Confidence Intervals for Injury Surveillance Data using Negative Binomial Regression

Steve Marshall, University of North Carolina at Chapel Hill

Jill Corlette, National Collegiate Athletic Association

Julie Agel, University of Minnesota

Randall Dick, National Collegiate Athletic Association

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

UNC Injury Prevention Research Center
IPRC

EXSS
Exercise and Sport Science
University of North Carolina at Chapel Hill

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Outline

- Non-independence in injury rates
- What Is Over-Dispersion & How Is It Identified?
- Dealing with Over-Dispersion
- Negative Binomial Distribution
- Negative Binomial Regression
- Example: NCAA Soccer Injury Data
- Summary

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Non-independence in injury rates

- **Often compute incidence rates from injury surveillance data**
  - Assumes counts follow independent Poisson process
- **THIS IS OFTEN NOT TRUE!**
  - mass fatality events resulting from natural disaster or terrorism
  - multiple fatalities in a car crash (FARS data)
  - mass homicide-suicide events stemming from intimate partner violence (NVDRS data)
  - multiple injuries to the same athlete (NCAA data)
- **More realistic to assume negative binomial process**

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

What Is Over-Dispersion & How Is It Identified?

- **Overdispersion**
  - Caused by positive correlation between the counts
  - Extra-Poisson variation
  - Small amounts of over-dispersion are okay

- **Poisson regression models:**
  - are over-dispersed if **Pearson chi-square/df > 1.5**
  - Betas (ln rates) are okay
  - Standard errors are too small

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Dealing with Over-Dispersion

1. PSCALE or DSCALE – fit the Poisson model, then Scale the SEs
   - simply scale-up the Standard Errors to account for the over-dispersion, as an additional step **after** the standard Poisson model-fitting process
   - Multiply the estimated variance from standard Poisson regression by either Pearson χ^2/df or deviance/df
   - $V' = V\phi$
   - ϕ = Pearson χ^2/df or deviance/df
   - In SAS, use GENMOD's PSCALE (χ^2) or DSCALE (deviance).
   - In STATA, use XTPOISSON with SCALE(X2) option .

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Dealing with Over-Dispersion

2. Negative Binomial regression with a **NB2** model
   - In **NB2**, the variance is obtained by combining the dispersion parameter α with a quadratic function of the mean
   - $V' = V\phi = \mu(1+\alpha\mu) = \mu + \alpha\mu^2$
   - Thus, the variance is obtained by scaling the mean (μ) by a function of the mean ($1+\alpha\mu$)
   - In STATA, use NBREG
   - In SAS, use PROC COUNTREG (experimental) or PROC GENMOD

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Dealing with Over-Dispersion

3. Negative Binomial regression with a **NB1** model
   - Multiply the variance by an over-dispersion parameter (α) estimated **during** the model-fitting process
   - Use the dispersion parameter (α) to scale the mean by a fixed quantity ($1+\alpha$)
   - $V' = V\phi = \mu(1+\alpha) = \mu + \alpha\mu$
   - In STATA, use NBREG
   - In SAS, use PROC COUNTREG (experimental)

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Binomial Distribution

Recall that the binomial distribution:

$$p(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

describes the number of successes (y) in n trials, where the trials are statistically independent and have a binary outcome (success or failure) with the constant probability of success (p) that is fixed between trials. The binomial distribution has mean np and variance $np(1-p)$.

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Negative Binomial Distribution

The negative binomial distribution describes a similar situation. Instead of describing the number of successes (*y*) in *n* trials, it describes the number of the trial on the which r^{th} success occurs, where each success is preceded by *y-1* failures. Specifically,

$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r}$$

with the probability of success (*p*) fixed between trials, where *y=r, r+1, r+2,…* etc. The negative binomial distribution has mean $\frac{r}{p}$ and variance $\frac{r(1-p)}{p^2}$

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Negative Binomial Regression

The negative binomial regression model:

- Accommodates heterogeneity in the counts
- Gamma mixture of Poisson random variables

$$\mathrm{E}\left[\lambda | x_1 \ldots x_k, \tau_i\right] = \mu_i \tau_i = e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon_i}$$

The terms in red (τ_i, ε_i) are included in the negative binomial regression model but are not included in Poisson regression. These are the terms that model over-dispersion.

