

Session #1278

Analysis of Data with Overdispersion using the SAS® System

Jorge G. Morel, Ph.D.

Principal Statistician

The Procter and Gamble Company

Overdispersion: To be or not to be

- Overdispersion is also known as Extra Variation
- Arises when Binary/Multinomial/Count data exhibit variances larger than those permitted by the Binomial/Multinomial/Poisson model
- Usually caused by clustering or lack of independence
- It might be also caused by a model misspecification

“In fact, some would maintain that over-dispersion is the norm in practice and nominal dispersion the exception.”

McCullagh and Nelder (1989, Pages 124-125)

Overdispersion: To be or not to be

Consequences of ignoring overdispersion:

- Standard errors of Naïve estimates are smaller than they should be.
- It results in inflated Type I Error Rates, i.e., False Positive Rates are larger than nominal ones.
- Furthermore, coverage probabilities of confidence intervals are lower than nominal levels.
- **Erroneous inferences !!!**

Overdispersion: To be or not to be.

Let us illustrate the consequences of ignoring overdispersion

- In a simulation 1000 datasets were generated
- Each dataset with $n=20$ subjects.
- Each subject with 5 correlated Bernoulli outcomes
- Probability of success $\pi=0.6$ and **Correlation=0.3, 0.5.**

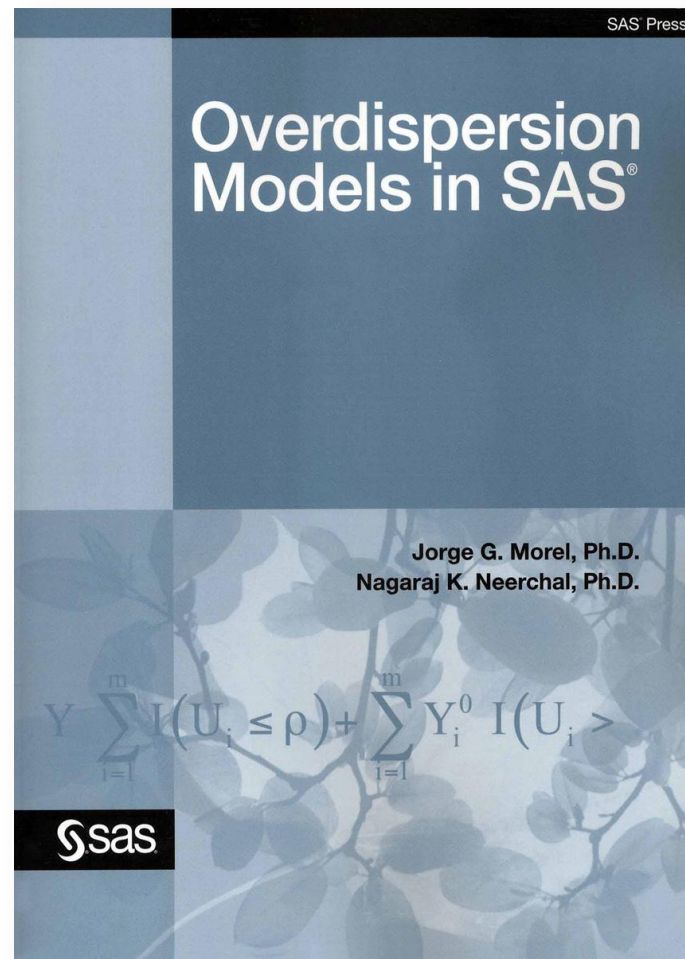
To test $H_0: \pi=0.6$

Actual Type I Error Rate at Nominal Level $\alpha = 0.05$

Correlation	Actual Type I Error Rate	Increase False Positive Rates (%)
0.3	0.160	220
0.5	0.197	294

Overdispersion: To be or not to be

**TAKEAWAY: Addressing
Overdispersion properly
is important**



Outline

- 1) Characterization of Overdispersion, Quasi-likelihoods and GEE Models
- 2) All Mice Are Created Equal, but Some Are More Equal
- 3) Overdispersion Models for Binomial of Data
- 4) All Mice Are Created Equal Revisited
- 5) Overdispersion Models for Count Data
- 6) Milk Does Your Body Good
- 7) Generalized Linear Mixed Overdispersion Models (GLOMM)
- 8) Final Remarks

Characterization of Overdispersion

- Usually characterized in terms of the first two moments

- Binomial distribution:**

$$E(T) = m\pi$$

$$\text{Var}(T) = m\pi(1-\pi)\{1 + (m-1)\rho^2\}$$

- Poisson distribution:**

$$E(Y) = \lambda$$

$$\text{Var}(Y) = \lambda(1 + \lambda\kappa) = \lambda + \lambda^2\kappa$$

$$\text{Var}(Y) = \lambda(1 + \kappa) = \lambda + \lambda\kappa$$

Quasi-likelihood and GEE Models

- **Quasi-likelihood Estimates (QLEs)** are based on the specification of the first two moments of the data.
- A true likelihood function may not exist.
- QLEs are obtained from a function known as quasi likelihood, (Wedderburn, 1974), which incorporates the relation between the mean and the variance of the observations that resembles a true likelihood function.
- These functions allow to analyze overdispersed data without full knowledge of the likelihood function.

Quasi-likelihood and GEE Models

It can be viewed as a two-step procedure: estimate first the systematic structure of the model via maximum likelihood procedures, and then inflate or deflate the variances by a suitable constant.
(see Cox and Snell, 1989)

Binomial-type of data: $E(T) = m\pi$

$$\text{Var}(T) = \phi m\pi(1 - \pi)$$

Count-type of data: $E(Y) = \lambda$

$$\text{Var}(Y) = \phi\lambda$$

The DEVIANCE/DF and the (PEARSON'S GOF)/DF are moment estimators of ϕ

Quasi-likelihood and GEE Models

- **Generalized Estimating Equations (GEE)** is a methodology adapted by Liang and Zeger (1986), and Zeger and Liang (1986) for Generalized Linear Models in a longitudinal setting.
- The methodology is related to Quasi-likelihood models and Generalized Linear Models [McCullagh and Nelder (1983)].
- GEE methodology was originally proposed for longitudinal data analysis. it has been applied to a wide variety of models for clustered data.

Quasi-likelihood and GEE Models

Three essential elements of GEE:

- a) Correct specification of the mean of the model via a Link Function
- b) A proposed structure of the correlation of the measurements that encompass the clusters (subjects in longitudinal studies) referred to as the Working Correlation Matrix
- c) An application of the Taylor series expansion to obtain the variance-covariance matrix of the estimates, usually known as Robust, Empirical, or Sandwich.

Example: All Mice Are Created Equal, but Some Are More Equal

Morel and Neerchal (1997), PROC FMM Documentation

Two-way factorial design with $n=81$ pregnant mice

- Purpose: to investigate synergistic effect of the anticonvulsant phenytoin (PHT) and trichloropropene oxide (TCPO) on the prenatal development of inbred mice
- Presence or absence of ossification at the phalanges of the forepaws is considered a measure of teratogenic effect
- Outcome: For simplicity presence or absence of ossification at the left middle third phalanx

Example: All Mice Are Created Equal, but Some Are More Equal

Ossification Data*

Group	Observations
Control	8/8, 9/9, 7/9, 0/5, 3/3, 5/8, 9/10, 5/8, 5/8, 1/6, 0/5, 8/8, 9/10, 5/5, 4/7, 9/10, 6/6, 3/5
Sham	8/9, 7/10, 10/10, 1/6, 6/6, 1/9, 8/9, 6/7, 5/5, 7/9, 2/5, 5/6, 2/8, 1/8, 0/2, 7/8, 5/7
PHT	1/9, 4/9, 3/7, 4/7, 0/7, 0/4, 1/8, 1/7, 2/7, 2/8, 1/7, 0/2, 3/10, 3/7, 2/7, 0/8, 0/8, 1/10, 1/1
TCPO	0/5, 7/10, 4/4, 8/11, 6/10, 6/9, 3/4, 2/8, 0/6, 0/9, 3/6, 2/9, 7/9, 1/10, 8/8, 6/9
PHT+TCPO	2/2, 0/7, 1/8, 7/8, 0/10, 0/4, 0/6, 0/7, 6/6, 1/6, 1/7

*Number of fetuses showing ossification / litter size.

