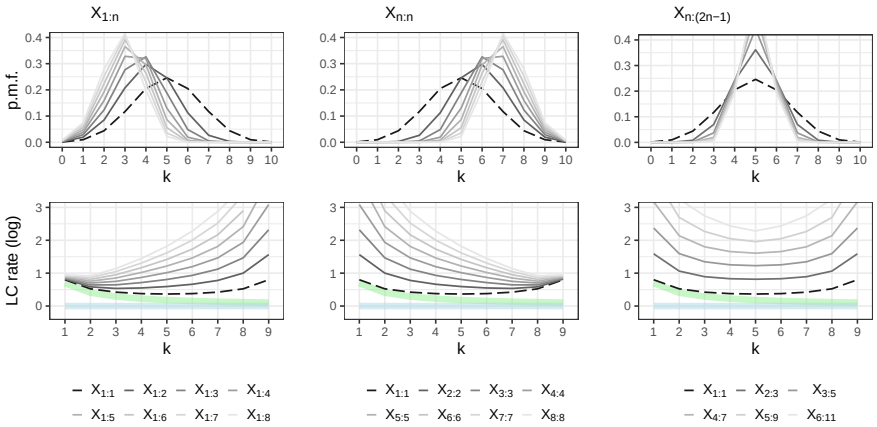

Figure S2: Binomial ($p = 0.5, n = 10$)

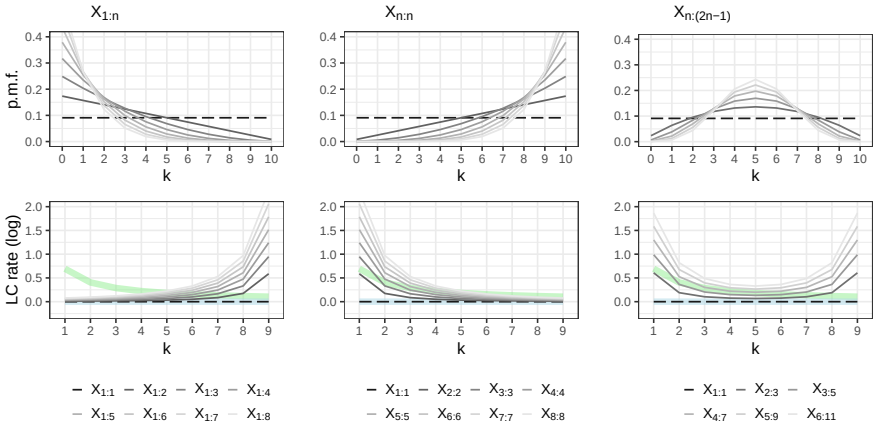

Figure S3: Discrete uniform ($a = 0, b = 10$)

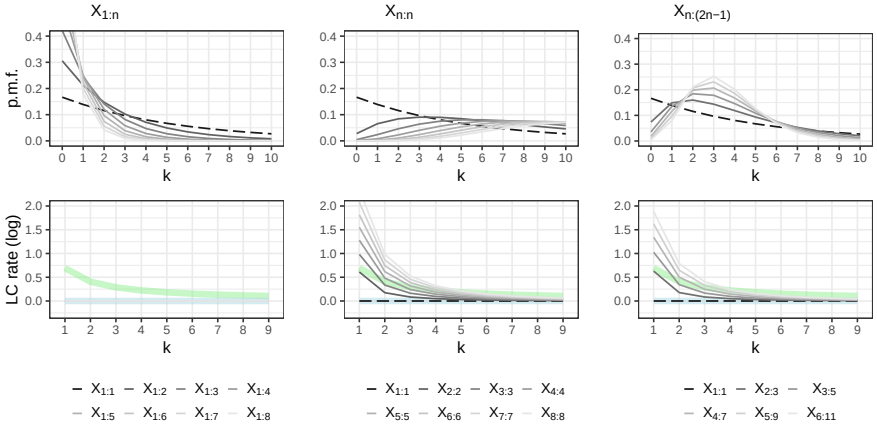

Figure S4: Geometric ($p = 1/6$)

