Analysis of Count Data – A Business Perspective

George J. Hurley
Sr. Research Manager
The Hershey Company
Milwaukee June 2013

Overview

- Count data
- Methods
- Conclusions

Count data

- Count data
    - Anything with a whole number response variable
    - Number of people in front of a person in a call center queue
    - Number of items purchased by a person in checking out in a store
    - Number of items purchased by a person entering a store
- Data is simulated for this talk

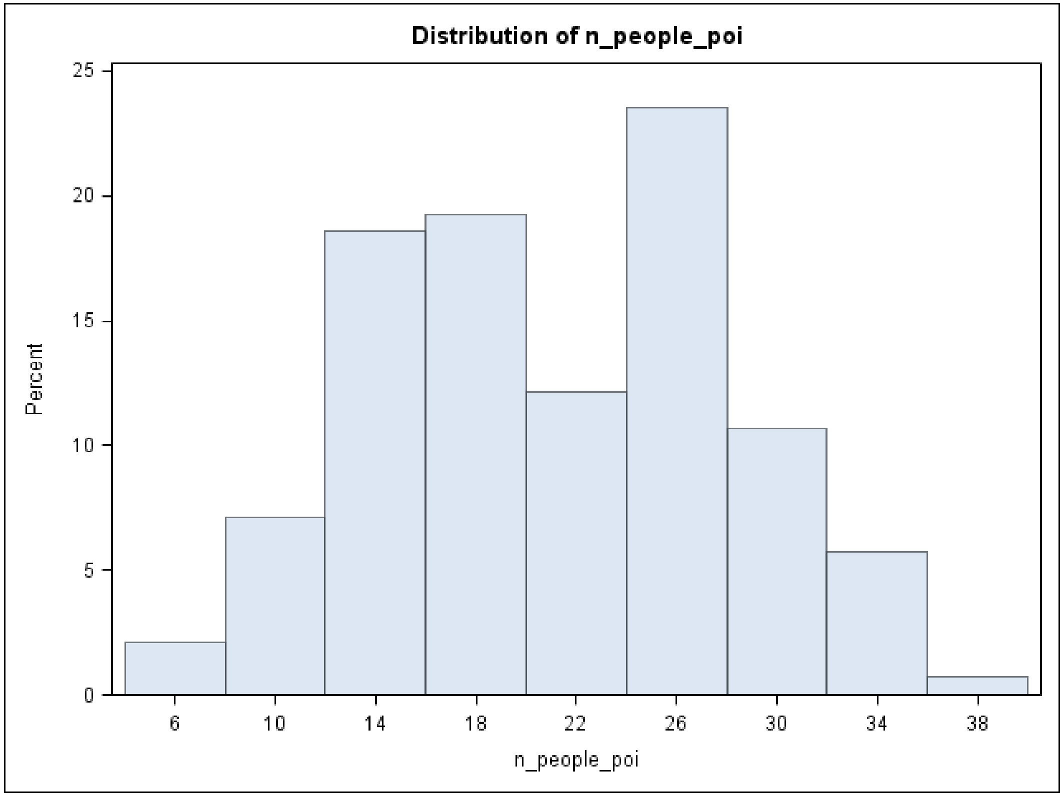
```
data dd1.poisson_data;
do i=1 to 40;
store_type="Big";
shelf_set="New";
n_people_poi=ranpoi(1978,27);
n_people_inf=round(ranpoi(1978,21)+sqrt(10)*rannor(1971),1);
if i<6 then n_people_zp=0;
else n_people_zp=n_people_poi;
output;
end;
do i=1 to 40;
store_type="Big";
shelf_set="Old";
n_people_poi=ranpoi(2009,23);
n_people_inf=round(ranpoi(2009,23)+sqrt(10)*rannor(2005),1);
if i<8 then n_people_zp=0;
else n_people_zp=n_people_poi;
output;
end;
do i=1 to 30;
store_type="Sml";
shelf_set="New";
n_people_poi=ranpoi(2006,17);
n_people_inf=round(ranpoi(2006,17)+sqrt(10)*rannor(2013),1);
if i<5 then n_people_zp=0;
else n_people_zp=n_people_poi;
output;
end;
do i=1 to 30;
store_type="Sml";
shelf_set="Old";
n_people_poi=ranpoi(1999,13);
n_people_inf=round(ranpoi(1999,13)+sqrt(10)*rannor(2012),1);
if i<7 then n_people_zp=0;
else n_people_zp=n_people_poi;
output;
end;

run;
```