Here, $\tau_i = e^{\varepsilon_i}$. If we assume that $E[\tau_i] = E[e^{\varepsilon_i}] = 1$, then the τ_i follow a gamma distribution with mean 1 and variance θ_i^{-1}, where $\theta > 0$.

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

NCAA Soccer Injury Data

- **Injury data from NCAA's Injury Surveillance System (ISS)**
  - Mens and womens soccer, 15 years (1988-89 to 2002-03)

- **Variables**
  - Sex (mens vs. womens soccer)
  - Year (linear trend)
  - Game vs. Practice exposure (GAMEPRAC)
  - Division (D1, D2, D3)
  - Time of Season (Pre-Season, Regular In-Season, Post-Season)

- **Outcome = injury rate.**
  - Numerator are number of injuries in each cell
  - Denominator are number of athlete-exposures (aka A-Es)

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

NCAA Soccer Injury Data

Tabular dataset

- One observation on dataset represents each cell in the m-way table formed by 2 sexes X 16 years X 2 exposure types X 3 Divisions X 3 time of season = 536 observations

- If this dataset were further disaggregated down to the school-level or athlete-level, there might be less over-dispersion.

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

NCAA Soccer Injury Data

- Model 1: Standard Poisson regression
  - Model 1 ignores over-dispersion
- Model 2: PSCALED Poisson regression
- Model 3: NB2 negative binomial regression
- Model 4: NB1 negative binomial regression
  - Models 2-4 account for over-dispersion

- Summary Table for Game Vs. Practice Variable

Model 1: Standard Poisson regression

```
proc genmod data=ncaa.ncaa8904_v6;
        /* Regression on Injury Rates for Historical NCAA data*/
        class division season gameprac sex ;
        model injuries = sex year gameprac division season / link=log
             dist=poisson offset=log_AEs;
        run ;
```

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	528	2248.6887	4.2589
Scaled Deviance	528	2248.6887	4.2589
Pearson Chi-Square	528	2062.2894	3.9059
Scaled Pearson X2	528	2062.2894	3.9059
Log Likelihood		83699.4450	

Over-dispersion is clearly present since $X^2/df >> 1$

Model 1: Standard Poisson regression

RESULTS

Sport_Name	Games_and_Practices_Combined	Rate Ratio	Lower Confidence Limit	Upper Confidence Limit	Confidence Limit Ratio	ChiSquare	Pr > ChiSq
Soccer	Exp(ln RR MSO vs WSO)	1.02	1.00	1.05	1.05	2.65	0.1036
Soccer	Exp(ln RR Av Ann 10Yr)	1.03	1.00	1.06	1.06	3.59	0.0581
Soccer	Exp(ln RR Game vs Prac)	5.56	5.40	5.73	1.06	12970.55	<.0001
Soccer	Exp(ln RR D2 vs D1)	0.96	0.92	0.99	1.07	6.44	0.0112
Soccer	Exp(ln RR D3 vs D1)	0.87	0.84	0.89	1.06	98.50	<.0001
Soccer	Exp(ln RR PreSeason vs InSeason)	2.61	2.53	2.69	1.06	3619.82	<.0001
Soccer	Exp(ln RR PostSeason vs InSeason)	0.69	0.63	0.75	1.19	70.01	<.0001

Interpretation: The rate of injury is 5.6 times higher in games than in practices (95%CI: 5.4, 5.7; CLR=1.1).

Problem: These CIs are falsely narrow because the model does not account for the over-dispersion.

Model 2: PSCALED Poisson regression

```
title1'PScaled Poisson Regression on Injury Rates for Historical NCAA data';
proc genmod data=ncaa.ncaa8904_v6;
        /* Regression on Injury Rates for Historical NCAA data*/
        class division season gameprac sex ;
        format division div. season sea. gameprac gp. ;
        model injuries = sex year gameprac division season
                  / pscale link=log dist=poisson offset=log_AEs
        run ;
```

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	528	2248.6887	4.2589
Scaled Deviance	528	575.7231	1.0904
Pearson Chi-Square	528	2062.2894	3.9059
Scaled Pearson X2	528	528.0000	1.0000
Log Likelihood		21429.2461	

