

# Intro to Categorical Data Analysis

Author: Nicholas Reich and Anna Liu

Course: Categorical Data Analysis (BIOSTATS 743)

## Where this course fits

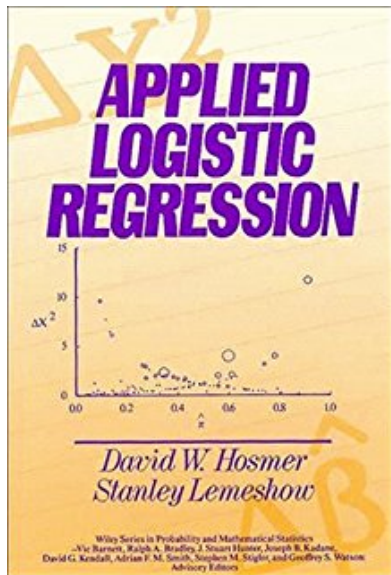
- ▶ Third course in Biostat “Methods Sequence”, after intro stats and linear regression.
- ▶ Good lead-in to random effects models, machine learning/classification models.
- ▶ Balance of traditional stat theory and application.
- ▶ Most applications will have biomedical/public health focus.

# Focus of the course

- ▶ Foundational concepts
  - ▶ Analysis of contingency tables
  - ▶ Generalized Linear Models (GLMs)
  - ▶ Discussion of Bayesian and frequentist approaches
- ▶ A taste of common, modern extensions to GLMs
  - ▶ Machine Learning classification methods
  - ▶ Longitudinal data (repeated measures)
  - ▶ Zero-inflated models, over-dispersion

# History of the course

- Taught since mid-1980s at UMass



# Course Introduction

- ▶ This course focuses on methods for categorical **response**, or **outcome** variables.
  - ▶ Binary, e.g.
  - ▶ Nominal, e.g.
  - ▶ Ordinal, e.g.
  - ▶ Discrete-valued (“interval”), e.g.
- ▶ **Explanatory**, or **predictor** variables can be any type
- ▶ Very generally, we are trying to build models

# Types of categorical variables

- ▶ The way that a variable is measured determines its classification
  - ▶ when measured as (public school, private school, home schooling), education is \_\_\_\_\_
  - ▶ when measured by highest degree attained (none, high school, barchelor's, master's, doctorate), it is \_\_\_\_\_
  - ▶ when measured by number of years of education completed using integers (0, 1, 2, 3, ...), it is \_\_\_\_\_
- ▶ The granularity of your data matters!
  - ▶ In terms of information per measured datapoint, discrete variables > ordinal variables > nominal variables
  - ▶ This has implications for study design and sample size.

## Distributions of categorical variables: Binomial

Let  $y_1, y_2, \dots, y_n$  denote observations from  $n$  **independent and identical** trials such that

$$P(Y_i = 1) = \pi \quad P(Y_i = 0) = 1 - \pi$$

The total number of successes (1s)  $Y = \sum_{i=1}^n Y_i$  has the **binomial distribution**, denoted by  $\text{bin}(n, \pi)$ . The probability mass function for the possible outcomes  $y$  for  $Y$  is

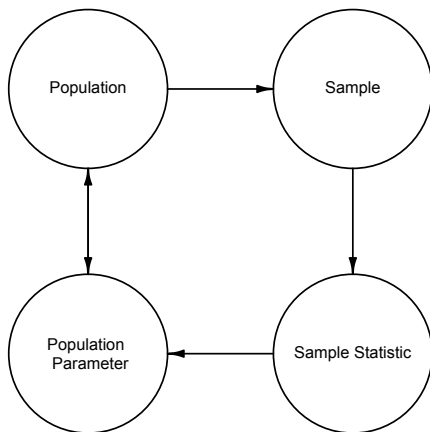
$$p(y) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}, y = 0, 1, \dots, n$$

with  $\mu = E(Y) = n\pi$  and  $\sigma^2 = \text{Var}(Y) = n\pi(1 - \pi)$ .

