



# STAT 426

**Author:** Wenxiao Yang

**Institute:** Department of Mathematics, University of Illinois at Urbana-Champaign

*All models are wrong, but some are useful.*

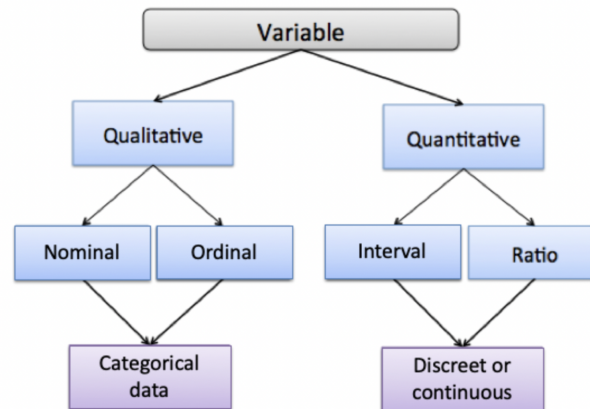
# Contents

<b>Chapter 1 Basic of Categorical Data</b>	<b>1</b>
1.1 Variable Measurement . . . . .	1
1.2 Statistical Inference for Categorical Data . . . . .	2
1.2.1 Maximum likelihood Estimation (MLE) . . . . .	2
1.2.2 Likelihood Inference (Wald, Likelihood-Ratio, Score) . . . . .	3
<b>Chapter 2 Association in Contingency Tables</b>	<b>6</b>
2.1 Association in Two-Way Contingency Tables . . . . .	6
2.1.1 Distribution . . . . .	6
2.1.2 Descriptive Statistics . . . . .	6
2.1.3 Sampling Models (Examples) . . . . .	7
2.1.4 Independent / Homogeneity . . . . .	7
2.1.5 Measuring Inhomogeneity . . . . .	7
2.1.6 Delta Method . . . . .	9
2.1.7 Testing Independence: $X^2$ and $G^2$ Test . . . . .	9
2.1.8 Testing Independence: Fisher's Exact Test . . . . .	11
2.2 Conditional Association in Three-Way Tables . . . . .	12
2.2.1 Conditional Association . . . . .	12
2.2.2 Simpson's Paradox . . . . .	13
2.2.3 Conditional Independence, Marginal Independence . . . . .	13
2.2.4 Homogeneous Association . . . . .	13
<b>Chapter 3 Generalized Linear Models</b>	<b>14</b>
3.1 Introduction . . . . .	14
3.1.1 Definition . . . . .	14
3.1.2 Fitting GLMs . . . . .	15
3.2 Binary and Binomial Responses . . . . .	15
3.2.1 Binary Regression . . . . .	15
3.2.2 Grouped Data: Binomial Response . . . . .	16
3.3 Count Responses . . . . .	17

3.3.1	Poisson Regression . . . . .	17
3.3.2	Rate Models . . . . .	18
3.4	Coefficient and Model Inferences . . . . .	19
3.4.1	Wald Inference . . . . .	19
3.4.2	Likelihood-Ratio Test . . . . .	21

# Chapter 1 Basic of Categorical Data

## 1.1 Variable Measurement



**Figure 1.1:** Variable Type

- a) Nominal: Categories do not have a natural order. Ex. blood type, gender.
- b) Ordinal: Categories have a natural order. Ex. low/middle/high education level
- c) Interval: There is a numerical distance (difference between two different values is meaningful) between any two values. Ex. blood pressure level, 100 blood pressure doesn't mean the double degree of 50 pressure.
- d) Ratio: An interval variable where ratios are valid (presence of absolute zero, i.e. zero is meaningful). Ex. weight, 4g is double degree of 2g, distance run by an athlete.

### Levels of measurements

A variable's level of measurement determines the statistical methods to be used for its analysis.

Variables hierarchy: Ratio > Interval > Ordinal > Nominal

Statistical methods applied to variables at a lower level can be used with variables at a higher level, but the contrary is not true.

## 1.2 Statistical Inference for Categorical Data

There is a distribution  $F(\beta)$  with p.d.f. (p.m.f.)  $f(x | \beta)$ , where  $\beta$  a generic unknown parameter and  $\hat{\beta}$  the parameter estimate.

### 1.2.1 Maximum likelihood Estimation (MLE)

Given a set of observations  $\vec{x} = (x_1, \dots, x_n)$ , the likelihood function of these observations with parameter  $\beta$  is  $l(\vec{x} | \beta)$ . We want to find parameter  $\hat{\beta}$  that maximizes the likelihood function,

$$\hat{\beta} = \arg \max_{\beta} l(\vec{x} | \beta)$$

which is also equivalent to maximizing the logarithm of the likelihood function  $L(\vec{x} | \beta) = \log(l(\vec{x} | \beta))$ ,

$$\hat{\beta} = \arg \max_{\beta} L(\vec{x} | \beta)$$

#### Definition 1.1 (score function)

The score function is

$$u(\beta, \vec{x}) = \nabla_{\beta} L(\vec{x} | \beta) = \frac{\nabla_{\beta} l(\vec{x} | \beta)}{l(\vec{x} | \beta)}$$



#### Lemma 1.1 (mean of score function)

The mean of score function is 0,

$$\mathbb{E}_{\vec{x}} u(\beta, \vec{x}) = 0$$



#### Proof 1.1

$$\begin{aligned} \mathbb{E}_{\vec{x}} u(\beta, \vec{x}) &= \int_{\vec{x}} l(\vec{x} | \beta) \frac{\nabla_{\beta} l(\vec{x} | \beta)}{l(\vec{x} | \beta)} d\vec{x} \\ &= \int_{\vec{x}} \nabla_{\beta} l(\vec{x} | \beta) d\vec{x} \\ &= \nabla_{\beta} \left( \int_{\vec{x}} l(\vec{x} | \beta) d\vec{x} \right) \\ &= \nabla_{\beta} 1 = 0 \end{aligned}$$

#### Lemma 1.2 (variance of score function)

The variance of the score function is

$$\text{Var}_{\vec{x}}(u(\beta, \vec{x})) = \mathbb{E}_{\vec{x}} (u(\beta, \vec{x}) u(\beta, \vec{x})^T)$$



#### Proof 1.2

Prove by the zero mean.

