Shanghai Archives of Psychiatry, 2014, Vol. 26, No. 4

•Biostatistics in psychiatry (22)•

Structural zeroes and zero-inflated models

Hua HE[1,2,3]*, Wan TANG[1], Wenjuan WANG[1], Paul CRITS-CHRISTOPH[4]

Summary: In psychosocial and behavioral studies count outcomes recording the frequencies of the occurrence of some health or behavior outcomes (such as the number of unprotected sexual behaviors during a period of time) often contain a preponderance of zeroes because of the presence of 'structural zeroes' that occur when some subjects are not at risk for the behavior of interest. Unlike random zeroes (responses that can be greater than zero, but are zero due to sampling variability), structural zeroes are usually very different, both statistically and clinically. False interpretations of results and study findings may result if differences in the two types of zeroes are ignored. However, in practice, the status of the structural zeroes is often not observed and this latent nature complicates the data analysis. In this article, we focus on one model, the zero-inflated Poisson (ZIP) regression model that is commonly used to address zero-inflated data. We first give a brief overview of the issues of structural zeroes and the ZIP model. We then given an illustration of ZIP with data from a study on HIV-risk sexual behaviors among adolescent girls. Sample codes in SAS and Stata are also included to help perform and explain ZIP analyses.

Keywords: count response, structural zeroes, random zeroes, zero-inflated models

[*Shanghai Arch Psychiatry*. 2014; **26**(4): 236-242. doi: http://dx.doi.org/10.3969/j.issn.1002-0829.2014.04.008]

1. Introduction

Count (or frequency) responses such as number of heart attacks, number of days of alcohol drinking, number of suicide attempts, and number of unprotected sexual encounters during a period of time arise quite often in biomedical and psychosocial research. Poisson-distribution based log-linear regression models are widely used when such count variables are treated as the dependent variable in an analysis. One major limitation of the Poisson model is that the mean is identical to the variance. In practice, heterogeneity in study populations due to data clustering or other factors often creates extra variability, resulting in variance that is larger than the mean. This renders the Poisson distribution inappropriate for modeling count data in such instances. Depending on the nature of the heterogeneity, there are different approaches to address this extra variability, or *overdispersion*. For example, overdispersion may occur if the length of the observation period varies across the subjects. We can use an offset in the log-linear model to remove the overdispersion if the length of the observation period is available for each subject. Otherwise, we can treat the length as a latent variable, which is equivalent to treating the mean of the Poisson distribution for each subject as a random variable. If such a random effect is modeled using a gamma distribution, this approach yields a negative binomial (NB) distribution. Compared with the Poisson, the NB distribution has an extra parameter to account for the additional variation beyond the Poisson, and hence is able to address the limitations of the Poisson model for over-dispersed count responses. The Poisson and NB log-linear models are implemented in most major statistical software packages such as SAS, SPSS, and Stata.

However, NB cannot address the overdispersion caused by an excessive number of zeroes in the count data, which is quite common in psychosocial and behavioral studies. Excessive zeroes are particularly evident in alcohol and substance abuse research. For

[1] Department of Biostatistics and Computational Biology, University of Rochester Medical Center, Rochester, NY, USA
[2] Veterans Integrated Service Network, Center of Excellence for Suicide Prevention, Canandaigua VA Medical Center, Canandaigua, NY, USA
[3] Department of Psychiatry, University of Rochester Medical Center, Rochester, NY, USA
[4] Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, USA
* correspondence: hua_he@urmc.rochester.edu

A full-text Chinese translation will be available at www.saponline.org on September 25, 2014.

example, in a substance abuse multicenter intervention study, 50% of patients (*n*=318) had zero days of drug use during the entire primary outcome phase (weeks 4 to 16).[1] In another multicenter trial involving over 1500 patients at 20 community sites, zero days of drug and alcohol use in the past 7 days was reported by over 80% of patients at each weekly assessment.[2] Moreover, a wide range of other types of alcohol research studies have reported zero-inflated data across various types of alcohol measures.[3-12] Like substance use outcomes, HIV risk behavior measures show excessive zeros. For example, in a HIV risk intervention study[13] with 102 subjects, nearly 50% had scores of zero on the HIV risk measure. Zero-inflated data were also evident in other HIV risk reduction studies.[14,15] Other common examples of zero-inflated data that could serve as outcome measures in intervention or prevention trials include counts of uncommon adverse events,[16] number of hospital stays,[17] number of arrests,[18] and number of traffic accidents.[19]

