Understanding Poisson Regression

Matthew J. Hayat, PhD; and Melinda Higgins, PhD

ABSTRACT

Nurse investigators often collect study data in the form of counts. Traditional methods of data analysis have historically approached analysis of count data either as if the count data were continuous and normally distributed or with dichotomization of the counts into the categories of occurred or did not occur. These outdated methods for analyzing count data have been replaced with more appropriate statistical methods that make use of the Poisson probability distribution, which is useful for analyzing count data. The purpose of this article is to provide an overview of the Poisson distribution and its use in Poisson regression. Assumption violations for the standard Poisson regression model are addressed with alternative approaches, including addition of an overdispersion parameter or negative binomial regression. An illustrative example is presented with an application from the ENSPIRE study, and regression modeling of comorbidity data is included for illustrative purposes. [*J Nurs Educ.* 2014;53(4):207-215.]

Received: May 16, 2013

Accepted: August 28, 2013

Posted Online: March 25, 2014

Dr. Hayat is Assistant Professor, College of Nursing, Rutgers University, Newark, New Jersey; and Dr. Higgins is Associate Professor, School of Nursing, Emory University, Atlanta, Georgia.

The authors thank Dr. Sandra B. Dunbar for use of her data from the ENSPIRE study.

The authors have disclosed no potential conflicts of interest, financial or otherwise.

Address correspondence to Matthew J. Hayat, PhD, Assistant Professor, College of Nursing, Rutgers University, 180 University Avenue, Office 322, Newark, NJ 07102; e-mail: matt.hayat@rutgers.edu.

doi:10.3928/01484834-20140325-04

Nursing-, health-, and education-related research often includes counting the number of times a behavior, symptom, condition, or event occurs in a fixed window of time. For example, the number of documented medication errors was recorded in an experimental study of the impact of clinical simulation in nursing education (Sears, Goldsworthy, & Goodman, 2010). Studies of risky sexual behavior often consist of count data, such as asking adolescents to recall the number of nonromantic sexual partners (Chen, Thompson, & Morrison-Beedy, 2010), interviewing patients with mental illness who at risk for contracting HIV about the number of sexual acts with and without a condom in the past 30 days (Johnson-Masotti, Laud, Hoffmann, Hayat, & Pinkerton, 2004), or surveying college students about the number of times they engaged in risky behaviors in the past 30 days (Schwartz et al., 2010). Clinicians in an intensive outpatient program for preschoolers diagnosed with disruptive behavior disorder recorded the number of temper tantrums occurring in 30-second time intervals (Eisbach et al., 2014). Workforce information is also often in the form of counts. Schreuder et al. (2011) conducted a prospective study of the number of sickness absences (individuals not coming to work due to illness) in a 1-year period. In another study, measures of nurse staffing were considered in relation to the number of assaults per patient day and the number of assaults resulting in injury per patient day in psychiatric units (Staggs, 2013).

Another growing area of nursing research includes health resource utilization studies. For example, in a study of hospitalizations for patients with mental illness, data were collected on the number of days of hospitalization, and a randomized clinical trial was conducted that aimed to reduce the length of stay (Xie, McHugo, Sengupta, Clark, & Drake, 2004). Health resource utilization was quantified with respect to outpatient psychiatric service use, which was included as part of the ACCESS (Access to Community Care and Effect Services and Support) study (Neelon, O'Malley, & Normand, 2010).

All of the applications described were analyzed using statistical methods that made use of the Poisson distribution. The Poisson distribution is a probability distribution that applies to count data. These type of data are quantified with a count variable that can take on discrete non-negative whole number val-

TABLE 1

Frequency Distribution, Probabilities, and Expected Counts, for Aggregate Charlson Comorbidity Index Unweighted Count Data

Number of Comorbidities	Observed Count	Probability	Expected Count[a]
0	27	0.204	23.9
1	40	0.324	37.9
2	23	0.258	30.2
3	11	0.137	16
4	14	0.054	6.4
5	1	0.017	2
6	1	0.005	0.5
Total	117	1	

[a] Expected counts calculated on expected rate of $\hat{\lambda} = 1.59$.

ues (0, 1, 2, 3,…) in a fixed time interval. Rare or infrequently recurring events may have a small mean with many values of 0 or 1. One common practice, perhaps due to lack of knowledge about alternative statistical approaches, is to dichotomize data of this nature into an occurred or did not occur event. This approach does not use all of the available data and results in loss of information, reduced statistical power, and misleading analytic results and inferences (Owen & Froman, 2005).

Another challenge that results from count data with a small mean is a positive skew. Skewed data can be challenging to analyze. Count data have also been addressed within the nursing literature by considering the discrete counts as continuous in nature and by analysis using classical statistical methods (i.e., ordinary least squares [OLS] regression; Hutchinson & Holtman, 2005). This approach is usually problematic because it involves violated assumptions in applying normal theory methods to skewed data. For example, transforming count data by applying the square root or some other mathematical function often does not adequately account for excess zeros and heteroscedasticity, which render data inappropriate for modeling with an ordinary least-squares regression model (Cohen, Cohen, West, & Aiken, 2003).