**Definition 1.2 (Fisher information)**

The (Fisher) information is

$$\iota(\beta) = -\mathbb{E}_{\vec{x}} [\nabla_{\beta}^2 L(\vec{x} | \beta)]$$

**Lemma 1.3**

The Fisher information is equal to the variance of score function.

$$\text{Var}_{\vec{x}}(u(\beta, \vec{x})) = \mathbb{E}_{\vec{x}}(u(\beta, \vec{x})u(\beta, \vec{x})^T) = -\mathbb{E}_{\vec{x}}[\nabla_{\beta}^2 L(\vec{x} | \beta)] = \iota(\beta)$$

**Proof 1.3**

$$\mathbb{E}_{\vec{x}}[\nabla_{\beta}^2 L(\vec{x} | \beta)] = \mathbb{E}_{\vec{x}}\left(\frac{\partial \frac{\nabla_{\beta} l(\vec{x} | \beta)}{l(\vec{x} | \beta)}}{\partial \beta}\right) = \mathbb{E}_{\vec{x}}\left(\frac{\nabla_{\beta}^2 l(\vec{x} | \beta)}{l(\vec{x} | \beta)} - \frac{\nabla_{\beta} l(\vec{x} | \beta) \nabla_{\beta} l(\vec{x} | \beta)^T}{(l(\vec{x} | \beta))^2}\right)$$

where  $\mathbb{E}_{\vec{x}}\left(\frac{\nabla_{\beta}^2 l(\vec{x} | \beta)}{l(\vec{x} | \beta)}\right) = \int_{\vec{x}} l(\vec{x} | \beta) \frac{\nabla_{\beta}^2 l(\vec{x} | \beta)}{l(\vec{x} | \beta)} d\vec{x} = \int_{\vec{x}} \nabla_{\beta}^2 l(\vec{x} | \beta) d\vec{x} = \nabla_{\beta}^2 \int_{\vec{x}} l(\vec{x} | \beta) d\vec{x} = \nabla_{\beta}^2 1 = 0$

Hence,

$$\mathbb{E}_{\vec{x}}[\nabla_{\beta}^2 L(\vec{x} | \beta)] = -\mathbb{E}_{\vec{x}}\left(\frac{\nabla_{\beta} l(\vec{x} | \beta) \nabla_{\beta} l(\vec{x} | \beta)^T}{(l(\vec{x} | \beta))^2}\right) = -\mathbb{E}_{\vec{x}}(u(\beta, \vec{x})u(\beta, \vec{x})^T)$$

**Proposition 1.1**

When the sample  $x$  is made up of i.i.d. observations, the covariance matrix of the maximum likelihood estimator  $\hat{\beta}$  is approximately equal to the inverse of the information matrix.

$$\text{Cov}(\hat{\beta}) \approx (\iota(\beta))^{-1}$$



Hence, the covariance matrix can be estimated as  $(\iota(\hat{\beta}))^{-1}$ . Similarly, SE is estimated by  $\sqrt{(\iota(\hat{\beta}))^{-1}}$ .

**1.2.2 Likelihood Inference (Wald, Likelihood-Ratio, Score)**

We want to test

$$H_0 : \beta = \beta_0 \quad H_a : \beta \neq \beta_0$$

or form a confidence interval (CI) for  $\beta$ .

**Definition 1.3 (Wald Test)**

The Wald statistic:

$$z_W = \frac{\hat{\beta} - \beta_0}{SE} = \frac{\hat{\beta} - \beta_0}{\sqrt{(\iota(\hat{\beta}))^{-1}}}$$

where  $SE = \sqrt{(\iota(\hat{\beta}))^{-1}}$ .

Usually, as  $n \rightarrow \infty$ ,  $z_W \xrightarrow{d} N(0, 1)$  under  $H_0 : \beta = \beta_0$ .



(1) We reject the  $H_0$  if  $|z_W| \geq z_{\frac{\alpha}{2}}$  for a two-sided level  $\alpha$  test.

(2) The  $(1 - \alpha)100\%$  Wald (confidence) interval is

$$\{\beta_0 : |z_W| = \frac{|\hat{\beta} - \beta_0|}{SE} < z_{\frac{\alpha}{2}}\} = (\hat{\beta} - z_{\frac{\alpha}{2}} SE, \hat{\beta} + z_{\frac{\alpha}{2}} SE)$$

(3) The Wald test also has a chi-squared form, using

$$z_W^2 = \frac{(\hat{\beta} - \beta_0)^2}{(\iota(\hat{\beta}))^{-1}} \sim \chi_1^2 \quad (\text{under } H_0)$$



#### Definition 1.4 (Likelihood Ratio Test)

Let

$$\Lambda = \frac{l(\vec{x} | \beta_0)}{l(\vec{x} | \hat{\beta})}$$

where  $l(\vec{x} | \hat{\beta}) = \max_{\beta} l(\vec{x} | \beta)$ , so the ratio  $\Lambda \in [0, 1]$ .