Count data

- It is always ideal to get an understanding of your data prior to any modeling

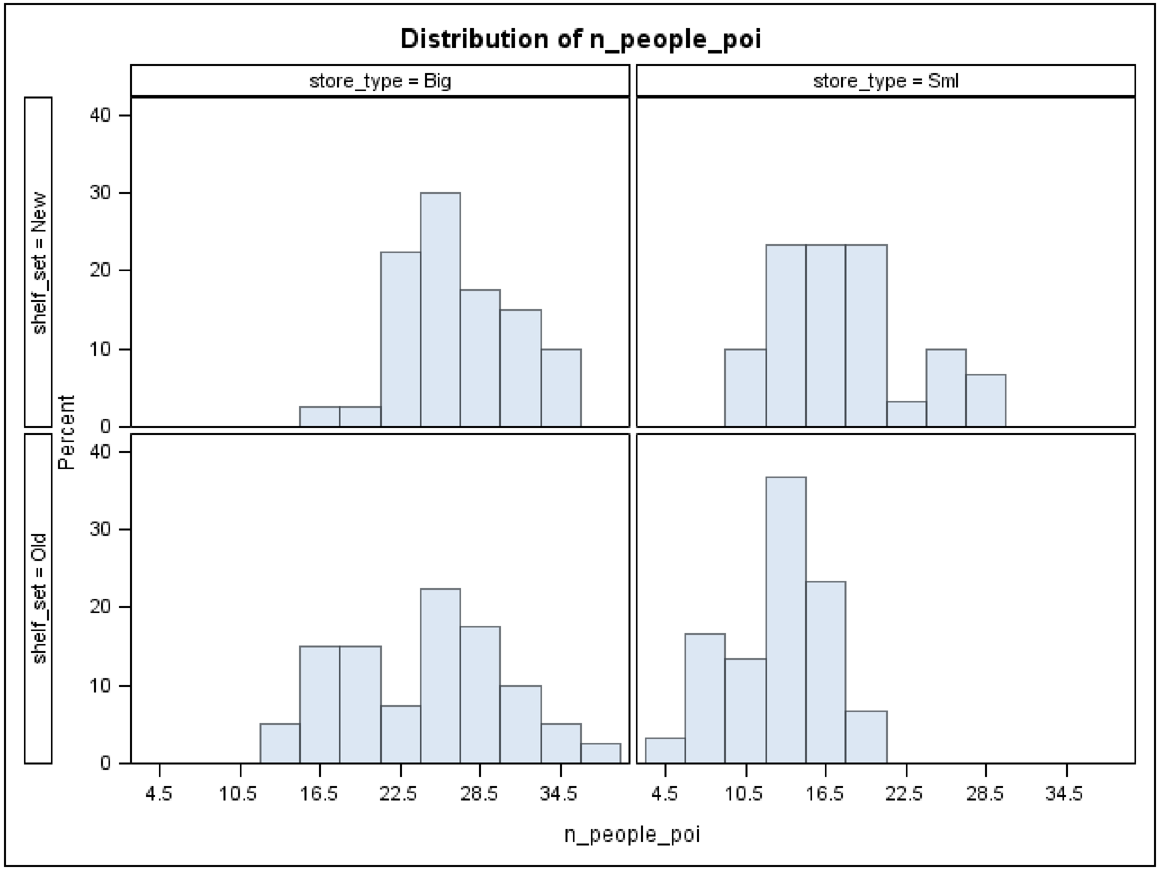
```
proc univariate data=dd1.poisson_data;
var n_people_poi n_people_inf n_people_zp;
histogram n_people_poi n_people_inf n_people_zp;
run;
```



