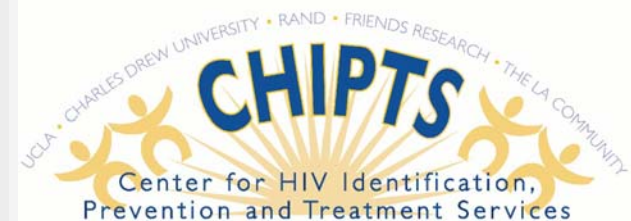


There will be Zer0s

W. Scott Comulada, Dr.P.H.
Senior statistician
Center for Community Health
scomulad@ucla.edu

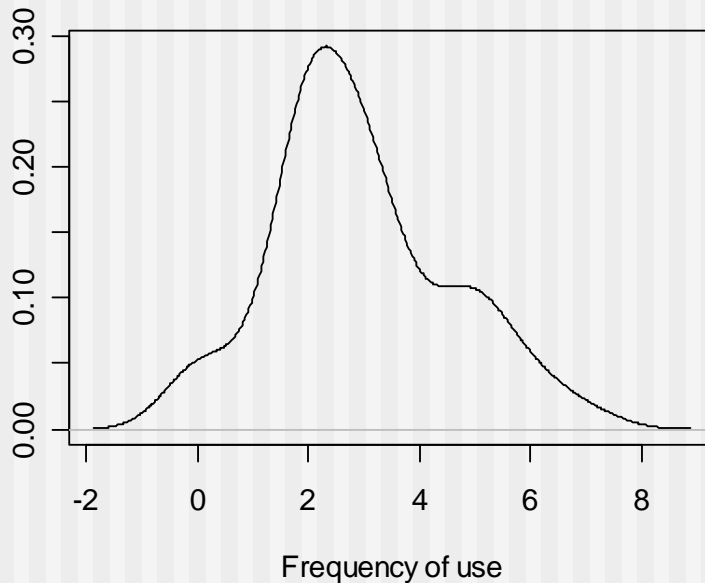


Count data

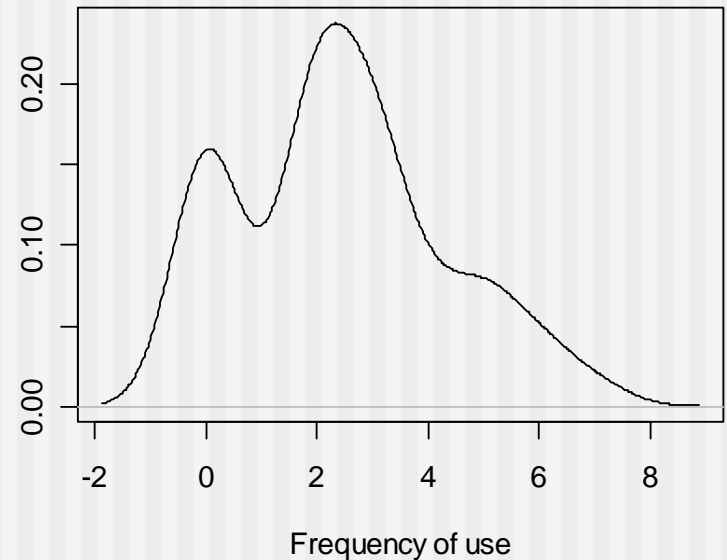
- Can contain larger proportion of zero responses than can be explained by a uni-modal distribution
 - i.e. zero-inflated count data
- Examples
 - Drug data – non-users
 - Sexual behavior – abstinent participants
 - Medical expenditures – healthy patients
 - Defects in manufacturing – Low defect rate

Hypothetical example: Frequency of drug use

What regression assumes



The reality



Disclaimer: Hypothetical example. Actual results may vary.

