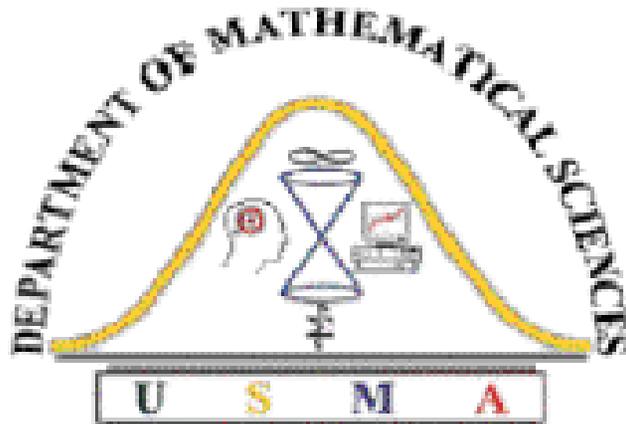


# Horseshoe Crabs, Supreme Court Justices and Shoulder Dislocations! Oh, My!

## Adventures in Poisson Regression or the Regression of Counts and Rates



**CDAS Presentation  
LTC Robert Burks  
8 October 2007**

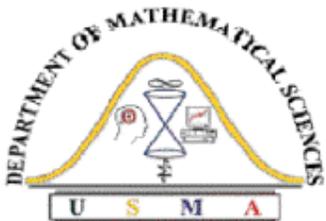


# Agenda

---



- ✓ **Overview of the Poisson Regression Model**
- ✓ **Understanding Rates and Ratios**
- ✓ **Review of CDAS Project - Shoulder Dislocation**
- ✓ **Questions**



# Poisson and the GLM



The poisson regression model is an example of a broad class of models known as generalized linear models (GLM); other examples include the logistics regression and linear regression.

There are three components to a GLM:

1. **Random Component** – refers to the probability distribution of the response variable ( $Y$ );

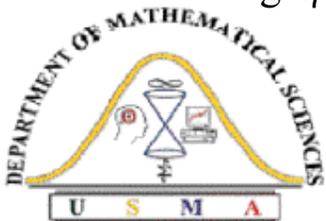
e.g. binomial distribution for  $Y$  in the binary logistic regression.

1. **Systematic Component** - refers to the explanatory variables  $(X_1, X_2, \dots, X_k)$  as a combination of linear predictors;

e.g.  $\beta_0 + \beta_1 x_1 + \beta_2 x_2$  as we have seen in logistic regression.

3. **Link Function,  $\eta$  or  $g(\mu)$**  - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables;

e.g.  $\eta = \text{logit}(\pi)$  for logistic regression.



# Poisson Regression



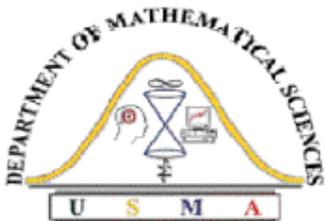
Poisson regression falls somewhere between linear and logistic regression.

- ✓ The dependent variable consists of counts following a Poisson distribution
- ✓ The equation relates a count (or rate) to a series of independent variables

It's three components are:

1. **Random Component** – The distribution of *counts* is Poisson
2. **Systematic Component** - X are discrete variables used in cross-classification
3. **Link Function,  $\eta$  or  $g(\mu)$**  – *Log Link*  $\text{Log } \eta = \log(\mu)$

**Note: As a GLM, Poisson Regression follows many of the same modeling procedures we all know (e.g., Parameter estimation , inference, model fitting)**



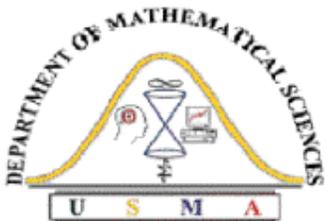
# Models for Count Data



**Count and Rate models are commonly found in the medical research or testing related fields**

**Examples:**

- **Study of nesting Horseshoe Crabs (Agresti, 2002)**
  - (1) **How does the number of satellites a female horseshoe crab has depend on the width of her back;**
  - (2) **What is the rate of satellites per unit width?**
- **Yearly Vacancies in U.S. Supreme Court Justices**
- **Number of cargo ships damaged by waves (McCullagh & Nadar, 1989)**
- **Daily homicide counts in California (Grogger, 1990)**
- **Founding of day care centers in Tronto (Baum & Oliver, 1992)**
- **Number of deaths due to SARs (Yu, Chan, and Fung, 2006)**