Excessive zeros when assessing these types of outcome measures are often due to the existence of a subpopulation of subjects who are not at risk for such a behavior during the study period. For example, the number of unprotected sexual occasions over a period of time is an important measure in HIV prevention research. But a specific study population may contain a subgroup of individuals who are not at risk at all of sexual activity and, thus, will always produce a zero outcome in the count variable. Such zeros are called *structural zeros*. On the other hand, subjects who are at risk of the behavior may still produce a zero outcome due to sampling variability; such zeros are called *random* (or *sampling*) *zeros*.

The concept of structural zeros or non-risk groups is very important in psychosocial studies, because the non-risk and at-risk groups may have very different health and demographic characteristics. For example, in alcohol studies days of alcohol use over a week may contain both structural and random zeros: structural zeros that come from the non-risk group who are abstinent from drinking, while random zeros from the at-risk group of subjects who, due to sampling variation, did not drink in the prior week. The two groups of subjects may have different health outcomes such as different rates of depression and anxiety. So it is critical to distinguish the structural zeros from random zeros when modeling a count response with structural zeros.

Neither the Poisson nor the NB has the capability to accommodate the difference between structural and random zeros. When structural zeros are present in a count response, the count response becomes a mixed distribution, a mixture of degenerate zeros from the non-risk group (structural zeros) and responses (positive or random zero outcomes) from the at-risk group. The inherent methodological problems with structural zeros have received a great deal of attention in the statistical literature.[19-25] One popular approach is to use the mixture distribution based on zero-inflated models such as the zero-inflated Poisson (ZIP) model, which has been applied to a diverse range of studies.[26-30]

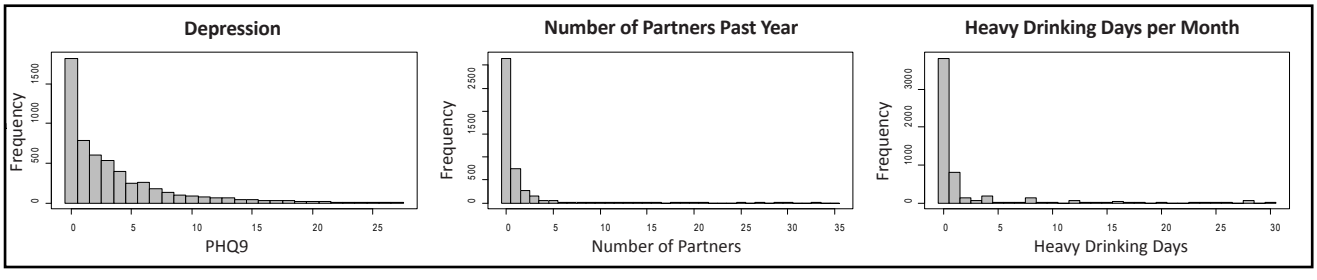
In this article, we first introduce some basic concepts about the mixture distribution and regression models for zero-inflated Poisson count responses and then use a real study example to illustrate the ZIP model. Sample codes in SAS and Stata and detailed explanations of the codes and output are provided. Finally, we discuss limitations of ZIP and related models and some newly developed methods that address such limitations.

2. Zero-inflated Poisson models

2.1 Zero-inflated Poisson distribution

In biomedical and psychosocial research the distribution of zeros often exceeds the expected frequency of zeros predicted by the Poisson model. Examples from the 2009-2010 National Health and Nutrition Examination Survey (NHANES) study are shown in Figure 1: the distribution scores on the 9-item Patient Health Questionnaire (PHQ-9), a popular screening test for depression; the number of sexual partners in the past year (with 1 subtracted from the original outcome for the married and living-together couples); and the days of heavy drinking per month in the past year. In this example, the presence of excessive zeros reflects a proportion of subjects who were not at risk for the health condition or behavior of interest. For example, in the case of the number of sexual partners in the past year, the non-risk group are individuals who never had extra sexual partners beyond their spouse or significant other; these non-risk individuals substantially inflate the number of zero results beyond what is under the Poisson distribution. The distributions of the results shown in the figure provide support for a mixed population consisting of an at-risk subgroup and a non-

Figure 1. Frequencies of scores on the 9-item Patient Health Questionnaire (PHQ-9), sexual partners in past year, and heavy drinking days per month



risk subgroup for each of the respective outcomes of interest: depression, sexually transmitted diseases, and alcohol-related health problems.