Count data

- It is always ideal to get an understanding of your data prior to any modeling

```
proc univariate data=dd1.poisson_data;
class shelf_set store_type;
var n_people_poi;
histogram n_people_poi;
run;
```



Methods: Model 1 – Simple Poisson Regression

- The simplest model for count data is Simple Poisson Regression
  - Dist=Poisson utilizes Poisson distribution to model data
  - Link=Log utilizes the log link function
    - Log is the canonical link function for the Poisson distribution
    - Essentially using a canonical link function provides the best estimate for β

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_poi=shelf_set / dist=poisson link=log;
lsmeans shelf_set / ilink;
run;
```

In the model statement, dist=Poisson indicates the Poisson distribution is to be used. Generally speaking, the link function used with the Poisson distribution is the log link, as it is the canonical link function. Since a link function is used, ilink is used in the lsmeans statement to produce means output back on the original scale.

Methods: Model 1 – Simple Poisson Regression

- Overdispersion is present in this model
  - Value/DF should be near 1 for Deviance and Pearson Chi-Square
  - Scaled Pearson and Deviance will be discussed in Model 3
- Poisson distribution has mean=variance, hence one parameter is estimated for both
  - Overdispersion is the case where the model underestimates the variance
  - A common cause is subject heterogeneity

Criterion	DF	Value	Value/DF
Deviance	138	345.1045	2.5008
Scaled Deviance	138	345.1045	2.5008
Pearson Chi-Square	138	337.9961	2.4492
Scaled Pearson X2	138	337.9961	2.4492
Log Likelihood		5866.8141	
Full Log Likelihood		-508.8216	
AIC (smaller is better)		1021.6433	
AICC (smaller is better)		1021.7309	
BIC (smaller is better)		1027.5266	

Methods: Model 2 – Simple Poisson Regression accounting for subject heterogeneity

- In Model 2, all relevant predictors are included
  - Little evidence of overdispersion

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_poi=store_type shelf_set store_type*shelf_set/ dist=poisson link=log;
lsmeans store_type*shelf_set /  pdiff ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	163.4923	1.2021
Scaled Deviance	136	163.4923	1.2021
Pearson Chi-Square	136	161.2446	1.1856
Scaled Pearson X2	136	161.2446	1.1856
Log Likelihood		5957.6202	
Full Log Likelihood		-418.0156	
AIC (smaller is better)		844.0311	
AICC (smaller is better)		844.3274	
BIC (smaller is better)		855.7977	

Methods: Model 2 – Simple Poisson Regression accounting for subject heterogeneity

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	2.5150	0.0519	2.4132	2.6168	2346.67	<.0001
store_type	Big		1	0.6515	0.0612	0.5315	0.7715	113.22	<.0001
store_type	Sml		0	0.0000	0.0000	0.0000	0.0000	.	.
shelf_set	New		1	0.3453	0.0679	0.2123	0.4783	25.90	<.0001
shelf_set	Old		0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Big	New	1	-0.2489	0.0813	-0.4083	-0.0895	9.37	0.0022
store_type*shelf_set	Big	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	New	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

store_type*shelf_set Least Squares Means

store_type	shelf_set	Estimate	Standard Error	z Value	Pr > |z|	Mean	Standard Error of Mean
Big	New	3.2629	0.03093	105.48	<.0001	26.1250	0.8082
Big	Old	3.1665	0.03246	97.55	<.0001	23.7250	0.7701
Sml	New	2.8603	0.04369	65.48	<.0001	17.4667	0.7630