# Understanding Rates and Ratios



## A Simple Multi-parameter Example:

We want to compare the Group 1 ( $j=1$ ) and Group 2 ( $j = 2$ ) mortality rates among the strata ( $i = 1, 2, \dots, 5$ )

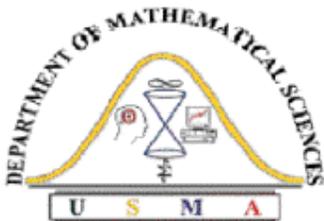
Stratum	Group 1			Group 2		
	Person - Years	Deaths	Rates ( $r_{i1}$ )	Person - Years	Deaths	Rates ( $r_{i1}$ )
1	1000	20	0.020	100	4	0.040
2	2000	60	0.030	200	12	0.060
3	3000	135	0.045	300	27	0.090
4	4000	360	0.090	400	72	0.180
5	5000	900	0.180	500	180	0.360
Total	15000	1475	0.098	1500	295	0.197

The basic *rate* model is:

$$\ln(r_{ij}) = \beta_0 + \beta_1 \text{Group} + \beta_2 S2 + \beta_3 S3 + \beta_4 S4 + \beta_5 S5$$

or

$$\text{rate}(r_{ij}) = e^{\beta_0 + \beta_1 \text{Group} + \beta_2 S2 + \beta_3 S3 + \beta_4 S4 + \beta_5 S5}$$



# A Multi-Parameter Example

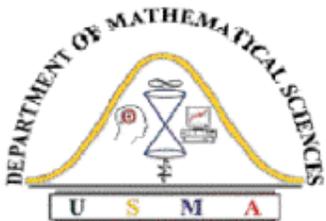


If we want to compare the two groups we need a ratio ...

Introducing the *standardized mortality ratio* (SMR):

$$\begin{aligned} SMR &= \frac{r_{i1}}{r_{i2}} \\ &= \frac{e^{\beta_0 + \beta_1(1) + \beta_2 S_2 + \beta_3 3 + \beta_4 4 + \beta_5 S_5}}{e^{\beta_0 + \beta_1(2) + \beta_2 S_2 + \beta_3 3 + \beta_4 4 + \beta_5 S_5}} \\ &= e^{\beta_0 + \beta_1(1) + \beta_2 S_2 + \beta_3 3 + \beta_4 4 + \beta_5 S_5 - (\beta_0 + \beta_1(2) + \beta_2 S_2 + \beta_3 3 + \beta_4 4 + \beta_5 S_5)} \end{aligned}$$

The ratio's title has become convention ...  
rate does not have to be a count of deaths.  
Also known as Incident Rate Ratio (IRR)



# A Multi-Parameter Example



Model output from your favorite stat software:

Deaths	Coefficient	Std. Err	z	P >  z	[95% Confidence Interval]	
S2	0.4054	0.2357	1.7200	0.0850	-0.0565	0.8674
S3	0.8109	0.2187	3.7080	0.0000	0.3822	1.2390
S4	1.5041	0.2097	7.1720	0.0000	1.0930	1.9151
S5	2.1972	0.2064	10.6470	0.0000	1.7927	2.6010
Group	0.6931	0.0638	10.8680	0.0000	0.5681	0.8181
Constant	-3.9123	0.2044	-19.1390	0.0000	-4.3126	-3.5114