PHT: phenytoin; TCPO: trichloropropene oxide.

- The two controls were combined. The experiment can be seen as a 2 x 2 factorial, with PHT and TCPO as the two factors.
- The levels of PHT are 60 mg/kg and 0 mg/kg, and the levels of TCPO are 100 mg/kg and 0 mg/kg.

Example: All Mice Are Created Equal, but Some Are More Equal

```
data ossi;
  length tx $8;
  input tx$ n @@;
  PHT = 0;
  TCPO = 0;
  do i=1 to n;
    Litters + 1;
    input t m @@;
    if tx = 'PHT' then PHT = 1;
    if tx = 'TCPO' then TCPO = 1;
    if tx = 'PHT+TCPO' then do;
      PHT = 1;
      TCPO = 1;
    end;
    output;
  end;
  drop n i;
  datalines;
```

Control	18	8	8	9	9	7	9	0	5	3	3	5	8	9	10	5	8	5	8	1	6	0	5
		8	8	9	10	5	5	4	7	9	10	6	6	3	5								
Control	17	8	9	7	10	10	10	1	6	6	6	1	9	8	9	6	7	5	5	7	9		
		2	5	5	6	2	8	1	8	0	2	7	8	5	7								
PHT	19	1	9	4	9	3	7	4	7	0	7	0	4	1	8	1	7	2	7	2	8	1	7
		0	2	3	10	3	7	2	7	0	8	0	8	1	10	1	1						
TCPO	16	0	5	7	10	4	4	8	11	6	10	6	9	3	4	2	8	0	6	0	9		
		3	6	2	9	7	9	1	10	8	8	6	9										
PHT+TCPO	11	2	2	0	7	1	8	7	8	0	10	0	4	0	6	0	7	6	6	1	6	1	7

```
;
```

Example: All Mice Are Created Equal, but Some Are More Equal

```
title "*** Quasi-likelihood Model using PROC LOGISTIC ***";
proc logistic data=ossi;
    model t/m = tcpo pht tcpo*pht / link=logit scale=d;
run;

title "*** Quasi-likelihood Model using PROC GENMOD ***";
proc genmod data=ossi;
    model t/m = tcpo pht tcpo*pht / dist=bin link=logit scale=p;
run;

title "*** Quasi-likelihood Model using PROC GLIMMIX ***";
proc glimmix data=ossi;
    class litters;
    model t/m = tcpo pht tcpo*pht / dist=bin link=logit s;
    random _residual_ / subject=litters;
run;
```

Example: All Mice Are Created Equal, but Some Are More Equal

*** Quasi-likelihood Model using PROC GLIMMIX ***

Fit Statistics	
-2 Res Log Pseudo-Likelihood	296.80
Generalized Chi-Square	252.08
Gener. Chi-Square / DF	3.27

Covariance Parameter Estimates		
Cov Parm	Estimate	Standard Error
Residual (VC)	3.2738	0.5276

Solutions for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	0.8323	0.2470	77	3.37	0.0012
TCPO	-0.8481	0.4051	77	-2.09	0.0396
PHT	-2.1094	0.4532	77	-4.65	<.0001
TCPO*PHT	1.0453	0.7431	77	1.41	0.1635

Example: All Mice Are Created Equal, but Some Are More Equal

```
title "*** GEE using PROC GENMOD ***";
proc genmod data=ossi;
  class litters tcpo pht / param=ref desc;
  model t/m = tcpo pht tcpo*pht / dist=binomial link=logit;
  repeated subject=litters / type=ind;
run;

title "*** GEE using PROC SURVEYLOGISTIC ***";
proc surveylogistic data=ossi;
  class litters tcpo pht / param=ref desc;
  model t/m = tcpo pht tcpo*pht / link=logit vadjust=morel;
  clusters litters;
run;

title "*** GEE using PROC GLIMMIX ***";
proc glimmix data=ossi empirical=mbn;
  class litters tcpo pht ;
  model t/m = tcpo pht tcpo*pht / dist=bin link=logit s;
  random _residual_ / subject=litters;
run;
```

Example: All Mice Are Created Equal, but Some Are More Equal

Beta Estimates and Standard Errors Ossification Data under Three Models				
		Standard Errors		
Parameter	Estimate	Naïve	Quasi-Likelihood	GEE
Intercept	0.8323	0.1365	0.2470	0.2472
TCPO	-0.8481	0.2239	0.4051	0.4063
PHT	-2.1094	0.2505	0.4532	0.3434
TCPO + PHT	1.0453	0.4107	0.7431	0.7628

TAKEAWAY: Naïve Standard Errors should be at least 1.7 times larger !!!

Overdispersion Models for Binomial of Data

Maximum-likelihood Estimation

- Some Distributions to Model Binomial Data with Overdispersion:
 - Beta-binomial
 - Random-clumped Binomial
 - Zero-inflated Binomial
 - Generalized Binomial

Overdispersion Models for Binomial Data: The Beta-binomial Distribution

Skellam(1948)

$$T | P \sim \text{Binomial}(P; m)$$

$$P\text{'s are i.i.d.} \sim \text{Beta}(a, b)$$

$$a = C\pi, \quad b = C(1 - \pi), \quad C = (1 - \rho^2) / \rho^2$$

Then the unconditional distribution of T is Beta-binomial

$$1) \quad E(T) = m\pi$$

$$2) \quad \text{Var}(T) = m\pi(1 - \pi) \{1 + (m - 1)\rho^2\} \text{ or,}$$

$$\text{Var}(T) = m\pi(1 - \pi) \{1 + (m - 1)/(C + 1)\}$$

Overdispersion Models for Binomial Data: The Random-clumped Binomial Distribution (aka Binomial Cluster)

(Morel and Nagaraj, 1993; Morel and Neerchal, 1997; Neerchal and Morel, 1998)

Results from an effort to model meaningfully the physical mechanism behind the overdispersion

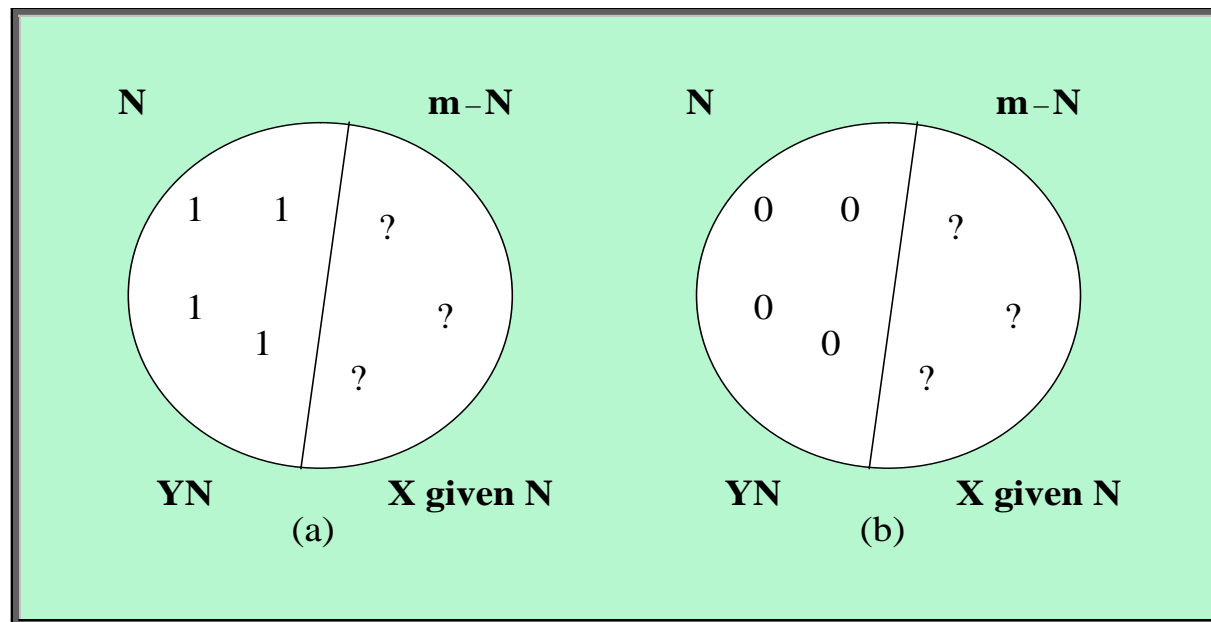
$$\mathbf{T} = \mathbf{Y}\mathbf{N} + (\mathbf{X} | \mathbf{N})$$

where $Y \sim \text{Bernoulli}(\pi)$

$N \sim \text{Binomial}(\rho; m)$, Y and N independent

$X | N \sim \text{Binomial}(\pi; m - N)$ if $N < m$

Overdispersion Models for Binomial Data: The Random-clumped Binomial Distribution (aka Binomial Cluster)



