Journal of Modern Applied Statistical Methods

Volume 11 | Issue 1 Article 2

5-1-2012

Underlying Distributions in Loglinear Models of Discrete Data

Tim Moses
Educational Testing Service, Princeton, NJ, tmoses@ets.org

Journal of Modern Applied Statistical Methods
May 2012, Vol. 11, No. 1, 2-23

1538 – 9472/12/$95.00

Regular Articles

Underlying Distributions in Loglinear Models of Discrete Data

Tim Moses
Educational Testing Service,
Princeton, NJ

The implications of loglinear models based on underlying uniform and binomial distribution are assessed with respect to modeling eight distributions. Regarding statistical selection of the loglinear models' parameterizations, results indicate that better fitting models are obtained when the distribution being modeled is dissimilar to the underlying distribution used. For loglinear models with predetermined numbers of parameters, results suggest that better fitting models can be obtained when the distribution being modeled is similar to the underlying distribution.

Key words: Loglinear models, uniform distribution, binomial distribution.

Introduction

Loglinear models are used to estimate the distributions of discrete data that occur in applied research involving political questionnaires, biomedical data and psychometric testing (Agresti, 2002; Bishop, Fienberg & Holland, 1975; Holland & Thayer, 2000; Kolen & Brennan, 2004). The choice of most interest when selecting a plausible loglinear model for a particular discrete distribution is usually the number of moments of the observed distribution to preserve in the modeled distribution. A less familiar choice pertains to the distribution that underlies the loglinear model, which is obtained when most or all of the loglinear model's parameters are set to zero. This study considers the implications of using different underlying distributions – specifically uniform and binomial distributions – for loglinear models of discrete distributions.

Tim Moses is a Senior Psychometrician at Educational Testing Service where he works on several testing programs. He completed his Ph.D. in Educational Psychology at the University of Washington. Please send correspondence regarding this manuscript to Tim Moses, Educational Testing Service, Rosedale Road MS 03-P, Princeton, NJ 08541. Email him at: tmoses@ets.org.

Loglinear Models of Discrete Distributions

Loglinear models of discrete distributions relate the log of a model's expected probabilities, ρ, to a linear function of a categorical variable's values, for example the scores of a psychometric test,

$$\log_e(\boldsymbol{\rho}) = \alpha + \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} \qquad (1)$$

where $\boldsymbol{\rho}$ is an *I*-by-1 column vector of the probabilities, α is a normalizing constant which ensures that the sum of the entries of $\boldsymbol{\rho}$ is 1, $\sum_i \rho_i = 1$, $\boldsymbol{\mu}$ is an *I*-by-1 column vector of known constants, $\mathbf{X}$ is an *I*-by-*K* design matrix containing *K* functions of categorical variable *X*, and $\boldsymbol{\beta}$ is a *K*-by-1 column vector of free parameters. The $k = 1$ to *K* columns of $\mathbf{X}$ give the first through K^{th} degrees of the *X* values that can be expressed as power functions, x_i^k, or as used in this study, the more numerically stable and less collinear orthogonal polynomials. When maximum likelihood estimation is used for model 1 then the estimation results in the first derivative of the log-likelihood being set to zero, or,

$$\sum_i \hat{\rho}_i x_i^k = \sum_i \frac{n_i}{N} x_i^k, \qquad (2)$$

where n_i is the observed frequency of the i^{th} category value of X and N is the total sample size. Equation 2 implies that the first K moments of X's observed distribution will be preserved in the loglinear model's distribution (Agresti, 2002; Holland & Thayer, 2000).

Loglinear Models' Underlying Distributions

Specific values of $\boldsymbol{\mu}$ can result in loglinear models such as model 1 reflecting different underlying distributions. When $\boldsymbol{\mu}$ is a vector of zeros or of any constant and $\boldsymbol{\beta}$ is also zero the loglinear model resolves into a uniform distribution, $\log_e(\boldsymbol{\rho}) = \alpha$. The loglinear model that produces a uniform distribution reflects the notion that the i-level probabilities are all equal and independent of X.

Another choice for the loglinear model's underlying distribution is available when $\boldsymbol{\mu}$ is defined as I constants that vary by the categories of X, μ_i. Holland and Thayer (2000, pp. 139-140) showed that when model 1 has a $\boldsymbol{\mu}$ with entries

$$\mu_i = \log_e \binom{x_I}{x_i},$$

where $\binom{x_I}{x_i}$ denotes the binomial coefficient, "x_I choose x_i," then a binomial distribution can be produced by defining $\boldsymbol{\beta}$ as β_1 and defining $\mathbf{X}$ as a single column of values for fitting the first degree of X, $(x_1^1, x_2^1, ..., x_I^1)^t$. With these definitions model 1 can be expressed as,

$$\rho_i = \binom{x_I}{x_i} \pi^{x_i} (1-\pi)^{x_I - x_i} \qquad (3)$$

where π is a function of the mean of X,

$$\pi = \frac{1}{x_I} \sum_i \rho_i x_i = \frac{1}{x_I} \bar{x},$$

and a function of β_1,

$$\pi = \frac{\exp(\beta_1)}{1 + \exp(\beta_1)}.$$

Thus, equation 3 implies that for fixed value, x_I, and a parameter based on the mean, π, the probability of obtaining a particular value of X is a variate from a binomial distribution based on x_I trials and success probability π.

Assessing the Role of the Loglinear Model's Underlying Distribution in Models of Population Distributions

The role of the underlying distribution used in loglinear models has not been extensively studied. The focus of loglinear modeling applications to psychometric test score distributions tends to be on fitting several of the observed distributions' moments – that is, more than three (Holland & Thayer, 2000; Kolen & Brennan, 2004); thus, the relatively simple uniform and binomial distributions underlying the fitted distributions have not received much attention. It is possible, however, that the uniform and binomial distributions underlying the models that fit observed distributions have subtle influences on the overall fit of the loglinear model.

To illustrate the influence of uniform and binomial distributions, consider how loglinear models based on each distribution fit eight different population distributions. Table 1 shows the eight population distributions of X variables with 10 categories. Six of the population distributions were obtained from Steele and Chasling's (2006) study: the decreasing, step, triangular, platykurtic and leptokurtic distributions. Two other population distributions are based on both considered underlying distributions (the uniform distribution and the binomial distribution with π = 0.5). An additional under-dispersed binomial distribution was created to be similar to the binomial distribution, but with a relatively small variance.

For each of the eight distributions, loglinear models similar to model 1 were fit based on the uniform distribution (K =0, $\mu_i = 0$), and on fitting K = 1, 2, 3 and 4 moments with the loglinear model based on the uniform

distribution (equation 1 where **μ** is a vector of zeros). Other loglinear models comparable to model 1 were fit based on the binomial distribution, $K = 1$, $\mu_i = \log_e \binom{x_I}{x_i}$, and on fitting $K = 2$, 3 and 4 moments with the loglinear model based on the binomial distribution (equation 1 where $\boldsymbol{\mu}$ has entries

$$\mu_i = \log_e \binom{x_I}{x_i}).$$

Table 1's population distributions and the fits of the loglinear models to the population distributions for a hypothetical sample size of $N = 100$ are illustrated in Figures 1-16. These figures show the fits of the considered models in terms of individual score values and each model's summarized likelihood ratio Chi-square statistic:

$$G^2 = 2\sum_i n_i \log_e \left(\frac{n_i}{N\hat{\rho}_i} \right).$$

Table 1: Population Distributions

X's Categories & Moments	Uniform	Decreasing	Step	Triangular	Platykurtic	Leptokurtic	Binomial	Under-Dispersed Binomial
1	0.10	0.32	0.05	0.17	0.04	0.05	0.01	0.00
2	0.10	0.13	0.05	0.13	0.11	0.05	0.04	0.03
3	0.10	0.10	0.05	0.10	0.11	0.05	0.12	0.13
4	0.10	0.08	0.05	0.07	0.12	0.05	0.21	0.22
5	0.10	0.07	0.05	0.03	0.12	0.30	0.25	0.26
6	0.10	0.07	0.15	0.03	0.12	0.30	0.21	0.21
7	0.10	0.06	0.15	0.07	0.12	0.05	0.12	0.12
8	0.10	0.06	0.15	0.10	0.11	0.05	0.04	0.04
9	0.10	0.05	0.15	0.13	0.11	0.05	0.01	0.00
10	0.10	0.05	0.15	0.17	0.04	0.05	0.00	0.00
Mean	5.50	3.86	6.75	5.50	5.50	5.50	5.00	5.00
Std. Dev.	2.87	2.91	2.59	3.41	2.51	2.06	1.57	1.41
Skew	0.00	0.69	-0.68	0.00	0.00	0.00	0.02	0.04
Kurtosis	1.78	2.17	2.56	1.38	1.91	3.35	2.75	2.40

Figures 1-16 suggest that model fit is a function of the number of moments fit in the model and also show how closely the underlying distribution reflects the distribution being modeled. Loglinear models that fit K = 3 and 4 moments tend to have better fits (lower G^2 values) compared to models that fit K = 0, 1 and 2 moments, however, the loglinear model's underlying distribution appears to moderate the influence of K.