The **likelihood-ratio test (LRT) chi-squared statistic**:

$$-2 \ln \Lambda = -2 \left( L(\beta_0) - L(\hat{\beta}) \right)$$

It has an approximate  $\chi_1^2$  distribution under  $H_0 : \beta = \beta_0$ , and otherwise tends to be larger.

(1) Thus, reject  $H_0$  if

$$-2 \ln \Lambda \geq \chi_1^2(\alpha)$$

(2) The  $(1 - \alpha)100\%$  likelihood-ratio (confidence) interval is

$$\{\beta_0 : -2 \ln \Lambda = -2 \left( L(\beta_0) - L(\hat{\beta}) \right) < \chi_1^2(\alpha)\}$$

Unlike Wald, this interval is not degenerate. (i.e., For general case, the interval does not have an explicit form.)



#### Definition 1.5 (Score Test)

The **score statistic**:

$$z_S = \frac{u(\beta_0)}{\sqrt{\iota(\beta_0)}}$$

As  $n \rightarrow \infty$ ,  $z_S \xrightarrow{d} N(0, 1)$  under  $H_0 : \beta = \beta_0$ . Otherwise, it tends to be further from zero.

(1) Thus, reject  $H_0$  if  $|z_S| \geq z_{\frac{\alpha}{2}}$  for a two-sided level  $\alpha$  test.

(2) The  $(1 - \alpha)100\%$  score (confidence) interval is

$$\{\beta_0 : |z_S| = \frac{|u(\beta_0)|}{\sqrt{\iota(\beta_0)}} < z_{\frac{\alpha}{2}}\}$$

Unlike Wald, it is not degenerate for some distributions.

(3) *There is also a chi-squared form:*

$$z_S^2 = \frac{u(\beta_0)^2}{\iota(\beta_0)} \sim \chi_1^2 \quad (\text{under } H_0)$$



We can also use P-value to measure the probability of the statistic is more extreme under the  $H_0$ . We can reject  $H_0$  if the P-value is  $\leq \alpha$ .

All three kinds tend to be “asymptotically equivalent” as  $n \rightarrow \infty$ . For smaller  $n$ , the likelihood-ratio and score methods are preferred.



## Chapter 2 Association in Contingency Tables

### 2.1 Association in Two-Way Contingency Tables

Consider joint observations of two categorical variables:  $X$  with  $I$  categories,  $Y$  with  $J$  categories.

We can summarize data in an  $I \times J$  **contingency table**:

		Y		
		1	...	J
X	1			
	$\vdots$			
	I			

Each **cell** contains a count  $n_{ij}$ .

#### 2.1.1 Distribution

If both  $X$  and  $Y$  are random, let

$$\pi_{ij} = P(X \text{ in row } i, Y \text{ in col } j)$$

be the **joint** distribution of  $X$  and  $Y$ .

The **marginal** distribution of  $X$  is defined by

$$\pi_{i+} = P(X \text{ in row } i)$$

and similarly for  $Y$ :

$$\pi_{+j} = P(Y \text{ in col } j)$$

The **conditional** distribution of  $Y$  given that  $X$  is in row  $i$  is defined by

$$\pi_{j|i} = P(Y \text{ in col } j \mid X \text{ in row } i) = \frac{\pi_{ij}}{\pi_{i+}}$$

#### 2.1.2 Descriptive Statistics

Let  $n_{ij}$  = count in row  $i$  and col  $j$  and  $n = \sum_i \sum_j n_{ij}$ .

The **margins** of the table:

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}$$

## Natural Estimation

1. Natural estimate of  $\pi_{ij}$ :  $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$
2. Similarly marginals:  $\hat{\pi}_{i+} = \sum_j \hat{\pi}_{ij} = \frac{n_{i+}}{n}$ ;  $\hat{\pi}_{+j} = \sum_i \hat{\pi}_{ij} = \frac{n_{+j}}{n}$
3. And conditionals:  $\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}} = \frac{n_{ij}}{n_{i+}}$

### 2.1.3 Sampling Models (Examples)

Possible joint distributions for counts in  $I \times J$  table:

1. Poisson (random total):  $Y_{ij}$  = count in cell  $(i, j)$ ,

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

and the  $Y_{ij}$ s are independent.

2. Multinomial (fixed total  $n$ ):  $N_{ij}$  = count in cell  $(i, j)$ ,

$$\{N_{ij}\} \sim \text{multinomial}(n, \{\pi_{ij}\})$$

3. Independent Multinomial: Assume  $n_{i+}$  (row totals  $n_i$ ) are fixed,

$$\left. \begin{aligned} \{N_{1j}\}_{j=1}^J &\sim \text{multinomial}(n_1, \{\pi_{j|1}\}_{j=1}^J) \\ &\vdots \\ \{N_{Ij}\}_{j=1}^J &\sim \text{multinomial}(n_I, \{\pi_{j|I}\}_{j=1}^J) \end{aligned} \right\}$$

(When  $J = 2$ , this is independent binomial sampling, for which we may just write  $\pi_i$  for  $\{\pi_{1|i}, \pi_{2|i}\}$ .)

### 2.1.4 Independent / Homogeneity

#### Definition 2.1 (independent)

If both  $X$  and  $Y$  are random, they are **independent** if

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \forall i, j$$

which implies  $\pi_{j|i} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}, \forall i, j$ . That is,  $\pi_{j|i}$  doesn't depend on  $i$  and is the same as the marginal distribution of  $Y$ . (Intuitively, knowing  $X$  tells nothing about  $Y$ .)



#### Definition 2.2 (homogeneity)

Even if  $X$  is not really random, the condition that  $\pi_{j|i} = \pi_{+j}, \forall i, j$  is called **homogeneity**. This might still be relevant in a situation where  $X$  is deliberately chosen and  $Y$  is observed as a response.



