

## Count Data

- We encounter count data when we are interested in how many as opposed to what proportion.
- A good example would be number born for a litter bearing species.
- Also used when we are looking at a small fraction of the whole as is typically the case for disease incidence rates.

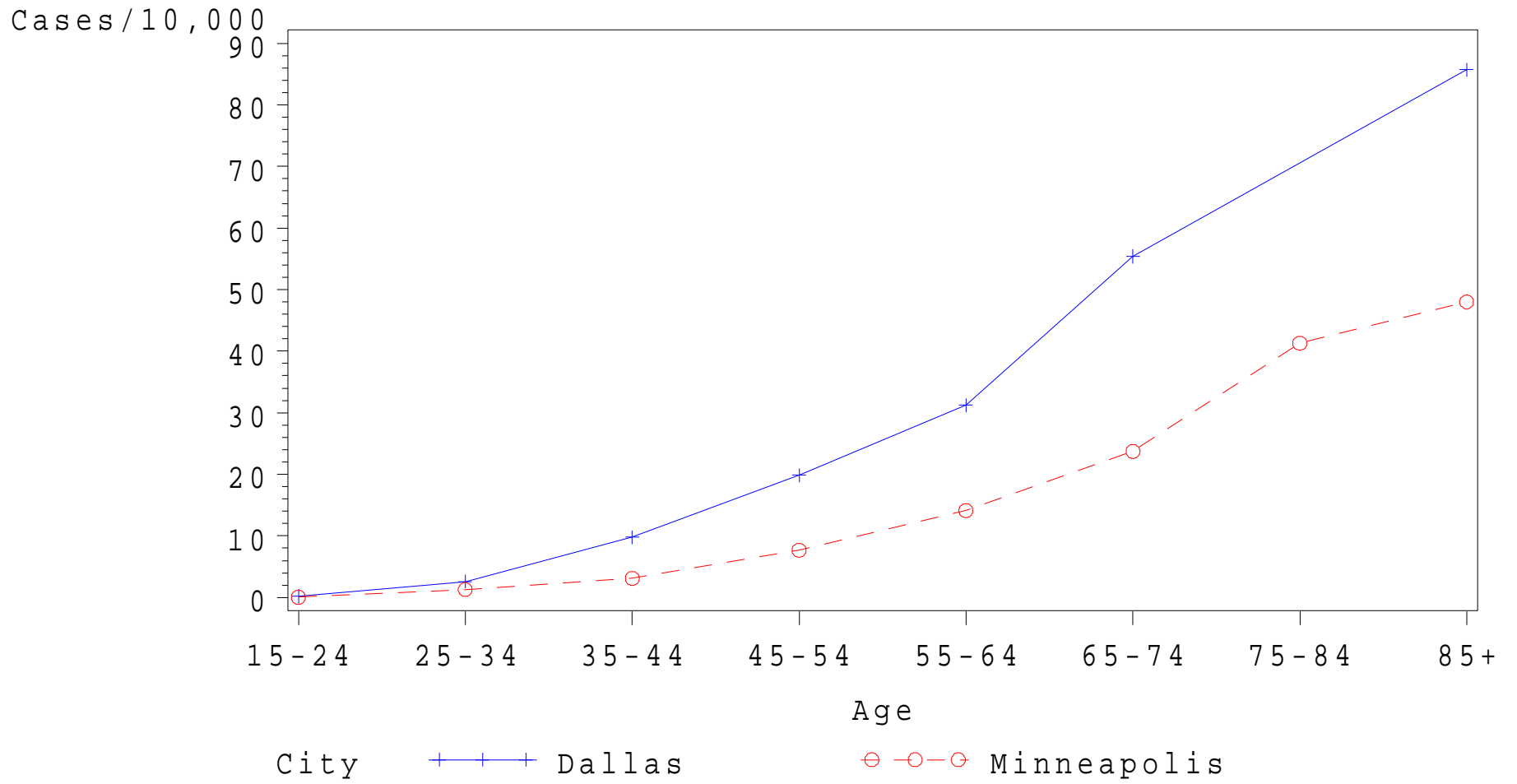
- Incidence of nonmelanoma skin cancer would be expected to increase with sun exposure.
- In the following data<sup>2</sup> from a study which compared nonmelanoma skin cancer incidence rates among women in Minneapolis-St. Paul and Dallas.

---

<sup>2</sup>Kleinbaum, D., Kupper, L., and Muller, K. (1989). Applied regression analysis and other multivariate methods. PWS-Kent, Boston, Massachusetts.

Hand, D. et al, (1994). A handbook of small data sets. Chapman and Hall, London.