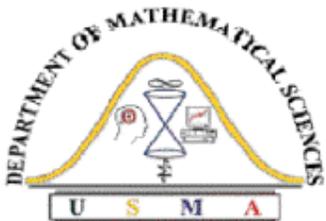
The mortality rate for the third strata of each group is:

$$\begin{aligned}\ln(r_{31}) &= -3.9120 + .6931(0) + .4054(0) + .8109(1) + 1.5040(0) + 2.1972(0) \\ &= -3.9120 + 0 + .8109 \\ &= -3.1011\end{aligned}$$

or

$$\begin{aligned}\text{rate}(r_{31}) &= e^{-3.1011} \\ &= .04500\end{aligned}$$

$$\begin{aligned}\text{rate}(r_{32}) &= e^{-2.4081} \\ &= .0900\end{aligned}$$



# A Multi-Parameter Example

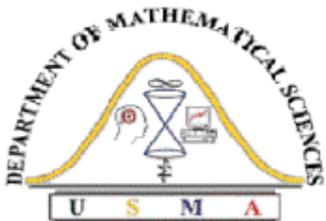


If we want to compare the two groups (Group 2 / Group 1):

The *standardized mortality ratio* (SMR):

$$\begin{aligned} SMR &= \frac{r_{32}}{r_{31}} = \frac{.0900}{.0450} \\ &= 2.0 \end{aligned}$$

This implies that the expected number of deaths in Group 2 (Strata 3) are double those in Group 1 (Strata 3)



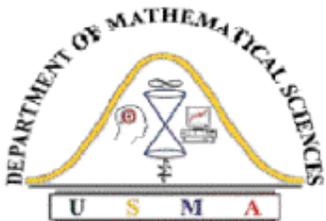
# A CDAS Case Study



## The Incidence of Shoulder Dislocation in Active Duty U.S. Military Personnel

**Background:** While shoulder dislocation is a common injury, few studies have determined the incidence rate among particular populations. We (KAH) are seeking to determine the incidence of shoulder dislocations among active duty U.S. service members.

**Objectives:** Determine if the incidence of shoulder dislocation vary significantly between males and females while controlling for demographic factors



# Collecting the Data



## Method:

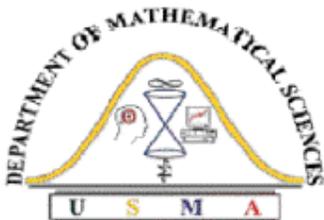
### Using the Defense Medical Epidemiology Database (DMED):

- A search was performed for International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code 831.00 (shoulder dislocation) among all service members between 1998-2006.
- The 831.00 code was further stratified by selection of only primary dislocation diagnosis in ambulatory patients.
- DMED was queried for this injury among the following demographic parameters:

Study Observation Categories and Sub-Categories				
Gender	Race	Age	Service	Rank
Male	White	< 20	Army	E1 - E4
Female	Black	20 - 24	Air Force	E5 - E9
		25 - 29	Navy	O1 - O3
		30 - 34	Marines	O4 - O9
		35 - 39		
		> 40		