Poisson regression is a statistical modeling technique that has been used in many recent publications in the nursing literature (Bowers & Crowder, 2012; Chang & Mark, 2011; Chen et al., 2010; da Cruz Ede et al., 2012; Krause, 2012; Li et al., 2011; Manojlovich, Sidani, Covell, & Antonakos, 2011; Ratner et al., 2010; Schreuder et al., 2011; Shang, Wenzel, Krumm, Griffith, & Stewart, 2012; Staggs, 2013; Staggs & Dunton, 2012; Theisen, Drabik, & Stock, 2012; van Gaal et al., 2011; Vitolo, Bortolini, Campagnolo, & Hoffman, 2012). This topic was considered important enough that a panel of statistics experts of a recent publication recommended that Poisson regression be included as a required core statistics education topic for doctoral nursing students (Hayat, Eckardt, Higgins, Kim, & Schmiege, 2013).

THE POISSON DISTRIBUTION

Data can be characterized by a Poisson distribution when observations are counted in whole numbers, when event occurrences are independent (one event occurrence does not affect the chance of another event occurring), and when the specifics of the observed time interval are known and are the same for each participant. Exact probabilities can be calculated with the probability mass function for the Poisson distribution, which is given by:

$$P(Y = \kappa) = \frac{\lambda^{\kappa} e^{-\lambda}}{\kappa!} \quad (1)$$

where Y is the variable of interest, κ is a specified count value ($\kappa = 0,1,2,\ldots$), λ (lambda) is the rate of occurrence (the mean number of events in the fixed time interval), e is the base of the natural logarithm, and $\kappa!$ is notation for κ factorial [$\kappa! = \kappa \times (\kappa-1) \times (k-2) \times \ldots \times 3 \times 2 \times 1$ and $0! = 1$]. Because the observed counts are non-negative whole numbers, λ will necessarily be non-negative. Unlike distributions characterized by two distinct parameters, such as the normal distribution described by its mean (μ) and variance (σ^2), the Poisson distribution is unique in that it is fully identified by the single parameter λ. The mean and variance for the Poisson distribution in Equation (1) above are the same and equal λ. In other words, to make inferences about the Poisson distribution, the single parameter λ needs to be estimated.

An example of how this is done can be seen in considering the aggregate count data of number of comorbidities displayed in **Table 1**. Background details about these data are provided in a later section. The first two columns display the number of comorbidities and the observed counts for each. The rate of occurrence (λ) in these data can be estimated as:

$$\hat{\lambda} = \frac{(0\times27)+(1\times40)+(2\times23)+(3\times11)+(4\times14)+(5\times1)+(6\times1)}{117} = \frac{186}{117} = 1.59. \quad (2)$$

We can calculate the exact probability of observing a specified number of comorbidities based on the observed data, as displayed in **Table 1**, using the probability mass function in Equation (1). For example, applying the Poisson distribution to the observed data (with $\hat{\lambda} = 1.59$), the probability of no comorbidities is estimated as:

$$P(X = 0) = \frac{1.59^0 e^{-1.59}}{0!} = e^{-1.59} = 0.204 \quad (3)$$

A similar calculation is applied to estimate the probabilities for each of the $\kappa = 1,2,\ldots6$ comorbidities. The expected number of comorbidities displayed in **Table 1** can be estimated as the product of the 117 observed total counts and the Poisson distribution probabilities for each of the $\kappa = 0,1,2,\ldots6$ comorbidities. For example, 27 participants were observed to have no comor-

bidities. The expected number of participants of the 117 with no comorbidities can be estimated as $117 \times 0.204 = 23.9$ (a count of ≈24), based on the Poisson distribution.

POISSON REGRESSION

Regression modeling techniques can be used to study the association of multiple variables of interest with count outcome data. However, classic linear regression is usually not plausible for count data due to a number of modeling assumption violations. The general linear model (GLM) is the mainstay of traditional statistics education and includes classic statistical techniques that assume a continuous and normally distributed dependent variable, such as the *t* test, analysis of variance, and linear regression. Count data on rare events, or occurrences, with a mean <10 are almost always skewed and non-normal, so the classic techniques that assume normality are not adequate. However, count data can be modeled within a larger framework of statistical techniques that do not require a normally distributed dependent variable. These techniques are referred to as generalized linear models (GzLM). The GzLM is similar to linear regression in that the model equation includes predictors as a function of the dependent variable. However, although linear regression models the actual observed value of the dependent variable, the GzLM instead models some function of the dependent variable. This function is referred to as the link function. For example, the well-known simple logistic regression model is a type of GzLM and uses the probability (p) of one of two possible outcomes and a logit-link function [$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$]. It can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 . \quad (4)$$

The regression coefficients (Betas, β) in equation (4) are interpreted on a log–odds scale. In other words, the value of $e^{\beta 1}$ is interpreted as an odds ratio. Poisson regression makes use of another link function, the log-link function, and models the natural log of the count outcome data (Y), expressed as:

$$\log(Y) = \beta_0 + \beta_1 X_1 . \quad (5)$$

It is important to note that the statistical models in Equations (4) and (5) are indeed linear models. The linearity is with respect to the link function of the observed outcome data. The interpretation of each of the regression coefficients (βs) is similar to that used in linear regression. For example, as in Equation (5), the intercept term (β_0) is the value of log (y) when $X_1 = 0$ and the slope term (β_1) is interpreted as the change in log (y), for a 1-unit increase in X_1. Unlike classic linear regression, the change here for the Poisson regression model is described in log values, which are not meaningful with regard to the raw count values (y). Count data refer to the number of times an event occurs. An interpretation of the regression coefficients in terms of the change in expected counts is possible with some mathematical manipulation. If we take the exponent of both sides of Equation (5) and apply the identity property [$(e^{\log(y)} = y$], this gives $y = e^{\beta'' + \beta_1 X_1}$. Further, applying properties of exponents gives $y = e^{\beta_0} e^{\beta_1 X_1}$. In other words, for a 1-unit change in the predictor X_1, we expect the dependent variable (y) to change by a multiple of e^{β_1}. The direction is analogous to the value of a regression coefficient in a classical linear regression analysis. For example, a negative slope (a negative β_1) results in decreased change in the expected counts because the multiplier e^{β_1} will be less than 1.