Electronic version: OzDASL - Australasian Data and Story Library  
<http://www.statsci.org/data/general/skin.html>



- A reasonable distribution for the number of cases would be a Poisson distribution.

$$y_{ij} \sim \text{Poi}(\mu_{ij})$$

$$\mathbf{E}(y_{ij}) = \mu_{ij}$$

$$\text{var}(y_{ij}) = \mu_{ij}$$

$$\ell(\boldsymbol{\mu}; \mathbf{y}) = c + \sum_{i=1}^2 \sum_{j=1}^{n_i} [y_{ij} \ln(\mu_{ij}) - \mu_{ij}]$$

where  $y_{ij}$  is the number of cases in age group  $j$  in city  $i$ .

- The linear predictor to be used is

$$\eta_{ij} = \mu + C_i + A_j$$

- $\mu$ : intercept
- $C_i$ : effect of city  $i$
- $A_j$ : effect of age group  $j$ .

- The canonical link for a Poisson distribution is a log link. The corresponding inverse link function is

$$\mu_{ij} = e^{\eta_{ij} + \ln(P_{ij})} = e^{\eta_{ij}} P_{ij}$$

- where  $P_{ij}$  are the number of women in city  $i$  who are in age group  $j$ .
- The  $\ln(P_{ij})$  is an offset so that we are modeling the rate as opposed to the number.
- The link function is also reasonable because it will enforce the requirement that the incidence rate must be non-negative.

To last component needed is  $\mathbf{H}$

$$\mathbf{H} = \text{Diag}(\mu_{ij}).$$

## ASReml

Skin Cancer Poisson Analysis

Cases

Town \* !A

Age \* !I

Population

LnPop !=Population !^0

skin2.dat !SKIP=1

!FCON

Cases !POISSON !OFFSET LnPop ~ mu Town Age

0 0 0

predict Town !TDIFF

predict Age !TDIFF

1	LogL= 8.31818	S2= 1.0000	6 df	1.000
2	LogL= 9.97219	S2= 1.0000	6 df	1.000
3	LogL= 10.4106	S2= 1.0000	6 df	1.000
4	LogL= 10.4379	S2= 1.0000	6 df	1.000
5	LogL= 10.4379	S2= 1.0000	6 df	1.000
6	LogL= 10.4379	S2= 1.0000	6 df	1.000

Final parameter values 1.0000  
 Deviance from GLM fit 6 5.21  
 Variance heterogeneity factor [Deviance/DF] 0.87

Source	Model	terms	Gamma	Component	Comp/SE	% C
Variance	15	6	1.00000	1.00000	0.00	0 F

Analysis of Variance	NumDF	DenDF_con	F_inc	F_con	M	P_con
6 mu	1	6.0	55929.14	55929.14	.	<.001
2 Town	1	6.0	84.75	204.54	A	<.001
3 Age	7	6.0	162.77	162.77	A	<.001

Notice: The DenDF values are calculated ignoring fixed/boundary/singular variance parameters using numerical derivatives.

## Predict File:

Ecode is E for Estimable, \* for Not Estimable

----- 1 -----

Predicted values of Cases

Age is averaged over fixed levels

The cells of the hypertable are calculated from all model terms constructed  
solely from factors in the averaging and classify sets.

Town	Logarithm_value	Stand_Error	Ecode	Retransformed_value	approx_SE
FortWorth	-7.4197	0.0730	E	0.0006	0.0000
StPaul	-6.5670	0.0663	E	0.0014	0.0001

SED: Standard Error of Difference: Min 0.0596 Mean 0.0596 Max 0.0596

Predicted values with t statistics

-7.420

-6.567 14.30

- From this analysis we can see that the nonmelanoma incidence rate among women is significantly lower in Minneapolis The estimated reduction is

$$\hat{C}_S - \hat{C}_F = -7.4197 - (-6.5670) = -0.8527$$

$$\text{Percent reduction } 100\%(1 - e^{-0.8527}) = 57.4\%.$$

- Approximate standard Errors: Delta Method

$$\exp(\hat{\eta}) \simeq \exp(\eta) + \frac{\partial \exp(\eta)}{\partial \eta}(\hat{\eta} - \eta)$$

$$= \mu + \mu(\hat{\eta} - \eta)$$

$$se_{\mu} \simeq \mu \times se_{\eta}$$

$$0.0006 \times 0.0730$$

Predicted values of Cases

Town is averaged over fixed levels  
 The cells of the hypertable are calculated from all model terms constructed  
 solely from factors in the averaging and classify sets.