### Study Population Consisted of :

- ✓ 11,680,893 man-year observations
- ✓ 19,730 cases of shoulder dislocation



# The First Model

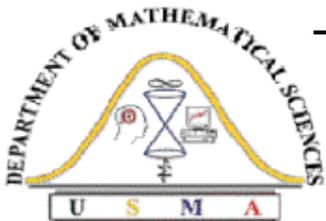


## Full 20 parameter Model (after collapsing Years):

	Shoulder Dislocations	Coefficient	[ 95% Conf. Interval ]		Std. Err	Chi-Square	P > ChiSq
Gender	Female	-0.6702	-0.7237	-0.6168	0.0446	603.53	< .0001
	Male	0	0	0			
Race	Black	-0.2239	-0.2626	-0.1851	0.0198	128.28	< .0001
	Other	-0.1385	-0.1872	-0.0897	0.0249	30.95	< .0001
	White	0	0	0			
Age	< 20	0.5559	0.4659	0.6460	0.046	146.37	< .0001
	20 - 24	0.5024	0.4218	0.5830	0.0411	149.31	< .0001
	25 - 29	0.4658	0.3887	0.5429	0.0393	140.24	< .0001
	30 - 34	0.2856	0.2065	0.3647	0.0403	50.1	< .0001
	35 - 39	0.1324	0.0527	0.2122	0.0407	10.59	0.0011
	> 40	0	0	0			
Service	Air Force	0.1454	0.0984	0.1924	0.024	36.74	< .0001
	Army	0.7511	0.7122	0.7899	0.1098	1434.82	< .0001
	Marines	0.5686	0.5205	0.6167	0.0245	536.91	< .0001
	Navy	0	0	0			
Rank	E1 - E4	0.4308	0.3351	0.5266	0.0489	77.74	< .0001
	E5 - E9	0.1823	0.0952	0.2694	0.0444	16.82	< .0001
	O1 - O3	0.0068	-0.0942	0.1077	0.0515	0.02	0.8957
	O4 - O9	0	0	0			
Constant	Intercept	-7.3668	-7.4541	-7.2794	0.0446	27300.4	< .0001

Why the Zeros?

Not Good



# Refining the Model



## Reduced 19 parameter Model (after collapsing Officer):

	Shoulder Dislocations	Coefficient	[ 95% Conf. Interval ]		Std. Err	Chi-Square	P > ChiSq
Gender	Female	-0.6702	-0.7237	-0.6167	0.0273	603.51	< .0001
	Male	0	0	0			
Race	Black	-0.2238	-0.2625	-0.1851	0.0198	128.29	< .0001
	Other	-0.1384	-0.1872	-0.0896	0.0249	30.94	< .0001
	White	0	0	0			
Age	< 20	0	0	0			
	20 - 24	-0.0535	-0.0998	-0.0073	0.0236	5.51	0.0232
	25 - 29	-0.0901	-0.1467	-0.0336	0.0289	9.75	0.0018
	30 - 34	-0.2705	-0.3390	-0.2020	0.0349	59.91	< .0001
	35 - 39	-0.4244	-0.4994	-0.3494	0.0383	123.04	< .0001
	> 40	-0.558	-0.6427	-0.4733	0.0432	166.58	< .0001
Service	Air Force	-0.6056	-0.6448	-0.5665	0.0200	917.78	< .0001
	Army	0	0	0			
	Marines	-0.1825	-0.2227	-0.1423	0.0205	79.14	< .0001
	Navy	-0.7511	-0.7899	-0.7122	0.0198	1434.81	< .0001
Rank	Enlisted	0	0	0			
	NCO	-0.2481	-0.2912	-0.2051	0.022	127.61	< .0001
	Officer	-0.4257	-0.4826	-0.3697	0.0291	214.72	< .0001
Constant	Intercept	-5.6289	-5.6734	-5.5844	0.0227	61427.2	< .0001

This generated some suspicions about the significance of age.

Testing for influence of each category variable in explaining risk of injury

Remind me to mention interactions

	Model	Log - Likelihood	Likelihood Ratio Test	Degrees of Freedom	p - value
	Full Model	89,080.72		13	
Variable Removed	Rank	88,960.45	240.54	11	< .0001
	Service	88,095.99	1,969.45	10	< .0001
	Age	88,939.02	283.39	8	< .0001
	Race	89,006.50	148.43	11	< .0001
	Gender	88,718.53	724.38	12	< .0001



# Rates and Ratios (an example)



Estimate the shoulder dislocation rate for a less than 20 year old white female junior enlisted in the Army (and her male counterpart):

## Female

$$\ln(r[< 20, white, female, Army, JE]) = -5.6289 + 0 + 0 - .6702 + 0 + 0 = -6.2991$$

$$r(< 20, white, female, Army, JE) = e^{-6.9197} = .00183796$$

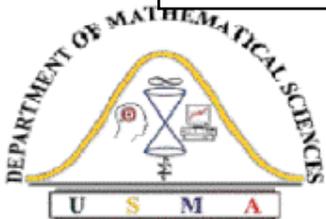
$$r(< 20, white, female, Army, JE) = 1.83796 \text{ shoulder dislocations / 1,000 person - years}$$