We use mixture distributions to model count responses with structural zeros. Within the current context, the mixture distribution is a mixture of two distributions, with one for the at-risk subgroup and other for the non-risk subgroup.

Let $f_R(y)$ be the distribution of the at-risk subpopulation, and $f_0(y)$ be a degenerate distribution at 0, [i.e., $f_0(y)=1$ if $y=0$ and $f_0(y)=0$ if $y \neq 0$] for the non-risk subpopulation. Suppose the mixture probabilities for the structural zeros (non-risk subgroup) and the at-risk subgroup are ρ and $1-\rho$. Then the mixture distribution can be expressed as

$$(1)\ f_{MIXTURE}(y)=\rho f_0(y)+(1-\rho)f_R(y),\ y=0, 1, \ldots$$

When a Poisson distribution with mean μ, $f_P(y|\mu)$, is applied to the at-risk subpopulation, we obtain the following zero-inflated Poisson (ZIP) distribution

$$(2)\ f_{ZIP}(y|\rho, \mu)=\rho f_0(y)+(1-\rho)f_P(y|\mu),\ y=0, 1, \ldots$$

More precisely, the distribution can be also expressed as

$$(3)\ f_{ZIP}(y|\rho, \mu)=\begin{cases}\rho+(1-\rho)f_P(0), & \text{if } y=0\\ (1-\rho)f_P(y|\mu), & \text{if } y>0\end{cases}$$

So, the probability of being zero, Pr(y=0), is inflated from $f_P(0|\mu)$ under the Poisson distribution by ρ to account for structural zeros. The mean and variance of $f_{ZIP}(y|\rho, \mu)$ are $(1-\rho)\mu$ and $(1-\rho)\mu+(1-\rho)\rho\mu^2$, respectively.

Thus, the variance is larger than the mean, confirming that overdispersion may occur if a Poisson distribution is applied in place of ZIP.

Depending on the nature of the data, other distributions may be more appropriate for the at-risk subpopulation. For example, if there is still overdispersion in the at-risk subgroup due to data clustering, we may use NB instead of Poisson for this group, and obtain a zero-inflated NB (ZINB) distribution.[20] For variables where the outcome is the sum of a very limited number of repeated trials, such as the number of days of any drinking over the last week, a binomial-like distribution may be a natural choice for the at-risk subgroup, which results in a zero-inflated binomial distribution.[22,23] In this paper, we restrict our considerations to ZIP and ZINB as these distributions are available in SAS and Stata.

2.2 Zero-inflated Poisson regression models

It is both conceptually and theoretically reasonable to model the outcomes from the two groups of subjects separately due to the heterogeneity of the study sample. The ZIP model has two components, one component is to model the probability of being the structural zeros ρ using the logistic regression and the other component is to model the Poisson mean μ. Specifically we have the ZIP model:

$$(4)\ \text{logit}(\rho_i)=U_i^T\beta_U,\ \log(\mu_i)=V_i^T\beta_V,$$

where the subscript i indicates the i^{th} subject, U and V (which may overlap) represent two sets of explanatory variables that will be linked to ρ and μ, respectively, in the ZIP model, and β_U and β_V are the vectors of parameters for the logistic and Poisson components.

