

Considerazioni statistiche sui limiti di incertezza nella classificazione delle aziende

di Alberto Baldasseroni
CeRIMP Regione Toscana

“Truth emerges more readily from error
than from confusion”

Francis Bacon (1562-1626)

Natura dei fenomeni osservati

Gli infortuni sono:

- Eventi discreti, ripetibili nello stesso soggetto
- Rari (37,4 per mille addetti nel 2008)
- Sono la realizzazione di un danno a partire da un evento incidentale, che ne rappresenta la condizione necessaria, ma non sufficiente.
- Emergono da una “massa a rischio” non direttamente conosciuta, ma stimata, soggetta a varie fonti di errore.

Incidente

- De Mauro: fatto inatteso, spec. negativo che viene a turbare lo svolgimento di un'attività o il normale corso degli eventi; disgrazia, sciagura: *i. sul lavoro, i. aereo, ferroviario, i. domestico*
- Vocabolario etimologico della lingua italiana di Ottorino Pianigiani

incidente = *lat. INCIDENTEM* part. pres. di **INCIDERE** *accadere, sopravvenire*, comp. della partic. **IN** *in* e **CIDERE** per **CADERE** *cadere* (v. q. voce).

Che accade o avviene di passaggio, ossia indirettamente, mentre si fanno altre cose; come *sost.* Evento, o Circostanza particolare di qualche evento.

Deriv. *Incidentale; Incidentemente; Incidenza.*

Infortunio

- DeMauro: incidente che provoca danni fisici più o meno gravi: *i. sul lavoro, subire un i., prevenire gli infortuni; assicurazione sugli, contro gli infortuni*
- Vocabolario etimologico della lingua italiana di Ottorino Pianigiani

infortúnio = *lat.* INFORTÚNIUM comp della part. IN per non e FORTÚNA *favo revole accidente* (v. *Fortuna*).

Tristo accidente sopravvenuto senza malvagità altrui o volontà propria.

Infortunio

- Evento dovuto a causa fortuita, violenta ed esterna (devono ricorrere tutti e tre questi requisiti), che produce lesioni oggettivamente constatabili, le quali abbiano per conseguenza la morte, una invalidità permanente oppure una inabilità temporanea.

(da una tesi sulle assicurazioni).

Infortunio

- SSI* : Incidente che ha provocato un danno biologico di rilevanza clinica ad una o più persone, con un brevissimo intervallo di tempo tra il momento in cui l'energia ha agito e quello in cui è iniziato il danno.

* Sbagliando S'Impara

There are a few in public health who believe that confidence intervals should not be used around estimates derived from 'population' statistics such as the death rate in a given population, because they believe there is no statistical uncertainty in such estimates. This belief is contrary to the statistical theory underlying confidence intervals, and the biological and random processes governing the occurrence of events such as deaths and illnesses (Brillinger,1986).

The Achilles Heel of Epidemiology...

problemi con i “piccoli numeri”

- Confidenzialità
 - Privacy (non ne parleremo, per fortuna, almeno stavolta)
- Incertezza statistica
 - Calcolo di tassi e LC
 - Problema della sovradisersione

Incertezza Statistica

- Eventi o addetti ?

Risposta: Eventi ! Lavorare sugli addetti espone a critiche circa il differente significato di considerare per esempio una ditta da 10 addetti nel settore “assicurativo-impiegatizio” e una ditta da 10 addetti nel settore edile. La prima non avrà un evento (grave !) nei successivi 20 o 30 anni, la seconda ne può aver avuti anche molteplici nel corso dei precedenti due tre anni ed è destinata ad averne anche nel futuro.

Incertezza Statistica

In termini più formali: dato che la vera “massa a rischio” è per definizione ignota (se ne ha una stima approssimata dal Numero di lavoratori equivalenti) e anche l'entità del rischio è solo stimata (attraverso il settore produttivo, per esempio) gli addetti non possono rappresentare un criterio sufficiente di cut-off.