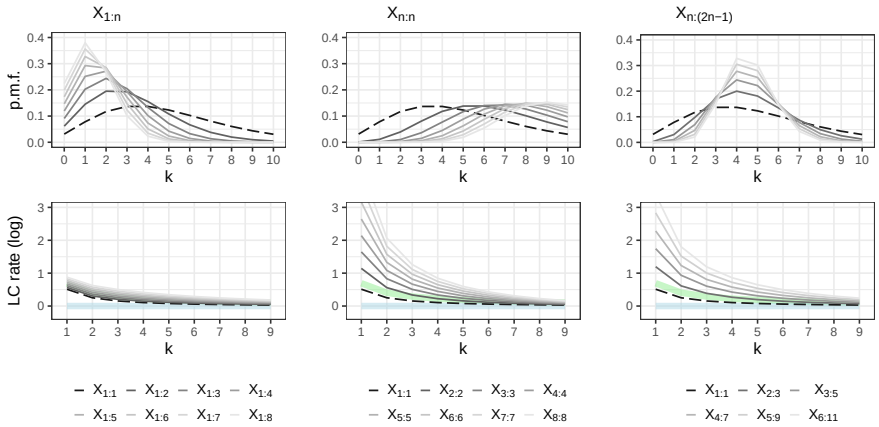

Figure S5: Negative binomial ($r = 5, p = 0.5$)

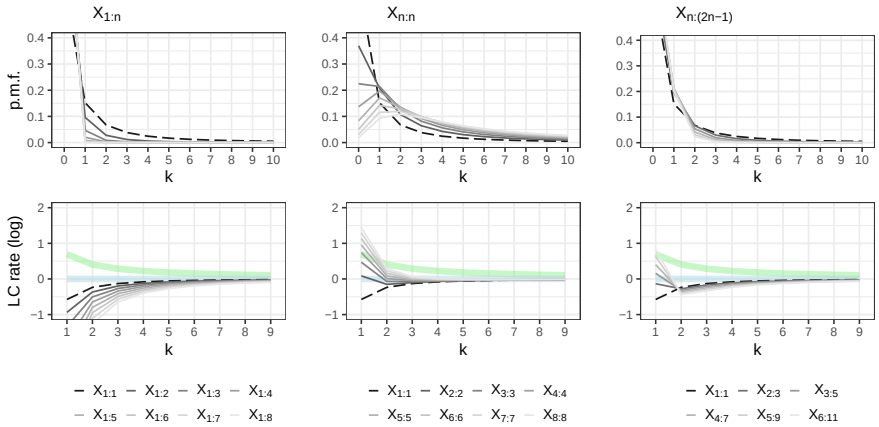

Figure S6: Zeta ($s = 2$)

CAPÍTOL 6

Valoracions i conclusions

En la present tesi s'han descrit les principals estratègies per a l'anàlisi de dades de recompte amb mesures repetides, basades en l'aplicació de models GLMM sota les visions marginal o condicional, incidint en les seves limitacions pràctiques.

En el model lineal mixt, ambdues propostes fan referència en realitat al mateix model, si bé la proposta del model condicional és aparentment més restrictiva. En un treball no publicat de Badiella and Brewer (2015), es comprova que donada una proposta de model marginal, mitjançant la parametrització adient, existeix una proposta de modelització amb factors aleatoris que dona lloc a la mateixa inferència. Aquesta relació implica que l'equivalència entre propostes és total.

En els models GLMM, donat que les relacions deixen de ser lineals, aquesta correspondència entre visions desapareix, esdevenint dues propostes diferents. La visió basada en el model marginal dona lloc a una proposta flexible, però amb eines de validació limitades, mentre que el model condicional és més vàlid i interpretable, però no sempre podrà capturar les subtileses de les dades i del disseny experimental.

En capítols anteriors s'han presentat casos pràctics on, a causa de la naturalesa dels experiments considerats, s'utilitzen diferents estratègies de modelització: el model condicional, el model marginal i estadístics d'ordre.

6.1 Valoracions de l'article: Effectiveness of a road traffic injury prevention intervention in reducing pedestrian injuries, Barcelona, 2002-2019

En aquest article es descriu l'experiment que es va dur a terme per a mesurar l'eficàcia de les Rutes Segures (SRTS) per a la prevenció i reducció de la sinistralitat viària a la ciutat de Barcelona. Es va considerar un disseny quasiexperimental pre-post, amb un grup de comparació aparellat, que incloïa seixanta-quatre escoles d'intervenció i seixanta-quatre escoles de control. Es van recollir dades de la sinistralitat al voltant de cada escola durant aproximadament vuit anys abans de la intervenció i vuit anys després de la mateixa. Per a cada any es disposava de dades de lesionats i col·lisions per grup d'edat.