YN might characterize the influence of a “leader” in stop-smoking or stop-drinking programs, or a genetic trait which is passed on with a certain probability to offspring of the same mother

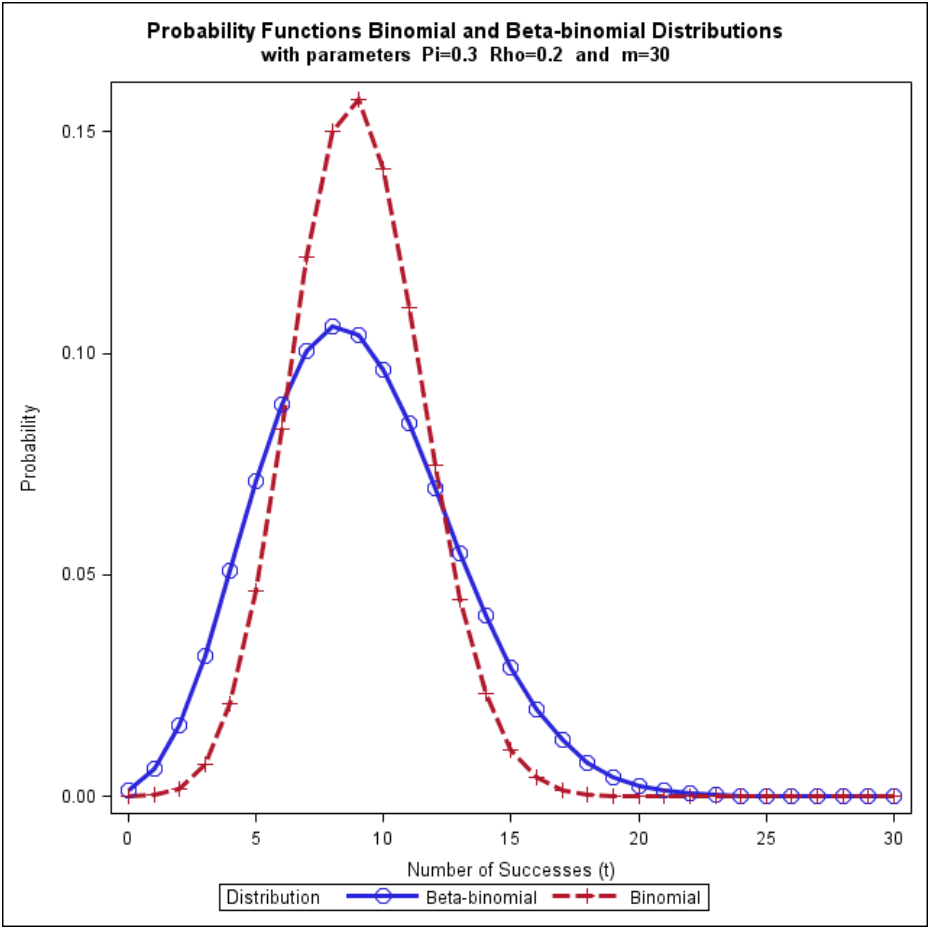
Overdispersion Models for Binomial Data: The Random-clumped Binomial Distribution (aka Binomial Cluster)

1) $\text{Pr ob}(T = t) = \pi \text{Pr}(X_1 = t) + (1 - \pi) \text{Pr}(X_2 = t), \quad t = 0, 1, \dots, m,$
where $X_1 \sim \text{Binomial}\{(1 - \rho)\pi + \rho; m\}, \quad X_2 \sim \text{Binomial}\{(1 - \rho)\pi; m\}$

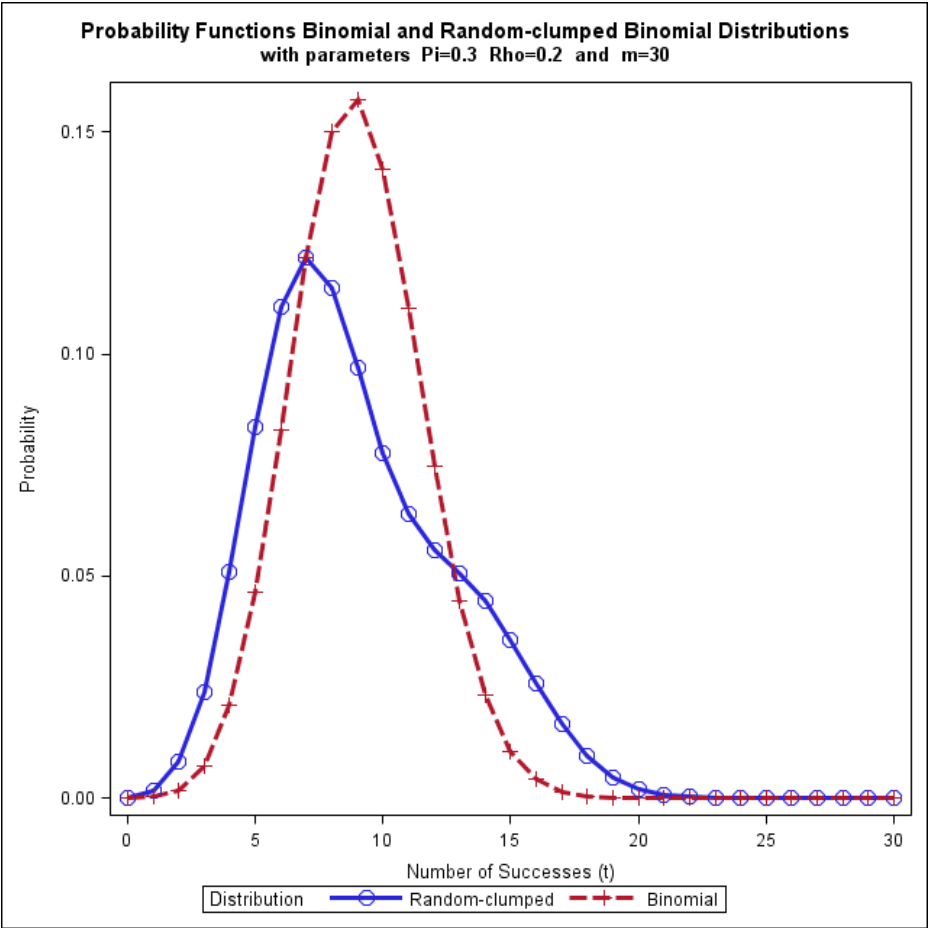
2) $E(T) = m\pi$

3) $\text{Var}(T) = m\pi(1 - \pi)\{1 + \rho^2(m - 1)\}$

Overdispersion Models for Binomial Data



Overdispersion Models for Binomial Data



All Mice Are Created Equal Revisited

Morel and Neerchal (1997), PROC FMM Documentation

Two-way factorial design with $n=81$ pregnant mice

- Purpose: to investigate synergistic effect of the anticonvulsant phenytoin (PHT) and trichloropropene oxide (TCPO) on the prenatal development of inbred mice
- Presence or absence of ossification at the phalanges of the forepaws is considered a measure of teratogenic effect
- Outcome: For simplicity presence or absence of ossification at the left middle third phalanx