Incertezza Statistica

Inoltre è noto che l'informazione in epidemiologia è completamente veicolata dal concreto realizzarsi dell'evento (infortunio, malattia, altra evenienza per la salute e il benessere degli individui), non dalle potenziali occasioni del suo verificarsi (massa a rischio).

Site Directory

- [DOH Home](#)
- [Health Data](#)
- [Health Alerts](#)

- [Publications](#)
- [Data Analysis Guidelines](#)



Data Guidelines

Guidelines for:

- [Address Matching and Geocoding Data](#)
- [Confidence Intervals for Public Health Assessment](#)
- [Human Subjects Review](#)
- [Population Denominators](#)
- [Racial and Ethnic Groups in Data Analyses](#)
- [Rates for Public Health Assessment](#)
- [Rural-Urban Classification Systems for Public Health Assessment](#)
- [Small Numbers](#)

General Purpose

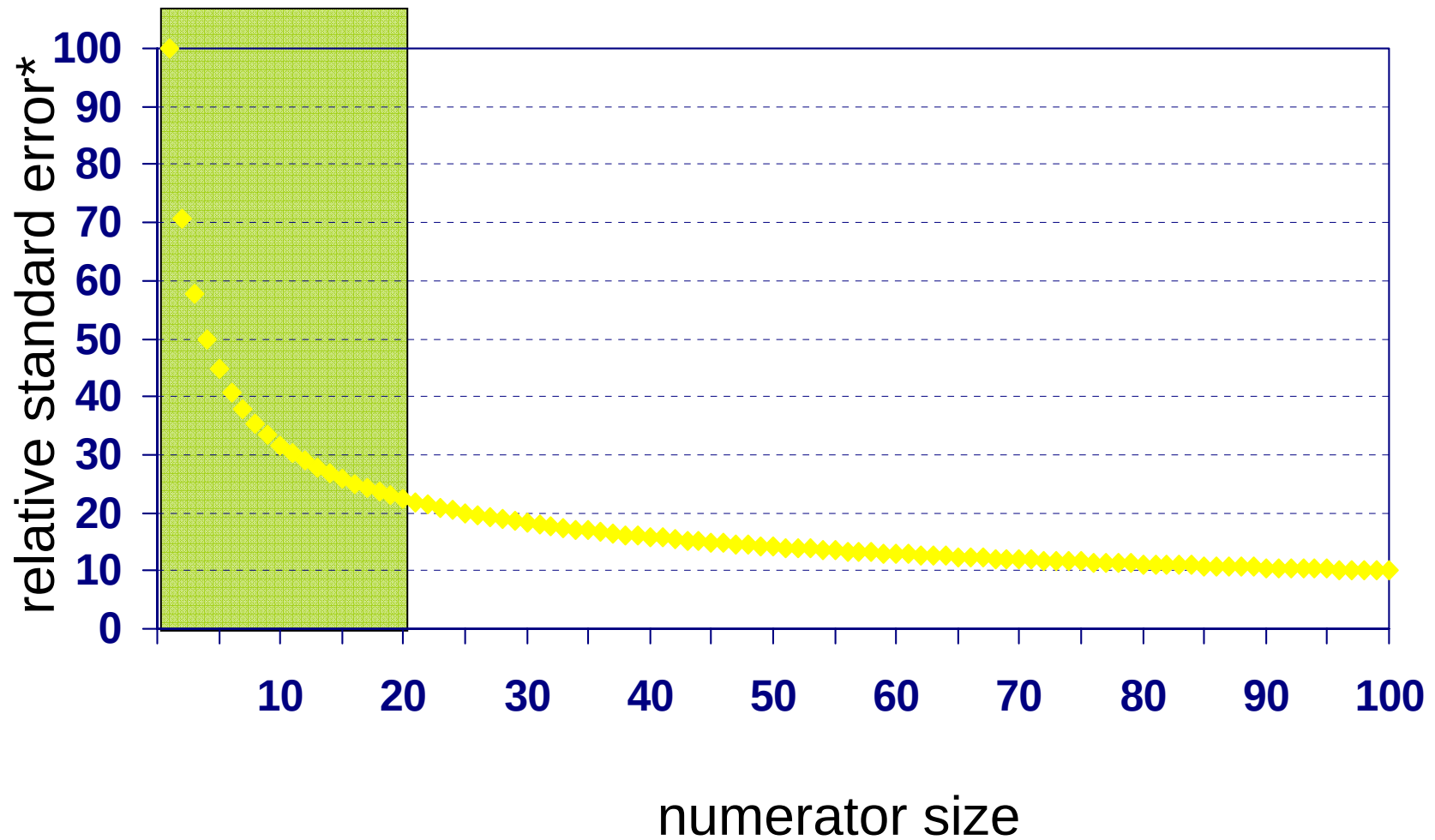
The Assessment Operations Group in the Washington State Department of Health is coordinating the development of guidelines for public health assessment. The guidelines are intended to promote good professional practice among staff involved in assessment activities within the Washington State Department of Health and Local Health Jurisdictions. Our purpose is to summarize what we consider best practice to encourage all those in public health to be better users of assessment information.