Estimation

There are two unknown parameters in the model statement in Equation (5). The intercept (β_0) and slope (β_1) are population parameters and need to be estimated. Statistical inference is the practice of drawing a random sample from the unknown and usually unobtainable population of interest and making inferences about the population based on the smaller obtainable sample. With respect to Poisson regression and count outcome data, this usually means estimating the unknown population parameters (βs) with some quantified degree of uncertainty, using the observed sample data. In classic linear regression, the method of OLS is used to estimate the population regression coefficients (βs). Because the mean is equal to the variance with the Poisson distribution, the variance increases with an increasing mean. A valid application of the method of OLS requires normally distributed outcome data, which is usually violated with count data. For these reasons, another estimation method, namely maximum likelihood estimation (MLE), is used to estimate the regression coefficients in Poisson regression. The method of MLE is also used in estimating the regression coefficients in other types of GzLM (e.g., logistic, gamma, ordinal logistic).

Selection and Goodness of Fit

When estimation is accomplished, interest may then be directed to assessing how well the model fits the data. Approaches to model selection and assessment are different for classic linear regression and GzLMs. The change in estimation approaches, from OLS to MLE, results in a need to use different statistics for making comparisons among competing models and assessing adequacy of model fit. The traditional coefficient of variation (r^2) measure is used in classic linear regression to quantify the proportion of variance of the dependent variable explained by the predictors in the model. However, Poisson regression models, as well as other GzLMs, do not have a direct analog to the r^2 measure. Instead, a measure called *deviance* can be used with MLE to assess model fit. The deviance statistic quantifies how much worst the fitted model is, compared with a perfectly fitting model (i.e., a model with as many predictors as participants, thus predicting all values perfectly; Coxe, West, & Aiken, 2009). Deviance values are relative and useful in comparing competing models; smaller deviance values are indicative of a better fitting model.

Model fit can be assessed with either a Pearson chi-square goodness-of-fit or an omnibus chi-square statistical test. The Pearson test is calculated as a ratio of the sum of squared differences between the observed and predicted values and the variance of the predicted values. The omnibus test is found by calculating the difference of the deviance for two competing nested models. A nested model means that all of the terms in a

smaller model are present in the larger model. A statistically significant result for the Pearson test suggests an adequate model fit, and a significant result for the omnibus statistical test suggests the additional predictors in the fuller model significantly contribute. In the event of non-nested models, such as comparing how well a Poisson regression model fits the data, compared to a normal linear regression model, another type of statistical test is needed. The scaled deviance and scaled Pearson chi-square tests are two goodness-of-fit measures that can be used with non-nested models. Each are ratios of the deviance chi square and Pearson chi square divided by their respective degrees of freedom (*df*). Values closer to 1 signify a better model fit, whereas values >1 indicate the presence of overdispersion (McCullagh & Nelder, 1989). An approach to model selection that works well with non-nested models is information criterion statistics, such as Akaike's information criterion (Akaike, 1974) or Bayesian information criterion (Raftery, 1995). McCullagh and Nelder (1989) provide a thorough exploration of these and other model selection statistics, as well as comprehensive coverage of regression diagnostics for assessing the impact of individual observations and outliers.

SPECIAL CONSIDERATIONS

Real-world data sometimes do not adequately meet the assumptions for a particular statistical technique. When this happens, it may be reasonable to consider one or more modifications to an existing statistical method. Several enhancements have been developed for the standard Poisson regression model to manage data that do not meet the necessary assumptions.

Overdispersion

The Poisson distribution is defined by a single parameter, λ (lambda), which represents its mean and variance. Sometimes this single parameter is too limiting and does not fully enable adequate characterization of the data at hand. Because the mean is equal to the variance, the result is a perfect positive correlation between the variance and mean. As the mean increases, the variance increases linearly. However, count data may be distributed, as its variance is larger than its mean, which is commonly referred to as overdispersion. When data exhibit this property, the overdispersion needs to be accounted for so that statistical inferences are valid. Otherwise, the standard error estimates (used to calculate test statistics and confidence intervals) will be understated, resulting in an overestimated number of statistically significant results.

Two approaches are commonly used to address the challenge of overdispersion (Cohen et al., 2003). Each handles the problem by introducing a second parameter to allow for the variance to take on a value larger than the mean. One approach is to use the Poisson regression with overdispersion model (also known as the quasi-Poisson regression model). In this modification of the standard Poisson regression model, a second non-negative parameter, φ (phi), is introduced, which is used as a multiplier of the mean of the count data so that the new variance is φμ (phi × Mμ; Ver Hoef & Boveng, 2007). A second approach to managing overdispersion is the use of the negative binomial regression model (also known as the gamma–Poisson regression model). Technical details of this approach may be found in the report by Ver Hoef and Boveng (2007), and the reader is referred to the work by McCullagh and Nelder (1989) for comprehensive overviews and technical details on the two mentioned approaches to handling overdispersion.