Des d'un punt de vista aplicat, l'estudi proporciona diferents contribucions rellevants a l'àmbit d'estudi. Per un costat s'obtenen evidències de l'eficàcia del programa SRTS per millorar la seguretat viària especialment entre els nens, quan a la literatura científica existeix certa controvèrsia en relació amb l'eficàcia d'aquest tipus d'intervencions. Per altra part, l'estudi elimina les limitacions de treballs anteriors mitjançant un disseny quasiexperimental, que permet un control dels principals factors de confusió, acompanyat d'una anàlisi estadística precisa, tenint en compte les particularitats de la variable resposta i incorporant els factors aleatoris necessaris i altres fonts de variació pertinents per a oferir una inferència vàlida.

Des de la perspectiva estadística, l'estudi analitza dades de recompte amb un disseny experimental complex, contemplant una estructura de diferents nivells experimentals jeràrquics:

- Parella: Efecte aleatori.
- Escola: Efecte aleatori.
- Escola: Pendents aleatoris per a tendències temporals.
- Any: Efecte aleatori creuat.

I al mateix temps per a cada any una sèrie de mesures repetides multivariants:

- Nombre de sinistres amb lesionats.

- Nombre de sinistres amb lesionats menors de setze anys.
- Nombre de sinistres amb lesionats vianants menors de setze anys.
- Nombre de lesionats.
- Nombre de lesionats menors de setze anys.
- Nombre de lesionats vianants menors de setze anys.

Aquestes mesures multivariants tenen correlació positiva, atès que part de les mesures són recomptes parcials de les altres. A partir de l'estadístic de Pearson dividit pels graus de llibertat del model, s'observà que les variables associades al nombre de lesionats donaven lloc a certa sobredispersió (presumiblement perquè hi ha una composició de variables, ja que en cas de sinistre, el nombre de lesionats no és constant) mentre que en les variables associades al recompte de sinistres això no succeïa. En aquest sentit, s'optà per analitzar cada variable de forma separada emprant un model GLMM-binomial negativa condicional per a les variables de recompte de lesions i un model GLMM-Poisson condicional per als sinistres. La proposta es validà comparant l'índex AIC de cadascun dels models.

Com a anàlisi de sensibilitat s'analitzaren també les dades agregades per períodes d'estudi (pre-post) simplificant el nombre de mesures repetides. A causa del fet que la suma de recomptes correlacionats dona lloc a recomptes sobredispersos, per a la validació s'utilitzaren models GLMM-binomial negativa condicional. Complementàriament, també es van ajustar aquestes dades amb models marginals obtenint conclusions compatibles.

L'ús de models condicionals està justificat pel fet que la interpretació de la intervenció és a escala individual, és a dir, a nivell d'escola. La interpretació poblacional podria mostrar biaixos perquè la mostra d'escoles analitzades no fou seleccionada a l'atzar.

S'espera poder donar continuïtat a aquest treball avaluant altres intervencions viàries mitjançant experiments amb disseny similar.

6.2 Valoracions de l'article: Influence of Red and Yellow cards on team performance in elite soccer

En aquest estudi s'analitzà el recompte de gols marcats segons diferents circumstàncies al llarg del temps en diferents partits de futbol. Des de la perspectiva aplicada, el treball proporciona diferents contribucions molt remarcables a l'àmbit d'estudi.

L'estudi planteja l'anàlisi d'un gran nombre de partits a partir de dades agregades en petits intervals de temps oferint una coherència temporal entre els gols marcats i les variables contextuals, entre les quals s'incorporen les quotes de les cases d'apostes. Aquesta visió elimina moltes de les limitacions de treballs anteriors. Per altra part, l'anàlisi estadística té en compte les particularitats de la variable resposta i incorpora els factors aleatoris i altres fonts de variació pertinents permetent estimar la taxa de gols al llarg del temps. En conseqüència, proporciona un marc original per quantificar l'impacte mitjà de l'expulsió d'un jugador.

En aquest sentit, la principal conclusió del treball vindria a ser que l'impacte de l'expulsió d'un jugador a trenta minuts pel final del partit és d'aproximadament 0,39 gols si l'expulsat és un visitant i de 0,5 gols si juga a l'equip local. Els models obtinguts en aquest treball s'han implementat en un aplicatiu web que permet calcular quotes d'apostes i probabilitats de marcar a partir de les quotes inicials indicant la situació en què es troba un partit en curs: temps de joc, marcador parcial i jugadors expulsats i amonestats Badiella, L. (2023c).

Des del punt de vista estadístic, el model té en compte una estructura de nivells experimentals jeràrquica i no jeràrquica:

- Partit: Efecte aleatori.
- Partit/Equip: Efecte aleatori.
- Partit/Equip: Pendents aleatoris per a tendències temporals.
- Equip: Efecte aleatori creuat.