The guidelines focus on issues commonly encountered in public health practice and where applicable, on issues unique to Washington state. The guidelines are intended for audiences with differing levels of public health assessment training. They assume a basic knowledge of epidemiology and biostatistics and are not intended to recreate basic texts and other sources of information. Some of these are listed under references and resources at the end of each topic.

$$\text{E.S.}(\lambda) = \lambda / \sqrt{\text{casi}}$$

Approssimazione Poissoniana

$$\text{E.S.R.} = (\text{E.S.} / \lambda) \times 100 = (\lambda / \sqrt{\text{casi}}) / \lambda \cdot 100 = 100 / \sqrt{\text{casi}}$$



How to address the statistical issues

Increase numerator size. In preparing a data table for dissemination, it is recommended that analysts first examine the counts in each cell of the table. If rates are desired and the numerator of any cell is less than 20, an effort should be made to increase the size of the numerator. (Use of 20 events as the threshold for reliability is consistent with standard CDC practice.) Techniques to accomplish this include the following:

- 1. Combine multiple years of data,**
- 2. Collapse data categories, and/or**
- 3. Expand the geographic area under consideration.**

How to address the statistical issues

Include confidence intervals. The inclusion of confidence intervals for rates is strongly recommended regardless of the number of health events, but it is especially important when the count is less than 20. Generally, rates with fewer than 20 events in the numerator have very wide confidence intervals. For example, an infant death rate of 10 per 1,000, based on 20 deaths out of a population of 2,000 live births, has a Poisson-based 95% confidence interval between 6 and 15. Clearly, this is not very precise information and users of the data need to know this. In instances where it is not feasible to incorporate confidence intervals into a data table it is recommended that analysts:

- 1. Always report the numerator on which the rate is based and**
- 2. Include a footnote indicating that rates based on fewer than 20 events are likely to be unstable and imprecise.**

How to address the statistical issues

- ***Suppress rates.*** Suppress rates based on very small numbers (i.e., fewer than 5 health events), reporting only the count (numerator). When rates are suppressed, tables should be constructed such that an indicator (e.g., asterisk) appears in the cell and a legend under the table explains the reason for suppression.

How to address the statistical issues

- ***Suppress confidence intervals.*** When rates based on very small numbers (i.e., fewer than 5 health events) are suppressed, confidence intervals should also be suppressed. When confidence intervals are suppressed, tables should be constructed such that an indicator (e.g., asterisk) appears in the cell and a legend under the table explains the reason for suppression.

Riassumendo ...

- > 20 eventi: calcolo di tassi e rispettivi LC
- >4 e < 20 eventi: calcolo dei tassi possibile, ma sconsigliato; LC molto ampi da enfatizzare. Warning sull'imprecisione delle stime.
- < 5 eventi: nessun calcolo di tassi. Warning sulla mancanza d'informatività.