All Mice Are Created Equal Revisited

Ossification Data*

Group	Observations
Control	8/8, 9/9, 7/9, 0/5, 3/3, 5/8, 9/10, 5/8, 5/8, 1/6, 0/5, 8/8, 9/10, 5/5, 4/7, 9/10, 6/6, 3/5
Sham	8/9, 7/10, 10/10, 1/6, 6/6, 1/9, 8/9, 6/7, 5/5, 7/9, 2/5, 5/6, 2/8, 1/8, 0/2, 7/8, 5/7
PHT	1/9, 4/9, 3/7, 4/7, 0/7, 0/4, 1/8, 1/7, 2/7, 2/8, 1/7, 0/2, 3/10, 3/7, 2/7, 0/8, 0/8, 1/10, 1/1
TCPO	0/5, 7/10, 4/4, 8/11, 6/10, 6/9, 3/4, 2/8, 0/6, 0/9, 3/6, 2/9, 7/9, 1/10, 8/8, 6/9
PHT+TCPO	2/2, 0/7, 1/8, 7/8, 0/10, 0/4, 0/6, 0/7, 6/6, 1/6, 1/7

*Number of fetuses showing ossification / litter size.

PHT: phenytoin; TCPO: trichloropropene oxide.

- The two controls were combined. The experiment can be seen as a 2 x 2 factorial, with PHT and TCPO as the two factors.
- The levels of PHT are 60 mg/kg and 0 mg/kg, and the levels of TCPO are 100 mg/kg and 0 mg/kg.

All Mice Are Created Equal Revisited

```

data ossi;
  length tx $8;
  input tx$ n @@;
  do i=1 to n;
    input t m @@;
    output;
  end;
  drop n i;
  datalines;
Control 18 8 8 9 9 7 9 0 5 3 3 5 8 9 10 5 8 5 8 1 6 0 5
          8 8 9 10 5 5 4 7 9 10 6 6 3 5
Control 17 8 9 7 10 10 10 1 6 6 6 1 9 8 9 6 7 5 5 7 9
          2 5 5 6 2 8 1 8 0 2 7 8 5 7
PHT      19 1 9 4 9 3 7 4 7 0 7 0 4 1 8 1 7 2 7 2 8 1 7
          0 2 3 10 3 7 2 7 0 8 0 8 1 10 1 1
TCPO     16 0 5 7 10 4 4 8 11 6 10 6 9 3 4 2 8 0 6 0 9
          3 6 2 9 7 9 1 10 8 8 6 9
PHT+TCPO 11 2 2 0 7 1 8 7 8 0 10 0 4 0 6 0 7 6 6 1 6 1 7
;
data ossi;
  set ossi;
  array xx{3} x1-x3;
  do i=1 to 3; xx{i}=0; end;
  pht = 0;
  tcpo = 0;
  if (tx='TCPO') then do;
    xx{1} = 1;
    tcpo = 100;
  end; else if (tx='PHT') then do;
    xx{2} = 1;
    pht = 60;
  end; else if (tx='PHT+TCPO') then do;
    pht = 60;
    tcpo = 100;
    xx{1} = 1; xx{2} = 1; xx{3}=1;
  end;
run;

```

All Mice Are Created Equal Revisited

```
title "Fitting a Beta-binomial in PROC NLMIXED";
proc nlmixed data=ossi;
  parms b0=0, b1=0, b2=0, b3=0, a0=0;
  linr = a0;
  linp = b0 + b1*x1 + b2*x2 + b3*x3;
  rho  = 1/(1+exp(-linr));
  c    = 1 / rho / rho - 1;
  pi   = 1/(1+exp(-linp));
  pic  = 1 - pi;
  z    = lgamma(m+1) - lgamma(t+1) - lgamma(m-t+1);
  ll   = z + lgamma(c) + lgamma(t+c*pi) + lgamma(m-t+c*pic)
        - lgamma(m+c) - lgamma(c*pi) - lgamma(c*pic);
  model t ~ general(ll);
  estimate 'Pi Control'      1/(1+exp(-b0));
  estimate 'Pi PHT+TCPO'    1/(1+exp(-b0-b1-b2-b3));
  estimate 'C' (1+exp(-a0))*(1+exp(-a0))-1;
run;
title;
```

Fitting a Beta-binomial in PROC NLMIXED

The NLMIXED Procedure

Specifications

Data Set	WORK.OSSI
Dependent Variable	t
Distribution for Dependent Variable	General
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Fit Statistics

-2 Log Likelihood	306.6
AIC (smaller is better)	316.6
AICC (smaller is better)	317.4
BIC (smaller is better)	328.5

Parameter Estimates

Parameter	Estimate	Standard Error	Pr > t	Lower	Upper
b0	0.7043	0.2341	0.0035	0.2385	1.1700
b1	-0.7822	0.4017	0.0549	-1.5814	0.01696
b2	-1.6917	0.4018	<.0001	-2.4912	-0.8923
b3	0.6770	0.6902	0.3296	-0.6963	2.0502
a0	0.3262	0.1915	0.0923	-0.05476	0.7071

Additional Estimates

Label	Estimate	Standard Error
Pi Control	0.6691	0.05182
Pi PHT+TCPO	0.2511	0.08831
C	1.9642	0.4758

All Mice Are Created Equal Revisited

```

title "Fitting a Beta-binomial in PROC FMM";
proc fmm data=ossi;
    model t/m = x1-x3 / dist=betabinomial;
run;

proc fmm data=ossi;
    class tcpo pht;
    model t/m = tcpo pht tcpo*pht / dist=betabinomial;
run;

```

Fitting a Beta-binomial in PROC FMM

The FMM Procedure

Model Information

Data Set	WORK.OSSI
Response Variable (Events)	t
Response Variable (Trials)	m
Type of Model	Homogeneous Regression Mixture
Distribution	Beta-Binomial
Components	1
Link Function	Logit
Estimation Method	Maximum Likelihood

Fit Statistics

-2 Log Likelihood	306.6
AIC (smaller is better)	316.6
AICC (smaller is better)	317.4
BIC (smaller is better)	328.5
Pearson Statistic	87.5379

Parameter Estimates for 'Beta-Binomial' Model

Effect	Estimate	Standard Error	z Value	Pr > z
Intercept	0.7043	0.2341	3.01	0.0026
x1	-0.7822	0.4017	-1.95	0.0515
x2	-1.6917	0.4018	-4.21	<.0001
x3	0.6770	0.6902	0.98	0.3267
Scale Parameter	1.9642	0.4758		

All Mice Are Created Equal Revisited

```
title "Fitting a Random-clumped Binomial in PROC FMM";
proc fmm data=ossi;
  model t/m = / dist=binomcluster;
  probmodel x1-x3;
run;

proc fmm data=ossi;
  class tcpo pht;
  model t/m = / dist=binomcluster;
  probmodel tcpo pht tcpo*pht;
run;
```

Fitting a Random-clumped Binomial in PROC FMM

The FMM Procedure

Model Information

Data Set	WORK.OSSI
Response Variable (Events)	t
Response Variable (Trials)	m
Type of Model	Binomial Cluster
Distribution	Binomial Cluster
Components	2
Link Function	Logit
Estimation Method	Maximum Likelihood

Fit Statistics

-2 Log Likelihood	305.1
AIC (smaller is better)	315.1
AICC (smaller is better)	315.9
BIC (smaller is better)	327.0
Pearson Statistic	89.2077
Effective Parameters	5
Effective Components	2

Parameter Estimates for 'Binomial Cluster' Model

Component	Effect	Estimate	Standard Error	z Value	Pr > z	Inverse Linked Estimate
1	Intercept	0.3356	0.1714	1.96	0.0503	0.5831

Parameter Estimates for Mixing Probabilities

Effect	Estimate	Standard Error	z Value	Pr > z
Intercept	0.6392	0.2266	2.82	0.0048
x1	-0.9457	0.3711	-2.55	0.0108
x2	-1.5291	0.3956	-3.87	0.0001
x3	0.6162	0.6678	0.92	0.3561