Applying regression

- Common ways to handle count data
 - Count → Poisson regression
 - Continuous → Linear regression
 - Dichotomize → Logistic regression
- Potential problems
 - Assume counts can be described by single distribution
 - Dichotomization may lead to less power

Two common solutions

- Zeros come from 1 process
 - e.g. medical expenditures
 - Patients only accrue expenses if they are sick

- Zeros come from 2 processes
 - E.g. Drug data
 - Zero can result from
 - Non-user
 - User who wasn't using when assessed

Two common solutions

- Zeros come from 1 process
 - Model for semicontinuous data
 - Duan et al. 1983; Olsen and Schafer 2001
 - Not discussed further here
- Zeros come from 2 processes
 - Zero-inflated Poisson (ZIP) model
 - Lambert 1992; Greene 1994; Hall 2000

ZIP model – Basic idea

- Model in two parts
- Logistic part
 - Probability observation in zero state
- Poisson part
 - Probability observation = 0 takes into account probability of being in zero state and probability of being a zero count from a Poisson distribution
 - Probability observation = positive value k is probability of not being in zero state and Poisson count with a value of k

Zero-inflated count data

Zero-inflated Poisson (ZIP) model

(Lambert 1992; Hall 2000)

$$y_{ijk} \sim \begin{cases} 0, & \text{with probability } p_{ijk}; \\ \text{Poisson}(\lambda_{ijk}), & \text{with probability } 1 - p_{ijk} \end{cases} \quad (2)$$

$$\text{logit}(p_{ijk}) = w'_{ijk}\beta_k$$

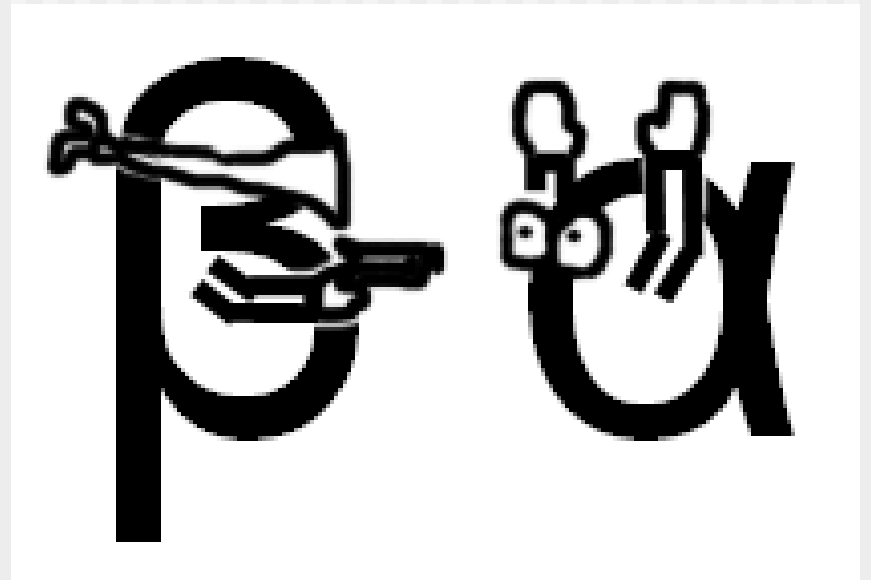
$$\log(\lambda_{ijk}) = z'_{ijk}\gamma_k + \eta_{ik}$$

Testing whether ZIP model provides better fit than Poisson model

- Cross sectional data
 - Vuong (1989) likelihood ratio
 - Van den Broek J (1995) score test
- Hierarchical data
 - Xiang et al. (2006) score test

Example: CLEAR study Messy drug data

When statistics
go bad



CLEAR study - description

- HIV-positive cohort (n = 175)
- Los Angeles, New York, San Francisco
- Entry criterion: substance using
- 16 – 29 years old
- 26% Black, 42% Latino
- 69% gay men
- 3 arms: In-person, Telephone, Control