Fenomeno della “sovera-dispersione”



American Journal of Epidemiology
© The Author 2009. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of
Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oxfordjournals.org.

Vol. 170, No. 10
DOI: 10.1093/aje/kwp265
Advance Access publication:
October 7, 2009

Practice of Epidemiology

Analyses of Injury Count Data: Some Do's and Don'ts

Ian Shrier*, Russell J. Steele, James Hanley, and Benjamin Rich

* Correspondence to Dr. Ian Shrier, Centre for Clinical Epidemiology and Community Studies, SMBD-Jewish General Hospital, 3755
Cote Sainte Catherine Road, Montreal, Quebec H3T 1E2, Canada (e-mail: ian.shrier@mcgill.ca).

Initially submitted April 16, 2009; accepted for publication August 5, 2009.

Downloaded from oem.bmj.com on 29 October 2009

Original article

Regression models for public health surveillance data: a simulation study

H Kim, D Kriebel

Fenomeno della “sovra-dispersione”

La distribuzione di Poisson assume che la struttura di errore nei dati prevede l'equivalenza fra media e varianza (cosiddetta “equidispersione”. Se esiste “sovradisersione”, cioè $\text{varianza} > \text{media}$, allora la misura di incertezza che ne deriva è sbagliata (troppo stretta).

Cause di sovradisersione:

- Non indipendenza degli eventi
 - Mancata osservazione di covariate influenti.
- Errata rappresentazione della massa a rischio (offset nel calcolo del tasso).

Poisson

La distribuzione di probabilità di Poisson prevede la “indipendenza statistica” delle unità di osservazione, ossia la equiprobabilità del verificarsi di un evento. Questo assunto è violato quando i soggetti sperimentino condizioni individuali di *pronesse* all'evento differenti fra di loro, determinate da covariate delle quali non siano registrati i valori nel data-base di partenza (nel nostro caso quello dell'INAIL)

Multiple admission

Per esempio, nel caso delle schede di dimissione ospedaliera è implicito che la probabilità di un singolo ricovero dipenderà dalla “storia” di ricoveri precedenti del soggetto interessato (es. dei “Revolving doors” in psichiatria). Ma anche nel caso degli infortuni sul lavoro covariate individuali possono produrre differenti probabilità di ripetere l'evento infortunio nel singolo soggetto.

Mancata osservazione di covariate influenti

- Sottonotifica differenziale.
 - Differente tendenza nelle ditte artigiane a segnalare le lesioni lievi e lievissime rispetto alle ditte dell'industria.
- Tendenza degli obesi ad avere più frequenti infortuni
- Tendenza in chi abusa di alcol ad avere più infortuni.
- Tendenza in chi presenta disturbi del sonno ad aumento del numero di errori e, quindi, di infortuni.
- “Risk of repeat accidents by economic activity in Italy”

Errata rappresentazione della massa a rischio.

- Numero di infermieri nel denominatore di un tasso di infortuni alla schiena in ospedale vs. n.di malati non autosufficienti.
- N.di dipendenti di una ditta di trasporto su strada nel denominatore del tasso di infortuni stradali.
- N.di dipendenti stimati sulla base del monte salario in regioni con alta o bassa evasione all'obbligo di dichiarazione del monte salari effettivo.