Sml	Old	2.5150	0.05192	48.44	<.0001	12.3667	0.6420

Methods: Model 3 – Response variable with inflated variance

- In Models 1 and 2, the response variable was generated by four Poisson distributions
- Model 3 examines a response variable with greater variance

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_inf=store_type shelf_set store_type*shelf_set/ dist=poisson link=log;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	259.0693	1.9049
Scaled Deviance	136	259.0693	1.9049
Pearson Chi-Square	136	243.9161	1.7935
Scaled Pearson X2	136	243.9161	1.7935
Log Likelihood		5693.7559	
Full Log Likelihood		-460.3821	
AIC (smaller is better)		928.7642	
AICC (smaller is better)		929.0605	
BIC (smaller is better)		940.5308	

Methods: Model 3 – Response variable with inflated variance

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	2.5284	0.0516	2.4273	2.6295	2403.68	<.0001
store_type	Big		1	0.5547	0.0617	0.4338	0.6756	80.85	<.0001
store_type	Sml		0	0.0000	0.0000	0.0000	0.0000	.	.
shelf_set	New		1	0.2316	0.0691	0.0963	0.3670	11.25	0.0008
shelf_set	Old		0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Big	New	1	-0.0225	0.0827	-0.1847	0.1396	0.07	0.7852
store_type*shelf_set	Big	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	New	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.0000	0.0000	1.0000	1.0000		

NOTE: The scale parameter was held fixed.

Methods: Model 4 – Response variable with inflated variance Scale=Deviance option

- One method to address the overdispersion in Model 3 is to use the Scale= option
  - This essentially scales the estimated variance “back up” to where it should be
  - Assumes roughly equal sample sizes

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_inf=store_type shelf_set store_type*shelf_set/ dist=poisson link=log
scale=deviance;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	259.0693	1.9049
Scaled Deviance	136	136.0000	1.0000
Pearson Chi-Square	136	243.9161	1.7935
Scaled Pearson X2	136	128.0453	0.9415
Log Likelihood		2988.9717	
Full Log Likelihood		-460.3821	
AIC (smaller is better)		928.7642	
AICC (smaller is better)		929.0605	
BIC (smaller is better)		940.5308	

Methods: Model 4 – Response variable with inflated variance Scale=Deviance option

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	2.5284	0.0712	2.3889	2.6679	1261.83	<.0001
store_type	Big		1	0.5547	0.0851	0.3878	0.7215	42.44	<.0001
store_type	Sml		0	0.0000	0.0000	0.0000	0.0000	.	.
shelf_set	New		1	0.2316	0.0953	0.0448	0.4184	5.90	0.0151
shelf_set	Old		0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Big	New	1	-0.0225	0.1142	-0.2463	0.2012	0.04	0.8435
store_type*shelf_set	Big	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	New	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale			0	1.3802	0.0000	1.3802	1.3802		

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

Methods: Model 5 – Negative Binomial Regression

- Another method to address the overdispersion of Model 3 is to use a distribution that estimates two parameters, such as the Negative Binomial distribution
  - Recall, NB has two parameters, k and μ, $E(Y)= \mu$ and $Var(Y)= \mu+\mu^2/k$
  - k^{-1} is the dispersion parameter
    - As $k^{-1} \rightarrow 0$ NB converges to the Poisson distribution
    - k^{-1} can then be used to quantify how much overdispersion was present in Poisson, yet captured in the Negative Binomial