CLEAR study - description

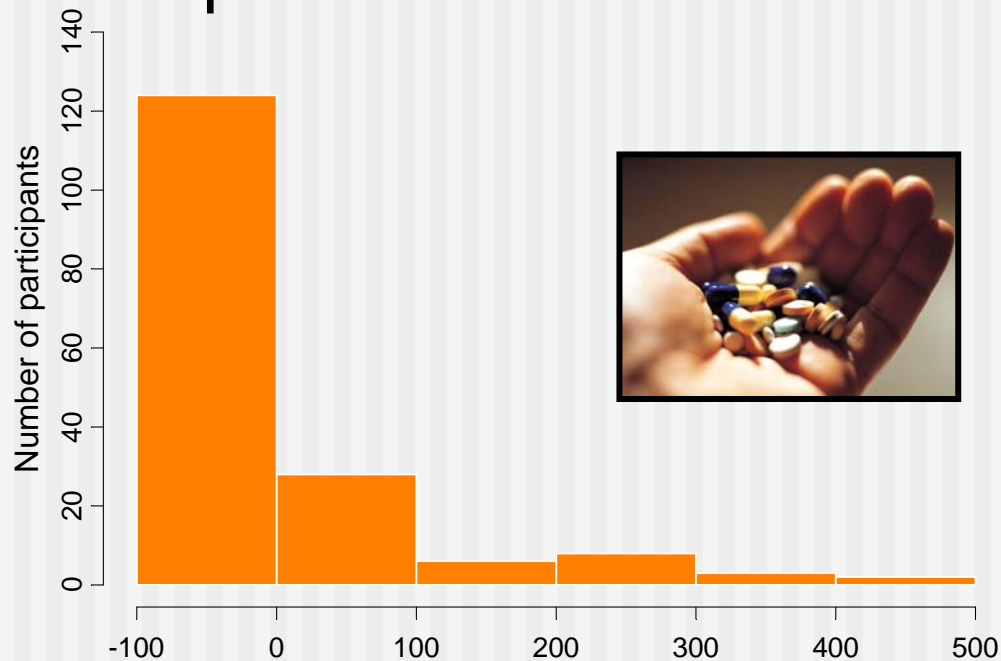
- Followed for 15 months
- Goal of intervention to reduce HIV-transmission risk behaviors
 - Risky sexual behavior
 - Substance use

CLEAR study – original analysis

- Rotheram-Borus et al. 2004
- Longitudinal regression models
- No impact of intervention on drug measures
 - Use / no use
 - Frequency of use

CLEAR study – what went wrong?

Observe frequency of past 3-month methamphetamine use at baseline



The leftmost bar represents the zero responses.

CLEAR study – what went wrong?

- High % counts = 0
- Two possibilities for zero counts
 - Non-users, participants in “zero state”
 - Users not using when interviewed

CLEAR study - reanalyzed

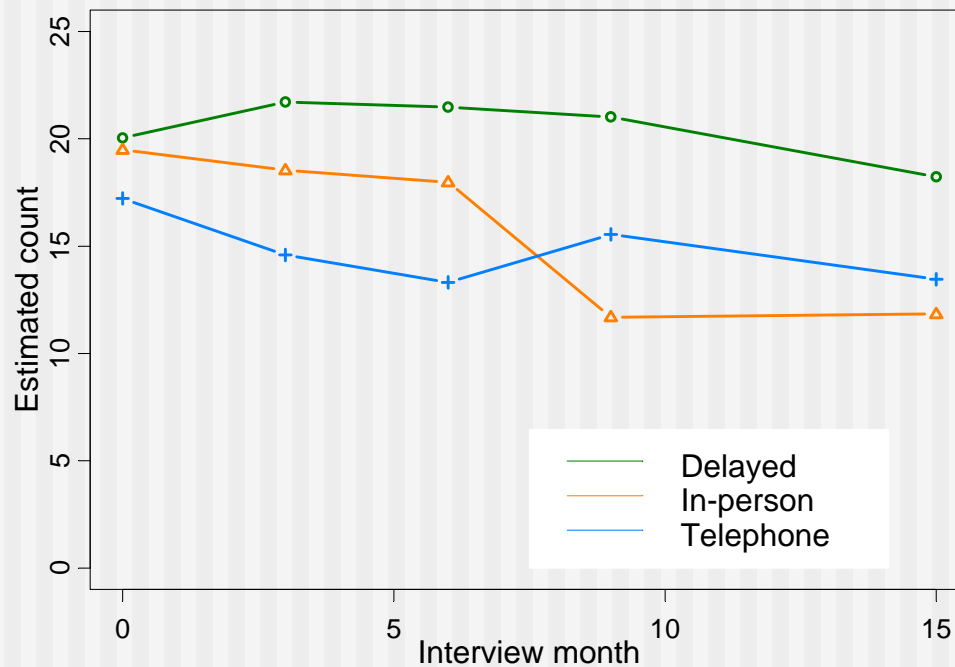
- Comulada et al. 2007
- Longitudinal ZIP models
 - Intervention reduced frequency of use
Across several substances
 - No impact on use/no use

Estimated changes in use over time

	In-person vs. Delayed	Telephone vs. Delayed
↓	Crack	Crack
↓		Heroin
↓	Inhalants	
↓	Marijuana	Marijuana
↓	Meth	Meth
↓		Stimulants
↑	Cocaine	Cocaine
↑	Heroin	

CLEAR study - reanalyzed

Estimated number of times using methamphetamines during the past three months at each time point by intervention condition.