### 2.1.5 Measuring Inhomogeneity

Homogeneity is the condition  $\pi_1 = \pi_2$ . We can measure inhomogeneity by three different measures:

$n_{11}$	$n_{12}$
$n_{21}$	$n_{22}$

$Y_1$	$n_1 - Y_1$
$Y_2$	$n_2 - Y_2$

where  $Y_i \sim \text{indep. binomial}(n_i, \pi_i)$ . This regards row totals as fixed.

### 1. difference of proportions:

$$\pi_1 - \pi_2$$

The estimation is

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}$$

The approx  $(1 - \alpha)100\%$  confidence interval is:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_2}}$$

(Problematic if  $\pi_1$  and  $\pi_2$  are near 0 or 1.)

### 2. relative risk:

$$RR = \frac{\pi_1}{\pi_2}$$

The estimation is

$$r = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{y_1/n_1}{y_2/n_2}$$

The approx  $(1 - \alpha)100\%$  confidence interval of  $\ln RR$  is:

$$\ln r \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1 - \hat{\pi}_1}{y_1} + \frac{1 - \hat{\pi}_2}{y_2}}$$

### 3. odds ratio:

$$\theta = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}$$

When  $\theta = 1$ , we can say there is no association.

The **odds** for a probability  $\pi$  is  $\Omega = \frac{\pi}{1 - \pi}$ . Note  $\pi = \frac{\Omega}{1 + \Omega}$ .

(In the multinomial model:  $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$  ("cross-product ratio"); in Poisson model:  $\theta = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}$ )

The usual (unrestricted) estimates

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The approx  $(1 - \alpha)100\%$  confidence interval for  $\ln \theta$  is

$$\ln \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Useful properties of odds ratio:

- (1) Interchanging rows (or cols) changes  $\theta$  to  $\frac{1}{\theta}$ .
- (2) Interchanging  $X$  and  $Y$  doesn't change  $\theta$ .
- (3) Multiplying a row (or col) by a factor doesn't change  $\hat{\theta}$ .
- (4) Relationship to relative risk:  $\theta = RR \cdot \frac{1-\pi_2}{1-\pi_1}$ . ( $\theta$  and  $RR$  are similar if both  $\pi_1$  and  $\pi_2$  are small.)

### 2.1.6 Delta Method

It is easy to obtain approximate CI for a mean based on a sample mean by using the Central Limit Theorem and a consistent estimate of standard error.

But the log Odds Ratio and log Relative Risk are transformed means. How were their CI's derived? And why take logs?

Suppose a statistic  $T_n$  and parameter  $\theta$  such that

$$\sqrt{n}(T_n - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, \sigma^2)$$

(e.g.  $T_n$  might be a sample mean from a sample of size  $n$  with population mean  $\theta$  and variance  $\sigma^2$ )

We want a CI for  $g(\theta)$ , for some smooth  $g$ .

The Taylor expand at  $T_n$  is

$$g(\theta) \approx g(T_n) + g'(T_n)(\theta - T_n)$$

So,

$$\sqrt{n}(g(T_n) - g(\theta)) \approx g'(T_n)\sqrt{n}(T_n - \theta) \xrightarrow[n \rightarrow \infty]{d} N(0, (g'(T_n))^2 \sigma^2)$$

(This is useful only if  $g'(T_n) \neq 0$ ) Hence, when  $n$  is large

$$\sqrt{n} \frac{g(T_n) - g(\theta)}{|g'(T_n)|\sigma} \sim N(0, 1)$$

which suggests this approximate CI for  $g(\theta)$ :

$$g(T_n) \pm z_{\frac{\alpha}{2}} \frac{|g'(T_n)|\sigma}{\sqrt{n}}$$

### 2.1.7 Testing Independence: $X^2$ and $G^2$ Test

Let  $\mu_{ij} = \mathbb{E}(N_{ij}) = n\pi_{ij}$ . Under  $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \forall i, j$

$$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

Under  $H_0$ , can show the MLEs are

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n \left( \frac{n_{i+}}{n} \right) \left( \frac{n_{+j}}{n} \right) = \frac{n_{i+}n_{+j}}{n}$$

(assuming no empty rows or cols)

**Residuals:**

1. Raw:  $n_{ij} - \hat{\mu}_{ij}$
2. Pearson:  $e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$ .  $X^2 = \sum_i \sum_j e_{ij}^2$ .
3. Standardized:  $r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - \hat{\pi}_{i+})(1 - \hat{\pi}_{+j})}}$

**Usage:** Look for Pearson or standardized residuals with absolute value *exceeding 2 or 3*. These suggest the reason for significant dependence.

**Remark:** Under independence, both Pearson and standardized residuals are asymp. normal, but only standardized has asymp. variance equal to 1.

**Definition 2.3 ( $X^2$  Test: Pearson  $\chi^2$  Test (Score Test))**

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \underset{H_0}{\sim} \chi^2_{(I-1)(J-1)}$$

*Note:*

$$\begin{aligned} (I-1)(J-1) &= (IJ-1) - ((I-1) + (J-1)) \\ &= \text{total \# params.} - \# \text{ params. under } H_0 \end{aligned}$$

*Reject  $H_0$  if*

$$X^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

*(or use P-value)*



**Definition 2.4 ( $G^2$  Test: Likelihood Ratio  $\chi^2$  Test)**

$$G^2 = 2 \sum_{ij} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}} \underset{H_0}{\sim} \chi^2_{(I-1)(J-1)}$$

*Reject  $H_0$  if*

$$G^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

*(or use P-value)*

*(Convention:  $0 \ln 0 = 0$ )*



**Comparison:**

1.  $X^2$  and  $G^2$  are asymptotically equivalent under  $H_0$
2. The  $X^2$  tends to be better.