For population distributions more similar to the uniform distribution (i.e., the uniform, decreasing, step, triangular and platykurtic population distributions), models based on an underlying uniform distribution can closely fit the population distributions with fewer moments than those required by models based on an underlying binomial distribution (Figures 1, 3, 5, 7 & 9 vs. Figures 2, 4, 6, 8 & 10). For population distributions similar to the binomial distribution (i.e., the leptokurtic, binomial and under-dispersed binomial population distributions), models based on an underlying binomial distribution can closely fit the population distributions with fewer moments than those required by models based on an underlying uniform distribution (Figures 11, 13 & 15 vs. Figures 12, 14 & 16).

Methodology

To better understand the implications of results shown in Figures 1-16, a series of simulations was conducted. For the simulations of interest, 1,000 datasets of sample sizes 30, 100 and 1,000 were drawn from each of Table 1's population distributions. For each of the randomly drawn datasets, loglinear models were fit based on an underlying uniform distribution with K = 0-4, and also based on an underlying binomial distribution with K = 1-4. For models reflecting one of the two underlying distributions, the K values were selected based on nested Chi-square tests for differences in models' G^2 statistics (Haberman, 1974) and also on minimizing models' *AIC* statistics (Akaike, 1981).

To consider the influence of the underlying distribution for situations similar to what might be encountered in psychometric testing practice, where the moments to be fit in a test score distribution might be predetermined rather than statistically selected, modeling results were also produced by always fitting K = 4 moments based on the both the uniform and the binomial distributions. The results of interest for each combination of sample size, underlying distribution and moment selection method were the percentages of datasets where specific K values were selected, the mean K values across all 1,000 datasets and the average model fit (i.e., mean G^2 values) across all 1,000 datasets.

Results

Simulation results are summarized in Tables 2-9. Each table presents the simulation results for one of Table 1's eight population distributions; rows show the simulation results for a specific combination of sample size (30, 100 or 1,000), underlying distribution (the uniform or binomial distribution) and selection method for K (G^2, *AIC*, or K = 4). Each row's results show the percentage of moments (K) selected in the 1,000 simulated datasets, the mean of the selected K's and the mean model fit (mean G^2). Because the percentages in Tables 2-9 are presented in rounded form, they do not always sum to exactly 100% within each row.

Some results shown in Tables 2-9 have been shown elsewhere (Moses & Holland, 2010). K selections based on the *AIC* result in larger K values than selections based on the G^2. Selections based on sample sizes of 1,000 result in larger K values than selections based on smaller sample sizes. Models with larger K values fit the sample distributions more closely; that is, they result in smaller G^2.

Tables 2-9 show that the influences of the G^2 and *AIC* selection strategies and the sample sizes are moderated by how closely the loglinear model's underlying distribution reflects the population distribution. For population distributions that closely reflect the uniform distribution (i.e., the uniform and decreasing population distributions, Tables 2-3), using the uniform distribution results in *AIC* and G^2 model selections with smaller mean K values and larger mean G^2 values than using the binomial distribution. These results were also partially obtained for the step population distribution (Table 4, N =30 and 100), the triangular population distribution (Table 5, N =30), and the leptokurtic population distribution (Table 7, N =30 and 100).