In (4), the logit link function is used to model the likelihood of structural zeros; we can also use other link functions such as probit and complementary log-log. Thus, the presence of structural zeros gives rise not only to a more complex distribution, but also creates an additional link function for modeling the effect of explanatory variables for the occurrence of such zeros. In other words, the ZIP model enables us to better understand the effect of covariates by distinguishing the effects of each specific covariate on structural zeros (likelihood for being non-risk) and on the count response (mean of Poisson for the at-risk subgroup). While the fact that the presence of excessive zeros itself is sufficient to justify the use of zero-inflated models, Vuong has developed a test to formally test whether a ZIP is superior to a Poisson regression.[31]

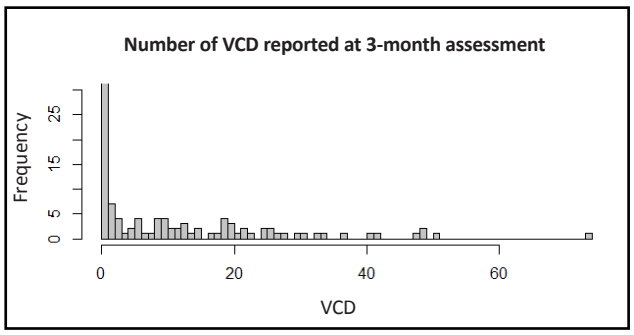
When there is still dispersion in the at-risk subgroup, we may use the ZINB model, which is identical to ZIP, except that the NB replaces the Poisson to account for overdispersion for modeling the count response from the at-risk subpopulation. Thus a ZINB regression model has one logistic regression for structural zeros and one NB log-linear for the count response for the at-risk subgroup, with the additional dispersion parameter α from the NB to account for overdispersion.

3. Example

We use a study of sexual behavior among adolescent girls to illustrate the application of zero-inflated models. In this controlled randomized study, 640 girls were randomized into either the intervention or a control condition (containing only nutritional materials) to evaluate the short and longer-term efficacy of a Human Immunodeficiency Virus (HIV) -prevention intervention for adolescent girls residing in a high-risk urban environment.[32] A primary outcome is the number of vaginal sex encounters using condoms (VCD) reported over the past 3-month period, which was assessed at 3, 6 and 12 months following the intervention in this longitudinal study. Since we restrict ourselves to cross-sectional analysis in the paper, we illustrate the models using this outcome from the 3-month assessment. The data can be downloaded from http://www.urmc.rochester.edu/biostat/people/faculty/Tang-He-Tu-Categorical-Book/sas5.html

Figure 2 presents the distribution of the frequency of VCD at the 3-month assessment. It is clear that there are excessive zeros in the distribution. The structural

Figure 2. Frequency of protected vaginal sex (VCD)



zeros represent those who were sexually abstinent. Since the status of structural zeros is not available, we apply zero-inflated models to analyze the data. For illustration purposes, we only present a simple model to examine the relationship between this outcome at 3 months and three potential covariates: score on the HIV Knowledge Questionnaire (HIVKQ, higher score means more informed about HIV knowledge), score on the Center for Epidemiological Depression scale (CESD, higher score means more depressed), and baseline number of vaginal sex encounters using condoms in the 3 months prior to intervention (VAGWCT1).

The ZIP model consisted of two components; the logistic model for the inflated zero component for VCD:

$$\text{logit}(\Pr(\text{VCD}=0)=\alpha_0+\alpha_1\text{VAGWCT1}+\alpha_2\text{HIVKQTOT}+\alpha_3\text{CESD}$$

and the component of Poisson loglinear model for the at-risk subgroup:

$$\text{logit}(\Pr(\text{VCD}|\text{at risk})=\alpha_0+\alpha_1\text{VAGWCT1}+\alpha_2\text{HIVKQTOT}+\alpha_3\text{CESD}$$

The Vuong test statistic was 4.858 (p-value <0.00001), which indicates that the ZIP model is better than the Poisson regression model.

Based on the estimates of the parameters in the count component (Table 1), both the baseline sexual behavior and HIV knowledge are highly associated with VCD. Subjects with higher baseline sexual behavior and higher HIV knowledge tend to have high VCD in the 3-month follow-up. Depression (CESD) is not significantly associated with VCD. Based on the estimates of the parameters in the zero inflation component (Table 2), only the baseline sexual behavior is associated with being a structural zero in VCD. That is, subjects with higher baseline sexual behavior tend to be less likely to be a structural zero in VCD in the 3-month follow-up.