Age	Logarithm_value	Stand_Error	Ecode	Retransformed_value	approx_SE	
20	-11.2657	0.4474 E		0.0000	0.0000	
30	-8.6367	0.1368 E		0.0002	0.0000	
40	-7.4201	0.0832 E		0.0006	0.0000	
50	-6.6719	0.0602 E		0.0013	0.0001	
60	-6.1793	0.0544 E		0.0021	0.0001	
70	-5.6200	0.0492 E		0.0036	0.0002	
80	-5.0626	0.0917 E		0.0063	0.0006	
90	-5.0900	0.0982 E		0.0062	0.0006	
SED: Standard Error of Difference: Min		0.0710	Mean	0.2034	Max	0.4675

Predicted values with t statistics

-11.27								
-8.637	5.62							
-7.420	8.46	7.66						
-6.672	10.19	13.26	7.43					
-6.179	11.30	16.84	12.74	6.26				
-5.620	12.55	20.92	18.98	13.93	7.87			
-5.063	13.56	21.37	18.51	14.18	10.12	5.18		
-5.090	13.49	21.18	18.28	13.89	9.82	4.88	-0.20	

# Over dispersion

- Over-dispersion is the situation where the variability is greater than expected.

$$\text{var}(y_i) > \nu(\mu_i)$$

- Can be due to:
  - Incorrect Linear Predictor
  - Incorrect Distribution
  - Incorrect Link Function
  - Additional Random Effects
  - Outliers

## Scale parameter

- The main advantage of the scale parameter approach is its simplicity.
- Revisiting the exponential family we see that it included a scale parameter  $\phi$ .

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}, \phi) &= \exp \left\{ \frac{\mathbf{y}'\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\phi^2} - c(\mathbf{y}, \phi) \right\} \\
&= \exp \left\{ \sum_{i=1}^N \frac{y_i \theta_i - b(\theta_i)}{\phi^2} - c(\mathbf{y}, \phi) \right\}
\end{aligned}$$

$$\ell(\boldsymbol{\theta}, \phi) = \frac{1}{\phi^2} \ell(\boldsymbol{\theta}, 1) - c(\mathbf{y}, \phi)$$

$$E(y_i) = \mu_i = \frac{\partial b(\theta_i)}{\partial \theta_i}$$

$$\text{var}(y_i) = \phi^2 \frac{\partial \mu_i}{\partial \theta_i} = \phi^2 \nu(\mu_i)$$

$$\nu(\mu_i) = \frac{\partial \mu_i}{\partial \theta_i} = \left( \frac{\partial \theta_i}{\partial \mu_i} \right)^{-1}$$

- The derivation of the estimating equations assumed  $\phi$  was known.
- The only place where  $\phi$  entered was in  $\mathbf{R} = \phi^2 \text{Diag}(\nu(\mu_i))$ .
- Factoring  $\phi$  out the estimating equations we obtain

$$[\mathbf{X}'\mathbf{H}\Sigma^{-1}\mathbf{H}\mathbf{X}] \boldsymbol{\beta}^{[i+1]} = \mathbf{X}'\mathbf{H}\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[i]} + \mathbf{H}\boldsymbol{\eta}^{[i]})$$

$$\Sigma = \text{Diag}(\nu(\mu_i))$$

$$\mathbf{i}(\boldsymbol{\beta}) = \frac{1}{\phi^2} [\mathbf{X}'\mathbf{H}\Sigma^{-1}\mathbf{H}\mathbf{X}]$$

- The estimates of the fixed effects does not depend on the scale parameter.
- It is the asymptotic variance and differences in the log-likelihood that depend on the scale parameter. Both of which are scaled by the value scale parameter.
- Now we will look at the estimation of  $\phi$ .

The scale parameter can be estimated using either a sum of squares approach or a likelihood approach. We would expect that

$$\mathbf{E} \left[ \frac{(\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}}))' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}}))}{n - \text{rank}(X)} \right] \simeq \phi^2$$
$$\mathbf{E} \left[ \frac{2[\ell(\boldsymbol{\mu} = \mathbf{y}) - \ell(h(\mathbf{X}\hat{\boldsymbol{\beta}}))]}{n - \text{rank}(X)} \right] \simeq \phi^2$$

From the sum of squares we get the Pearson  $\chi^2$  estimator

$$(\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}}))' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}})) \sim \phi^2 \chi^2(n - \text{rank}(X))$$

$$\sqrt{\frac{(\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}}))' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - h(\mathbf{X}\hat{\boldsymbol{\beta}}))}{n - \text{rank}(X)}} = \hat{\phi}_P.$$

From the likelihood we get the deviance estimator

$$2[\ell(\boldsymbol{\mu} = \mathbf{y}) - \ell(h(\mathbf{X}\hat{\boldsymbol{\beta}}))] \sim \phi^2 \chi^2(n - \text{rank}(X))$$

$$\sqrt{\frac{2[\ell(\boldsymbol{\mu} = \mathbf{y}) - \ell(h(\mathbf{X}\hat{\boldsymbol{\beta}}))]}{n - \text{rank}(X)}} = \hat{\phi}_D$$