Figure 1: Uniform Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

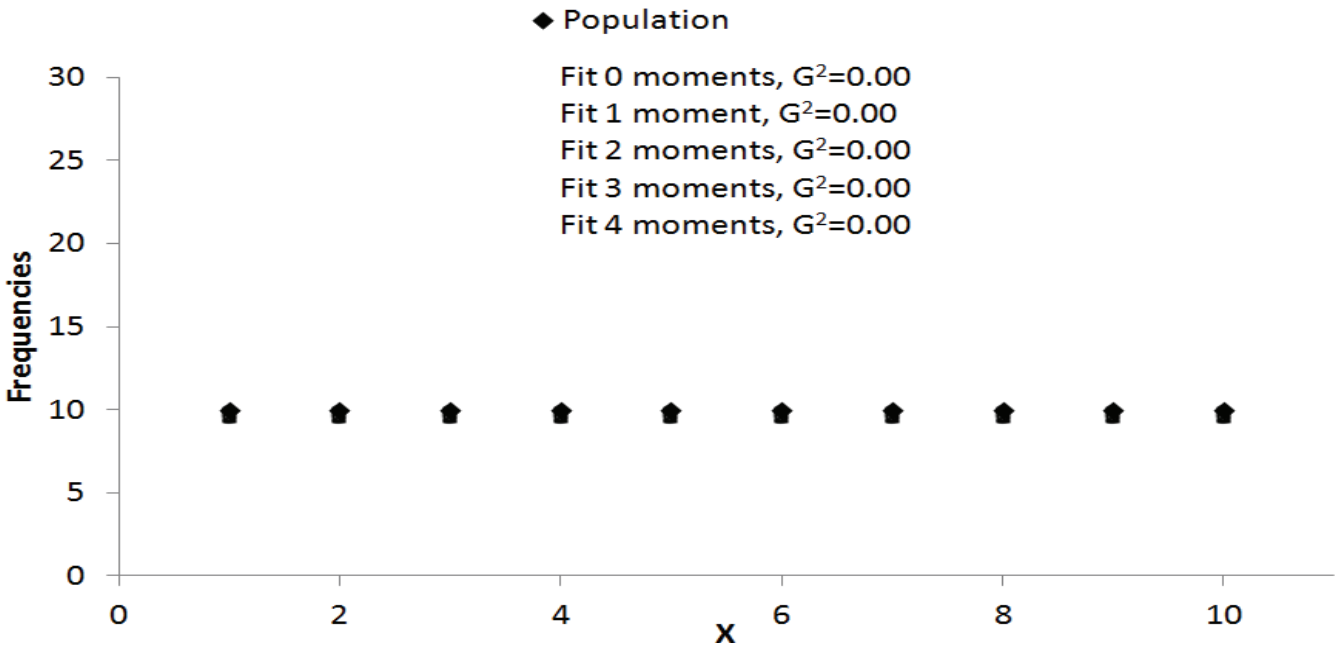

Figure 2: Uniform Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

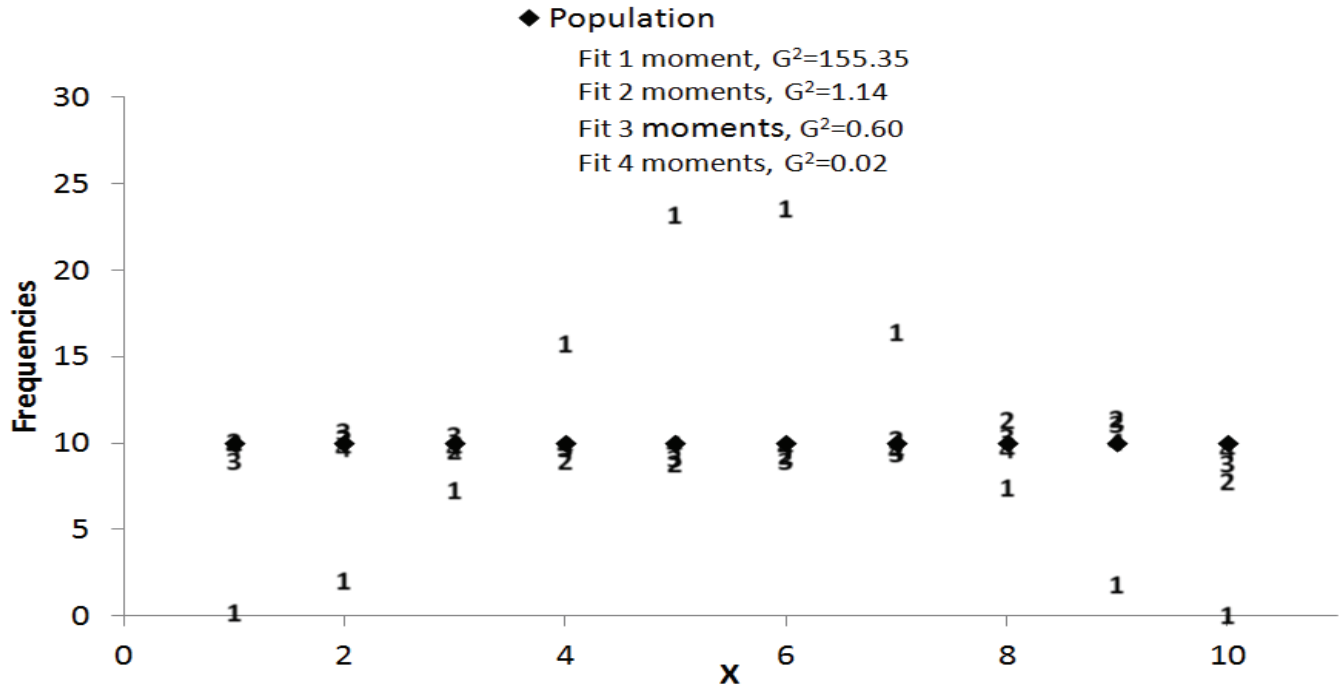

Figure 3: Decreasing Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

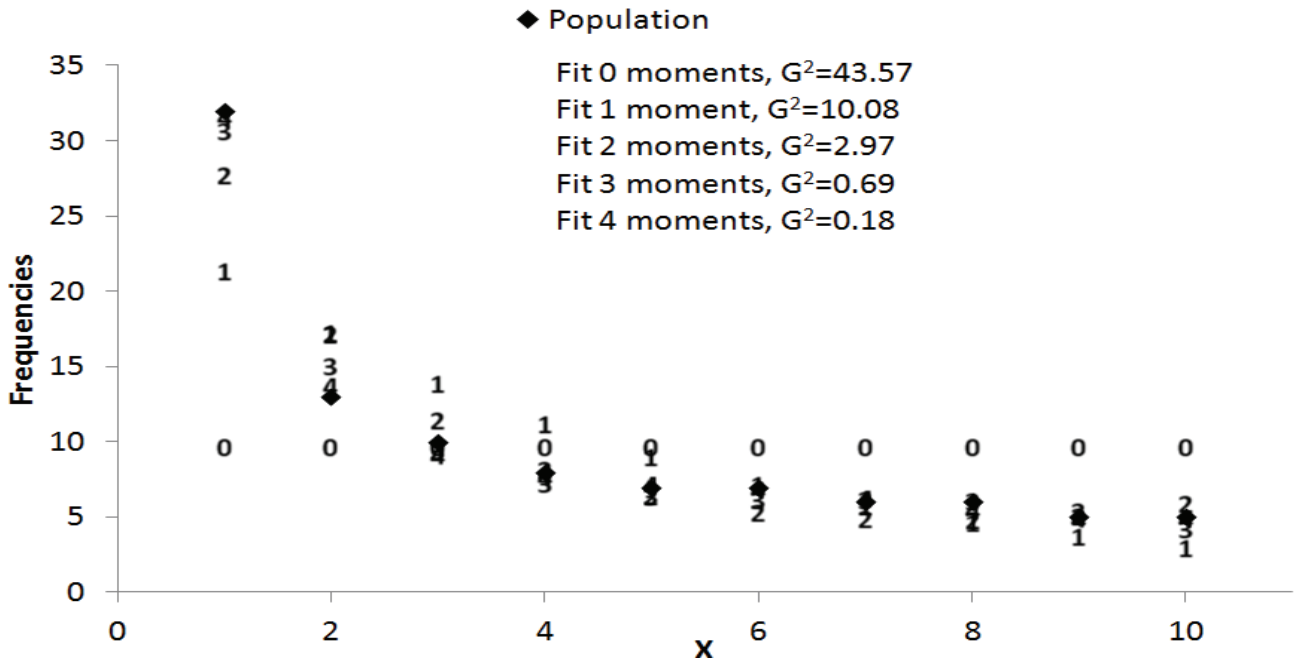

Figure 4: Decreasing Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

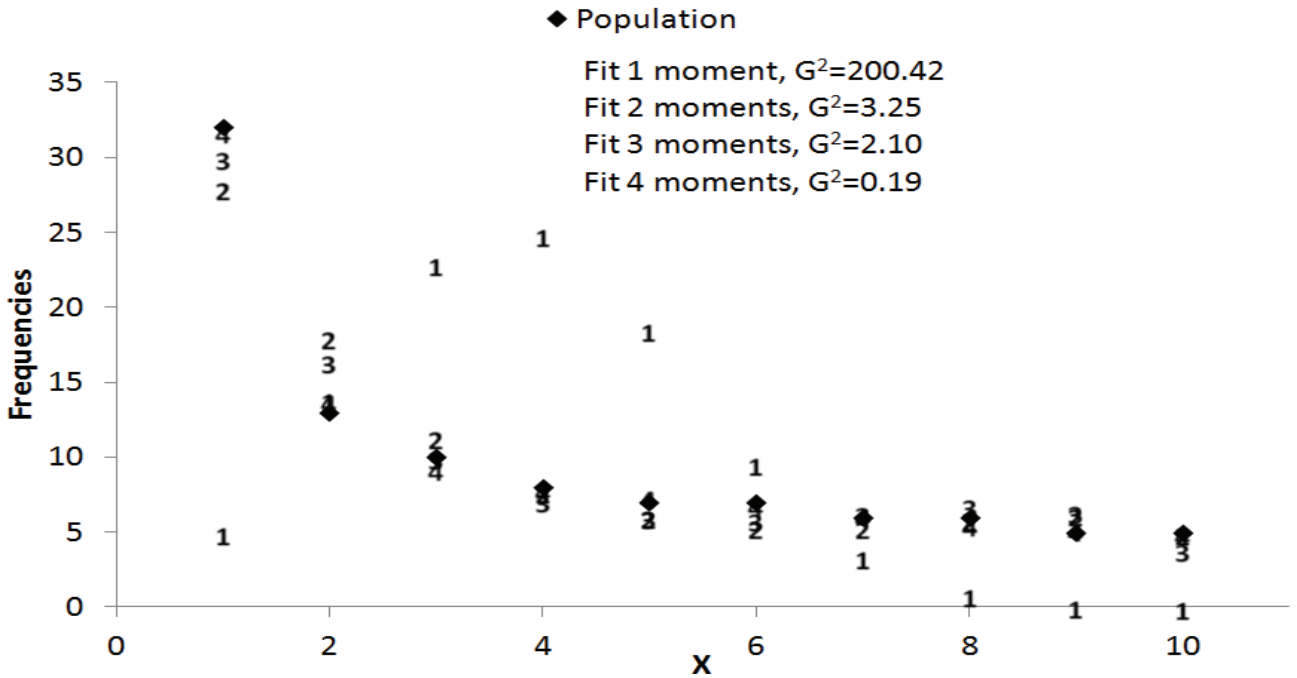