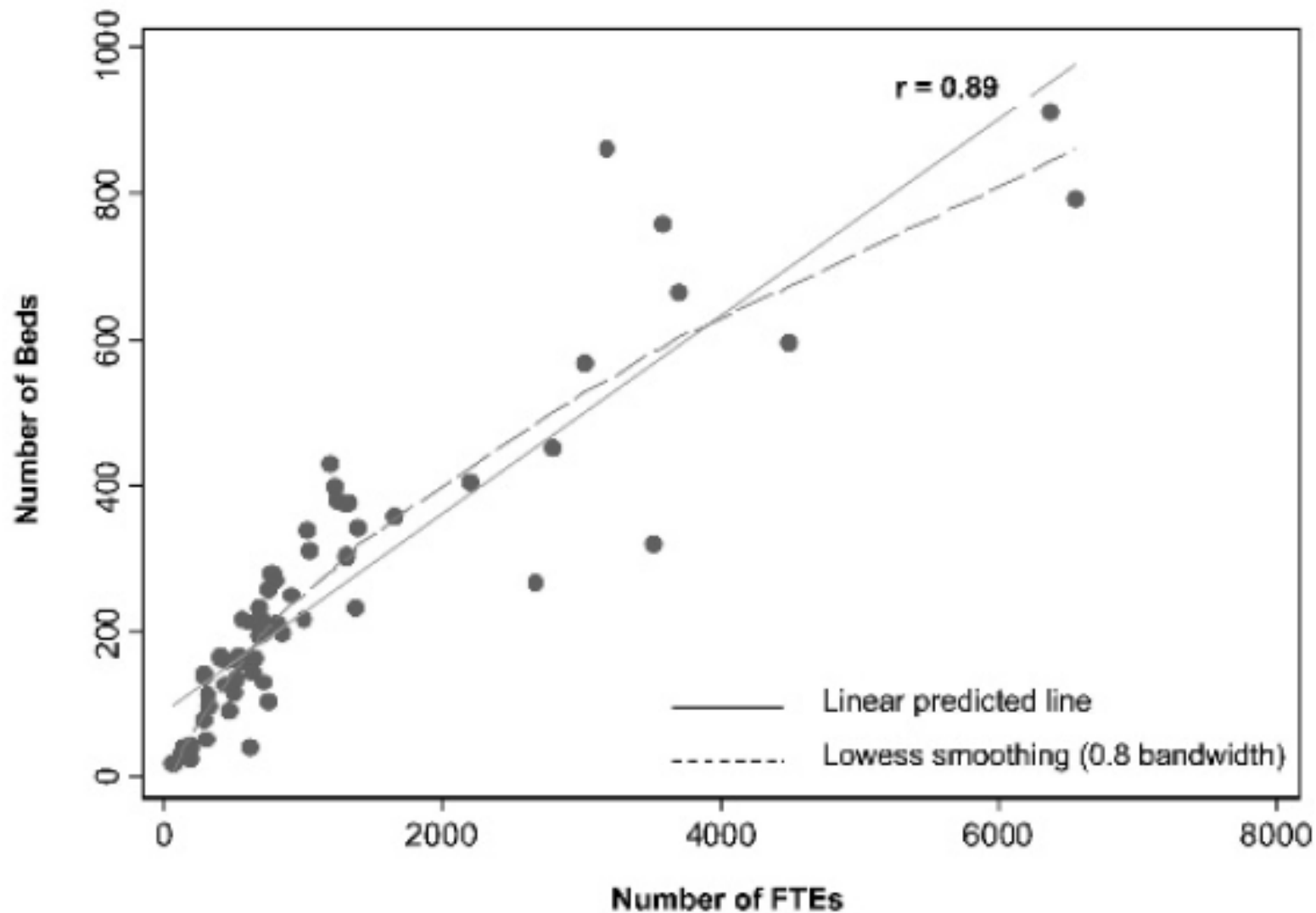


Figure 1 Comparing denominators (offsets) for sharps injury rates. Number of beds versus number of FTEs among 64 Massachusetts hospitals, 2002–2003. FTE, full-time equivalent worker.