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_inf=store_type shelf_set store_type*shelf_set/ dist=nb link=log;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	162.3178	1.1935
Scaled Deviance	136	162.3178	1.1935
Pearson Chi-Square	136	147.0568	1.0813
Scaled Pearson X2	136	147.0568	1.0813
Log Likelihood		5704.3629	
Full Log Likelihood		-449.7751	
AIC (smaller is better)		909.5502	
AICC (smaller is better)		909.9980	
BIC (smaller is better)		924.2584	

Methods: Model 5 – Negative Binomial Regression

- So, 0.0368 indicates at a predicted μ_hat, the estimated variance is μ_hat+0.0368 μ_hat^2, compared to μ_hat estimated via Poisson Regression

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept			1	2.5284	0.0623	2.4062	2.6506	1644.86	<.0001
store_type	Big		1	0.5547	0.0772	0.4035	0.7059	51.69	<.0001
store_type	Sml		0	0.0000	0.0000	0.0000	0.0000	.	.
shelf_set	New		1	0.2316	0.0850	0.0650	0.3982	7.43	0.0064
shelf_set	Old		0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Big	New	1	-0.0225	0.1055	-0.2294	0.1843	0.05	0.8308
store_type*shelf_set	Big	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	New	0	0.0000	0.0000	0.0000	0.0000	.	.
store_type*shelf_set	Sml	Old	0	0.0000	0.0000	0.0000	0.0000	.	.
Dispersion			1	0.0368	0.0115	0.0200	0.0679		

NOTE: The negative binomial dispersion parameter was estimated by maximum likelihood.

Methods: Model 6 – Zero-inflated data with Standard Poisson Regression

- Zero-inflated data arises when a structural event generates zeros in the response variable
  - Consider a dichotomous process
    - Consumer walks into retail outlet and decides to buy or not buy
    - Consumer then decides how many items to buy

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp=store_type shelf_set store_type*shelf_set/ dist=poisson link=log;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	948.1527	6.9717
Scaled Deviance	136	948.1527	6.9717
Pearson Chi-Square	136	605.4346	4.4517
Scaled Pearson X2	136	605.4346	4.4517
Log Likelihood		4646.1529	
Full Log Likelihood		-757.9261	
AIC (smaller is better)		1523.8523	
u			
BIC (smaller is better)		1535.6189	

Methods: Model 7 – Zero-inflated data with Standard Negative Binomial Regression

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp=store_type shelf_set store_type*shelf_set/ dist=nb link=log;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	136	185.0223	1.3605
Scaled Deviance	136	185.0223	1.3605
Pearson Chi-Square	136	52.6985	0.3875
Scaled Pearson X2	136	52.6985	0.3875
Log Likelihood		4869.1641	
Full Log Likelihood		-534.9149	
AIC (smaller is better)		1079.8299	
AICC (smaller is better)		1080.2776	
BIC (smaller is better)		1094.5381	

Methods: Model 8 – ZIP Model

- Since the data is generated by a structural process, why not consider the following?

$$\Pr(y_i = 0) = \pi_i + (1 - \pi_i)e^{-\lambda_i}$$

$$\Pr(y_i = h_i) = (1 - \pi_i)\frac{\lambda_i^{h_i} e^{-\lambda_i}}{h_i!}, h_i \geq 1$$

where the outcome variable y_i has any non-negative integer value; λ_i is the expected Poisson count for the i^{th} individual; π_i is the probability of extra zeros. [1]

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp=store_type shelf_set store_type*shelf_set/ dist=zip link=log;
zeromodel store_type shelf_set / link=logit;
lsmeans store_type*shelf_set / ilink;
run;
```