## Male

$$\ln(r[< 20, white, male, Army, JE]) = -5.6289 + 0 + 0 + 0 + 0 + 0 = -5.6289$$

$$r(< 20, white, male, Army, JE) = e^{-5.6289} = .00359252$$

$$r(< 20, white, male, Army, JE) = 3.59252 \text{ shoulder dislocations / 1,000 person - years}$$



# Rates and Ratios (an example)



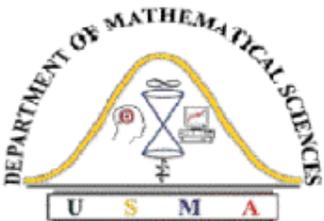
Estimate the SMR (Female/Male) for shoulder dislocation rate for a less than 20 year old white female (and male) junior enlisted in the Army:

$$\begin{aligned} SMR &= \frac{1.8380}{3.5925} \\ &= .5166 \end{aligned}$$

Reversing the ratio (Male/Female)  
provides an SMR of 1.9546

With a SMR = .5116 and a 95% Confidence Interval of [.4850, .5397], we can say that:

< 20 year old junior enlisted Army women are generally less likely to suffer an ICD-9 code 831.00 shoulder dislocation than their male counterparts.



# Always Check for Interaction



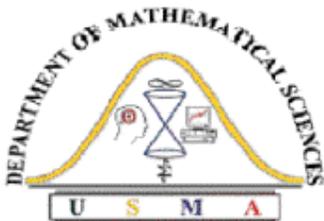
We (At Krista's insistence) conducted additional analysis to check for interaction among the five main effects and unfortunately found several (9) significant interactions:

Significant Interaction Pairs		
1	Gender	Race
2	Gender	Rank
3	Age	Rank
4	Service	Rank
5	Age	Service
6	Race	Service
7	Gender	Service
8	Race	Age
9	Gender	Age

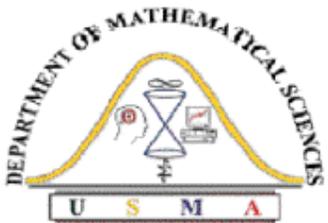
Without interaction terms,  
(male/female) IRR was 1.9546

Comparison of Main Effects with a Significant Difference				
Comparison		IRR	95% CI	
Male	Female	1.8852	1.7113	2.0766
White	Black	1.3780	1.2775	1.4864
White	Other	1.2761	1.1476	1.4191
JE	NCO	1.4367	1.1207	1.8418
JE	OFF	1.4747	1.4260	1.9032
Army	AF	1.9377	1.7739	2.1167
Marines	AF	1.6608	1.4464	1.9071
Army	Marines	1.1667	1.0295	1.3223
Army	Navy	2.0000	1.8338	2.1812
Marines	Navy	1.7142	1.4944	1.9663
20 - 24	25 - 29	1.1274	1.0217	1.2440
20 - 24	30 - 34	1.1976	1.0563	1.3578
20 - 24	35 - 39	1.4553	1.2371	1.7118
20 - 24	> 40	1.6368	1.0797	2.4814
25 - 29	35 - 39	1.2908	1.1010	1.5133
30 - 34	35 - 39	1.2151	1.0247	1.4409
< 20	35 - 39	1.5025	1.1760	1.9195

We updated the Poisson model with these nine pairs of interactions and recalculated the shoulder dislocation incidence ratios



# Questions?



# Simple Linear Regression



Models how mean expected value of a *continuous response variable* depends on a set of explanatory variables.

$$E(Y_i) = \beta_0 + \beta x_i + \varepsilon_i$$

It's three components are:

1. **Random Component** –  $Y$  is a response variable and has a normal distribution, and generally we assume  $e_i \sim N(0, \sigma^2)$
2. **Systematic Component** -  $X$  is the explanatory variable (can be continuous, discrete, or both) and are linear in the parameters  $\beta_0 + \beta x_i$
3. **Link Function,  $\eta$  or  $g(\mu)$**  - Identity Link  $\eta = g(E(Y_i)) = E(Y_i)$