Table 1. Analysis of maximum likelihood parameter estimates for count component of VCD

Parameter	degrees of freedom	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	p-value
Intercept	1	2.1249	0.1831	1.7659	2.4838	134.63	<0.0001
VAGWCT1	1	0.0132	0.0019	0.0096	0.0169	50.61	<0.0001
HIVKQTOT	1	0.0339	0.0103	0.0136	0.0541	10.77	0.0010
CESD	1	0.0009	0.0044	-0.0078	0.0096	0.04	0.8398

VCD, number of vaginal sex encounters using condoms in 3 months after enrollment
VAGWCT1, number vaginal sex encounters using condoms in 3 months prior to enrollment
HIVKQ, HIV Knowledge Questionnaire score
CESD, Center for Epidemiological Studies of Depression scale score

Table 2. Analysis of maximum likelihood zero inflation parameter estimates for inflated zero component of VCD

Parameter	degrees of freedom	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	p-value
Intercept	1	3.1852	1.7207	-0.1874	6.5577	3.43	0.0642
VAGWCT1	1	-0.3960	0.1190	-0.6292	-0.1628	11.08	0.0009
HIVKQTOT	1	-0.1234	0.0886	-0.2972	0.0503	1.94	0.1638
CESD	1	-0.0606	0.0479	-0.1545	0.0333	1.60	0.2060

VCD, number of vaginal sex encounters using condoms in 3 months after enrollment
VAGWCT1, number vaginal sex encounters using condoms in 3 months prior to enrollment
HIVKQ, HIV Knowledge Questionnaire score
CESD, Center for Epidemiological Studies of Depression scale score

In this example, application of zero-inflated models enables us to ascertain the exact effect of the educational intervention. HIV knowledge is associated with VCD, but mainly through its effect on the count for the at-risk subgroup.

4. Statistical Software

The ZIP and ZINB regression models have been implemented in some popular statistical software packages, including SAS and Stata. However, they are not yet available in SPSS. In SAS, one may use either PROC GENMOD or PROC COUNTREG for ZIP and ZINB models. Here are the sample codes for the example described above using PROC GENMOD:

```
PROC GENMOD DATA=path.Sex;
MODEL VCD=VAGWCT1 HIVKQTOT CESD/d=zip;
ZEROMODEL VAGWCT1 HIVKQTOT CESD/link=logit;
RUN;
```

In this sample code, text in italic can be modified to specify the data source and models, while the text not in italics are SAS key words and must be entered exactly as they appear. The Poisson component for the count response is specified in the statement 'MODEL', following the common format for generalized linear models ('response = <effects>'). The option 'd=zip' is used to indicate a ZIP model as desired. One may replace this with 'd=zinb' to fit a ZINB model.

The structural zero component is specified by the 'ZEROMODEL' statement. In the sample codes above, the logit link is used, yielding the logistic regression. However, as in modeling binary outcomes, other commonly used link functions such as probit and complementary log-log may also be used, which are both available in SAS.

In SAS, one may also use the COUNTREG procedure to fit a ZIP or ZINB model. Below is the sample codes for using this procedure. Although the statements are quite similar to PROC GENMOD, it is important to note the difference in specifying the "ZEROMODEL" statement. In particular, the extra '*VCD~*' statement is required to indicate the count variable whose structural zeros are modeled by the logistic regression.

```
PROC COUNTREG DATA = path.Sex;
MODEL VCD = VAGWCT1 HIVKQTOT CESD/d=zip;
ZEROMODEL VCD ~ VAGWCT1 HIVKQTOT CESD/link=logistic;
RUN;
```

Like PROC GENMOD, one may also use the COUNTREG procedure to fit the ZINB as well as use different link functions.

Vuong's test is not available from PROC GENMOD, but a SAS macro program is available and can be downloaded from http://support.sas.com/kb/42/514.html. In addition to testing whether the ZIP is a better fit to the data at hand than the Poisson, the test may also be used to compare the ZIP and ZINB to see which one fits the data better.

In Stata, one may use the 'zip' or 'zinb' commands to fit ZIP or ZINB models. For example, we may apply the following Command for the ZIP analysis for the example in the previous section:

```
zip vcd vagwct1 hivkqtot cesd,
    inflate(vagwct1 hivkqtot cesd) vuong
```

The first variable after the command 'zip' is the count response, followed by all the predictors for the count component in italics until the comma. The predictors for the zero component are specified in the parenthesis after the 'inflate' statement. The optional 'vuong' statement is specified to use the Vuong test to compare the ZIP and Poisson models. One may change the Command from zip to zinb to fit the ZINB regression model.