Kim, Kriebel, 2009

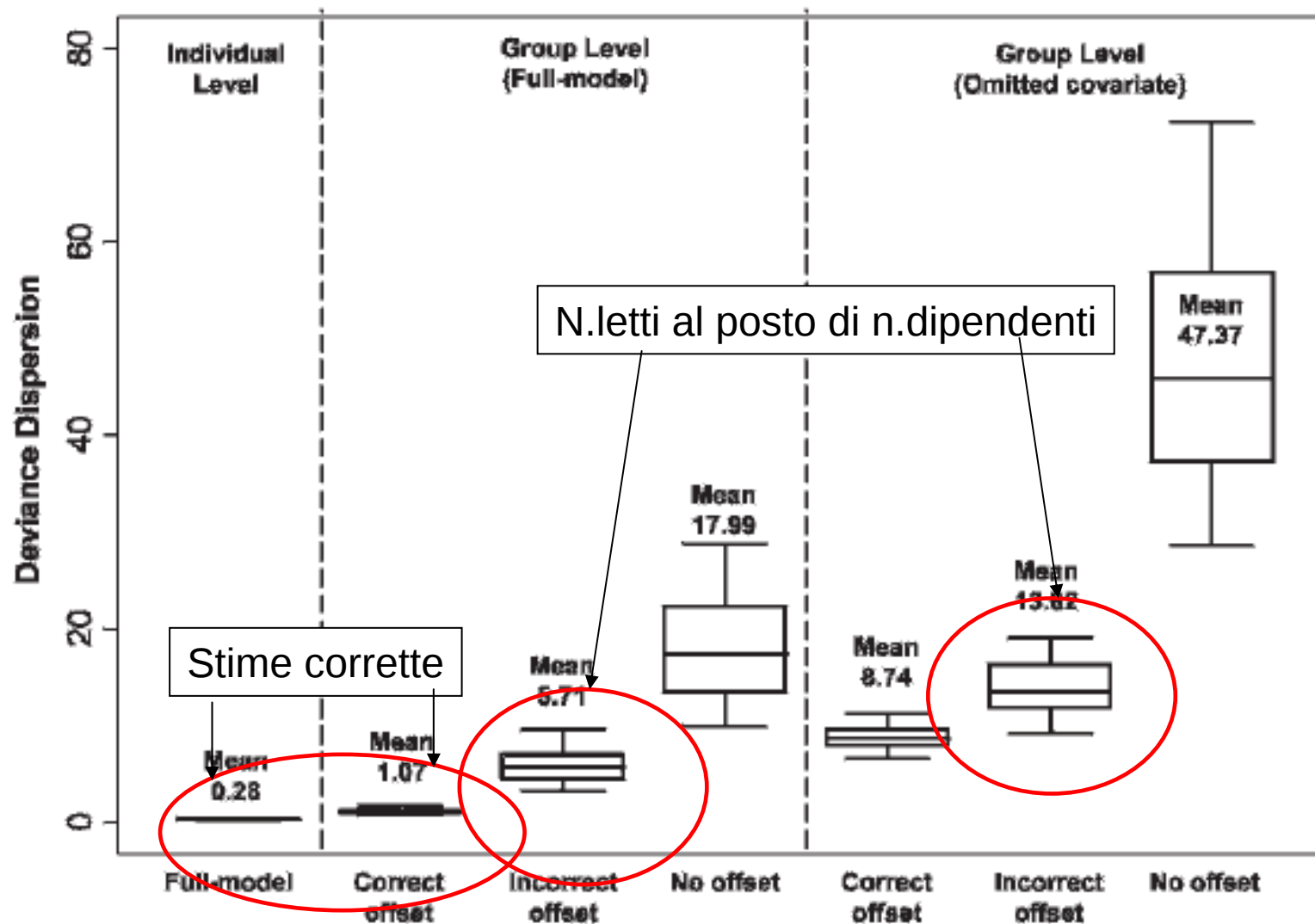


Figure 3 Mean Poisson deviance dispersion statistics for the 100 simulated datasets and six alternative models. Box and whisker plots indicate mean, inter-quartile range and minimum and maximum values.

Fenomeno della “sovra-dispersione”

- Sperimentare l'uso della distribuzione Binomiale Negativa al posto della distribuzione di Poisson.
- Se del caso inserire apposito algoritmo nel programma Epiwork.