I al mateix temps per a cada interval les mesures repetides multivariants associades als dos equips:

- Partit/Interval/Equip Local.
- Partit/Interval/Equip Visitant.

El disseny experimental contempla una jerarquia de nivells que s'incorporen en l'anàlisi mitjançant factors aleatoris.

Al marge del disseny, s'observà que les dades mostraven certa infradispersió, associada pressumiblement al fet que la taxa de gols d'ambdós equips en intervals concrets té correlació negativa. Per tal de tenir en compte aquestes correlacions negatives fou necessari, però, emprar un model marginal.

El model marginal proporciona interpretacions des d'un punt de vista poblacional en relació amb els nivells dels factors aleatoris que considera. En aquest sentit, el model permet comparar la mitjana de gols obtinguts pels equips en què s'expulsà un jugador enfront dels equips sense aquesta incidència.

Com a anàlisi de sensibilitat, la proposta basada en intervals de 5 minuts es validà considerant intervals de diferent durada, assolint resultats compatibles. Es va detectar que a mesura que els intervals considerats eren de major durada, la sobredispersió era també major fet associat a la correlació positiva d'aquestes mesures.

S'espera poder donar continuïtat a aquest treball avaluant l'impacte de les targetes grogues en la taxa de gols eliminant la confusió provocada per la intensitat del joc i incloent l'impacte indirecte provocat pel risc de rebre una doble amonestació.

6.3 Valoracions de l'article: Ultra log-concavity of discrete order statistics

Aquest treball conté dues innovacions importants relacionades amb els estadístics d'ordre de distribucions discretes i les seves propietats. Per un costat, es proposa una formulació original per als estadístics d'ordre discrets mitjançant una relació recursiva. Per altra part, es demostra que els estadístics d'ordre discrets preserven les propietats relatives a la log-concavitat i ultra log-concavitat de la distribució original. En particular, aquest resultat permet concloure que els estadístics d'ordre de la distribució de Poisson són infradispersos.

Els estadístics d'ordre (generalment el màxim, el mínim o la mediana) poden ser un recurs interessant per a simplificar dissenys experimentals que contemplen subrèpliques. Com que els estadístics d'ordre donen lloc a una reducció de la dispersió original, el soroll associat als errors de mesura es redueix tot mantenint la natura discreta de les dades.

Per exemple, en el cas d'un experiment on diferents observadors duen a terme el recompte de lesions de les mateixes imatges mèdiques. Es tracta efectivament d'un procés de recompte amb mesures repetides on cada imatge és avaluada en diferents ocasions. De fet, les diferències entre observadors són atribuïbles únicament a errors de detecció: falsos negatius o falsos positius. Condicionat a les variables explicatives i efectes aleatoris (és a dir, condicionat al recompte real, però desconegut, de lesions) la variabilitat residual de les dades és exclusivament deguda a l'error de mesura, que podria ser arbitràriament petit. A més, els efectes aleatoris no segueixen una distribució normal, de fet es tractaria d'una distribució de recompte. La situació descrita també es produeix en l'àmbit de l'estudi de l'abundància de fauna aviària niadora. Diferents observadors en diferents dies, duen a terme recomptes d'observació d'ocells seguint determinats transsectes. De nou es tracta d'un procés de recompte amb mesures repetides on cada transsecte és mesurat en diferents ocasions. En aquest àmbit se sol considerar que les diferències són degudes tan sols a falsos negatius. De nou, es tractaria d'un disseny amb presència d'un factor aleatori i, de la mateixa forma que abans, les suposicions en què es basa el model GLMM-Poisson no són vàlides. Per exemple, si els errors de mesura són prou petits, les observacions condicionades als efectes fixos i aleatoris mostrarien una infradispersió remarcable i aleshores la distribució de Poisson no seria vàlida.

En ambdós casos, si el nombre de mesures repetides en cada unitat és homogeni, una estratègia interessant consistiria a agregar d'alguna manera la informació. En processos amb dades quantitatives en què l'error és additiu, és habitual considerar aquesta intervenció a partir de la mitjana de les dades recollides per a la mateixa unitat. Quan el nombre de mesures de cada unitat és idèntic, l'anàlisi (i la inferència) d'aquestes mitjanes és equivalent al model que contemplaria les unitats com a factor aleatori. En els casos exposats, donat que es tracta de recomptes, considerar la mitjana per a agregar les dades no és una bona solució atès que, per un costat, s'altera la distribució natural de les dades (deixarien de ser recomptes discrets) i, per altre, l'error en aquest tipus de dades no és homogeni.