The Zero Problem

Sometimes participants in a study do not experience a behavior, symptom, or event of interest. In other words, data for some individuals in a study with a count outcome measure may include zero counts. Coxe et al. (2009) provided a detailed description of the different underlying mechanisms that may result in zero counts. For example, some participants may never exhibit an occurrence of an event, such as in a nursing education experiment of recording medication errors in a simulated clinical experience. An application with zero counts might also occur when asking a nonsmoker to report the number of cigarettes he or she smoked in the past 7 days. These structural zeroes can often be avoided with careful study planning. In the data analysis, it may be reasonable to exclude such participants when describing a count outcome. Another reason for zero counts is simply the absence of an event. For example, a chart review of patients undergoing electroconvulsive therapy may include quantifying the number of occurrences of delirium. Some patients plausibly may not have any occurrences of delirium, resulting in some zero counts. Modifications to the standard Poisson regression model have included a framework of techniques termed *zero-inflated models*. An overview of zero-inflated Poisson and negative binomial regression models can be found in the studies by Lambert (1992) and Hall (2000).

Correlated Data

Hayat et al. (2013) listed multilevel modeling as an important topic to be included in statistics education for doctoral nursing students. Although not the focus of this illustrative article on Poisson regression, assuming independent observations, multilevel, cluster, or repeated measures, count data are sometimes collected in nursing and health studies. For example, multiple count measures may be taken from one individual, from the same hospital in a multisite study, or from members of the same family. Clustering by individual, site, or family results in correlated data because measurements taken from the same child will be more similar than those taken from another child. A nontechnical overview of statistical models for correlated data is provided in the study by Hayat & Hedlin (2012). The generalized linear mixed model with the Poisson distribution and a log-link function may be used to model repeated measures count data. Because studies often involve a time component that entails following participants over time, correlated data occur frequently in nursing and health research. For more details on this advanced topic, refer to the studies by Diggle, Heagerty, Liang, and Zeger (2002) and Skondral and Rabe-Hasketh (2004).

DISCUSSION OF THE ENSPIRE INTERVENTION STUDY AND CHARLSON COMORBIDITY MEASURE

Background

The following data are presented from the ENSPIRE (EducatioN and SuPport Interventions to Improve Self CaRE) study ti-

TABLE 2

Summary Statistics and Distribution Estimates for Unweighted Charlson Comorbidity Index Count Data by Race

	Black (*n* = 68)		White (*n* = 49)		Overall (*n* = 117)	
Comorbidity	***n***	**%**	***n***	**%**	***n***	**%**
0	20	29.4	7	14.3	27	23.1
1	22	32.4	18	36.7	40	34.2
2	14	20.6	9	18.4	23	19.7
3	5	7.4	6	12.2	11	9.4
4	7	10.3	7	14.3	14	12
5	0	0	1	2	1	0.9
6	0	0	1	2	1	0.9
Summary Statistics	**Black (*n* = 68)**		**White (*n* = 49)**		**Overall (*n* = 117)**	
Mean	1.37		1.9		1.59	
Median	1.00		1.00		1.00	
Standard deviation	1.27		1.48		1.38	
Variance	1.61		2.18		1.90	
Overdispersion[a]	1.18		1.15		1.19	
Interquartile range	[0, 2]		[1, 3]		[1, 2]	
Range	[0, 4]		[0, 6]		[0, 6]	
Distribution Estimates	**Black (*n* = 68)**		**White (*n* = 49)**		**Overall (*n* = 117)**	
Normal	$\hat{\mu} = 1.37$, $\hat{\sigma} = 1.61$		$\hat{\mu} = 1.90$, $\hat{\sigma} = 2.18$		$\hat{\mu} = 1.59$, $\hat{\sigma} = 1.90$	
Poisson	$\hat{\lambda} = 1.37$		$\hat{\lambda} = 1.90$		$\hat{\lambda} = 1.59$	
Negative binomial	$\hat{r} = 7.35$, $\hat{p} = 0.84$		$\hat{r} = 14.71$, $\hat{p} = 0.89$		$\hat{r} = 7.87$, $\hat{p} = 0.83$	

[a] *Overdispersion parameter (φ) estimated as (variance/mean).*

tled "A Family Partnership Intervention for Heart Failure," which tested a patient–family partnership intervention designed to improve both dietary sodium reduction and medication adherence in patients with heart failure (Dunbar et al., 2013). Comorbidity information was collected using the Charlson Comorbidity Index (CCI) (Charlson, Pompei, Alex, & MacKenzie, 1987).

The CCI was developed in 1987 as a simple and readily applicable method of classifying comorbidity, which is valid for estimating risk of death from comorbid disease. The instrument assesses the presence of 19 comorbidities and results in an unweighted count of the number of comorbid diseases and a weighted index (sum) that takes into account both the number (frequency) and seriousness (severity) of comorbid conditions. All enrolled participants ($N = 117$) had congestive heart failure; therefore, the comorbidity count presented here describes the sum total of 18 comorbid conditions (all participants had the 19th comorbid condition—congestive heart failure). For the illustrative purposes of this article, only CCI unweighted count measures and the race demographic variable were analyzed.