Figure 5: Step Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

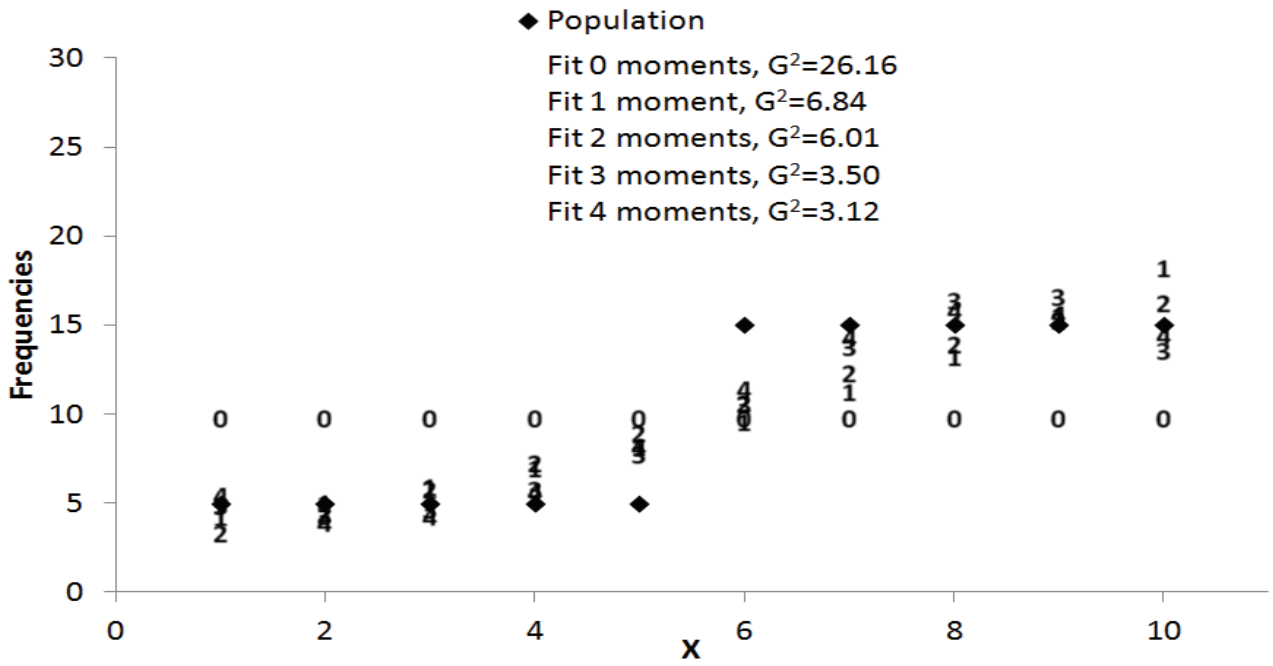

Figure 6: Step Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

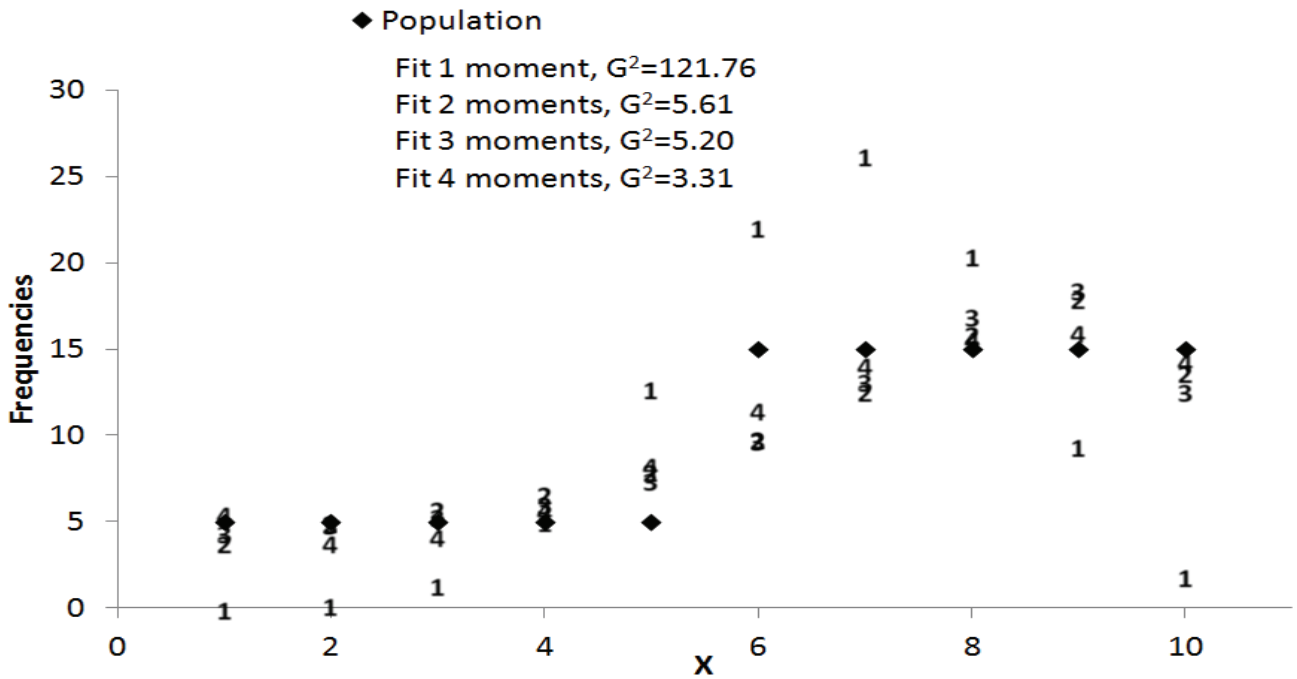

Figure 7: Triangular Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

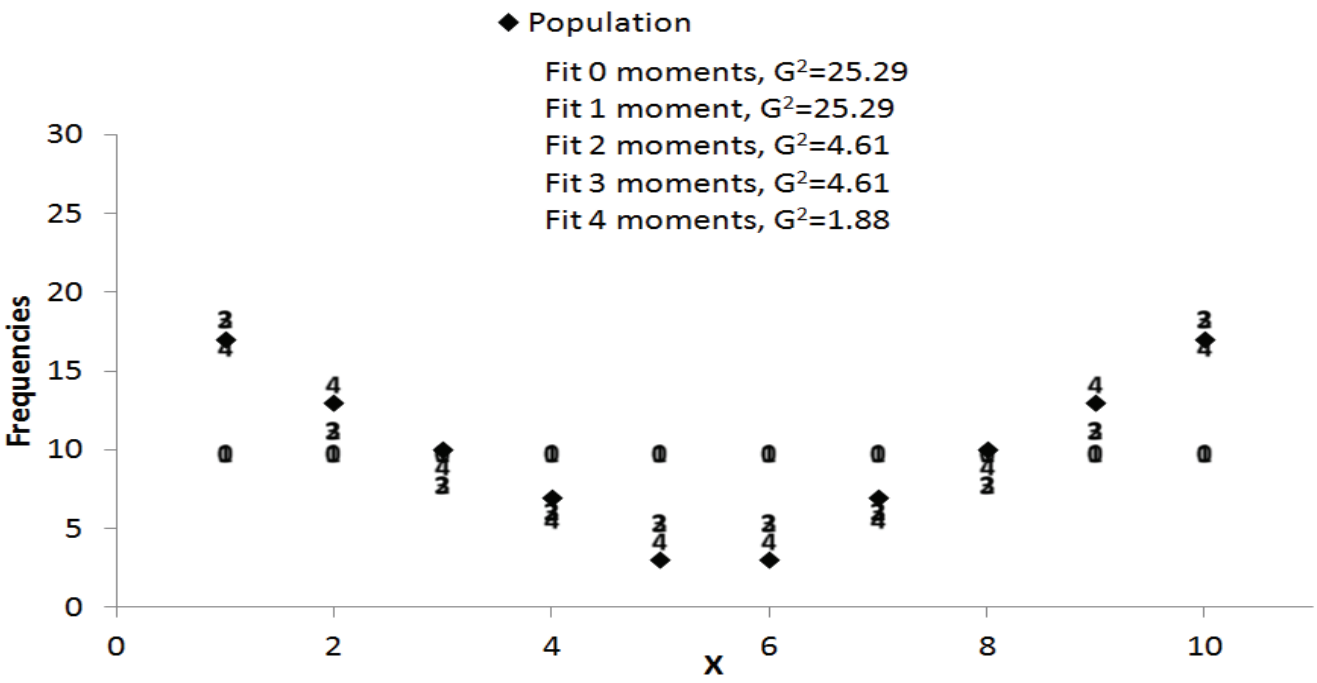

Figure 8: Triangular Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

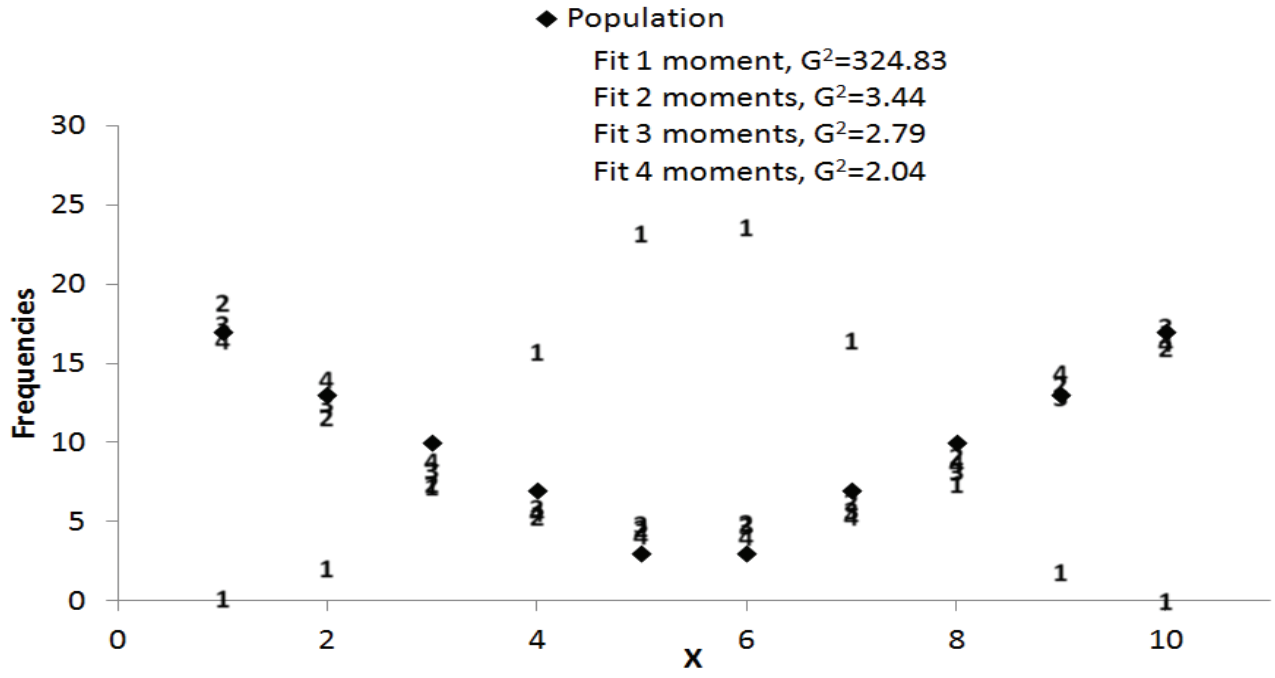