Alternativament, una altra operació d'agregació de dades que eliminaria les limitacions plantejades es basa en l'ús d'estadístics d'ordre. De fet, en l'àmbit de la imatge mèdica i el recompte de cèl·lules de diferents tipus és habitual repetir els recomptes i reportar la mediana d'aquests (Lesko et al., 2013; Nielson et al., 1991). En canvi, en el cas de l'abundància aviària, és usual considerar el valor màxim dels recomptes observats (Chamberlain et al., 2009; Hartill et al., 2011).

En resum, l'ús d'estadístics d'ordre per a agregar i resumir la informació de recomptes replicats pot esdevenir una estratègia pràctica per a minimitzar el soroll associat als errors de mesura.

Aquesta visió basada en els estadístics d'ordre, parteix de l'expressió de la funció de densitat d'aquests estadístics d'ordre emprant la funció beta incompleta:

$$P(X_{r:n} = k) = I_{F_k}(r, n+1-r) - I_{F_{k-1}}(r, n+1-r)$$

on $I_x(a,b) = \int_0^x t^{a-1}(1-t^{b-1}dt/B(a,b)$ i $B(a,b)$ és la funció Beta.

El present article utilitza que r i n són enters ≥ 1, ja que es tracta d'estadístics d'ordre. S'espera poder donar continuïtat al treball estudiant en detall el cas $r = n+1-r$ sense la restricció que siguin enters. Aparentment, es tractaria d'un model per a dades de recompte admetent sobredispersió i infradispersió.

6.4 Conclusions

Les dades de recompte sovint tenen particularitats que cal mirar d'incorporar en la seva modelització: heterogeneïtat entre individus, zero-inflació o deflació, presència d'errors de mesura, *outliers*, taxes d'esdeveniments no constants, etc. Per altra part, de vegades els dissenys experimentals contemplen mesures repetides, provocant que les observacions deixin de ser independents i requerint que els models ho tinguin en compte per a una correcta inferència.

En els apartats anteriors, tant en l'exposició més teòrica com en els exemples aplicats presentats s'han descrit els procediments per a l'anàlisi de dades de recompte amb me-

sures repetides emprant models lineals generalitzats mixtos sota les visions condicional i marginal.

Com s'ha exposat, existeix una certa controvèrsia oberta sobre la preferència d'una visió o una altra. Tenint en compte els avantatges i inconvenients d'ambdues propostes, sembla prou clar que des del punt de vista de la modelització estadística és preferible considerar models condicionals sempre que sigui viable. De fet, la necessitat d'haver de recórrer a models marginals és un símptoma de la mancança de propostes vàlides que proporcionin models de dades adients.

Una solució intermèdia per a aquelles situacions on l'experiment contempla subrèpliques pot consistir a agregar la informació disponible de les mesures repetides sota les mateixes condicions experimentals simplificant l'experiment. La mitjana d'aquestes subrèpliques no és una bona funció d'agregació, ja que distorsiona la distribució de les dades (les mesures podrien deixar de ser enteres). La suma de les subrèpliques és una possibilitat, però pot provocar un increment de la dispersió, i requeriria que el nombre de subrèpliques fos idèntic per a totes les observacions. Finalment, els estadístics d'ordre sí tenen la capacitat de resumir la informació de les subrèpliques, mantenint la naturalesa de les dades i controlant els errors de mesura.

CAPÍTOL 7

Referències

Albanese, A., De Meyere, A., Vanruymbeke, W. and Baert, S. (2020). Player dismissal and full-time results in the UEFA champions league and Europa league. International Journal of Sport Finance, 15(1), 27–38.

Alimohammadi, M., Alamatsaz, M. H. and Cramer, E. (2015). Discrete strong unimodality of order statistics. Statistics and Probability Letters, 103, 176–185.

Anders, A. and Rotthoff, K. W. (2011). Yellow cards: Do they matter? Journal of Quantitative Analysis in Sports, 7, 1–12.

Antoine, R., Perera, F. and Thiel, H. (2018). Tensor products and regularity properties of Cuntz semigroups (Vol. 251, No. 1199). American Mathematical Society.

Antoine, R., Perera, F., Robert, L. and Thiel, H. (2022). C*-algebras of stable rank one and their Cuntz semigroups. Duke Mathematical Journal, 171(1), 33-99.

Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (2008). A first course in order statistics. SIAM

Badiella, L. and Brewer, M. (2015). Equivalència entre els models mixtos condicionals i marginals. Materials de treball del Servei d'Estadística Aplicada UAB.