ENSPIRE Study Method

The ENSPIRE study had an enrollment of 58% Black (68 of 117) and 42% White (49 of 117) participants, with the White participants having more comorbidities than the Black participants (**Table 2**). For comparative purposes, five regression approaches, or steps, were considered for studying race as a predictive factor of a function of the number of comorbidities:

- Linear regression. CCI unweighted count data are treated as a continuous dependent variable and modeled directly (normal distribution, identity-link function).
- Logistic regression. CCI count data are dichotomized into has a comorbid condition (≥ 1) or does not have a comorbid condition and the logit of the probability of has a comorbid condition is the dependent variable (binomial distribution, logit-link function).
- Poisson regression. A generalized linear model with a log-link function is used to regress race on the natural log of the CCI unweighted counts (Poisson distribution, log-link function, scale parameter fixed at 1).
- Poisson regression with overdispersion. A scale parameter is added to the Poisson regression analysis in the Poisson regression approach (above) to regress race on the natural log of the CCI unweighted counts (Poisson distribution, log-link function, scale parameter estimated).
- Negative binomial regression. A dispersion parameter is added to the Poisson regression analysis approach (above) to

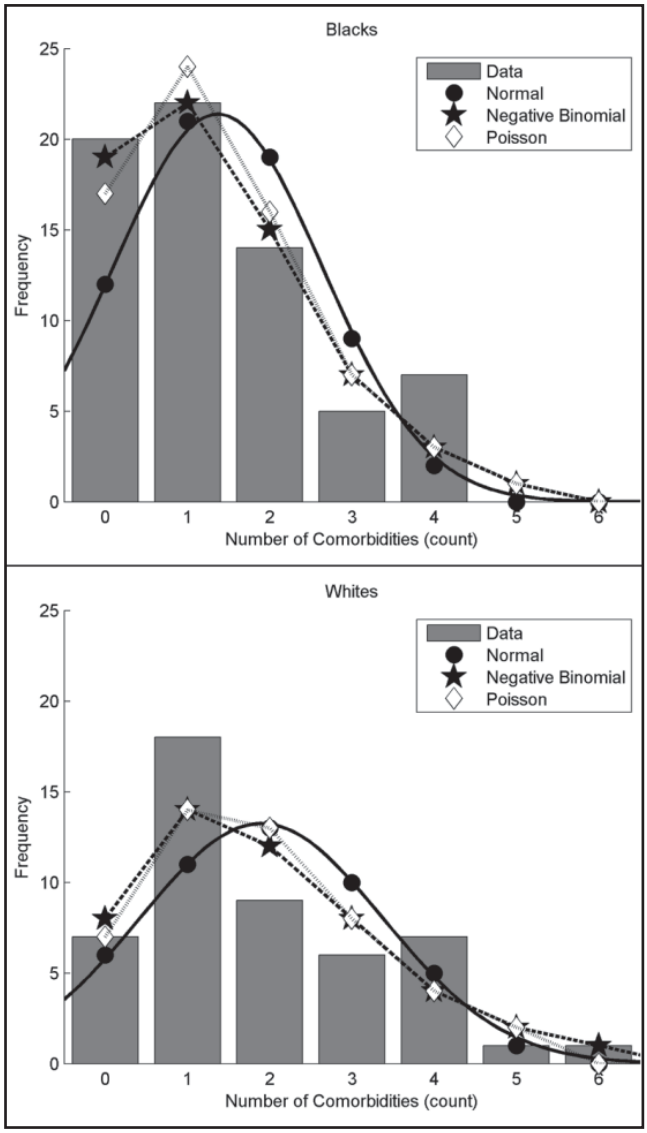

Figure. Histograms of unweighted Charlson Comorbidity Index data by race, with overlays of predicted frequencies for three different regression models. Predicted frequencies are identical for the standard Poisson and Poisson with overdispersion regression models.

regress race on the natural log of the CCI unweighted counts (negative binomial distribution, log-link function).

The predetermined level of significance in this study was set at $\alpha = .05$. SPSS® version 20.0.0 software and the GENLIN procedure were used for regression analyses and estimation of goodness-of-fit statistics. MATLAB version 7.12.0.635 software was used with the distribution fitting tool in the statistics toolbox to estimate the parameters for each underlying assumed distribution and to create the **Figure**.

ENSPIRE Study Results

Frequency distributions and summary statistics for number of comorbidities by race are displayed in **Table 2**. The counts are largely concentrated around none or a few, with a few participants reporting three or more comorbidities. White participants had a slightly larger mean than Black participants (1.90 versus 1.37). As is characteristic of count data, the variance for White participants is also slightly larger than for Black participants (2.18 versus 1.61). Overdispersion is estimated as the ratio of the variance to the mean (i.e., a ratio greater than 1 indicates overdispersion, and a ratio equal to 1 indicates no overdispersion). Because the estimate of overdispersion is greater than 1, it appears that both races had a slight amount of overdispersion (1.15 for White participants, 1.18 for Black participants). Distribution estimates are also displayed in **Table 2**. Histograms displaying the distribution for number of comorbidities by race are displayed in the **Figure**. The distribution for each race is non-normal and positively skewed to the right.