Figure 9: Platykurtic Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

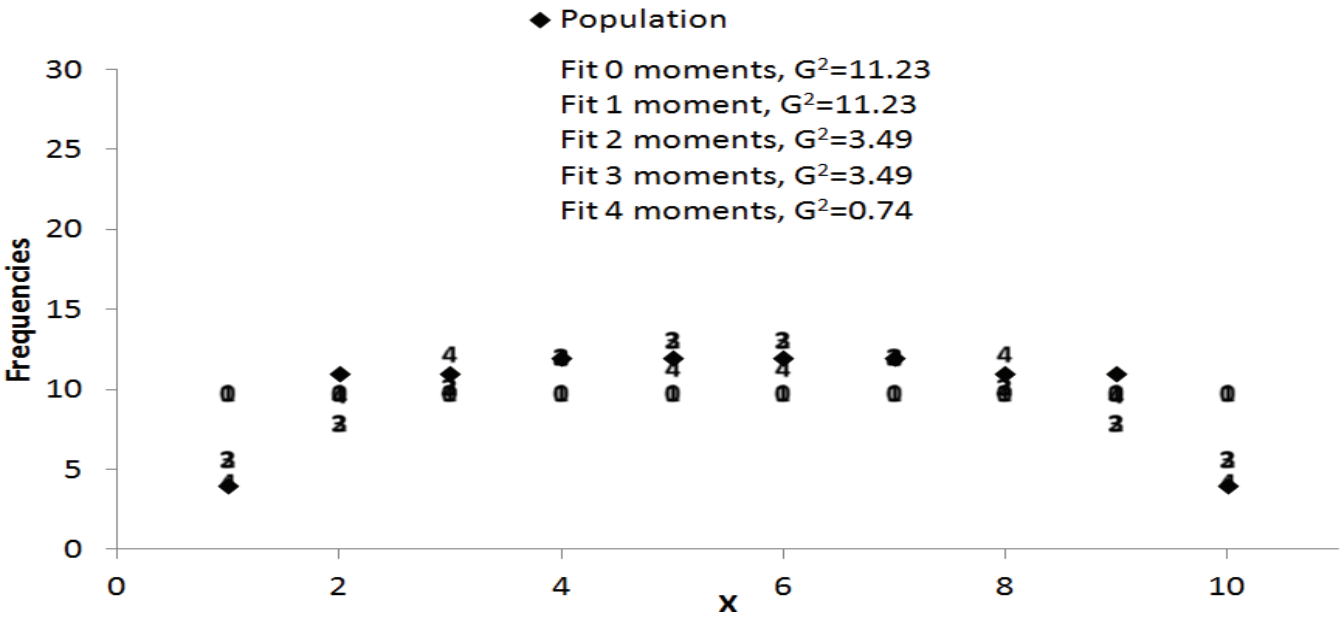

Figure 10: Platykurtic Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

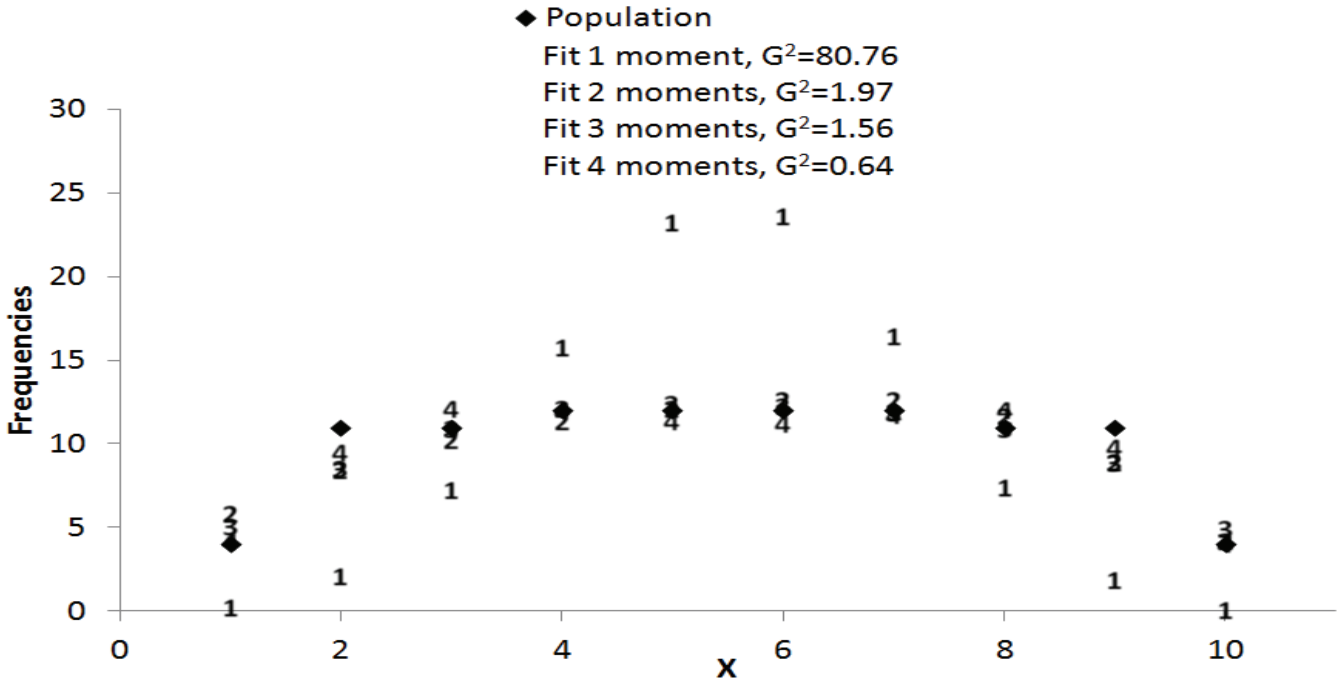

Figure 11: Leptokurtic Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

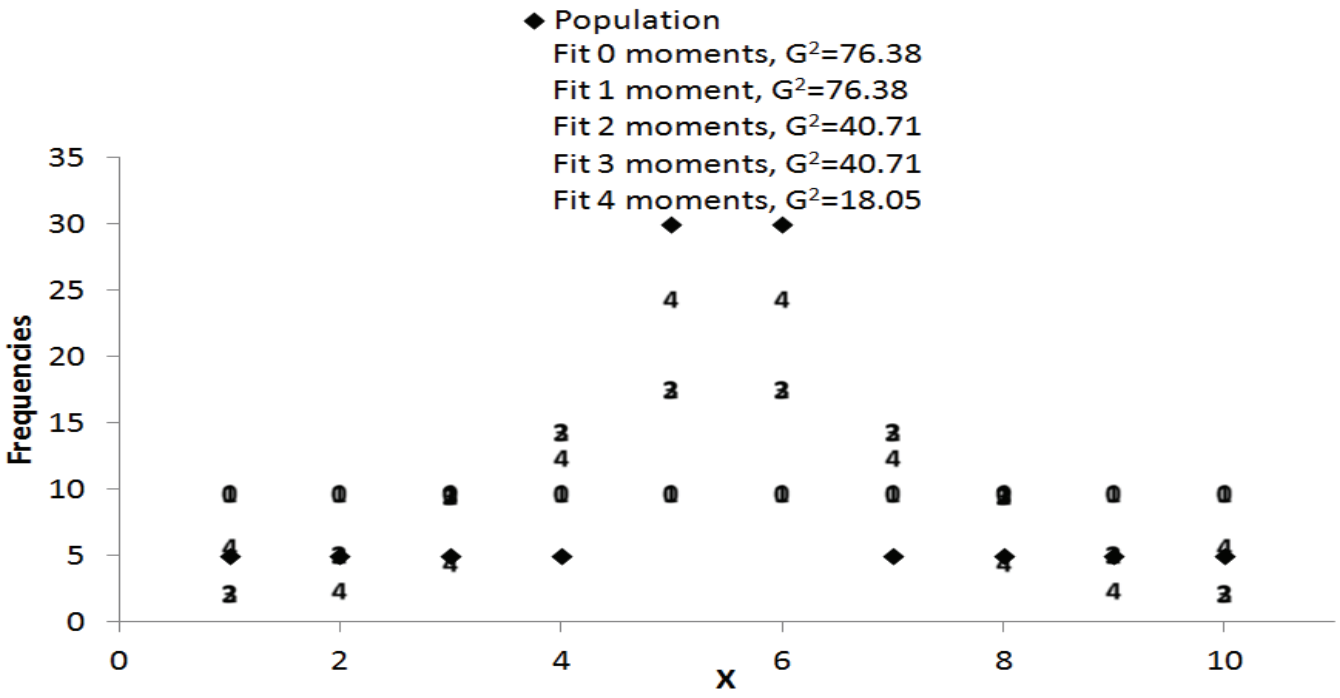

Figure 12: Leptokurtic Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

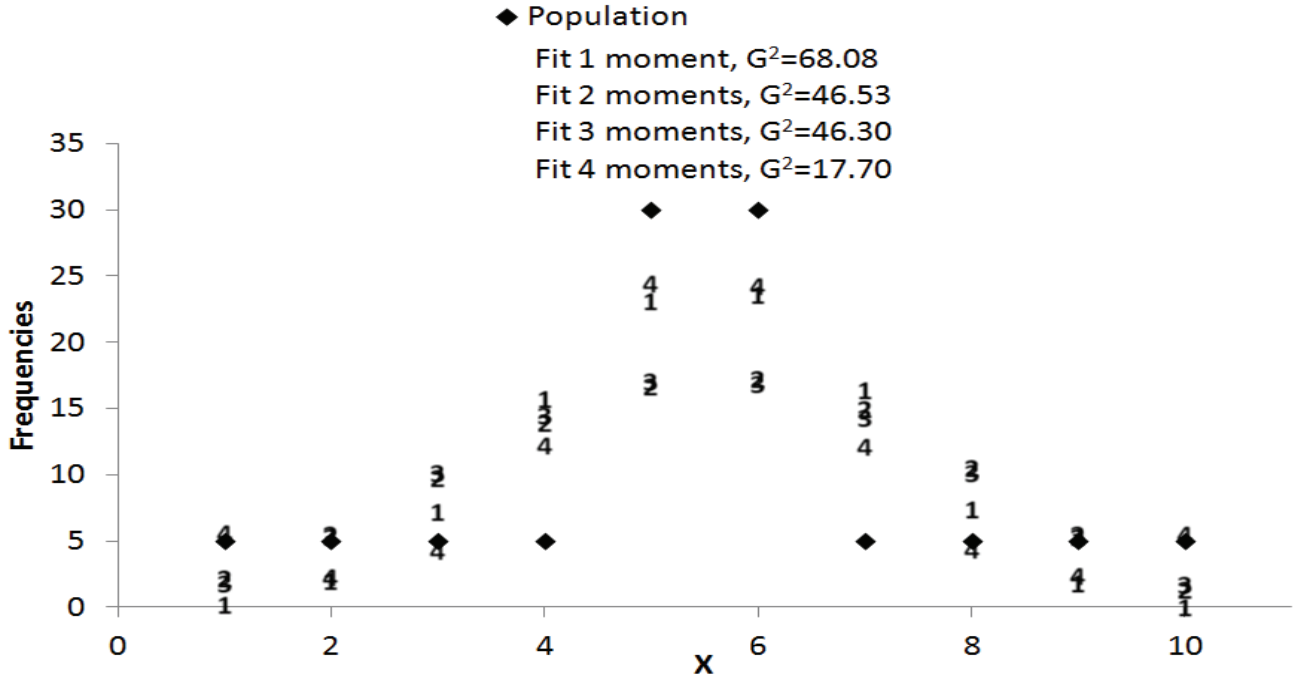

Figure 13: Binomial Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