Badiella, L., Puig, P., Lago-Peñas, C. and Casals, M. (2023). Influence of Red and Yellow cards on team performance in elite soccer. Annals of Operations Research, 325(1), 149-165.

Badiella, L., del Castillo, J. and Puig, P. (2023). Ultra log-concavity of discrete order statistics. Statistics & Probability Letters, 109900.

Badiella, L. (2023). Dynamic Soccer Predictions. `https://servei-estadistica-uab2.shinyapps.io/SoccerPredictions/`

Bar-Eli, M., Sachs, S., Tenenbaum, G., Pie, J. S. and Falk, B. (1996). Crisis-related observations in competition: A case study in basketball. Scandinavian Journal of Medicine and Science in Sports, 6, 313–321.

Bar-Eli, M. and Tenenbaum,G. (1989). A theory of individual psychological crisis in competitive sport. Applied Psychology, 38, 107–120.

Bar-Eli, M., Tenenbaum, G. and Geister, S. (2006). Consequences of players' dismissal in professional soccer: A crisis-related analysis of group-size effects. Journal of Sports Sciences, 24, 1083–1094.

Bornn, L., Cervone, D. and Fernandez, J. (2018). Soccer analytics: Unravelling the complexity of the "beautiful game". Significance, 15, 26–29.

Breslow, N. E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88, 9-25.

Caliendo, M. and Radic, D. (2006). Ten do it better, do they?: An empirical analysis of an old football myth. IZA Discussion Paper 2158. Institute for the Study of Labor: Bonn.

Carling, C. and Bloomfield, J. (2010). The effect of an early dismissal on player work-rate in a professional soccer match. Journal of Science Medicine in Sport, 2010(13), 126–28.

Carmichael, F. and Thomas, D. (2005). Home-field effect and team performance evidence from English premiership football. Journal of sports Economics, 6, 264–281.

Cervený, J., van Ours, J. C. and van Tuijl, M. A. (2018). Effects of a red card on goal-scoring in World Cup football matches. Empirical Economics, 55, 883–903.

Chamberlain, D. E., Glue, D. E. and Toms, M. P. (2009). Sparrowhawk Accipiter nisus presence and winter bird abundance. Journal of Ornithology, 150(1), 247-254.

Chowdhury, A. (2015). Can ten do it better? Impact of red card in the English Premier League. No. 2015-01. Marquette University, Center for Global and Economic Studies and Department of Economics.

Chriqui, J. F., Taber, D. R., Slater, S. J., Turner, L., Lowrey, K. M. G. and Chaloupka, F. J. (2012). The impact of state safe routes to school-related laws on active travel to school policies and practices in US elementary schools. Health Place, 18(1), 8–15.

Clarke, R. D. (1946). An application of the Poisson distribution. Journal of the Institute of Actuaries, 72(3), 481.

Cleveland,W. S. (1979). Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association, 74(368), 829–836.

Consul, P. C. (1989). Generalized Poisson Distribution: Properties and Applications. Marcel Dekker, New York.

Conway, R. W. and Maxwell, W. L. (1962). Network dispatching by the shortest-operation discipline. Operations Research, 10(1), 51-73.

del Castillo, J., Pérez-Casany, M. (2005). Overdispersed and underdispersed Poisson generalizations. Journal of Statistical Planning and Inference, 134(2), 486-500.

Diggle, P., Liang, K. Y., and Zeger, S. L. (1994). Longitudinal data analysis. New York: Oxford University Press, 5, 13.

DiMaggio, C. and Li, G. (2013) Effectiveness of a safe routes to school program in preventing school-aged pedestrian injury. Pediatrics, 131(2), 290–296.

DiMaggio, C., Brady, J. and Li, G. (2015) Association of the Safe Routes to School program with school-age pedestrian and bicyclist injury risk in Texas. Injury Epidemiology, 2(1), 15.

DiMaggio, C., Frangos, S. and Li, G. (2016) National Safe Routes to School program and risk of school-age pedestrian and bicyclist injury. Annals of Epidemiology, 26(6), 412–417.

Emparanza, I. D. and Núnez-Antón, V. (2010) . On the use of simulation methods to compute probabilities: application to the first division Spanish soccer league. SORT-Statistics and Operations Research Transactions, 34, 181-200.

Erlang, A. K. (1909). Sandsynlighedsregning og Telefonsamtaler Nyt Tidsskrift for Matematik, 20(B), 33–39.

Fieberg, J., Rieger, R. H., Zicus, M. C., and Schildcrout, J. S. (2009). Regression modelling of correlated data in ecology: subject-specific and population averaged response patterns. Journal of Applied Ecology, 46(5), 1018-1025.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character, 222(594-604), 309-368.