If the apparent skewness and non-normality of the count data are ignored, and the data are incorrectly assumed to be normally distributed, an independent samples *t* test can be used to compare the mean number of comorbidity counts between White participants and Black participants. The result of this *t* test is statistically significant ($t[115] = 2.083$, $p = 0.039$). This is similar to the *p* value for the race regression coefficient from the normal linear regression (**Table 3**). In fact, these are equivalent statistical tests, each having a different presentation of the general linear model, with the mean number of counts as a dependent variable and race as the single predictor. Because race was coded as Black participants = 0 and White participants = 1, the intercept coefficient for the normal linear regression model is interpreted as the predicted mean counts for Black participants, and the slope coefficient for race for the normal linear regression model is interpreted as the predicted mean counts for White participants. This is, in essence, applying the same test of group mean differences as accomplished with an independent samples *t* test. To calculate the predictive mean counts for White participants, the two coefficients can be added together. For example, the mean number of comorbidities for Black participants, as indicated by the intercept term, was estimated to be 1.37. The mean number of comorbidities for White participants was estimated to be 1.90, which is calculated as the sum of the intercept and race coefficients ($1.37 + 0.53 = 1.90$).

Model results displayed in **Table 3** include parameter estimates for five regression models, including scale and dispersion parameter estimates, where appropriate, and corresponding model fit statistics. The logistic regression analysis resulted from dichotomizing the CCI unweighted count data into the categories of had at least one comorbidity or had no comorbidities and predicting the logit of the probability of had at least one comorbidity. This naïve approach requires consolidating the raw data and discarding the magnitude of CCI unweighted count data for participants with one or more comorbidities. As a result, predictions with logistic regression are limited to the two collapsed categories and are likely inadequate for quantifying the occurrence of comorbidities. The normal linear regression model assumes that the CCI unweighted counts are continuous and normally distributed. The histograms displayed in the **Figure** show the count data to be highly skewed, suggesting the normal assumption to be unreasonable with these data. From a conceptual standpoint, the Poisson regression model results,

TABLE 3

Regression Coefficient Results and Model Statistics for Five Types of Regression Models

Variable	Logistic Regression (Logit Link)	Normal Linear Regression (Identity Link)	Poisson Regression (Log Link)	Poisson Regression With Overdispersion (Log Link)	Negative Binomial Regression (Log Link)
Intercept					
Estimate (*SE*)	–0.88 (0.27)	1.37 (0.16)	0.31 (0.10)	0.31 (0.11)	0.31 (0.11)
95% CI	–1.40, –0.35	1.05, 1.69	0.10, 0.51	0.09, 0.53	0.09, 0.53
p value	0.001	<0.001	0.003	0.005	0.005
Race[a]					
Estimate (*SE*)	–0.92 (0.49)	0.53 (0.25)	0.33 (0.15)	0.33 (0.16)	0.33 (0.16)
95% CI	–1.87, 0.04	0.03, 1.03	0.04, 0.62	0.02, 0.64	0.02, 0.64
p value	0.060	0.036	0.025	0.038	0.038
Scale					
Estimate	1	NA	1	1.16[b]	1
Negative binomial dispersion					
Estimate (*SE*)	NA	NA	NA	NA	0.096[c] (0.10)
Model fit statistics					
df	NA	115	115	115	114
Deviance	NA	212.3	151.72	151.72	134.47
Deviance/*df*[d]	NA	1.85	1.32	1.32	1.18
Pearson χ^2	NA	212.3	133.88	133.88	116.23
Pearson χ^2/df[e]	NA	1.85	1.16	1.16	1.02
Log likelihood	–4.07	–200.87	–188.8	–188.8	–188.24
Omnibus test $\chi^2_{(1)}$	3.83	4.33	4.97	4.27	4.22
p value	0.050	0.037	0.026	0.039	0.040
AIC	12.15	407.74	381.6	381.6	382.48
BIC	17.67	416.03	387.13	387.13	390.76

Note. SE = *standard error; CI = confidence interval; NA = not applicable;* df = *degrees of freedom; AIC = Akaike information criterion; BIC = Bayesian information criterion.*

[a] *Race coded Black = 0, White = 1.*

[b] *Estimated using Pearson* χ^2.

[c] *Estimated using maximum likelihood estimation.*

[d] *Scaled deviance.*

[e] *Scaled Pearson* χ^2.

with or without overdispersion, and the negative binomial regression model use all of the available data and assume correctly that the counts are discrete in nature and observed over a finite time interval.

Slope estimates for race are statistically significant for all, except for the logistic regression model. This is expected because a simple chi-square test of independence comparing the percentages of non-zero comorbidities between the two races was also not statistically significant (70.6% Black participants versus 85.7% White participants, $\chi^2_{(df=1)} = 3.671$, $p = 0.055$). The MLE parameter estimates for the intercept and race coefficients for the three discrete regression models are identical. The intercept coefficient value is 0.31, which is calculated as the natural log of the mean for Black participants [log(1.37) = 0.31]. The race coefficient is estimated to be 0.33, which is the difference between the natural log of the means for each race [log(1.90) – log(1.37) = 0.33]. A race coefficient estimate of 0.33 is interpreted in terms of a multiplicative factor of $e^{0.33}$ = 1.39. In other words, the predicted mean number of comorbidities is 1.39 more for White participants than Black participants. Of note, the normal linear regression estimate, based on ordinary least squares regression, was 0.53, which is the expected average difference between the races, as described. Of importance, although the parameter estimates for the intercept and race term in the Poisson regression models have the same value, the standard error estimates are not the same. The standard er-

ror is larger for the Poisson regression with overdispersion and negative binomial regression models. Because the outcome is different for the Poisson regression models (log-link function), normal linear regression (identity-link function), and logistic regression (logit-link function), the standard error values are not directly comparable among models. The scaled Pearson chi square goodness-of-fit measure for the standard Poisson regression model is estimated to be 1.16, which is the same value found with the Poisson with overdispersion regression model. The value of 1.16 is the weighted sum of the overdispersion parameters of the two races as reported in **Table 2** [(68/117) * 1.18 + (49/117) * 1.15 = 1.16]. An estimate of 1.16 would likely be interpreted as evidence of a lack of overdispersion.