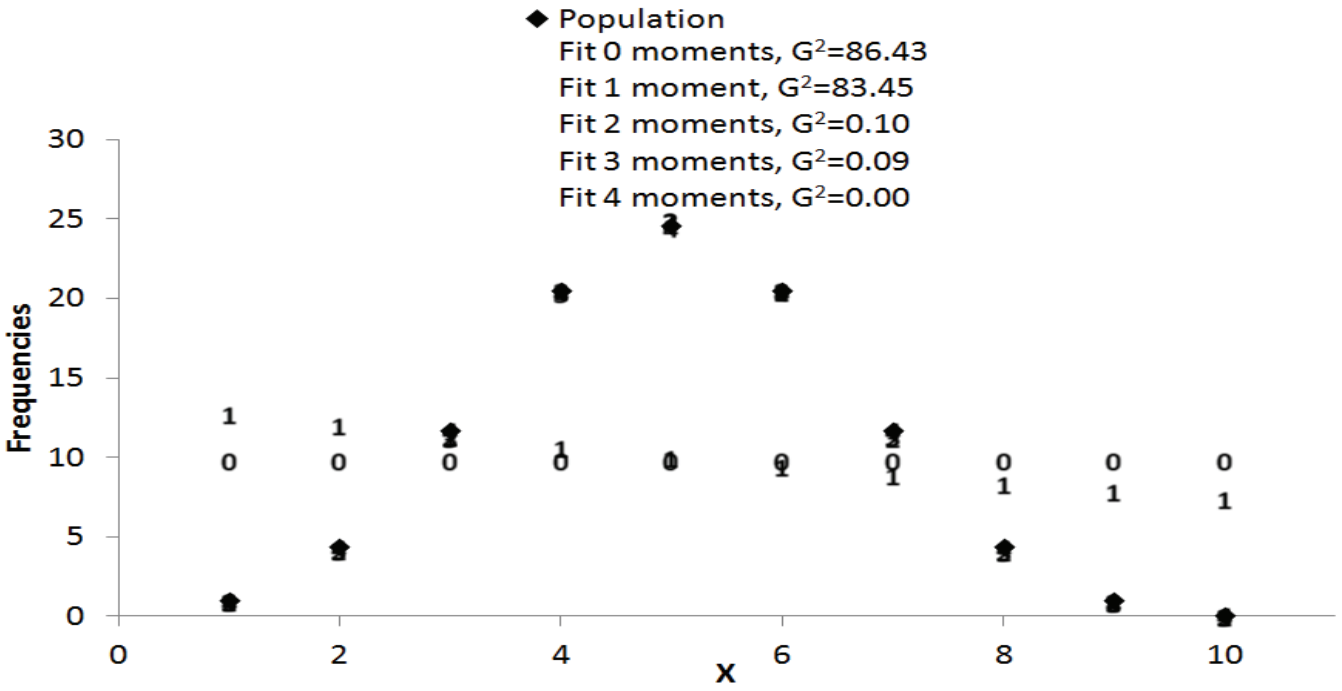

Figure 14: Binomial Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

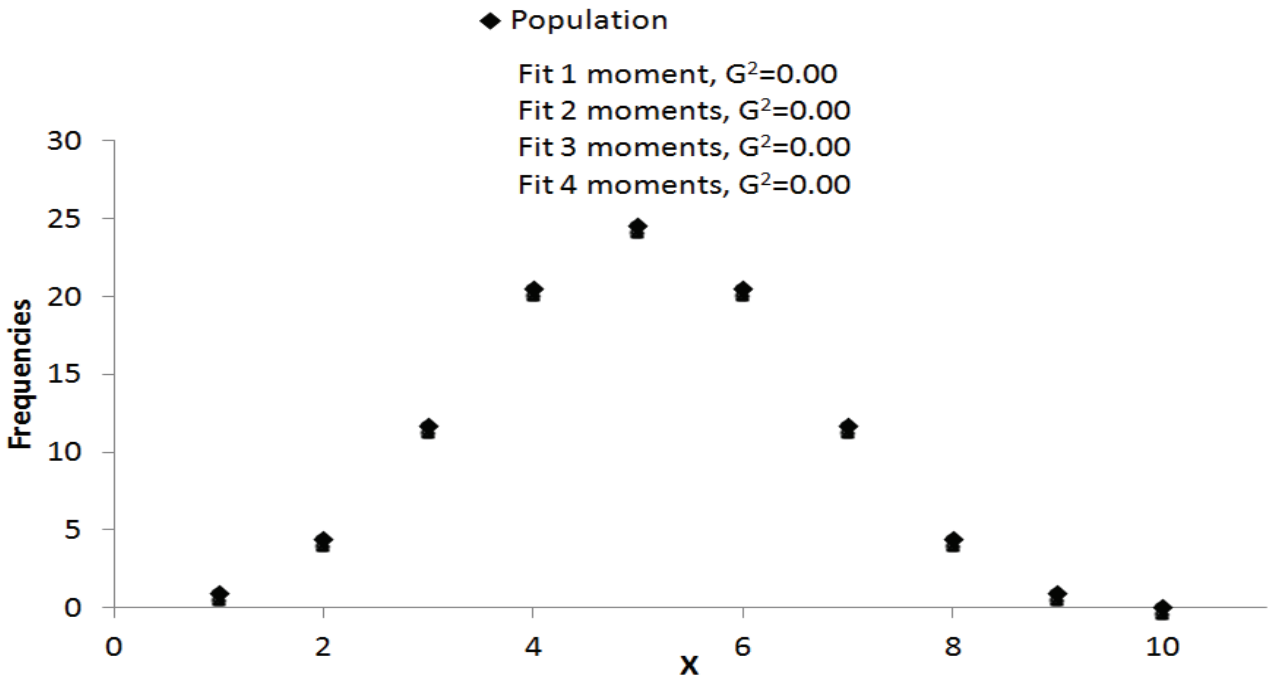

Figure 15: Under-Dispersed Binomial Population Distribution Modeling Results
Based on an Underlying Uniform Distribution

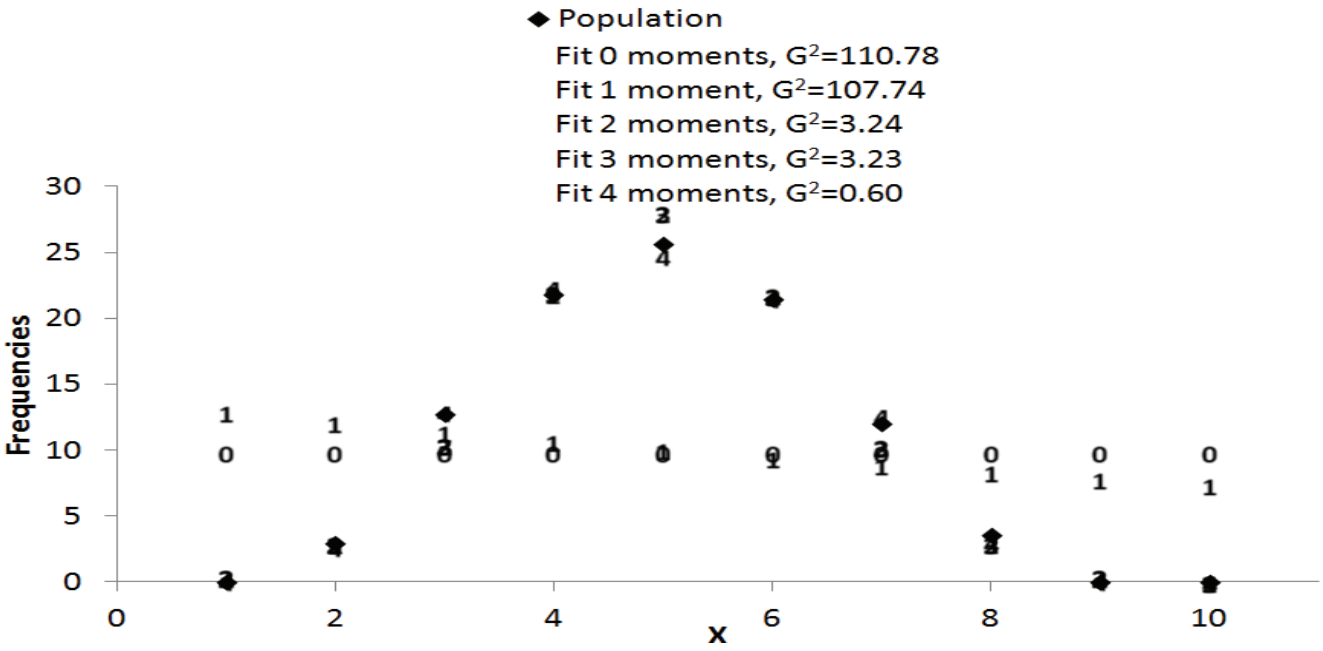

Figure 16: Under-Dispersed Binomial Population Distribution Modeling Results
Based on an Underlying Binomial Distribution