Methods: Model 8 – ZIP Model

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		829.1066	
Scaled Deviance		829.1066	
Pearson Chi-Square	133	145.6394	1.0950
Scaled Pearson X2	133	145.6394	1.0950
Log Likelihood		4989.5257	
Full Log Likelihood		-414.5533	
AIC (smaller is better)		843.1066	
AICC (smaller is better)		843.9551	
BIC (smaller is better)		863.6981	

store_ type	shelf_set	Estimate	Standard Error	z Value	Pr > |z|	Mean	Standard Error of Mean
Big	New	3.2592	0.03313	98.37	<.0001	26.0286	0.8624
Big	Old	3.1538	0.03597	87.68	<.0001	23.4242	0.8425
Sml	New	2.8731	0.04663	61.62	<.0001	17.6923	0.8249
Sml	Old	2.5357	0.05745	44.14	<.0001	12.6250	0.7253

Methods: Model 9 – ZINB Model

- Much like ZIP, but uses NB rather than Poisson

```
proc genmod data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp=store_type shelf_set store_type*shelf_set/ dist=zinb link=log;
zeromodel store_type shelf_set / link=logit;
lsmeans store_type*shelf_set / ilink;
run;
```

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance		828.0976	
Scaled Deviance		828.0976	
Pearson Chi-Square	133	140.9873	1.0601
Scaled Pearson X2	133	140.9873	1.0601
Log Likelihood		-414.0488	
Full Log Likelihood		-414.0488	
AIC (smaller is better)		844.0976	
AICC (smaller is better)		845.1968	
BIC (smaller is better)		867.6307	

store_ type	shelf_set	Estimate	Standard Error	z Value	Pr > |z|	Mean	Error of Mean
Big	New	3.2592	0.03592	90.74	<.0001	26.0286	0.9349
Big	Old	3.1538	0.03870	81.49	<.0001	23.4242	0.9066
Sml	New	2.8731	0.04933	58.25	<.0001	17.6923	0.8727
Sml	Old	2.5357	0.05984	42.37	<.0001	12.6249	0.7555

Methods: Model 9-2 – ZINB Model, using PROC FMM

- ZIP and ZINB are special cases of mixture models
  - Proc FMM can be used to model mixture models

```
proc fmm data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp = store_type shelf_set store_type*shelf_set / dist=nb;
model + / dist=constant;
run;
```

Fit Statistics

-2 Log Likelihood	829.0
AIC (smaller is better)	841.0
AICC (smaller is better)	841.7
BIC (smaller is better)	858.7
Pearson Statistic	141.1
Effective Parameters	6
Effective Components	2

Parameter Estimates for Mixing Probabilities

	----------------Linked Scale---------------				
Effect	Estimate	Standard Error	z Value	Pr > |z|	Probability
Intercept	1.6796	0.2322	7.23	<.0001	0.8429

In our simulated data, about 15.7% were zeros

Methods: Model 10 – Poisson Hurdle Model

- Hurdle model uses a mixture of two distributions, like ZIP and ZINB
  - However, Hurdle model uses truncated Poisson rather than Poisson
  - Effectively, Hurdle model assumes all of the zeros are structural
    - Eg. Two groups of people, "purchasers" and "non-purchasers"
  - ZIP assumes some zeros are from the Poisson process
    - Eg. Two groups of people, "non-purchasers" and "may purchase"

```
proc fmm data=dd1.poisson_data;
class store_type shelf_set;
model n_people_zp = store_type shelf_set store_type*shelf_set / dist=tpoisson;
model + / dist=constant;
run;
```

Fit Statistics

-2 Log Likelihood	830.0
AIC (smaller is better)	840.0
AICC (smaller is better)	840.5
BIC (smaller is better)	854.8
Pearson Statistic	145.6
Effective Parameters	5
Effective Components	2

Parameter Estimates for Mixing Probabilities

	----------------Linked Scale---------------				
Effect	Estimate	Standard Error	z Value	Pr > |z|	Probability
Intercept	1.6796	0.2322	7.23	<.0001	0.8429

Methods: Model 11(a) – Determining how many Poissons to mix

- Consider Model 1, what if we know very little about the independent variables
- Perhaps Model 1 can be considered using a mix of Poisson distributions?
  - First, how many should be used?

```
proc fmm data=dd1.poisson_data criterion=PEARSON;
       class shelf_set;
    model n_people_poi = shelf_set/ dist=poisson kmin=1 kmax=7;
run;
```

The 1-component model is Simple Poisson Regression

Component Evaluation for Mixture Models

Model ID	Number of Components Total	Number of Components Eff.	Number of Parameters Total	Number of Parameters Eff.	-2 Log L	AIC	AICC	BIC	Pearson	Max Gradient
1	1	1	2	2	1017.64	1021.64	1021.73	1027.53	338.00	0.00047
2	2	2	5	5	931.55	941.55	942.00	956.26	139.49	0.00082
3	3	3	8	8	926.29	942.29	943.39	965.82	136.26	0.00178
4	4	4	11	11	924.96	946.96	949.02	979.32	134.21	0.00619
5	5	5	14	14	924.96	952.96	956.32	994.14	134.21	0.00029
6	6	6	17	17	924.96	958.96	963.97	1008.97	134.15	0.00947
7	7	7	20	20	924.96	964.96	972.02	1023.79	134.21	0.00547