**Example 2.1** Testing independence is equivalent to testing homogeneity in the indep. binomial model:

$$H_0 : \pi_1 = \pi_2$$

Can show

$$X^2 = z^2$$

where

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})(1/n_1 + 1/n_2)}} \quad \hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}$$

**Remark:** The  $X^2$  and  $G^2$  tests are not necessarily compatible with the Wald CIs. For example,

$$\text{reject } H_0 \nleftrightarrow \text{odds ratio } \theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1 \text{ not in Wald CI}$$

### 2.1.8 Testing Independence: Fisher's Exact Test

When cell counts are small, the  $X^2$  and  $G^2$  independence tests are not recommended: The  $\chi^2$  approximations are poor. In this section we introduce a *Fisher's Exact Test*.

Consider a  $2 \times 2$  table with row and col totals fixed:

	Y		
	$N_{11}$	$N_{12}$	$n_{1+}$
X	$N_{21}$	$N_{22}$	$n_{2+}$
	$n_{+1}$	$n_{+2}$	$n$

**Note:** Any cell count, say  $N_{11}$ , determines the whole table.

Can show that, under  $H_0$  : independence,  $N_{11}$  is (conditionally) hypergeometric:

$$P_{H_0}(N_{11} = t) = \frac{\binom{n_{1+}}{t} \binom{n_{2+}}{n_{+1} - t}}{\binom{n}{n_{+1}}}$$

In terms of odds ratio  $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ , independence is

$$H_0 : \theta = 1$$

Possible alternatives:

$$H_\alpha : \theta > 1 \Rightarrow N_{11} \text{ tends larger}$$

$$H_\alpha : \theta < 1 \Rightarrow N_{11} \text{ tends smaller}$$

$$H_\alpha : \theta \neq 1 \Rightarrow N_{11} \text{ tends larger or smaller}$$

For  $H_\alpha : \theta > 1$ , the (one-sided)  $p$ -value is  $P_{H_0}(N_{11} \geq t_0)$ , where  $t_0 = n_{11}$  is the observed value of  $N_{11}$ .

**Remarks:** Could use mid  $p$ -values instead; Implemented in R function `fisher.test()`; Can be extended to  $I \times J$  tables (with some computational difficulty).

## 2.2 Conditional Association in Three-Way Tables

Add a third categorical variable  $Z$ .

**Example 2.2** Is a drug more effective at curing a disease among younger patients than among older?  $X$  = drug or placebo;  $Y$  = disease cured or not;  $Z$  = age group (young, old).

### 2.2.1 Conditional Association

$Z$  may be called a **stratification variable**. We are interested in the distribution of  $(X, Y)$  *conditional* on  $Z$ .

#### Definition 2.5 (partial table)

Each  $Z$  category defines a **partial table** for  $X$  and  $Y$ .



**Example 2.3** When  $Z = 1, 2$  and  $X, Y$  are binary ( $2 \times 2 \times 2$  table):

				$Y$	
				$n_{111}$	$n_{121}$
				$n_{211}$	$n_{221}$
$Z = 1 :$	$X$				

				$Y$	
				$n_{112}$	$n_{122}$
				$n_{212}$	$n_{222}$
$Z = 2 :$	$X$				

These represent **conditional associations**.

#### Definition 2.6 (marginal table)

The **marginal table** sums the partial tables:



				$Y$	
				$n_{11+}$	$n_{12+}$
				$n_{21+}$	$n_{22+}$
$X$					

This represents the **marginal association** (ignoring  $Z$ ).

In general, let  $\mu_{ijk}$  = *expected count in row  $i$ , col  $j$ , table  $k$* .

The **conditional odds ratios**,

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

which are estimated by

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

The **marginal odds ratio**

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

is estimated from the marginal table.



### 2.2.2 Simpson's Paradox

Some counter-intuitive but possible situations:

1. There are conditional associations ( $\theta_{XY(k)} \neq 1$ ) but no marginal association ( $\theta_{XY} = 1$ )
2. There is a marginal association ( $\theta_{XY} \neq 1$ ) but no conditional associations ( $\theta_{XY(k)} = 1$ )
3. **Simpson's paradox:** The conditional associations are in the opposite direction from the marginal, e.g.  
 $\theta_{XY(k)} > 1, \theta_{XY} < 1$

	Full Population, N = 52			Men (M), N = 20			Women (¬M), N = 32		
	Success (S)	Failure (¬S)	Success Rate	Success	Failure	Success Rate	Success	Failure	Success Rate
Treatment (T)	20	20	50%	8	5	≈ 61%	12	15	≈ 44%
Control (¬T)	6	6	50%	4	3	≈ 57%	2	3	≈ 40%

TABLE 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).

**Figure 2.1:** Simpson's paradox

### 2.2.3 Conditional Independence, Marginal Independence

#### Definition 2.7 (conditionally independent given Z, marginal independent)

We also call  $X$  and  $Y$  are **conditionally independent given**  $Z = k$  if  $\theta_{XY(k)} = 1$ . If this is true for all  $k$ ,  $X$  and  $Y$  are **conditionally independent given**  $Z$ . Not the same to " $X$  and  $Y$  are **marginal independent** if  $\theta_{XY} = 1$ ".



#### Proposition 2.1

For multinomial sampling, can show that conditional independence is

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad \forall i, j, k$$



### 2.2.4 Homogeneous Association

#### Definition 2.8

Let  $Z$  have  $K$  categories.  $X$  and  $Y$  have **homogeneous association** over  $Z$  if

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

(Conditional independence is a special case.)