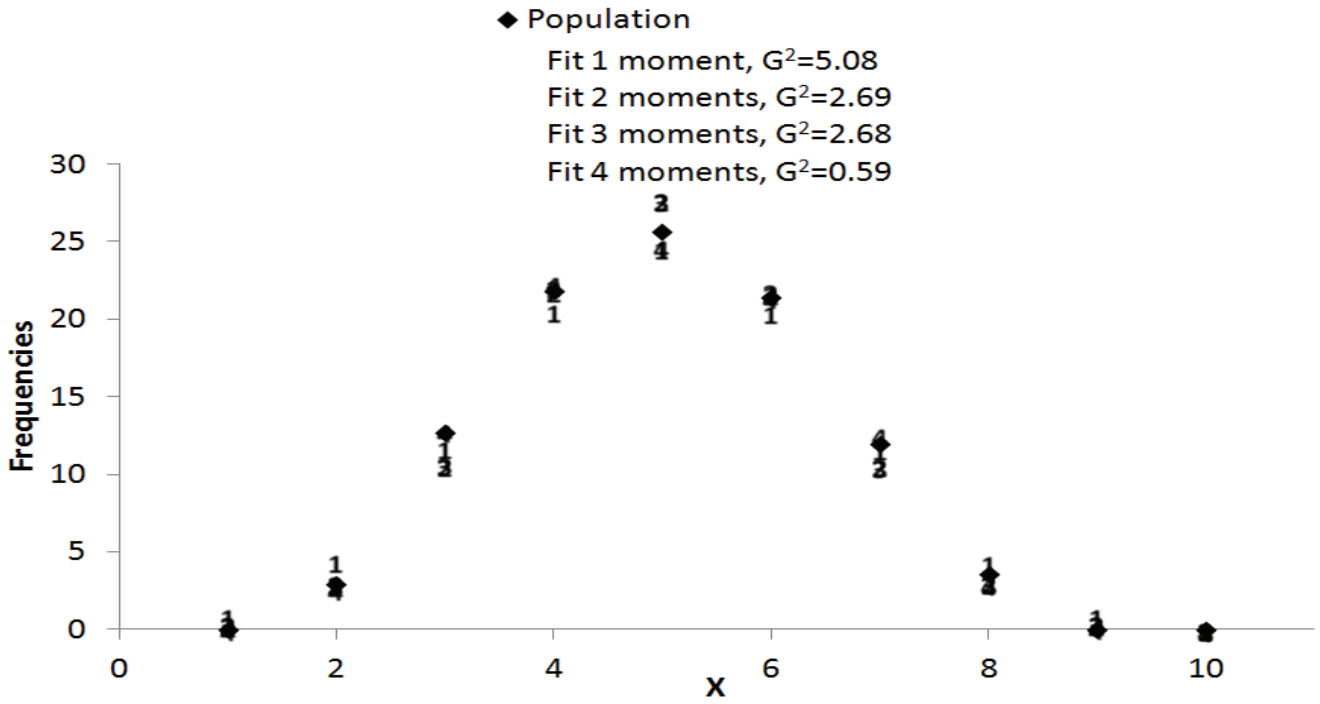

Table 2: Simulation Results for the Uniform Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	94%	1%	1%	2%	2%	0.16	9.37
	Binomial		0%	94%	3%	3%	2.08	7.75
100	Uniform	96%	1%	1%	1%	1%	0.11	8.97
	Binomial		0%	91%	4%	5%	2.15	7.56
1,000	Uniform	96%	2%	0%	1%	1%	0.08	8.38
	Binomial		0%	26%	25%	50%	3.24	7.11
AIC Selections of *K*								
30	Uniform	74%	12%	6%	4%	4%	0.52	8.35
	Binomial		0%	72%	16%	12%	2.41	6.80
100	Uniform	73%	13%	8%	4%	3%	0.51	7.84
	Binomial		0%	62%	21%	18%	2.56	6.28
1,000	Uniform	74%	13%	5%	5%	3%	0.49	7.28
	Binomial		0%	5%	13%	82%	3.77	5.22
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	5.78
	Binomial	0%	0%	0%	0%	100%	4.00	5.78
100	Uniform	0%	0%	0%	0%	100%	4.00	5.22
	Binomial	0%	0%	0%	0%	100%	4.00	5.24
1,000	Uniform	0%	0%	0%	0%	100%	4.00	4.80
	Binomial	0%	0%	0%	0%	100%	4.00	4.96

Table 3: Simulation Results for the Decreasing Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	19%	62%	13%	5%	1%	1.08	10.62
	Binomial		0%	92%	5%	4%	2.12	8.21
100	Uniform	0%	35%	46%	16%	3%	1.86	9.55
	Binomial		0%	78%	9%	13%	2.35	8.17
1,000	Uniform	0%	0%	1%	69%	30%	3.29	8.61
	Binomial		0%	1%	3%	96%	3.96	6.84
AIC Selections of *K*								
30	Uniform	3%	36%	33%	21%	8%	1.96	7.58
	Binomial		0%	65%	22%	13%	2.49	7.09
100	Uniform	0%	6%	35%	40%	19%	2.72	6.35
	Binomial		0%	44%	21%	35%	2.91	6.38
1,000	Uniform	0%	0%	0%	27%	73%	3.73	6.84
	Binomial		0%	0%	0%	100%	4.00	6.67
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	6.04
	Binomial	0%	0%	0%	0%	100%	4.00	6.04
100	Uniform	0%	0%	0%	0%	100%	4.00	5.43
	Binomial	0%	0%	0%	0%	100%	4.00	5.43
1,000	Uniform	0%	0%	0%	0%	100%	4.00	6.59
	Binomial	0%	0%	0%	0%	100%	4.00	6.67

Table 4: Simulation Results for the Step Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	45%	44%	4%	6%	3%	0.78	11.57
	Binomial		1%	91%	3%	5%	2.13	8.95
100	Uniform	2%	72%	7%	17%	3%	1.48	12.34
	Binomial		0%	81%	4%	15%	2.33	10.90
1,000	Uniform	0%	0%	0%	69%	30%	3.30	37.72
	Binomial		0%	2%	1%	97%	3.95	38.01
AIC Selections of *K*								
30	Uniform	10%	50%	16%	17%	8%	1.63	8.65
	Binomial		0%	68%	16%	16%	2.48	7.91
100	Uniform	0%	35%	15%	34%	16%	2.32	9.68
	Binomial		0%	54%	12%	34%	2.81	9.50
1,000	Uniform	0%	0%	0%	30%	70%	3.70	36.16
	Binomial		0%	0%	0%	100%	4.00	37.84
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	6.83
	Binomial	0%	0%	0%	0%	100%	4.00	6.89
100	Uniform	0%	0%	0%	0%	100%	4.00	8.32
	Binomial	0%	0%	0%	0%	100%	4.00	8.52
1,000	Uniform	0%	0%	0%	0%	100%	4.00	35.89
	Binomial	0%	0%	0%	0%	100%	4.00	37.84

Table 5: Simulation Results for the Triangular Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	46%	2%	44%	2%	7%	1.21	10.83
	Binomial		0%	93%	4%	3%	2.10	8.56
100	Uniform	1%	0%	75%	1%	22%	2.43	9.69
	Binomial		0%	88%	6%	6%	2.19	9.66
1,000	Uniform	0%	0%	0%	0%	100%	4.00	23.55
	Binomial		0%	14%	19%	66%	3.52	26.68
AIC Selections of *K*								
30	Uniform	17%	2%	50%	11%	19%	2.13	8.13
	Binomial		0%	71%	16%	12%	2.41	7.62
100	Uniform	0%	0%	46%	7%	46%	3.00	7.98
	Binomial		0%	56%	21%	23%	2.66	8.20
1,000	Uniform	0%	0%	0%	0%	100%	4.00	23.55
	Binomial		0%	2%	6%	92%	3.91	25.23
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	6.50
	Binomial	0%	0%	0%	0%	100%	4.00	6.55
100	Uniform	0%	0%	0%	0%	100%	4.00	7.08
	Binomial	0%	0%	0%	0%	100%	4.00	7.24
1,000	Uniform	0%	0%	0%	0%	100%	4.00	23.55
	Binomial	0%	0%	0%	0%	100%	4.00	25.12

Table 6: Simulation Results for the Platykurtic Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	74%	0%	17%	0%	9%	0.71	10.38
	Binomial		2%	91%	1%	6%	2.11	8.00
100	Uniform	26%	0%	50%	1%	24%	1.96	9.44
	Binomial		0%	86%	4%	10%	2.24	7.94
1,000	Uniform	0%	0%	0%	0%	100%	3.99	12.24
	Binomial		0%	16%	8%	76%	3.60	12.41
AIC Selections of *K*								
30	Uniform	44%	3%	29%	5%	19%	1.53	8.22
	Binomial		0%	70%	13%	16%	2.46	6.93
100	Uniform	7%	0%	41%	5%	47%	2.85	6.95
	Binomial		0%	60%	15%	26%	2.66	6.72
1,000	Uniform	0%	0%	0%	0%	100%	4.00	12.22
	Binomial		0%	2%	3%	96%	3.94	11.25
Always Fit $K = 4$								
30	Uniform	0%	0%	0%	0%	100%	4.00	5.88
	Binomial	0%	0%	0%	0%	100%	4.00	5.84
100	Uniform	0%	0%	0%	0%	100%	4.00	5.78
	Binomial	0%	0%	0%	0%	100%	4.00	5.67
1,000	Uniform	0%	0%	0%	0%	100%	4.00	12.22
	Binomial	0%	0%	0%	0%	100%	4.00	11.18