All Mice Are Created Equal Revisited

Parameter	Distribution					
	Binomial (Naïve)		Beta-binomial		Binomial Cluster	
	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error
Intercept ($\hat{\beta}_0$)	0.8323	0.1365	0.7043	0.2341	0.6392	0.2266
TCPO ($\hat{\beta}_1$)	-0.8481	0.2239	-0.7822	0.4017	-0.9457	0.3711
PHT ($\hat{\beta}_2$)	-2.1094	0.2505	-1.6917	0.4018	-1.5291	0.3956
TCPO + PHT ($\hat{\beta}_3$)	1.0453	0.4107	0.6770	0.6902	0.6162	0.6678
Overdispersion [\hat{C} or $\hat{\alpha} = \text{logit}(\hat{\rho})$]	--	--	1.9642	0.4758	0.3356	0.1714
- 2 * Log Likelihood	401.8	--	306.6	--	305.1	--

TAKEAWAY: Beta-binomial and Binomial Cluster now available in PROC FMM

All Mice Are Created Equal Revisited

```
title "Fitting a Zero-inflated Binomial in PROC FMM";  
proc fmm data=ossi;  
    model t/m = x1-x3 / dist=binomial;  
    model      +      / dist=Constant;  
run;
```

```
title "Fitting an Arbitrary Mixture of Two Binomials in PROC FMM";  
proc fmm data=ossi;  
    model t/m = x1-x3 / k=2;  
run;  
*--- Interpretation might be difficult!!!;
```

All Mice Are Created Equal Revisited

Parameter Estimates for 'Binomial' Model					
Component	Effect	Estimate	Standard Error	z Value	Pr > z
1	Intercept	1.6876	0.2049	8.23	<.0001
	x1	-0.7364	0.3324	-2.22	0.0267
	x2	-2.5593	0.3644	-7.02	<.0001
	x3	4.3154	1.1270	3.83	0.0001
2	Intercept	-1.6757	0.4668	-3.59	0.0003
	x1	-0.4363	0.6838	-0.64	0.5234
	x2	-0.6293	0.9055	-0.70	0.4870
	x3	-0.1100	1.1947	-0.09	0.9267

Parameter Estimates for Mixing Probabilities					
Effect	Linked Scale				Probability
	Estimate	Standard Error	z Value	Pr > z	
Intercept	0.5289	0.2690	1.97	0.0493	0.6292

All Mice Are Created Equal Revisited (GOF)

Omnibus tests are designed to test if a specific distribution fits the data well. (The Shapiro-Wilk test of normality is an example)

- **Neerchal and Morel (1998)** proposed an extension of the traditional Pearson's Chi-square statistic

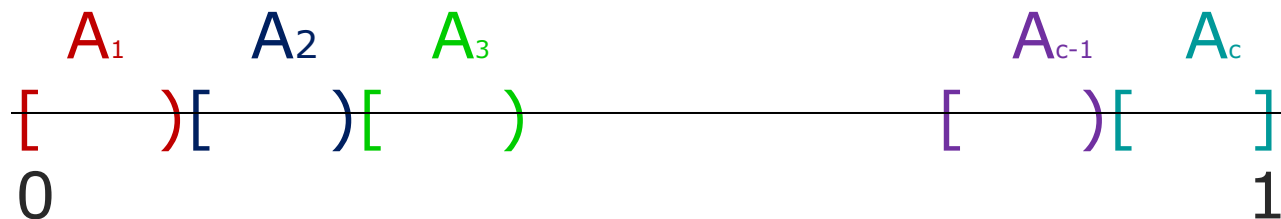
$$X^2 = \sum_{s=1}^c (O_s - E_s)^2 / E_s$$

when the clusters sizes are allowed to be different and/or covariates are present in the model

- Asymptotic properties of this test have been investigated by **Sutradhar et al. (2008)**.

All Mice Are Created Equal Revisited (GOF)

Divide the $[0,1]$ interval into C mutually exclusive intervals:



Compute $\frac{t_j}{m_j}$ for $j = 1, 2, \dots, n$

Then get

O_s : Observed number of $\frac{t_j}{m_j}$'s in the s^{th} interval, $s = 1, 2, \dots, c$

E_s : Expected number of $\frac{t_j}{m_j}$'s in the s^{th} interval, $s = 1, 2, \dots, c$

All Mice Are Created Equal Revisited (GOF)

Properties of the GOF Test:

- 1) $GOF \sim \chi^2_{df}$
- 2) Degrees of freedom (df) of GOF is between:
 $C - 1 - (\text{Number of Parameters Estimated in the Model})$ and $C - 1$
(see chapter 30 of Kendall, Stuart, and Ord, 1991)
- 3) If needed, underlying DF and P-value can be obtained via Parametric Bootstrapping
- 4) GOF is also applicable when cluster sizes are not the same
and/or covariates are present

All Mice Are Created Equal Revisited (GOF)

	Results Omnibus Goodness-of-fit Tests			
Distribution	GOF-Stat	Degrees of Freedom		P-Value
Binomial	56.94	Lower Bound	4	< 0.01
		Upper Bound	8	< 0.01
Beta-binomial	9.79	Lower Bound	3	0.02
		Upper Bound	8	0.28
Binomial Cluster	6.81	Lower Bound	3	0.08
		Upper Bound	8	0.56

All Mice Are Created Equal Revisited (GOF)

Parametric Bootstrapping Results Based on 5,000 Replications		
Distribution	Parameter	Estimate
Beta-binomial	Degrees of Freedom	5.83
	P-value	0.11
Random-clumped Binomial	Degrees of Freedom	5.79
	P-value	0.31

- a) Both distributions fit the data, however, the RCB seems to provide a better fit than the BB
- b) Since in this example the RCB provides a clear mechanism on how the offspring inherit the genetic trait, I prefer the RCB over the BB

All Mice Are Created Equal Revisited (GOF)

TAKEAWAY: An Omnibus Goodness-of-test is available. See Morel and Neerchal (2012) “Overdispersion Models in SAS®”

It can be applied to Binomial, Beta-binomial, Random-clumped Binomial (aka Binomial Cluster), Zero-inflated Binomial Distributions, and to arbitrary mixtures of Binomials

Overdispersion Models for Count Data

Maximum-likelihood Estimation

- Some Distributions to Model Count Data with Overdispersion:
 - Negative-binomial
 - Zero-inflated Poisson
 - Zero-inflated Negative-binomial
 - Hurdle Poisson
 - Hurdle Negative-binomial
 - Generalized Poisson

Overdispersion Models for Count Data: The Negative-binomial

- Count data can exhibit overdispersion.
- It occurs when the variance is greater than the mean.
- The Negative-binomial distribution is a benchmark model to account for overdispersion on count responses.