The log-likelihood produced by GENMOD is **incorrect** in SAS ver9.X

Model 2: PSCALED Poisson regression

Sport_Name	Games_and_Practices_Combined	Rate Ratio	Lower Confidence Limit	Upper Confidence Limit	Confidence Limit Ratio	ChiSquare	Pr > ChiSq
Soccer	Exp(ln RR MSO vs WSO)	1.02	0.97	1.07	1.11	0.68	0.4102
Soccer	Exp(ln RR Av Ann 10Yr)	1.03	0.97	1.09	1.13	0.92	0.3377
Soccer	Exp(ln RR Game vs Prac)	5.56	5.25	5.90	1.12	3320.80	<.0001
Soccer	Exp(ln RR D2 vs D1)	0.96	0.89	1.02	1.15	1.65	0.1991
Soccer	Exp(ln RR D3 vs D1)	0.87	0.82	0.92	1.12	25.22	<.0001
Soccer	Exp(ln RR PreSeason vs InSeason)	2.61	2.45	2.78	1.13	926.77	<.0001
Soccer	Exp(ln RR PostSeason vs InSeason)	0.69	0.58	0.82	1.41	17.93	<.0001

Interpretation: The rate of injury is 5.6 times higher in games than in practices (95%CI: 5.3, 5.9; CLR=1.1).

Problem: These CIs are wider, and more realistic, than the CIs from standard Poisson regression.

Model 3: NB2 negative binomial regression

```
proc genmod data=ncaa.ncaa8904_v6;
        /* Regression on Injury Rates for Historical NCAA data*/
        class division season gameprac sex ;
        model injuries = sex year gameprac division season / link=log
             dist=negbin offset=log_AEs;
        run ;
```

…is the same as…

```
proc countreg data=ncaa.ncaa8904_v6 type=negbin ;
        /* Regression on Injury Rates for Historical NCAA data*/
        format gameprac gp. ;
        model injuries = sex year GamePrac Division21 Division32
                         PreVsReg PostVsReg / offset=log_AEs ;
        run ;
```

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	528	657.4756	1.2452
Scaled Deviance	528	657.4756	1.2452
Pearson Chi-Square	528	560.4033	1.0614
Scaled Pearson X2	528	560.4033	1.0614
Log Likelihood		84156.9558	

NB2 model has done a great job of addressing over-dispersion

Model 3: NB2 negative binomial regression

Analysis Of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Season	2(ref)	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion		1	0.1119	0.0122	0.0880	0.1357		

Alpha (over-dispersion)=0.11

Note: The over-dispersion parameter (alpha) = 0.11. This parameter is much larger than its SE (0.01) and is greater than zero according the Wald test, supporting the idea that this is better model for the data than the standard Poisson regression.

Model 3: NB2 negative binomial regression

Sport_Name	Games_and_Practices_Combined	Rate Ratio	Lower Confidence Limit	Upper Confidence Limit	Confidence Limit Ratio	ChiSquare	Pr > ChiSq
Soccer	Exp(ln RR MSO vs WSO)	1.04	0.96	1.12	1.16	0.93	0.3362
Soccer	Exp(ln RR Av Ann 10Yr)	1.01	0.93	1.10	1.19	0.07	0.7880
Soccer	Exp(ln RR Game vs Prac)	4.85	4.49	5.24	1.17	1580.75	<.0001
Soccer	Exp(ln RR D2 vs D1)	0.96	0.87	1.05	1.20	0.98	0.3231
Soccer	Exp(ln RR D3 vs D1)	0.84	0.77	0.92	1.19	15.52	<.0001
Soccer	Exp(ln RR PreSeason vs InSeason)	2.22	2.05	2.42	1.18	355.03	<.0001
Soccer	Exp(ln RR PostSeason vs InSeason)	0.71	0.63	0.80	1.26	33.31	<.0001

Interpretation: The rate of injury is 4.9 times higher in games than in practices (95%CI: 4.5, 5.2; CLR=1.2).

Problem: These CIs are even wider, suggesting that this model does a better job of accounting for over-dispersion than the PSCALE method. However, the rate ratio point estimate is now 4.9 instead of 5.6!