Football-Data.co.uk. (2020). Available online: http://www.football-data.co.uk. Accessed on July 2023.

Gómez-Déniz, E., Cárdenes, N. D. and Sánchez Pérez, J. M. (2019). A probabilistic model for explaining the points achieved by a team in football competition. Forecasting and regression with applications to the Spanish competition. SORT-Statistics and Operations Research Transactions, 43, 95–112.

Greenberg, A. (2015). The red card cliché. Significance, 12, 30–33.

Gross, J. L. Mansour, T., Tucker, T. W. and Wang, D. G. (2015). Log-concavity of combinations of sequences and applications to genus distributions. SIAM Journal on Discrete Mathematics, 29(2), 1002-1029.

Guikema, S. D. and Coffelt, J. P. (2008). A Flexible Count Data Regression Model for Risk Analysis. Risk Analysis, 28, 213–223.

Hagel, B. E., Macpherson, A., Howard, A. et al.(2019). The built environment and active transportation safety in children and youth: a study protocol. BMC Public Health, 19(1), 728.

Harremoës, P. (2001). Binomial and Poisson distributions as maximum entropy distributions. IEEE Transactions on Information Theory, 47(5), 2039-2041.

Hartill, B. W., Watson, T. G. and Bian, R. (2011). Refining and applying a maximum-count aerial-access survey design to estimate the harvest taken from New Zealand's largest recreational fishery. North American Journal of Fisheries Management, 31(6), 1197-1210.

Hoelscher, D. M., Ganzar, L. A., Salvo, D. et al. (2022) Effects of large-scale municipal safe routes to school infrastructure on student active travel and physical activity: design, methods, and baseline data of the Safe Travel Environment Evaluation in Texas Schools (STREETS) natural experiment. International Journal of Environmental Research and Public Health, 19(3), 1810.

Jiang, R. (2010). Discrete competing risk model with application to modeling bus-motor failure data Reliability Engineering and System Safety, 95(9), 981-988.

Johnson, O. (2007). Log-concavity and the maximum entropy property of the Poisson distribution. Stochastic Processes and their Applications, 117(6), 791-802.

Keilson, J. and Gerber, H. (1971). Some results for discrete unimodality. Journal of the American Statistical Association, 66(334), 386-389.

Kang, B. (2019). Identifying street design elements associated with vehicle-to-pedestrian collision reduction at intersections in New York City. Accident Analysis and Prevention, 122, 308–317.

Kim, B., Kim, J. and Lee, S. (2018). Strong unimodality of discrete order statistics. Statistics and Probability Letters, 140, 48-52.

Lago-Peñas, C., Gómez-Ruano, M. A., Owen, A. L. and Sampaio, J. (2016). The effects of a player dismissal on competitive technical match performance. International Journal of Performance Analysis in Sport, 16, 792–800.

Lee, Y. and Nelder, J.A. (2004). Likelihood Inference for Models with Unobservables: Another View. Statistical Science, 24(3), 255-269.

Lee, Y., Ronnegard, L. and Noh, M. (2017). Data analysis using hierarchical generalized linear models with R. CRC Press.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2018). Generalized Linear Models with Random effects, 2nd Edition. Chapman and Hall/CRC.

Lesko, C. R., Cole, S. R., Zinski, A., Poole, C. and Mugavero, M. J. (2013). A systematic review and meta-regression of temporal trends in adult CD4+ cell count at presentation to HIV care, 1992–2011. Clinical infectious diseases, 57(7), 1027-1037.

Lindsey, J.K., and Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. Statistics in medicine, 17(4), 447-469.

Litière, S., Alonso, A. and Molenberghs, G. (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Statistics in medicine, 27(16), 3125-3144.

Liu, L. and Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. Statistics in medicine, 27(16), 3105-3124.

Lizarazo, C. G., Hall, T. and Tarko, A. (2021). Impact of the Safe Routes to School Program: comparative analysis of infrastructure and noninfrastructure measures in Indiana. Journal of Transportation Engineering, Part A: Systems, 147(1).

Mattera, R. (2021). Forecasting binary outcomes in soccer. Annals of Operations Research, 1–20.

McGarry, T., O'Donoghue, P. and de Eira Sampaio, A. J. (2013). Routledge handbook of sports performance analysis. London: Routledge.

Mechtel, M., Bäker, A., Brändle, T. and Vetter, K. (2011). Red cards: not such bad news for penalized guest teams. Journal of Sports Economics, 12, 621–646.