## Chapter 3 Generalized Linear Models

### 3.1 Introduction

A linear model  $Y = \alpha + \sum_{i=1}^p \beta_i x_i + \varepsilon$  is usually not appropriate if  $Y$  is binary or a count.

#### 3.1.1 Definition

We seek to model independent observations  $Y_1, \dots, Y_n$  of a **response variable**, in terms of corresponding vectors  $\vec{x}_i = (x_{i1}, \dots, x_{ip})$ ,  $i = 1, \dots, n$  of values of  $p$  **explanatory variables**.

- (1) **Random component:** density of  $Y_i$  from a **natural exponential family**

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i))$$

where  $Q(\theta_i)$  is the **natural parameter**.

(**Fact:** Since  $Y_i$  is from a natural exponential family, its distribution is completely determined by its mean  $\mu_i$ . In particular,  $\text{Var}(Y_i)$  is a function of  $\mu_i$ .)

- (2) **Systematic component:** the **linear predictor**

$$\eta_i = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

with parameters  $\alpha, \beta_1, \dots, \beta_p$  (**coefficients**)

$Y_i$  will depend on  $\vec{x}_i$  only through  $\eta_i$ .

- (3) **Link function:** monotonic, differentiable  $g$  such that

$$g(\mu_i) = \eta_i \quad \text{where } \mu_i = \mathbb{E}(Y_i)$$

(Note: Ordinary linear models use the identity link:  $g(\mu) = \mu$ , which means  $\mu_i = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ .)

#### Definition 3.1 (Canonical Link)

The **canonical link** satisfies

$$g(\mu_i) = Q(\theta_i)$$

which means  $Q(\theta_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$ .



Let  $F$  be a continuous and invertible c.d.f. on the real line. A reasonable link might be

$$g(\pi) = F^{-1}(\pi)$$

since it transforms interval  $(0, 1)$  to the whole real line.

**Definition 3.2 (Probit Regression)**

Using the c.d.f.  $\Phi$  for a standard normal is called **probit regression**.



### 3.1.2 Fitting GLMs

Usually by maximum likelihood: find

$$\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p$$

maximizing

$$\prod_{i=1}^n f(y_i; \theta_i)$$

Explicit solutions exist only in special cases, so need numerical methods: e.g. Newton-Raphson, Fisher Scoring.

## 3.2 Binary and Binomial Responses

### 3.2.1 Binary Regression

**Example 3.1 (Binary Regression)**

$$Y_i \sim \text{Bernoulli}(\pi_i) \quad (\theta_i = \pi_i)$$

$$\begin{aligned} f(y_i; \pi_i) &= \begin{cases} 1 - \pi_i & y_i = 0 \\ \pi_i & y_i = 1 \end{cases} \\ &= \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \\ &= (1 - \pi_i) \left( \frac{\pi_i}{1 - \pi_i} \right)^{y_i} \\ &= (1 - \pi_i) \exp \left( y_i \ln \left( \frac{\pi_i}{1 - \pi_i} \right) \right) \end{aligned}$$

So  $a(\pi) = 1 - \pi$ ,  $b(y) = y$ , and

$$Q(\pi) = \ln \left( \frac{\pi}{1 - \pi} \right) = \text{logit}(\pi)$$

The natural parameter is the **log odds**.

Note:  $\mu_i = E(Y_i) = \pi_i$ . Hence, we can write  $\pi_i(\vec{x}_i)$  as a response to

- **Identity Link:**

$$\pi(\vec{x}_i) = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

- **Log Link:**

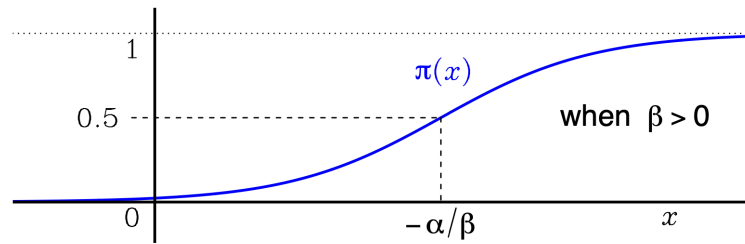
$$\ln(\pi(\vec{x}_i)) = \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

• **Canonical link: (logistic regression)**

$$\text{logit}(\pi(\vec{x}_i)) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Specifically, when  $p = 1$ ,

$$\begin{aligned} \text{logit}(\pi(x)) = \alpha + \beta x &\Leftrightarrow \text{odds}(\pi(x)) = e^{\alpha + \beta x} \\ &\Leftrightarrow \pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} \end{aligned}$$



**Figure 3.1:**  $\pi(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}$

### 3.2.2 Grouped Data: Binomial Response

If several observations have the same  $\vec{x}$  ("replications"), then they have the same  $\pi(\vec{x})$ .

Summing binary (0/1) observations with the same  $\vec{x}$  gives **grouped** data:

$$Y_i \sim \text{binomial}(n_i, \pi(\vec{x}_i))$$

where " $i$ " now refers to the  $i^{\text{th}}$  group (of  $n_i$  binary obs.).

*Note:* Both  $Y_i$  and  $n_i$  (or  $n_i - Y_i$ ) must be included in the data.

*Remarks:*

1. Whether data are grouped or ungrouped, fitting with maximum likelihood gives the same results.
2. Technically, the binomial GLM should use  $\bar{Y}_i = \frac{Y_i}{n_i}$  as the responses, and use an *exponential dispersion family form* for the density.

#### For $2 \times 2$ Tables

$x = 1$	$Y_1$	$n_1 - Y_1$
$x = 0$	$Y_2$	$n_2 - Y_2$

Note: Can regard as grouped data with two groups.