5. Discussion

This article discusses structural zeros in count outcomes and how to use zero-inflated models to address this issue. Zero-inflated models are the natural approach when the status of structural zeros are unknown, that is, when structural zeros cannot be distinguished from random zeros. In cases where this distinction is known, we may take advantage of the additional information and apply hurdle models.[20] Like the ZIP, hurdle models have two components, one for the count response and the other for the structural zeros. Again, the Poisson and NB may be used for modeling the count response, and logistic regression may be applied for the structural zeros. However, since the status of structural zero is known, no mixture distribution is needed and the Poisson (or NB) and logistic regression of the hurdle model are essentially two separate models. Thus, no new software is needed for fitting the hurdle model.

As illustrated by the real data example presented, zero-inflated models have both conceptual and analytics advantages when there are excessive zeros. The zero-inflated models not only correct the overdispersion arising from the existence of structural zeros, but also allow for the distinction of different risk groups, providing better understanding of the data.

We limited ourselves to parametric models and cross-sectional data analysis because of the availability in common software packages. However, parametric approaches are prone to distribution misspecification, potentially yielding bias in estimates. For example, if the count response for the at-risk subgroup in a study data does not follow the NB distribution, assuming and fitting a ZINB model may yield biased estimates. Another problem is that cross-sectional models cannot be applied to investigate temporal changes from repeated assessments in longitudinal studies. Some new methods have been developed to address both of these limitations,[22,23] but they have not yet been included in popular statistical software packages such as SAS and Stata.

We have only discussed the structural zero issue when zero-inflated count variables are used as the response. The issue is also present when such variables serve as predictors in regression analyses. Indeed, using such variables as predictors and failing to distinguish structural and random zeros results in biased inference and makes it quite difficult to interpret estimates.[25] However, these issues have not yet been addressed in popular statistical software packages.

Conflict of interest

The authors decalre no conflict of interest.

Funding

This research was supported in part by NIH grant R33 DA027521 and a Novel Biostatistical and Epidemiologic Methods grants from the University of Rochester Medical Center Clinical and Translational Science Institute Pilot Awards Program.

结构性零和零膨胀模型

贺华 , Wan TANG, Wenjuan WANG, Paul CRITS-CHRISTOPH

概述：在社会心理学和行为学的研究中，记录某些健康或行为结果发生频率的计数中（如在一段时间内无防护措施的性行为的次数）往往含有大量的零，这是因为当某些对象对于某种研究行为没有危险时就会产生“结构性零”。不像随机零（结果可以是大于零，但是也可能由于样本变异性而成为零），结构性零在统计和临床上通常是非常不同的。如果两种类型零的差异被忽略，就可能会导致对结果和研究发现的错误解释。然而在实践中，结构性零经常会没有被观察到而这种潜在性使数据分析复杂化了。在这篇文章中，我们专注于一种模式，即通常用于解决零膨胀数据的零膨胀泊松（Zero-inflated Poisson，ZIP）回归模型。首先，我们对结构性零和 ZIP 模型做一个简要概述。然后我们以一项青春期少女艾滋病高危性行为的研究数据来阐述 ZIP 模型。文中还附有 SAS 和 Stata 的示例代码，以帮助运行和解释 ZIP 分析。