McCullagh, P. and Nelder, J. (1983). Generalized Linear Models. Boca Raton, FL: Chapman and Hall/CRC.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Statistical Science, 26, 388–402.

Molenberghs G. and Verbeke G. (2005). Models for discrete longitudinal data. Springer, New York

Molenberghs, G., Verbeke, G. and Demétrio, C. G. (2017). Hierarchical models with normal and conjugate random effects: a review. Statistics and Operations Research Transactions, 41(2), 191-253.

Muennig, P. A., Epstein, M., Li, G. and DiMaggio, C. (2014) The cost-effectiveness of New York City's Safe Routes to School program. American Journal of Public Health, 104(7), 1294–1299.

Muff, S., Held, L. and Keller, L. F. (2016). Marginal or conditional regression models for correlated non-normal data?. Methods in Ecology and Evolution, 7(12), 1514-1524.

Nakagawa, S. and Holger, S. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133–142.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society Series A: Statistics in Society, 135(3), 370-384.

Newcomb, S. (1860). Notes on the theory of probabilities The Mathematical Monthly 2, 134-140.

Nielson, L., Smyth, G. and Greenfield, P. (1991). Hemacytometer cell count distributions: implications of non-Poisson behavior. Biotechnology progress, 7(6), 560-563.

Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D. and Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific Data, 6, 1–15.

Pérez, K., Santamariña-Rubio, E., Ferrando, J., López, M. J. and Badiella, L. (2023). Effectiveness of a Road Traffic Injury Prevention Intervention in Reducing Pedestrian

Injuries, Barcelona, Spain, 2002–2019. American journal of public health, 113(5), 495-499.

Poisson, S. (1837). Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile Prècédées des Règles Générales du Calcul des Probabilités Bachelier, Paris

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics, 4, 12–35.

Prékopa, A. (2013). Stochastic programming (Vol. 324). Springer Science and Business Media.

Ridder, G., Cramer, J. S. and Hopstaken, P. (1994). Down to ten: Estimating the effect of a red card in soccer. Journal of the American Statistical Association, 89, 1124–1127.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. Annual Review of Statistics and Its Application, 4, 395-421.

Rutherford, E., Geiger, H. and Bateman, H. (1910). LXXVI. The probability variations in the distribution of α particles. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 20(118), 698-707.

Sapp, R. M., Spangenburg, E. E. and Hagberg, J. M. (2019). Markers of aggressive play are similar among the top four divisions of English soccer over 17 seasons. Science and Medicine in Football, 3, 125–130.

Saumard, A. and Wellner, J. A. (2014). Log-concavity and strong log-concavity: a review. Statistics surveys, 8, 45.

Saurí, E., Sintes, E. and Truñó, M. (2017). Avaluació Del Programa Camí Escolar, Espai Amic. Available at: https://institutinfancia.cat/es/mediateca/informe-davaluacio-programa-cami-escolarespai-amic. Accessed August 7, 2022.

Sellers, K. F., Borle, S. and Shmueli, G. (2012). The COM-Poisson model for count data: a survey of methods and applications. Applied Stochastic Models in Business and Industry, 28(2), 104-116.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution. Journal of the Royal Statistical Society Series C: Applied Statistics, 54(1), 127-142.

Stewart, O., Moudon, A. V. and Claybrooke, C. (2014) Multistate evaluation of safe routes to school programs. American Journal of Health Promotion, 28(3 suppl), S89–S96.

a student. (1907). On the error of counting with a haemacytometer Biometrika, 5(3), 351–360.

Titman, A., Costain, D., Ridall, P. and Gregory, K. (2015). Joint modelling of goal and bookings in association football. Journal of the Royal Statistical Society: Series A, 178(3), 659–683.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Computational Statistics and Data Analysis, 23(4), 541-556.

von Bortkewicz L. (1898). Das Gesetz der Kleinen Zahlen. Teubner: Leipzig.

Williams, C. B. (1944). Some applications of the logarithmic series and the index of diversity to ecological problems. The Journal of Ecology, 1-44.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48,233–243.

Wu, H., Gao, L. and Zhang, Z. (2014). Analysis of crash data using quantile regression for counts. Journal of transportation engineering, 140(4), 04013025.

Wunderlich, F. and Memmert, D. (2018). The betting odds rating system: Using soccer forecasts to forecast soccer. PloS One, 13, e0198668.

Xekalaki, E. (2014). Under-and overdispersion. Wiley StatsRef: Statistics Reference Online, 1-9.

Yu, C. Y. (2015). How differences in roadways affect school travel safety. Journal of the American Planning Association, 81(3), 203–220.