The model fit statistics shown in **Table 3** provide information to assess the adequacy and performance of the different models. As previously mentioned, the deviance for a Poisson regression model does not change with an additional overdispersion parameter. The negative binomial model has a smaller deviance than the Poisson regression model, with or without overdispersion (134.47 versus 151.72), suggesting a better fitting model. Similarly, the Pearson chi-square fit statistic was 133.88 for the two Poisson regression models and 116.23 for the negative binomial model. The omnibus chi-square test results are statistically significant for all but the logistic regression model. A significant result suggests that race is a statistically significant addition and should be included in the model. The Akaike's information criterion results for the three Poisson regression models are similar, with this statistic having a slightly larger value for the negative binomial model, compared with the Poisson regression models, with or without overdispersion (382.48 versus 381.60). The Bayesian information criterion results were also similar (390.76 versus 387.13). Predicted probabilities of comorbidity count occurrences by race for the normal, Poisson, and negative binomial regression models are depicted as curves overlaying the histograms in the **Figure**. The normal linear regression model is a continuous line, and the curves for the Poisson and negative binomial regression models are interpolated lines passing through the predicted probabilities for the seven discrete number of comorbidity values (0,1,…6). The Poisson and negative binomial models perform considerably better than the normal linear regression model in predicting counts that are close to the actual observed count values.

DISCUSSION

Count data frequently occur in nursing and health studies. It is desirable to align statistical techniques with the type of data collected in a research project. One of the strengths of using the Poisson distribution is its ability to realistically describe, quantify, and predict count data. Added to this strength are the statistical software capabilities to implement this statistical method. Computing power has grown tremendously in the past decade, and Poisson regression is now readily available and accessible in modern software. For example, the current SPSS software version 20 includes menu-driven options for fitting generalized linear models, including the standard Poisson regression model, Poisson regression with overdispersion, and negative binomial regression. Procedures are also readily available for fitting the generalized linear mixed model. Although few documented limitations to using this statistical technique exists, the hands-on fitting of this type of statistical model may present technical challenges with regard to model estimation, selection, and interpretation. The regression coefficients are interpreted in terms of a multiplicative change in the expected number of counts, which may not be as intuitive as the additive interpretation used with classic linear regression. The immediate accessibility of these advanced statistical techniques needs to be considered when developing statistics education curriculum and course content for nursing students. The inclusion of Poisson regression in a broader framework as a generalized linear model lends to grouping this technique with other modeling techniques in a statistical modeling curriculum. Course content for graduate nursing students includes one or more statistics courses. Although it may be unrealistic to teach all of the technical details related to statistical modeling, such as model estimation, prediction, inference, and diagnostics, in a few statistics courses, a thorough introduction to the generalized linear model framework is needed to prepare students to critically read the literature.

CONCLUSION

An overview of the Poisson distribution has been provided, including numerous statistical modeling techniques that make use of this discrete distribution. Poisson regression is a statistical method that has been growing in use in the nursing literature. This is encouraging, as outdated approaches to modeling count data are replaced with more adequate techniques. The approaches of either dichotomizing count data or assuming it is continuous and modeling using normal linear regression are limited and often do a poor job of describing count data. Knowledge of the Poisson distribution is needed so students can be fully prepared to read and critically evaluate the nursing and health literature. The standard Poisson regression model has been fully developed to allow for special considerations when modeling assumptions are not fully met. Methods have been described for extending the Poisson regression model to manage overdispersion, inflated zero counts, and correlated data. The current state of statistical software, including SPSS, makes Poisson regression modeling accessible and manageable. This useful method should be included as a core component of statistics education for doctoral nursing students.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Bowers, L., & Crowder, M. (2012). Nursing staff numbers and their relationship to conflict and containment rates on psychiatric wards—A cross sectional time series Poisson regression study. *International Journal of Nursing Studies, 49*(1), 15-20.

Chang, Y., & Mark, B. (2011). Effects of learning climate and registered nurse staffing on medication errors. *Nursing Research, 60,* 32-39.

Charlson, M.E., Pompei, P., Ales, K.L., & MacKenzie, C.R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases, 40,* 373-383.

Chen, A.C., Thompson, E.A., & Morrison-Beedy, D. (2010). Multi-system influences on adolescent risky sexual behavior. *Research in Nursing and Health, 33,* 512-527.

Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Coxe, S., West, S.G., & Aiken, L.S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. *Journal of Personality Assessment, 91,* 121-136.

da Cruz Ede, P., Toporcov, T.N., Rotundo, L.D., Biazevic, M.G., Brasileiro, R.S., de Carvalho, M.B., . . . Antunes, J.L. (2012). Food restrictions of patients who are undergoing treatment for oral and oropharyngeal cancer. *European Journal of Oncology Nursing, 16,* 253-257.

Diggle, P.J., Heagerty, P., Liang, K.Y., & Zeger, S.L. (2002). *Analysis of longitudinal data* (2nd ed.). New York, NY: Oxford University Press.