关键词：计数反应，结构性零，随机零，零膨胀模型

本文全文中文版从 2014 年 9 月 25 日起在 www.saponline.org 可供免费阅览下载

References

1. Ball SA, Martino S, Nich C, Frankforter TL, van Horn D, Crits-Christoph P, et al. Site matters: multisite randomized trial of motivational enhancement therapy in community drug abuse clinics. *J Consult Clin Psychol.* 2007; **75**(4): 556-567. doi: http://dx.doi.org/10.1037/0022-006X.75.4.556
2. Crits-Christoph P, Ring-Kurtz S, McClure B, Temes C, Kulaga A, Gallop R, et al. A randomized controlled study of a web-based performance improvement system for substance abuse treatment providers. *J Subst Abuse Treat*. 2010; **38**(3): 251-262. doi: http://dx.doi.org/10.1016/j.jsat.2010.01.001
3. Neal DJ, Sugarman DE, Hustad JT, Caska CM, Carey KB. It's all fun and games... or is it? Collegiate sporting events and celebratory drinking. *J Stud Alcohol Drugs*. 2005; **66**(2): 291-294
4. Pardini D, White HR, Stouthamer-Loeber M. Early adolescent psychopathology as a predictor of alcohol use disorders by young adulthood. *Drug Alcohol Depend*. 2007; **88**: S38-S49. doi: http://dx.doi.org/10.1016/j.drugalcdep.2006.12.014
5. Hagger-Johnson G, Bewick BM, Conner M, O'Connor DB, Shickle D. Alcohol, conscientiousness and event-level condom use. *Br J Health Psychol*. 2011; **16**(4): 828-845. doi: http://dx.doi.org/10.1111/j.2044-8287.2011.02019.x
6. Connor JL, Kypri K, Bell ML, Cousins K. Alcohol outlet density, levels of drinking and alcohol-related harm in New Zealand: a national study. *J Epidemiol Community Health*. 2011; **65**(10): 841-846. doi: http://dx.doi.org/10.1136/jech.2009.104935
7. Buu A, Johnson NJ, Li R, Tan X. New variable selection methods for zero-inflated count data with applications to the substance abuse field. *Stat Med.* 2011; **30**(18): 2326-2340
8. Fernandez AC, Wood MD, Laforge R, Black JT. Randomized trials of alcohol-use interventions with college students and their parents: lessons from the Transitions Project. *Clin Trials*. 2011; **8**(2): 205-213. doi: http://dx.doi.org/10.1177/1740774510396387
9. Cranford JA, Zucker RA, Jester JM, Puttler LI, Fitzgerald HE. Parental alcohol involvement and adolescent alcohol expectancies predict alcohol involvement in male adolescents. *Psychol Addict Behav*. 2010; **24**(3): 386-396. doi: http://dx.doi.org/10.1037/a0019801
10. Hildebrandt T, McCrady B, Epstein E, Cook S, Jensen N. When should clinicians switch treatments? An application of signal detection theory to two treatments for women with alcohol use disorders. *Behav Res Ther*. 2010; **48**(6): 524-530. doi: http://dx.doi.org/10.1016/j.brat.2010.03.001
11. Hernandez-Avila CA, Song C, Kuo L, Tennen H, Armeli S, Kranzler HR. Targeted versus daily naltrexone: secondary analysis of effects on average daily drinking. *Alcohol Clin Exp Res*. 2006; **30**(5): 860-865. doi: http://dx.doi.org/10.1111/j.1530-0277.2006.00101.x
12. Witkiewitz K, van der Maas HL, Hufford MR, Marlatt GA. Nonnormality and divergence in posttreatment alcohol use: reexamining the Project MATCH data "another way". *J Abnorm Psychol*. 2007; **116**(2): 378-394. doi: http://dx.doi.org/10.1037/0021-843X.116.2.378
13. Carey MP, Braaten LS, Maisto SA, Gleason JR, Forsyth AD, Durant LE, et al. Using information, motivational enhancement, and skills training to reduce the risk of HIV infection for low-income urban women: a second randomized clinical trial. *Health Psychol*. 2000; **19**(1): 3-11. doi: http://dx.doi.org/10.1037/0278-6133.19.1.3