A binomial regression model (with  $x = 0$  or  $1$ ) is equivalent to the independent binomial model:

$$\left. \begin{aligned} Y_1 &\sim \text{binomial}(n_1, \pi_1 = \pi(1)) \\ Y_2 &\sim \text{binomial}(n_2, \pi_2 = \pi(0)) \end{aligned} \right\} \text{independent}$$

For logistic regression:

$$\text{logit}(\pi(x)) = \alpha + \beta x$$

so the odds ratio is

$$\begin{aligned}\theta &= \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \exp(\text{logit}(\pi_1) - \text{logit}(\pi_2)) \\ &= \exp(\alpha + \beta \cdot 1 - (\alpha + \beta \cdot 0)) = e^\beta\end{aligned}$$

So  $\beta$  is the **log odds ratio**.

### 3.3 Count Responses

For binomial data, the maximum possible count is known (for each observation). What if there are no known maximum counts? Counts of independently-occurring incidents (without any maximum) are often modeled using the Poisson distribution.

#### 3.3.1 Poisson Regression

##### Example 3.2 (Poisson Regression)

$$Y_i \sim \text{Poisson}(\mu_i) \quad (\theta_i = \mu_i)$$

Note:  $\mu_i = E(Y_i) = \text{Var}(Y_i)$

$$\begin{aligned}f(y_i; \mu_i) &= \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \\ &= e^{-\mu_i} \frac{1}{y_i!} \exp(y_i \ln \mu_i)\end{aligned}$$

So  $a(\mu) = e^{-\mu}$ ,  $b(y) = \frac{1}{y!}$

$$Q(\mu) = \ln \mu$$

The natural parameter is the log-mean.

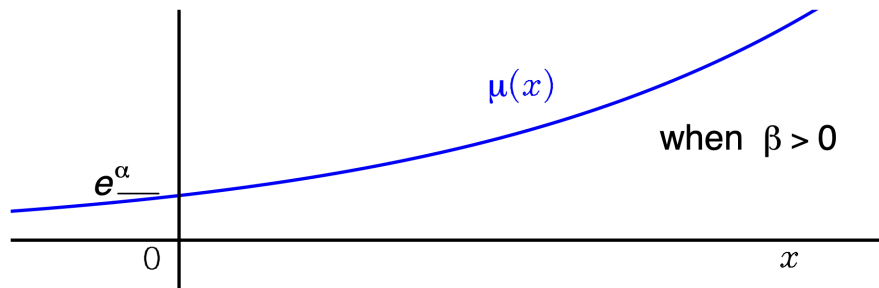
**Canonical link:**

$$\ln \mu(\vec{x}_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

which gives the **(Poisson) loglinear model**.

Specifically, when  $p = 1$ ,

$$\begin{aligned}\ln \mu(x) &= \alpha + \beta x \\ \Leftrightarrow \mu(x) &= e^{\alpha + \beta x} = e^\alpha (e^\beta)^x\end{aligned}$$



**Figure 3.2:**  $\mu(x) = e^\alpha(e^\beta)^x$

### 3.3.2 Rate Models

$E(Y_i) = \mu_i$  is sometimes expected to be proportional to another observed variable  $t_i > 0$  :

$$\mu_i = \lambda_i t_i$$

e.g.

$Y_i$  = cases of rare disease in nation  $i$

$t_i$  = national population (known)

$\lambda_i$  = disease **rate** (unknown)

( $t$  could alternatively be a temporal or spatial extent)

**Canonical link:**

$$\begin{aligned} \ln \mu_i &= \ln \lambda_i + \ln t_i \\ &= \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \ln t_i \end{aligned}$$

where  $\lambda_i$  works as linear predictor,  $\ln \lambda_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

Note:  $\ln t_i$  has no coefficient. We call  $\ln t_i$  an **offset**.

### For $2 \times 2$ Tables

	$x_2 = 1$	$x_2 = 0$
$x = 1$	$Y_{11}$	$Y_{12}$
$x = 0$	$Y_{21}$	$Y_{22}$

$$\{Y_{ij}\} \sim \text{indep. Poisson } (\{\mu_{ij}\})$$

The full loglinear regression model can be parameterized as

$$\ln \mu_{ij} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

(Can solve for  $\alpha, \beta_1, \beta_2, \beta_3$  in terms of the  $\mu_{ij}$ .)

Recall relation to multinomial:

$$\{Y_{ij}\} \mid \sum_{ij} Y_{ij} = n \sim \text{multinomial}(n, \{\pi_{ij}\})$$

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}}$$

Recall odds ratio:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}$$

Can show  $\theta = 1$  (i.e., no association) is equivalent to  $\beta_3 = 0$  (i.e., no interaction term):

$$\ln \mu_{ij} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

## 3.4 Coefficient and Model Inferences

### Matrix Forms

We can write the **linear predictor** of GLM

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, \dots, N$$

in vector form:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

where  $\boldsymbol{\eta} = [\eta_1, \dots, \eta_N]^T$ ,  $\boldsymbol{\beta} = [\alpha, \beta_1, \dots, \beta_p]^T$ , and the model matrix  $\mathbf{X}$  has  $i^{\text{th}}$  row  $[1, x_{i1}, \dots, x_{ip}]$ .