Extensions to the ZIP model

Hierarchical ZIP model

- Longitudinal substance use data
 - Zero-state for non-use across all time points and substances
 - Zero-state for non-use at time point and particular substance
- Hierarchy to zeros
- Fit a ZIP model with two zero states

Hierarchical ZIP model

■ CLEAR data

- Model probability of men being in zero state across all time points and substances
- Model probability of being in zero state over time for each substance
- Model intervention effect on frequency of using each substance over time

HZIP model results for CLEAR amphetamine (Amph), cocaine, and methamphetamine (Meth) count outcomes

Parameter	Est.	S.E.	Est.	S.E.	Est.	S.E.
Logistic						
Subject-level						
Intercept	-.43	.21				
Male	-1.03	.41				
Poisson						
	Amph		Cocaine		Meth	
Intercept	.44	.23	.21	.21	-.66	.22
Time	.056	.027	-.024	.024	.048	.022
Base	1.13	.60	.17	.45	1.64	.44
Base*Intv	.20	1.12	-.78	.56	-.62	.59
Follow	.014	.59	-.070	.45	1.30	.44
Follow*Time	.0032	.0074	-.074	.0074	-.0036	.0025
Follow*Intv	-2.58	1.13	2.22	.57	-.61	.60
Follow*Intv*Time	.0054	.015	-.11	.015	-.054	.0050

ZINB model

- Zero-inflated data contains extreme values
 - E.g. zero-inflated drug data with heavy users in sample
- Replace Poisson part of model with negative binomial distribution
- Score test to compare ZIP and ZINB models (Ridout et al. 2001)

Markov models

- Longitudinal ZIP and ZINB models
 - Probability being in zero state same regardless of whether event occurred in previous time point
 - e.g. Drug use data, probability of being zero state for non-use same whether or not person was in zero state in previous time point

Markov models

- ZIP or ZINB with hidden markov chain
 - Estimate transition probabilities for moving in and out of zero state over time
 - E.g. foreign direct investments over time (Wang and Alba 2006)

References

Comulada WS, Weiss RE, Cumberland W, Rotheram-Borus MJ (2007). Reductions in Drug Use Among Young People Living with HIV. *American Journal of Drug and Alcohol Abuse* 33:493-501.

Duan N, Manning WG, Morris CN, Newhouse JP (1983). A comparison of alternative models for the demand for medical care. *Journal of Business and Economic Statistics* 1:115-126.

Hall D (2000). Zero-inflated Poisson and binomial regression with random effects: A case study. *Biometrics* 56:1030–1039.

Lambert D (1992). Zero-inflated Poisson regression with an application to defects in manufacturing. *Technometrics* 34:1–14.

Mullahy J (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* 33:341-365.

Olsen MK, Schafer JL (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 96:730-745.

Ridout M, Hinde J, Demetrio CGB (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics* 57:219-223.

References

Rotheram-Borus MJ, Swendeman D, Comulada WS, et al. (2004). Prevention for substance using HIV positive young people: Telephone and in-person delivery. *J Acquir Immune Defic Syndr* 37:S68–S77.

Van den Broek J (1995). A score test for zero inflation in a poisson distribution. *Biometrics* 51:738-743.

Vuong QH (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 57:307-333.

Wang P, Alba JD (2006). A zero-inflated negative binomial regression model with hidden Markov chain. *Economics Letters* 92:209-213.

Xiang L, Lee AH, Yau KKW, McLachlan GJ (2006). A score test for zero-inflation in correlated count data. *Statistics in Medicine* 25:1660-1671.

Thank you very much,

Scott Comulada

scomulad@ucla.edu