14. Garfein RS, Golub ET, Greenberg AE, Hagan H, Hanson DL, Hudson SM, et al. A peer-education intervention to reduce injection risk behaviors for HIV and hepatitis C virus infection in young injection drug users. *AIDS.* 2007; **21**(14): 1923-1932. doi: http://dx.doi.org/10.1097/QAD.0b013e32823f9066
15. Xia Y, Morrison-Beedy D, Ma J, Feng C, Cross W, Tu X. Modeling count outcomes from HIV risk reduction interventions: a comparison of competing statistical models for count responses. *AIDS Res Treat*. 2012; **2012**: 593569. doi: http://dx.doi.org/10.1155/2012/593569
16. Rose CE, Martin SW, Wannemuehler KA, Plikaytis BD. On the use of zero-inflated and hurdle models for modeling vaccine adverse event count data. *J Biopharm Stat*. 2006; **16**(4): 463-481. doi: http://dx.doi.org/10.1080/10543400600719384
17. Yau KK, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biom J*. 2003; **45**(4): 437-452. doi: http://dx.doi.org/10.1002/bimj.200390024
18. Johnso JE, O'Leary CC, Striley CW, Abdallah AB, Bradford S, Cottler LB. Effects of major depression on crack use and arrests among women in drug court. *Addiction*. 2011; **106**(7): 1279-1286. doi: http://dx.doi.org/10.1111/j.1360-0443.2011.03389.x
19. Chin HC, Quddu, MA. Modeling count data with excess zeroes: an empirical application to traffic accidents. *Sociol Methods Res*. 2003; **32**(1): 90-116. doi: http://dx.doi.org/10.1177/0049124103253459
20. Tang W, He H, Tu XM. *Applied Categorical and Count Data Analysis*. FL: Chapman & Hall/CRC; 2012.
21. Welsh A, Cunningham RB, Donnelly CF, Lindenmayer DB. Modeling the abundance of rare species: statistical-models for counts with extra zeros. *Ecol Modell*. 1996; **88**: 297-308. doi: http://dx.doi.org/10.1016/0304-3800(95)00113-1
22. Hall DB. Zero-Inflated Poisson and binomial regression with random effects: a case study. *Biometrics*. 2000; **56**: 1030-1039. doi: http://dx.doi.org/10.1111/j.0006-341X.2000.01030.x
23. Yau KW, Lee AH. Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme. *Stat Med*. 2001; **20**: 2907-2920. doi: http://dx.doi.org/10.1002/sim.860
24. Yu Q, Chen R, Tang W, He H, Gallop R, Crits-Christoph P, et al. Distribution-free models for longitudinal count responses with overdispersion and structural zeros. *Stat Med*. 2012; **32**(14): 2390-2405. doi: http://dx.doi.org/10.1002/sim.5691
25. He H, Wang W, Crits-Christoph P, Gallo, R, Tang W, Chen D, et al. On the implication of structural zeros as independent variables in regression analysis: applications to alcohol research. *J Data Sci*. 2013; in press
26. Crepon B, Duguet E. Research and development, competition and innovation pseudo-maximum likelihood and simulated maximum likelihood methods applied to count data models with heterogeneity. *J Econom.* 1997; **79**: 355-378
27. Miaou SP. The relationship between truck accidents and geometric design of road sections Poisson versus negative binomial regressions. *Accid Anal Prev*. 1994; **26**: 471-482
28. Gurmu S, Trivedi P. Excess zeros in count models for recreational trips. *J Bus Econ Stat*. 1996; **14**: 469-477. doi: http://dx.doi.org/10.1080/07350015.1996.10524676
29. Shonkwiler J, Shaw W. Hurdle count-data models in recreation demand analysis. *Aust J Agric Resour Econ*. 1996; **21**: 210-219
30. Cheung YB. Zero-infated models for regression analysis of count study of growth and development. *Stat Med*. 2002; **21**: 1461-1469. doi: http://dx.doi.org/10.1002/sim.1088
31. Vuong Q. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. 1989; **57**(2): 307-333
32. Morrison-Beedy D, Carey M, Crean H, Jones S. Risk behaviors among adolescent girls in an hiv prevention trial. *West J Nurs Res*. 2011; **33**(5): 690-711. doi: http://dx.doi.org/10.1177/0193945910379220

Dr. He is an Assistant Professor of Biostatistics in the Department of Biostatistics and Department of Psychiatry at the University of Rochester, as well as a researcher at the Center of Excellence for Suicide Prevention in Canandaigua, New York. Her research interests are in ROC analysis, semi-parametric and non-parametric inference, missing data modeling, causal inference, social network analysis, count data analysis and applications of statistical methods to psychosocial research. Dr. He received her PhD in Statistics from the Department of Biostatistics and Computational Biology at University of Rochester in 2007.