Table 7: Simulation Results for the Leptokurtic Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	11%	0%	42%	7%	41%	2.68	13.94
	Binomial		21%	16%	9%	54%	2.97	13.21
100	Uniform	0%	0%	3%	0%	97%	3.94	23.57
	Binomial		0%	1%	0%	99%	3.98	23.12
1,000	Uniform	0%	0%	0%	0%	100%	4.00	185.28
	Binomial		0%	0%	0%	100%	4.00	181.76
AIC Selections of *K*								
30	Uniform	3%	0%	21%	10%	66%	3.35	11.86
	Binomial		8%	11%	8%	73%	3.47	11.65
100	Uniform	0%	0%	1%	0%	99%	3.99	23.41
	Binomial		0%	0%	0%	100%	4.00	23.05
1,000	Uniform	0%	0%	0%	0%	100%	4.00	185.28
	Binomial		0%	0%	0%	100%	4.00	181.76
Always Fit $K = 4$								
30	Uniform	0%	0%	0%	0%	100%	4.00	11.17
	Binomial	0%	0%	0%	0%	100%	4.00	11.07
100	Uniform	0%	0%	0%	0%	100%	4.00	23.40
	Binomial	0%	0%	0%	0%	100%	4.00	23.05
1,000	Uniform	0%	0%	0%	0%	100%	4.00	185.28
	Binomial	0%	0%	0%	0%	100%	4.00	181.76

Table 8: Simulation Results for the Binomial Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	0%	0%	96%	1%	4%	2.08	6.09
	Binomial		94%	1%	1%	4%	1.15	6.86
100	Uniform	0%	0%	95%	1%	4%	2.09	6.89
	Binomial		94%	2%	1%	3%	1.13	7.62
1,000	Uniform	0%	0%	92%	1%	7%	2.14	7.46
	Binomial		96%	1%	1%	2%	1.10	7.69
AIC Selections of *K*								
30	Uniform	0%	0%	74%	9%	17%	2.43	5.11
	Binomial		72%	12%	6%	10%	1.53	5.82
100	Uniform	0%	0%	77%	10%	13%	2.36	6.12
	Binomial		75%	12%	6%	7%	1.46	6.70
1,000	Uniform	0%	0%	68%	11%	21%	2.53	6.31
	Binomial		76%	12%	7%	5%	1.42	6.78
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	4.01
	Binomial	0%	0%	0%	0%	100%	4.00	3.99
100	Uniform	0%	0%	0%	0%	100%	4.00	4.99
	Binomial	0%	0%	0%	0%	100%	4.00	4.96
1,000	Uniform	0%	0%	0%	0%	100%	4.00	5.13
	Binomial	0%	0%	0%	0%	100%	4.00	5.08

Table 9: Simulation Results for the Under-Dispersed Binomial Population Distribution

N	Underlying Distribution	Percentage of *K* Moments Selected (out of 1,000 replications)					Mean Moments (*K*)	Mean G^2
		0	1	2	3	4		
G^2 Selections of *K*								
30	Uniform	0%	0%	91%	0%	9%	2.19	4.70
	Binomial		85%	5%	0%	10%	1.36	5.70
100	Uniform	0%	0%	79%	0%	21%	2.42	5.80
	Binomial		64%	18%	0%	18%	1.73	7.13
1,000	Uniform	0%	0%	0%	0%	100%	4.00	8.49
	Binomial		0%	0%	0%	100%	4.00	8.37
AIC Selections of *K*								
30	Uniform	0%	0%	71%	1%	27%	2.56	3.72
	Binomial		57%	24%	1%	18%	1.81	4.40
100	Uniform	0%	0%	49%	1%	50%	3.01	4.26
	Binomial		23%	40%	1%	36%	2.49	4.85
1,000	Uniform	0%	0%	0%	0%	100%	4.00	8.49
	Binomial		0%	0%	0%	100%	4.00	8.35
Always Fit *K* = 4								
30	Uniform	0%	0%	0%	0%	100%	4.00	2.56
	Binomial	0%	0%	0%	0%	100%	4.00	2.55
100	Uniform	0%	0%	0%	0%	100%	4.00	3.20
	Binomial	0%	0%	0%	0%	100%	4.00	3.18
1,000	Uniform	0%	0%	0%	0%	100%	4.00	8.49
	Binomial	0%	0%	0%	0%	100%	4.00	8.35

For population distributions that closely reflect the binomial distribution (i.e., the binomial and under-dispersed binomial population distributions, Tables 8-9), using the binomial distribution results in *AIC* and G^2 model selections with smaller mean *K* values and larger mean G^2 values than using the uniform distribution. Model selection results for the platykurtic population distribution were mixed (Table 6).

When loglinear models were fit with each underlying distribution using a predetermined $K = 4$, rather than a statistically selected *K* value, the results depended on how closely the underlying distribution reflected the population distribution. For the uniform, decreasing, step and triangular population distributions (Tables 2-5), the use of a predetermined *K* value of 4 resulted in slightly smaller mean G^2 values with the uniform distribution than the binomial distribution. For the platykurtic, leptokurtic, binomial and under-dispersed binomial population distributions (Tables 6-9), the use of a predetermined *K* value of 4 resulted in slightly smaller mean G^2 values with the binomial distribution than the uniform distribution.

Conclusion

The results of this study show that the loglinear model's underlying distribution has a small – but real – effect in modeling discrete distributions. This effect depends on the population distribution and on how the number of moments is determined. For models fitting a predetermined number of moments, this study shows that the use of a binomial underlying distribution can result in better fits for distributions that are similar to the binomial distribution, whereas the use of a uniform underlying distribution can result in better fits for distributions similar to the uniform distribution.

For the statistical selection of loglinear models in sample distributions, results suggest that using a distribution less similar to the distribution being modeled (e.g., using the binomial as the underlying distribution for uniformly distributed populations and samples) results in more moments being chosen and slightly better model fit. The implications for modeling distributions more likely to resemble binomial distributions than uniform distributions (e.g., psychometric tests) are that better fitting models can be statistically selected when using an underlying uniform distribution, and better fitting models for a predetermined number of moments can be obtained using an underlying binomial distribution.

Results obtained herein are useful replications and extensions of other studies that have assessed statistical power for detecting departures from uniform distributions (Choulakian, Lockhart & Stephens, 1994; Pettitt & Stephens, 1977; Steele & Chaseling, 2006). This study also showed that the likelihood ratio (G^2) selection strategy had relatively moderate power levels and the *AIC* selection strategies had relatively high power levels compared to the strategies considered in Steele and Chaeseling's study (i.e., the Kolmogorov-Smirnov, nominal Kolmogorov-Smirnov, Cramer-von Mises, Anderson-Darling, Pearson Chi-square and Watson's tests). Similar to the prior studies, this study found that power is higher when the underlying distribution is less similar to the distribution being modeled.

This study extends prior power studies by considering Type I error, where this study shows that Type I error rates were closer to 5% for the G^2 selection strategy than the *AIC* selection strategy. The more controlled Type I error rates of the G^2 selection strategy were observed both for the uniform distribution (underlying and population, Table 2) and also for the binomial distribution (underlying and population, Table 8). This study's findings that the G^2 and *AIC* selection strategies' Type I error and power tendencies for assessing binomial distributions are similar to those for assessing uniform distributions also extend prior studies that have primarily focused on detecting departures from uniform distributions. Results suggest that future studies considering loglinear models' underlying distributions would be useful for comparing other distributions and statistical selection strategies.

References

Agresti, A. (2002). *Categorical data analysis* (*2nd Ed*.). New York, NY: Wiley.

Akaike, H. (1981). Likelihood of a model and information criteria. *Journal of Econometrics*, *16*, 3-14.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

Choulakian, V., Lockhart, R. A., & Stephens, M. A. (1994). Cramer-von Mises statistics for discrete distributions. *The Canadian Journal of Statistics*, *22*, 125-137.

Haberman, S. J. (1974). Log-linear models for frequency tables with ordered classifications. *Biometrics*, *30*, 589-600.

Holland, P. W., & Thayer, D. T. (2000). Univariate and bivariate loglinear models for discrete test score distributions. *Journal of Educational and Behavioral Statistics*, *25*, 133-183.

Kolen, M. J., & Brennan, R. J. (2004). *Test equating, scaling and linking* (*2nd Ed*.). New York, NY: Springer-Verlag.

Moses, T., & Holland, P. W. (2010). A comparison of statistical selection strategies for univariate and bivariate log-linear models. *British Journal of Mathematical and Statistical Psychology*, *63*, 557-574.

Pettitt, A. N., & Stephens, M. A. (1977). The Komogorov-Smirnov goodness-of-fit statistic with discrete and grouped data. *Technometrics*, *19*, 205-210.

Steele, M., & Chaseling, J. (2006). Powers of discrete goodness-of-fit test statistics for a uniform null against a selection of alternative distributions. *Communications in Statistics-Simulation and Computation*, *35*, 1067-1075.