Model 4: NB1 negative binomial regression

NB1 is not currently supported in GENMOD, so use experimental PROC COUNTREG

```
proc countreg data=ncaa.ncaa8904_v6 type=negbin1 ;
        /* Regression on Injury Rates for Historical NCAA data*/
        format gameprac gp. ;
        model injuries = sex year GamePrac Division21 Division32
                         PreVsReg PostVsReg / offset=log_AEs ;
        run ;
```

Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	-9.368285	6.663497	-1.41	0.1598
sex	-0.031520	0.029092	-1.08	0.2786
year	0.001001	0.003335	0.30	0.7641
GamePrac	1.681534	0.031083	54.10	<.0001
Division21	-0.037795	0.038847	-0.97	0.3306
Division31	-0.153204	0.032000	-4.79	<.0001
PreVsReg	0.922246	0.033212	27.77	<.0001
PostVsReg	-0.097082	0.077732	-1.25	0.2117
_Alpha	4.002387	0.344519	11.62	<.0001

Over-dispersion parameter alpha=4. This is a linear function of the mean and is not comparable to the quadratic NB2 function (alpha=0.11)

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Model 4: NB1 negative binomial regression

There is no ESTIMATE statement in COUNTREG, but we can readily compute rate ratios and CIs in Excel:

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Rate Ratio	LCL	UCL	CLR
Intercept	-9.368285	6.663497	-1.41	0.1598				
sex	-0.03152	0.029092	-1.08	0.2786	0.97	0.92	1.03	1.12
year	0.001001	0.003335	0.3	0.7641	1.00	0.99	1.01	1.01
GamePrac	1.681534	0.031083	54.1	<.0001	5.37	5.06	5.71	1.13
Division21	-0.037795	0.038847	-0.97	0.3306	0.96	0.89	1.04	1.16
Division31	-0.153204	0.032	-4.79	<.0001	0.86	0.81	0.91	1.13
PreVsReg	0.922246	0.033212	27.77	<.0001	2.51	2.36	2.68	1.14
PostVsReg	-0.097082	0.077732	-1.25	0.2117	0.91	0.78	1.06	1.36
_Alpha	4.002387	0.344519	11.62	<.0001				

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Summary Table for Game Vs. Practice Variable

Model	#	Rate Ratio	95% Confidence Interval		Confidence Limit Ratio
Standard Poisson	1	5.56	5.40	5.73	1.06
Scaled Poisson	2	5.56	5.25	5.90	1.12
NegBin NB2	3	4.85	4.49	5.24	1.17
NegBin NB1	4	5.37	5.06	5.71	1.13

- Different models give very different rate ratios and 95%CIs !!
- Prefer models 2 & 4 (NB1 and scaled Poisson)
  - CI wide; RR credible
  - CI too narrow in model 1 (standard Poisson)
  - RR biased downwards in model 4 (NB2)
  - NB2 is sensitive wide variations in denominator counts

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Summary

- **Negative binomial models are preferred to Poisson models when:**
  - there is significant over-dispersion in the data
  - the betas from negative binomial model track closely to the betas from Poisson model
  - The SEs from negative binomial are larger than the SEs from Poisson
- **Negative binomial provides an SE that reflects extra-Poisson variation**
  - larger than the SE from standard Poisson regression
- **Scaled Poisson regression (PSCALE or DSCALE Poisson)**
  - Often performs as well as negative binomial

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Other Approaches

- Mixed Poisson …. school-level random intercepts
- Compound Poisson process

Epidemiologic Perspectives & Innovations

BioMed Central

Methodology

Open Access

Applying the compound Poisson process model to the reporting of injury-related mortality rates

Scott R Kegler*

Abstract

Injury-related mortality rate estimates are often analyzed under the assumption that case counts follow a Poisson distribution. Certain types of injury incidents occasionally involve multiple fatalities, however, resulting in dependencies between cases that are not reflected in the simple Poisson model and which can affect even basic statistical analyses. This paper explores the compound Poisson process model as an alternative, emphasizing adjustments to some commonly used interval estimators for population-based rates and rate ratios. The adjusted estimators involve relatively simple closed-form computations, which in the absence of multiple-case incidents reduce to familiar estimators based on the simpler Poisson model. Summary data from the National Violent Death Reporting System are referenced in several examples demonstrating application of the proposed methodology.

UNC SCHOOL OF PUBLIC HEALTH
DEPARTMENT OF EPIDEMIOLOGY

Why bother to get it right?

"Knowledge is not a loose-leaf notebook of facts. Above all, it is a responsibility for the integrity of what we are, primarily of what we are as ethical creatures."

"Every judgment in science stands on the edge of error and is *personal*. Science is a tribute to what we can know *although* we are fallible."

- Jacob Bronowski, 1908-1974