Overdispersion Models for Count Data: The Negative-binomial

$$Y \mid u \sim \text{Poisson}(u)$$

$$U \sim \text{Gamma}(\alpha, \beta)$$

The unconditional distribution of Y is Negative-binomial

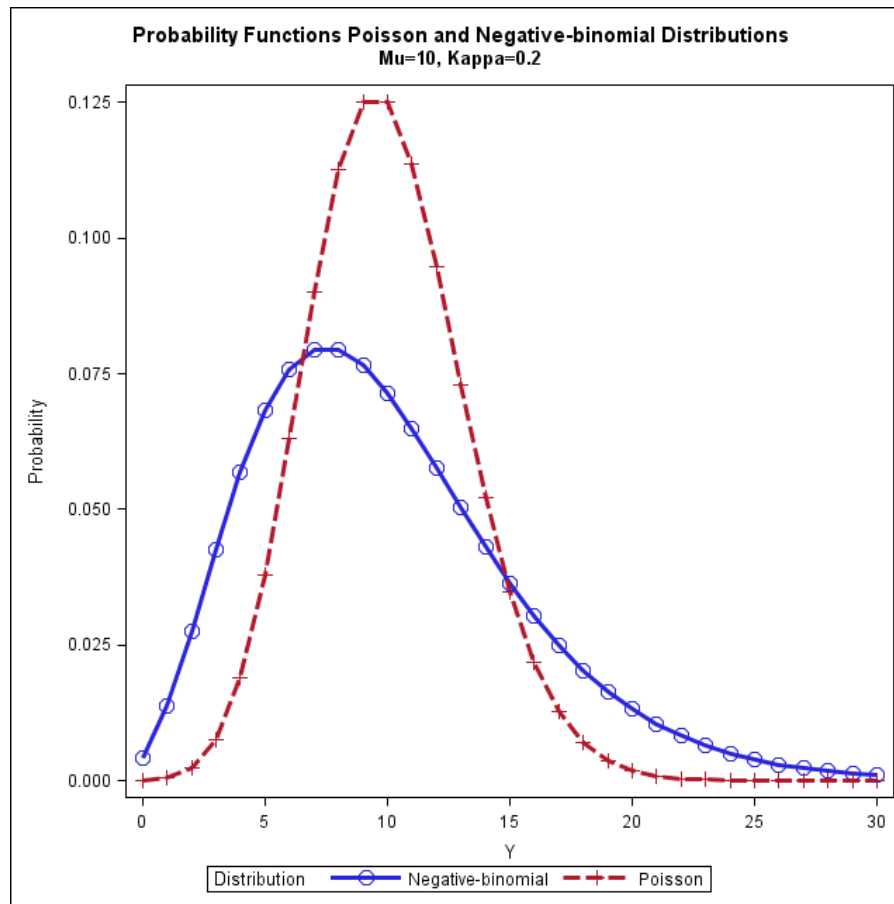
Under the parameterization $\mu = \alpha\beta$ and $\kappa = \frac{1}{\alpha}$

$$E(Y) = \mu$$

$$\text{Var}(Y) = \mu + \kappa\mu^2 = \mu(1 + \kappa\mu)$$

Overdispersion Models for Count Data: The Negative-binomial

Mean Poisson = 10
Mean NB = 10
Variance NB = 30



Overdispersion Models for Count Data: The Zero-inflated Poisson (ZIP)

A Zero-inflated Poisson distribution (**ZIP**)

$$\Pr(Y = y) = \omega I(y = 0) + (1 - \omega) \{ \Pr(X = y) \}, y = 0, 1, \dots$$

where $I(\cdot)$ is an indicator function and $X \sim \text{Poisson}(\mu)$

$$E(Y) = (1 - \omega)\mu$$

$$\text{Var}(Y) = (1 - \omega)(\mu + \omega\mu^2)$$

Since, $\text{Var}(Y) > E(Y)$,

an excess of zeros in the Poisson distribution implies overdispersion

Overdispersion Models for Count Data: The Zero-inflated Negative-binomial (ZINB)

A Zero-inflated Negative-binomial distribution (**ZINB**)

can be defined in a similar way.

$$E(Y) = (1 - \omega)\mu$$

$$\text{Var}(Y) = (1 - \omega)\{\mu(1 + \mu\kappa) + \omega\mu^2\}$$

Since, $\text{Var}(Y) > E(Y)$,

an excess of zeros in the NB distribution implies overdispersion

Overdispersion Models for Count Data: Hurdle Poisson and Negative-binomial

A Hurdle Model is interpreted as a two – part model:

$$\Pr(Y = y) = \begin{cases} \Pr(Y_1 = 0) & \text{if } y=0 \\ \{1 - \Pr(Y_1 = 0)\} \frac{\Pr(Y_2 = y)}{\{1 - \Pr(Y_2 = 0)\}} & \text{if } y \geq 1 \end{cases}$$

$Y_1 \sim \text{Bernoulli}$ with $\pi_1 = \Pr(Y_1 = 0)$, Y_2 taking values $0, 1, 2, \dots$

Note that $\frac{\Pr(Y_2 = y)}{\{1 - \Pr(Y_2 = 0)\}}$, $y = 1, 2, \dots$ represents

the probability function of a Zero-truncated Y_2 random variable

Overdispersion Models for Count Data: Hurdle Poisson and Negative-binomial

- As pointed out by Mullahy (1986), “the idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome of whether a count variable has a zero or a positive realization.”
- He then indicated that if the realization is positive, “the ‘hurdle’ is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model.”

Hurdle Models can accommodate Overdispersion and Under-Dispersion !!!

Overdispersion Models for Count Data: Hurdle Poisson and Negative-binomial

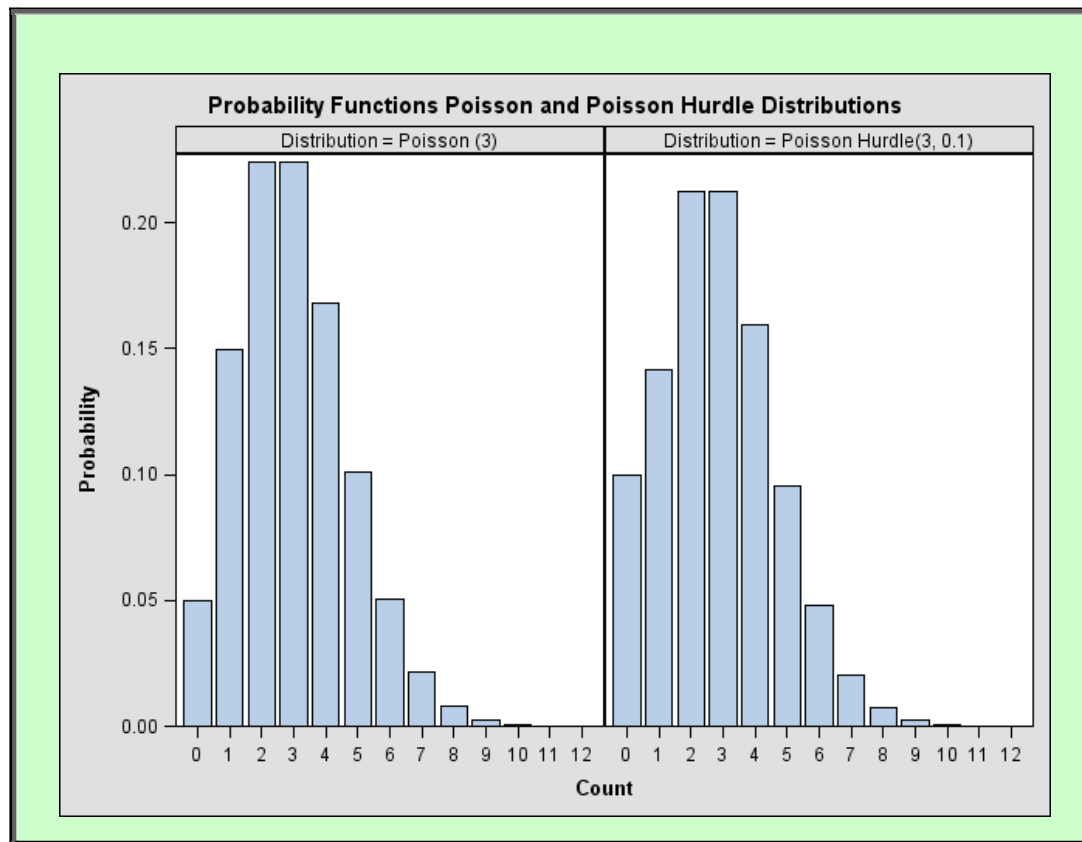
Y_1 is Bernoulli with $\pi_1 = \Pr(Y_1 = 0)$