Methods: Model 11(b) – A Mixture of Two Poisson Distributions

```
proc fmm data=dd1.poisson_data criterion=PEARSON;
        class shelf_set;
    model n_people_poi = shelf_set/ dist=poisson k=2;
run;
```

Fit Statistics

-2 Log Likelihood	931.6
AIC (smaller is better)	941.6
AICC (smaller is better)	942.0
BIC (smaller is better)	956.3
Pearson Statistic	139.5
Effective Parameters	5
Effective Components	2

Parameter Estimates for 'Poisson' Model

Component	Effect	shelf_set	Estimate	Standard Error	z Value	Pr > |z|
1	Intercept		3.2541	0.05234	62.18	<.0001
1	shelf_set	New	0.02272	0.06003	0.38	0.7051
1	shelf_set	Old	0	.	.	.
2	Intercept		2.5973	0.05728	45.34	<.0001
2	shelf_set	New	0.3002	0.08008	3.75	0.0002
2	shelf_set	Old	0	.	.	.

Parameter Estimates for Mixing Probabilities

	----------------Linked Scale---------------				
Effect	Estimate	Standard Error	z Value	Pr > |z|	Probability
Intercept	-0.1042	0.3188	-0.33	0.7437	0.4740

The mixing probability is reasonable considering the "unknown" independent variable is divided 57%-43% across its levels

Methods: Proc Countreg

- Proc Countreg can also be used to perform count data analyses
- Consider Model 2

```
proc countreg data=dd1.poisson_data;
   class store_type shelf_set;
   model n_people_poi=store_type shelf_set store_type*shelf_set / dist=poisson;
run;
```

Parameter Estimates

<table>
<tr><th>Parameter</th><th>DF</th><th>Estimate</th><th>Standard Error</th><th>t Value</th><th>Approx Pr > |t|</th></tr>
<tr><td>Intercept</td><td>1</td><td>2.515005</td><td>0.051917</td><td>48.44</td><td><.0001</td></tr>
<tr><td>store_type Big</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td></tr>
<tr><td>store_type Sml</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td></tr>
<tr><td>shelf_set New</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td></tr>
<tr><td>shelf_set Old</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td></tr>
<tr><td>store_type*shelf_set Big New</td><td>1</td><td>0.747888</td><td>0.060435</td><td>12.38</td><td><.0001</td></tr>
<tr><td>store_type*shelf_set Big Old</td><td>1</td><td>0.651525</td><td>0.061230</td><td>10.64</td><td><.0001</td></tr>
<tr><td>store_type*shelf_set Sml New</td><td>1</td><td>0.345290</td><td>0.067851</td><td>5.09</td><td><.0001</td></tr>
<tr><td>store_type*shelf_set Sml Old</td><td>0</td><td>0</td><td>.</td><td>.</td><td>.</td></tr>
</table>

Conclusions and Notes

- Overdispersion has been a focus of this paper
  - Methods such as plotting studentized residuals vs. predicated should also be considered when evaluating model fit

- For n_people_poi, the author prefers Model 2, which addresses overdispersion and correctly specifies the independent variables

- For n_people_inf, the author prefers Model 5, which uses the Negative Binomial distribution to allow a second parameter to estimate variance. While Model 4 was useful, Model 5 had superior AICC and BIC

- For n_people_zf, the author would utilize either Model 8, the ZIP model in most cases. There was no demonstrable need to use a ZINB model and utilize an extra parameter. From a business perspective, this author believes the ZIP assumptions might better fit consumer behavior relative to the Hurdle assumptions. However, if there was reason to believe the Hurdle assumptions were better from a business perspective, it would be the proper model to choose

Review

- Count data

- Methods

- Conclusions

Author Info

George J. Hurley
The Hershey Company
19 E Chocolate Ave.
Hershey, PA 17033

ghurley@hersheys.com