Dunbar, S.B., Clark, P.C., Reilly, C.M., Gary, R.A., Smith, A., McCarty, F., . . . Ryan, R. (2013). A trial of family partnership and education interventions in heart failure. *Journal of Cardiac Failure, 19,* 829-841.

Eisbach, S.S., Cluxton-Keller, F., Harrison, J., Krall, J.R., Hayat, M.J., & Gross, D. (2014). Characteristics of temper tantrums in preschoolers with disruptive behavior in a clinical setting. *Journal of Psychosocial Nursing and Mental Health Services.* Advance online publication. doi:10.3928/02793695-20140110-02

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

Hayat, M.J., Eckardt, P., Higgins, M., Kim, M., & Schmiege, S. (2013). Teaching statistics to nursing students: An expert panel consensus. *Journal of Nursing Education, 52,* 330-334.

Hayat, M.J., & Hedlin, H. (2012). Modern statistical modeling approaches for analyzing repeated-measures data. *Nursing Research, 61,* 188-194.

Hutchinson, M.K., & Holtman, M.C. (2005). Analysis of count data using Poisson regression. *Research in Nursing and Health, 28,* 408-418.

Johnson-Masotti, A.P., Laud, P.W., Hoffmann, R.G., Hayat, M.J., & Pinkerton, S.D. (2004). A Bayesian approach to net health benefits: An illustration and application to modeling HIV prevention. *Medical Decision Making, 24,* 634-653.

Krause, M.R. (2012). Director of nursing current job tenure and past experience and quality of care in nursing homes. *Health Care Management Review, 37,* 98-108.

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

Li, J., Galatsch, M., Siegrist, J., Müller, B.H., Hasselhorn, H.M., & European NEXT Study Group. (2011). Reward frustration at work and intention to leave the nursing profession—Prospective results from the European longitudinal NEXT study. *International Journal of Nursing Studies, 48,* 628-635.

Manojlovich, M., Sidani, S., Covell, C.L., & Antonakos, C.L. (2011). Nurse dose: Linking staffing variables to adverse patient outcomes. *Nursing Research, 60,* 214-220.

McCullagh, P., & Nelder, J.A. (1989). *Generalized linear models* (2nd ed.). London, United Kingdom: Chapman & Hall.

Neelon, B.H., O'Malley, A.J., & Normand, S.T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Statistical Modeling, 10,* 421-439.

Owen, S.V., & Froman, R.D. (2005). Why carve up your continuous data? *Research in Nursing and Health, 28,* 496-503.

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111-196.

Ratner, P.A., Spinelli, J.J., Beking, K., Lorenzi, M., Chow, Y., Teschke, K., . . . Dimich-Ward, H. (2010). Cancer incidence and adverse pregnancy outcome in registered nurses potentially exposed to antineoplastic drugs. *BMC Nursing, 9,* 9-15.

Schreuder, J.A., Plat, N., Magerøy, N., Moen, B.E., van der Klink, J.J., Groothoff, J.W., & Roelen, C.A. (2011). Self-rated coping styles and registered sickness absence among nurses working in hospital care: A prospective 1-year cohort study. *International Journal of Nursing Studies, 48,* 838-846.

Schwartz, S.J., Forthun, L.F, Ravert, R.D., Zamboanga, B.L., Umaña-Taylor, A.J., Filton, B.J., . . . Hudson, M. (2010). Identity consolidation and health risk behaviors in college students. *American Journal of Health Behavior, 34,* 214-224.

Sears, K., Goldsworthy, S., & Goodman, W.M. (2010). The relationship between simulation in nursing education and medication safety. *Journal of Nursing Education, 49,* 52-55.

Shang, J., Wenzel, J., Krumm, S., Griffith, K., & Stewart, K. (2012). Who will drop out and who will drop in: Exercise adherence in a randomized clinical trial among patients receiving active cancer treatment. *Cancer Nursing, 35,* 312-322.

Staggs, V.S. (2013). Nurse staffing, RN mix, and assault rates on psychiatric units. *Research in Nursing and Health, 36,* 26-37.

Staggs, V.S., & Dunton, N. (2012). Hospital and unit characteristics associated with nursing turnover include skill mix but not staffing level: An observational cross-sectional study. *International Journal of Nursing Studies, 49,* 1138-1145.

Theisen, S., Drabik, A., & Stock, S. (2012). Pressure ulcers in older hospitalised patients and its impact on length of stay: A retrospective observational study. *Journal of Clinical Nursing, 21,* 380-387.

van Gaal, B.G., Schoonhoven, L., Mintjes, J.A., Borm, G.F., Hulscher, M.E., Defloor, T., . . . van Achterberg, T. (2011). Fewer adverse events as a result of the SAFE or SORRY? programme in hospitals and nursing homes. Part I: Primary outcome of a cluster randomised trial. *International Journal of Nursing Studies, 48,* 1040-1048.

Ver Hoef, J.M., & Boveng, P.L. (2007). Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology, 88,* 2766-2772.

Vitolo, M.R., Bortolini, G.A., Campagnolo, P.D., & Hoffman, D.J. (2012). Maternal dietary counseling reduces consumption of energy-dense foods among infants: A randomized controlled trial. *Journal of Nutrition Education and Behavior, 44,* 140-147.

Xie, H., McHugo, G., Sengupta, A., Clark, R., & Drake, R. (2004). A method for analyzing longitudinal outcomes with many zeros. *Mental Health Services Research, 6,* 239-246.