## Deviance Formulas

- Binomial (proportions)

$$\ell(\mu_i; y_i) = \ln \left[ \binom{n_i}{n_i y_i} \right] + n_i y_i \ln [\mu_i] + n_i(1 - y_i) \ln[1 - \mu_i]$$

$$\ell(y_i; y_i) = \ln \left[ \binom{n_i}{n_i y_i} \right] + n_i y_i \ln [y_i] + n_i(1 - y_i) \ln[1 - y_i]$$

$$d_i = 2n_i \left[ y_i \ln \left( \frac{y_i}{\mu_i} \right) + (1 - y_i) \ln \left( \frac{1 - \mu_i}{1 - y_i} \right) \right]$$

- Poisson

$$\ell(\mu_i; y_i) = y_i \ln(\mu_i) - \mu_i - \ln(y_i!)$$

$$\ell(y_i; y_i) = y_i \ln(y_i) - y_i - \ln(y_i!)$$

$$d_i = 2 \left[ y_i \ln \left( \frac{y_i}{\mu_i} \right) - (y_i - \mu_i) \right]$$

- Deviance

$$d = \sum_{i=1}^N d_i$$

## The Deviant Carrot Example

- Data from a study examining the effect of insecticide dose on insect damage to carrot<sup>3</sup>
- Dependent Variable: Proportion of damaged carrots in a plot.
- Experimental design: Randomized Complete Block Design with 3 blocks
- The treatments were 7 dose of insecticide (log dose  $x = 1.52, 1.64, \dots, 2.36$ ).

---

<sup>3</sup>McCullagh and Nelder (1989) *Generalized Linear Models*. pg 409.

# Model

- Distribution: Binomial
- Link function: Logit
- Linear Predictor:

$$\eta_{ij} = \mu + Bl_j + b x_i$$

## ASReml

Carrot Analysis

Block !I

LogDose

Damage

N

carrot.txt !SKIP=1

Damage !BINOMIAL !LOGIT !TOTAL N ~ mu Block LogDose

0 0 0

Final parameter values 1.0000  
 Deviance from GLM fit 20 39.98  
 Variance heterogeneity factor [Deviance/DF] 2.00  $\leftarrow \tilde{\phi}_D^2$

Source	Model	terms	Gamma	Component	Comp/SE	% C
Variance	24	20	1.00000	1.00000	0.00	0 F

Analysis of Variance	NumDF	DenDF	F_inc	Prob
5 mu	1	20.0	240.27	<.001
1 Block	2	20.0	5.91	0.010
2 LogDose	1	20.0	27.93	<.001

Notice: The DenDF values are calculated ignoring fixed/boundary/singular variance parameters using numerical derivatives.

	Estimate	Standard Error	T-value	T-prev
2 LogDose				
1	-1.81740	0.343870	-5.29	
1 Block				
2	0.300882	0.199099	1.51	
3	-0.542390	0.231796	-2.34	-3.73
5 mu				
1	2.02265	0.650122		3.11

## ASReml Pearson Scale

Carrot Analysis

Block !I

LogDose

Damage

N

carrot.txt !SKIP=1

Damage !BINOMIAL !LOGIT !TOTAL N !DISP ~ mu Block LogDose

0 0 0

## Results

Deviance from GLM fit                    20            39.98  
 Variance heterogeneity factor [Deviance/DF]            2.00

Source	Model	terms	Gamma	Component	Comp/SE	% C
Variance	24	20	1.00000	2.05874	3.16	0 U

$\tilde{\phi}_P^2$

Analysis of Variance	NumDF	DenDF	F_inc	Prob
5 mu	1	20.0	116.71	<.001
1 Block	2	20.0	2.87	0.080
2 LogDose	1	20.0	13.57	0.001

Notice: The DenDF values are calculated ignoring fixed/boundary/singular variance parameters using numerical derivatives.

	Estimate	Standard Error	T-value	T-prev
2 LogDose				
1	-1.81740	0.493396	-3.68	
1 Block				
2	0.300882	0.285673	1.05	
3	-0.542390	0.332589	-1.63	-2.60
5 mu				
1	2.02265	0.932816	2.17	

## Differences

- F-values

Source	UnScaled	Formula	Scaled
mu	240.27	$240.27/2.05874$	116.71
Block	5.91	$5.91/2.05874$	2.87
LogDose	27.93	$27.93/2.05874$	13.57

- Standard Errors

Source	UnScaled	Formula	Scaled
LogDose	0.343870	$0.343870 \times \sqrt{2.05874}$	0.49396

- t-values

Source	UnScaled	Formula	Scaled
LogDose	-5.29	$-5.29 / \sqrt{2.05874}$	3.68