Let the MLE of  $\boldsymbol{\beta}$  be

$$\hat{\boldsymbol{\beta}} = [\hat{\alpha}, \hat{\beta}_1, \dots, \hat{\beta}_p]^T$$

### 3.4.1 Wald Inference

#### 3.4.1.1 (Fisher) Information Matrix

##### Definition 3.3 ((Fisher) Information Matrix)

The (Fisher) information matrix for  $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$  is a  $(p+1) \times (p+1)$  matrix

$$\mathcal{J}$$

with element  $(h, j)$  being

$$\mathbb{E} \left( -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_h \partial \beta_j} \right)$$



For a GLM, the information matrix becomes

$$\mathcal{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$



where  $\mathbf{W} = \text{diag}(w_1, \dots, w_N)$  with

$$w_i = \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \cdot \frac{1}{\text{var}(Y_i)}$$

Recall  $\mu_i = \mathbb{E}(Y_i)$ .

### Example 3.3 Logistic Regression

$$\mu_i = E(Y_i) = n_i \pi_i \quad \text{var}(Y_i) = n_i \pi_i (1 - \pi_i)$$

$$\eta_i = \text{logit}(\pi_i) = \ln \pi_i - \ln(1 - \pi_i)$$

Then

$$\begin{aligned} \frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \pi_i} \cdot \frac{\partial \pi_i}{\partial \eta_i} = n_i \cdot \left( \frac{\partial \eta_i}{\partial \pi_i} \right)^{-1} \\ &= n_i \left( \frac{1}{\pi_i} + \frac{1}{1 - \pi_i} \right)^{-1} = n_i \pi_i (1 - \pi_i) \end{aligned}$$

Thus

$$w_i = (n_i \pi_i (1 - \pi_i))^2 \cdot \frac{1}{n_i \pi_i (1 - \pi_i)} = n_i \pi_i (1 - \pi_i)$$

#### 3.4.1.2 Wald Inference

Under regularity conditions, as  $N \rightarrow \infty$ , the distribution of  $\hat{\beta}$  is approximately multivariate normal with mean vector  $\beta$  and covariance matrix  $\mathcal{J}^{-1}$  (a proposition in MLE part):

$$\hat{\beta} \sim N(\beta, \mathcal{J}^{-1})$$

So the asymptotic covariance of  $\hat{\beta}$  is

$$\mathcal{J}^{-1} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$

which is estimated as

$$\widehat{\text{cov}}(\hat{\beta}) = (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}$$

where  $\hat{\mathbf{W}}$  is  $\mathbf{W}$  estimated using  $\hat{\beta}$  for  $\beta$ .

In particular, the element  $\hat{\beta}_j$  of  $\hat{\beta}$  is asymptotically normal with asymptotic variance

$$\widehat{\text{var}}(\hat{\beta}_j) = (j+1) \text{ st diagonal element of } \widehat{\text{cov}}(\hat{\beta})$$

The Wald  $z$  statistic for testing  $H_0 : \beta_j = \beta_{j0}$  is

$$z_W = \frac{\hat{\beta}_j - \beta_{j0}}{SE(\hat{\beta}_j)} \underset{H_0}{\sim} N(0, 1)$$

where  $SE(\hat{\beta}_j) = \sqrt{\widehat{\text{var}}(\hat{\beta}_j)}$ .

Also Wald CIs:

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot SE(\hat{\beta}_j)$$

### 3.4.2 Likelihood-Ratio Test

#### 3.4.2.1 Deviance and Goodness of Fit

Then it can be shown that  $\mu = \mathbf{y}$  maximizes  $L$ . It follows that

$$L(\mathbf{y}; \mathbf{y}) \geq L(\hat{\mu}; \mathbf{y})$$

where  $\hat{\mu}$  is the MLE of  $\mu$  (when it exists) for the GLM. The unrestricted case, in which each observation has its own mean, is called the **saturated model**.

#### Definition 3.4 (Deviance)

The *deviance* of the GLM is

$$D(\mathbf{y}; \hat{\mu}) = -2(L(\hat{\mu}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y}))$$

Note:  $D(\mathbf{y}; \hat{\mu})$  is the likelihood-ratio test (LRT) chi-squared statistic for

- $H_0$  : the GLM is correct
- $H_a$  : the GLM is incorrect (but the saturated model is correct)



The deviance is associated with degrees of freedom

$$\begin{aligned} \text{df} &= \# \text{ means in sat. model} - \# \text{ params. in GLM} \\ &= N - (p + 1) \quad (\text{usually}) \end{aligned}$$

#### 3.4.2.2 Goodness of Fit Test / Likelihood-Ratio Test

Under certain asymptotic conditions,

$$D(\mathbf{y}; \hat{\mu}) \underset{H_0}{\overset{\sim}{\sim}} \chi_{\text{df}}^2$$

and tends larger under  $H_a$ .

So reject correctness of the GLM if

$$D(\mathbf{y}; \hat{\mu}) > \chi_{\text{df}}^2(\alpha)$$

(or use a  $P$ -value).

Warning: Chi-squared approximation can be poor.

**Example 3.4 Poisson Case** For a (Poisson) loglinear model,  $L(\mu; \mathbf{y}) = \sum_i \ln \left( \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i} \right)$ , we can show

$$\begin{aligned} D(\mathbf{y}; \hat{\mu}) &= 2 \sum_i \left( y_i \ln \frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i \right) \\ &= 2 \sum_i y_i \ln \frac{y_i}{\hat{\mu}_i} \end{aligned}$$

**Remark:**

- The chi-squared approximation (under  $H_0$ ) is adequate if all  $\mu_i$  are sufficiently large.
- These formulas also apply to loglinear rate models (with rate variable  $t_i$ ), for which  $\mu_i = \lambda_i t_i$ ,  $\hat{\mu}_i = \hat{\lambda}_i t_i$ , where  $\hat{\lambda}_i$  is the MLE of rate  $\lambda_i$ .

**Example 3.5 Binomial Case**  $Y_i \sim \text{binomial}(n_i, \pi_i)$