so $(1 - \pi_1)$ is the Probability of Crossing the Hurdle

A Hurdle Poisson Model arises when $Y_2 \sim \text{Poisson}(\mu)$

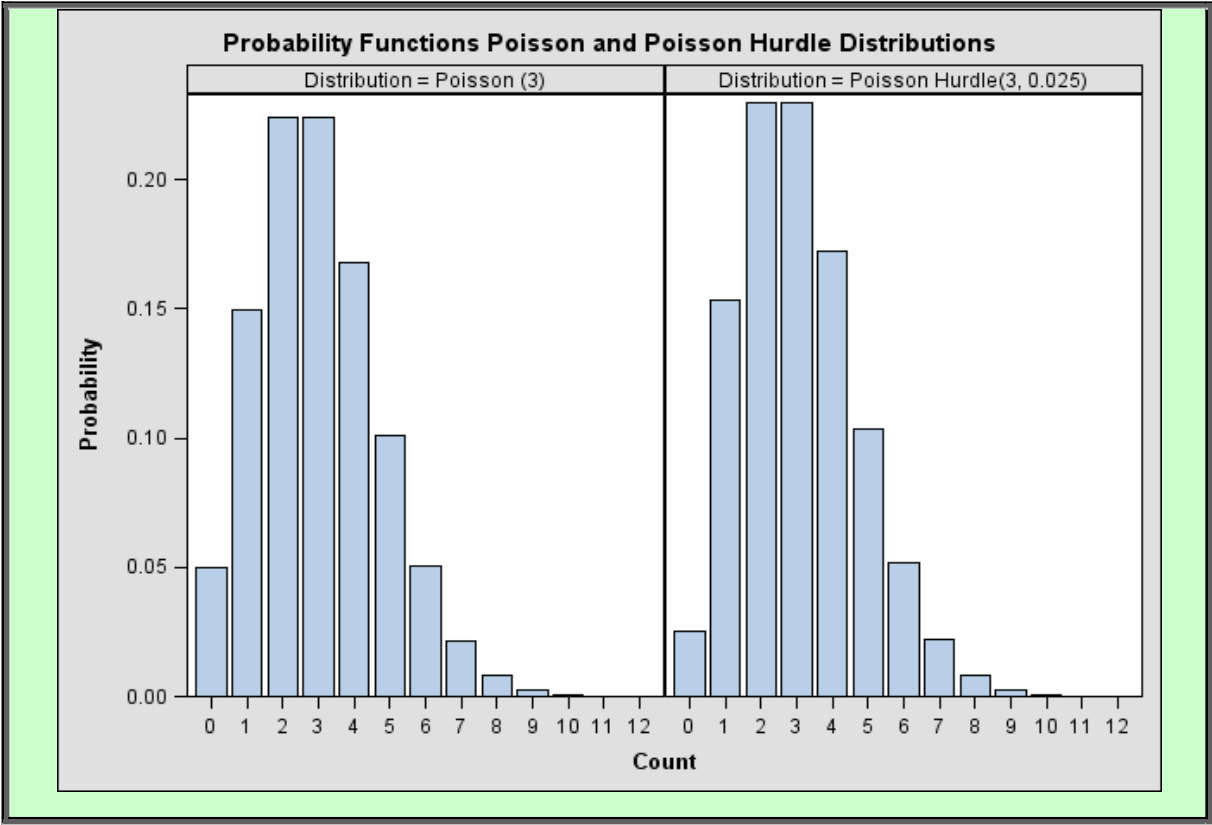
Similarly, a Hurdle Negative-binomial Model arises
when $Y_2 \sim \text{Negative-binomial}(\mu, \kappa)$

Overdispersion Models for Count Data: Hurdle Poisson and Negative-binomial



Hurdle Poisson has an excess of zeros and exhibits overdispersion: Mean=2.84, Var=3.29

Overdispersion Models for Count Data: Hurdle Poisson and Negative-binomial



Hurdle Poisson has a “deficit” of zeros and exhibits under-dispersion: Mean=3.08, Var=2.84

Milk Does Your Body Good

- Consider a survey aimed to investigate the daily milk consumption (in number of glasses)
- n=1,900 subjects: 1980 WAVE II of the National Survey of Personal Health Practices and Consequences (as cited in Cameron and Trivedi, 2008)
- The number of subjects reporting no consumption of milk might come from two different sources:
 - 1) those who do not drink milk at all
 - 2) those who drink milk but did not drink any around the time the survey was conducted

Milk Does Your Body Good

Milk Consumption (in glasses)	Frequency Distribution
0	767
1	557
2	333
3	142
4	62
5	23
6	16
TOTAL	1900

About 40% of the subjects indicated that they did not consume any milk. There seems to be an excess of zeros, and it is very likely that these data might exhibit overdispersion.

Milk Does Your Body Good

```
data Milk_Consumption;  
  input Y freq;  
  label y = ' (Number of Glasses of Milk)';  
  datalines;
```

0	767
1	557
2	333
3	142
4	62
5	23
6	16

```
;  
title "ZIP Model using PROC COUNTREG";  
proc countreg data=Milk_Consumption;  
  model y = / dist=poisson;  
  zeromodel y ~ / link=logistic;  
  freq freq;  
  
run;
```

```
title "ZIP Model using PROC GENMOD";  
proc genmod data=Milk_Consumption;  
  model y = / dist=zip;  
  zeromodel / link=logit ;  
  freq freq;  
  
run;
```

```
title "ZIP Model using PROC FMM";  
proc fmm data=Milk_Consumption;  
  model y = / dist=poisson;  
  model + / dist=Constant;  
  freq freq;  
  
run;
```

Milk Does Your Body Good

```
title "Poisson Hurdle Model using PROC FMM";  
proc fmm data=Milk_Consumption;  
model y = / dist=TRUNCPOISSON ;  
model + / dist=Constant;  
freq freq;  
run;
```

Milk Does Your Body Good

Truncated
Poisson
Results

Probability of
Crossing the Hurdle

Prob = 0.5963

Poisson Hurdle Model -- FMM

Parameter Estimates for Truncated Poisson Model						
Component	Effect	Estimate	Standard Error	z Value	Pr > z	Inverse Linked Estimate
1	Intercept	0.3388	0.02957	11.46	<.0001	1.4032

Parameter Estimates for Mixing Probabilities					
Component	Mixing Probability	Linked Scale			
		Logit(Prob)	Standard Error	z Value	Pr > z
1	0.5963	0.3901	0.04676	8.34	<.0001
2	0.4037	-0.3901			

Milk Does Your Body Good

```
title "Negative-binomial Hurdle Model -- FMM";  
proc fmm data=Milk_Consumption;  
    model y =          / dist=TRUNCNEGBIN ;  
    model          +    / dist=Constant;  
    freq freq;  
run;
```

Milk Does Your Body Good

Truncated
Negative-binomial
Results
Mu=1.2168
Kappa=0.2520

Probability of
Crossing the Hurdle

Prob = 0.5963

Negative-binomial Hurdle Model -- FMM						
Parameter Estimates for Truncated Negative Binomial Model						
Component	Effect	Estimate	Standard Error	z Value	Pr > z	Inverse Linked Estimate
1	Intercept	0.1962	0.05606	3.50	0.0005	1.2168
1	Scale Parameter	0.2520	0.08143			

Parameter Estimates for Mixing Probabilities					
Component	Mixing Probability	Linked Scale			
		Logit(Prob)	Standard Error	z Value	Pr > z
1	0.5963	0.3901	0.04676	8.34	<.0001
2	0.4037	-0.3901			

Milk Does Your Body Good

Pearson's GOF Chi-square statistic

$$X^2 = \sum_{s=0}^6 (O_s - E_s)^2 / E_s$$

Value of Y	Observed Frequencies
0	767
1	557
2	333
3	142
4	62
5	23
>=6	16
TOTAL	1900

Distribution	Chi-square	DF	P-Value
Poisson Hurdle	38.04	4	<.0001
Negative-binomial Hurdle	2.06	3	0.5600

Conclusion: A Negative-binomial Hurdle Model is a good fit to the milk consumption data

Generalized Linear Overdispersion Mixed Models

- Analogous to Generalized Linear Models (GLM), the overdispersion models (Beta-binomial, Random-clumped, ZIB, ZINB,...) allow for link functions on the main parameters, including the overdispersion parameter.
- They can be referred as **Generalized Linear Overdispersion Models (GLOM)**
- These models can be extended to a larger family referred as **Generalized Linear Overdispersion Mixed Models (GLOMM)**

Examples arise in longitudinal studies, cross-over studies where the outcome at each occasion exhibits overdispersion

Generalized Linear Overdispersion Mixed Models

Simulated data under a Poisson Hurdle GLOMM

```
data poisson_hurdle_glomm;
  beta      = log(10);
  p1        = 0.20;
  n         = 200;
  m         = 12;
  seed      = 16670;
  sigma     = sqrt(1.5);
  do id=1 to n;
    z = sigma*normal(seed);
    do t=1 to m;
      u = uniform(seed);
      if u < p1 then do;
        y = 0;
        output;
      end;
      else do; *--- Crossing the hurdle;
        mu = exp(beta + z);
        do until (y>0);
          y = ranpoi(seed,mu);
        end;
        output;
      end;
    end;
  end;
  keep id y;
run;
```

Generalized Linear Overdispersion Mixed Models

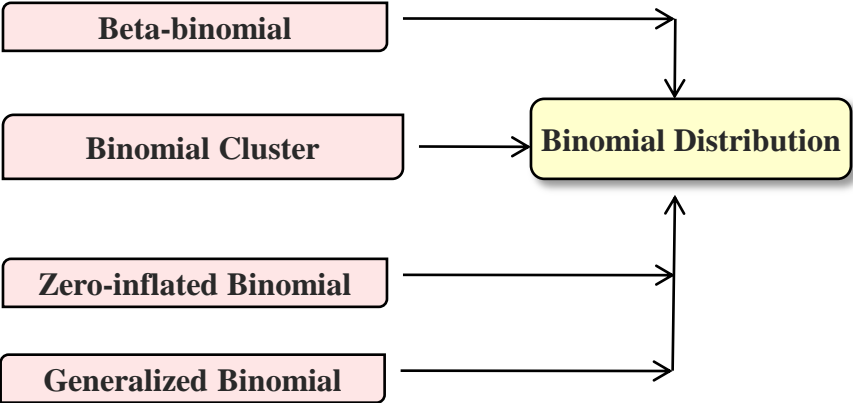
```
title "Poisson Hurdle Model GLMM";
proc nlmixed data=poisson_hurdle_glomm;
  parms alpha=0 beta=0 sigma=1;
  p1 = 1/(1+exp(-alpha));
  mu = exp(beta + u);
  p2 = exp(-mu);
  if y=0 then
    ll = log(p1);
  else ll = log((1-p1)/(1-p2)) + y * log(mu)
    - lgamma(y+1) - mu;
  model y ~ general(ll);
  random u ~ normal(0, sigma*sigma) subject=id;
  estimate "1-P1" 1/(1+exp(alpha));
  estimate "Mu" exp(beta);
  estimate "Sigma**2" sigma*sigma;
run;
```

Generalized Linear Overdispersion Mixed Models

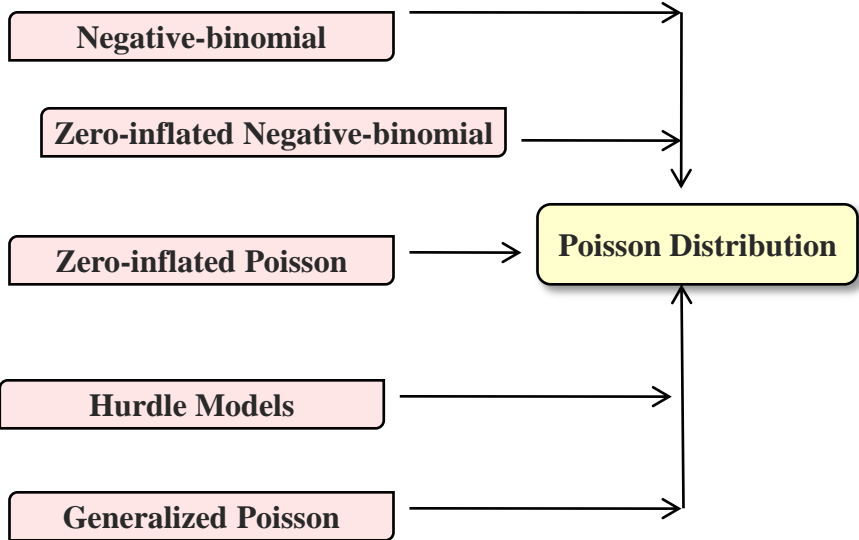
Poisson Hurdle Model GLMM								
The NLMIXED Procedure								
Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
1-P1	0.7988	0.008184	199	24.59	<.0001	0.05	0.7826	0.8149
Mu	10.6414	0.9428	199	11.29	<.0001	0.05	8.7822	12.5006
Sigma**2	1.5469	0.1587	199	9.75	<.0001	0.05	1.2339	1.8598

Final Remarks

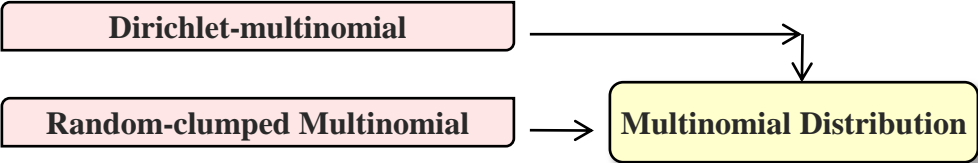
1) Binomial Distribution



2) Poisson Distribution



3) Multinomial Distribution



TAKEAWAY: The family of Generalized Linear Overdispersion Models (GLOM) can be extended to GLOMM

Final Remarks

- Addressing Overdispersion is important
- Beta-binomial, Binomial Cluster, Zero-inflated Binomial and arbitrary mixtures of Binomials are available in PROC FMM
- An Omnibus Test to assess the Goodness-of-fit was presented
- Distributions such as Negative-binomial, Zero-inflated Poisson and Zero-inflated Negative-binomial are available in several PROCs. Hurdle Poisson and Hurdle Negative-binomial are available in PROC FMM.

Thanks for your attention! 😊

References

1. Cameron, A.C. and Trivedi, P.K. (2008), **Regression analysis of count data**. 7th printing. Cambridge University Press, New York, N. Y.
2. Cox, D.R. and Snell, E.J. (1989), **Analysis of Binary Data**. Second Ed., New York: Chapman and Hall.
3. Liang, K.Y. and Zeger, S.L. (1986) , “Longitudinal data analysis using generalized linear models,” *Biometrika*, 73, 13-22.
4. McCullagh, P. and Nelder, J.A. (1983), **Generalized Linear Models**. First Ed., London: Chapman and Hall.
5. McCullagh, P. and Nelder, J.A. (1989), **Generalized Linear Models**. Second Ed., London: Chapman and Hall.
6. Morel, J.G. and Nagaraj, N.K. (1993), “A finite mixture distribution for modelling multinomial extra variation,” *Biometrika*, 80, 363-371.
7. Morel, J.G. and Neerchal, N.K. (1997), “Clustered binary logistic regression in teratology data using a finite mixture distribution,” *Statistics in Medicine*, 16, 2843-2853.
8. Morel, J. G. and Neerchal, N. K. (2012), **Overdispersion Models in SAS®**. SAS Press
9. Mullahy, J. (1986), "Specification and testing of some modified count data models," *Journal of Econometrics*, 33, 341-365.
10. Neerchal, N.K. and Morel, J.G. (1998), “Large cluster results for two parametric multinomial extra variation models,” *Journal of the American Statistical Association*, 93, 1078-1087.
11. Skellam, J.G. (1948), “A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials,” **Journal of the Royal Statistical Society, Series B**, 10, 257-261.
12. Sutradhar, S. C., Neerchal, N. K. and Morel, J. G. (2008), “A Goodness-of-fit Test for Overdispersed Binomial or Multinomial Models,” *Journal of Statistical Planning and Inference*, 138, 1459-1471.
13. Wedderburn, R.W.M. (1974), “Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method,” *Biometrika*, 61, 439-447.
14. Zeger, S.L. and Liang, K.Y. (1986), “Longitudinal data analysis for discrete and continuous outcomes,” *Biometrics*, 42, 121-130.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.