- ▶ The binomial distribution converges to normality as  $n$  increases, for fixed  $\pi$ , the approximation being reasonable when  $n[\min(\pi, 1 - \pi)]$  is as small as 5.
- ▶ **Interactive binomial distribution**

# Statistical inference

Inference is the use of sample data to **estimate** unknown parameters of the population. One method we will focus on is **maximum likelihood estimation (MLE)**.





## Statistical inference: maximum likelihood

- ▶ The **likelihood function** is the likelihood (or probability in the discrete case) of the sample of your data  $X_1, \dots, X_n$ , given the unknown parameter(s)  $\beta$ . Denoted as  $l(\beta|X_1, \dots, X_n)$  or simply  $l(\beta)$ .
- ▶ The MLE of  $\beta$  is defined as

$$\hat{\beta} = \sup_{\beta} l(\beta) = \sup_{\beta} L(\beta)$$

where  $L(\beta) = \log(l(\beta))$ . The MLE is the parameter value under which the data observed have the highest probability of occurrence.

## Statistical inference: MLE (con't)

- ▶ MLE have desirable properties: under weak regularity conditions, MLE have large-sample normal distributions; they are **asymptotically consistent**, converging to the parameter as  $n$  increases; and they are **asymptotically efficient**, producing large-sample standard errors no greater than those from other estimation methods.

# Covariance matrix of the MLE

Let  $\text{cov}(\hat{\beta})$  denote the asymptotic covariance matrix of  $\hat{\beta}$ , where  $\beta$  is a multidimensional parameter.

- Under regularity conditions,  $\text{cov}(\hat{\beta})$  is the inverse of the **information matrix**, which is

$$[I(\beta)]_{ij} = -E \left( \frac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j} \right)$$

- The standard errors are the square roots of the diagonal elements for the inverse of the information matrix. The greater the curvature of the log likelihood function, the smaller the standard errors.

# Statistical inference for Binomial parameter

- ▶ The binomial log likelihood function is

$$\begin{aligned}L(\pi) &= \log[\pi^y(1 - \pi)^{(n-y)}] \\&= y \log(\pi) + (n - y) \log(1 - \pi)\end{aligned}$$

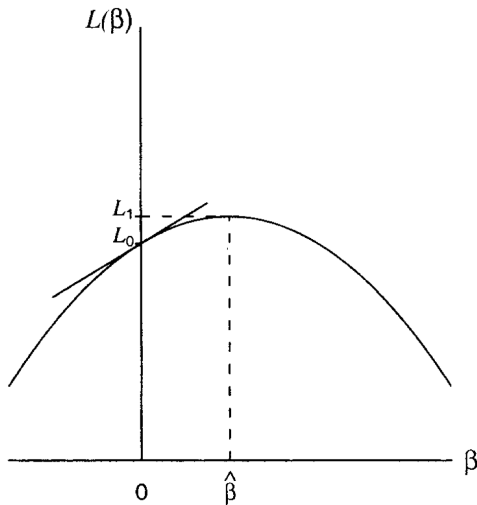
- ▶ Differentiating wrt  $\pi$  and setting it to 0 gives the MLE  $\hat{\pi} = y/n$ .
- ▶ The **Fisher information** is

$$I(\pi) = n/[\pi(1 - \pi)]$$

- ▶ The asymptotic distribution of the MLE  $\hat{\pi}$  is  $N(\pi, \pi(1 - \pi)/n)$ .

# Statistical inference for Binomial parameter

The score, Wald, and likelihood ratio tests use different information from this curve to draw inference about  $\pi$ .



# Wald test

Consider the hypothesis

$$H_0 : \beta = \beta_0 \quad H_1 : \beta \neq \beta_0$$

The **Wald test** defines a test statistic

$$z = (\hat{\beta} - \beta_0)/SE, \text{ where } SE = 1/\sqrt{I(\hat{\beta})} = n/[\hat{\pi}(1 - \hat{\pi})]$$

Under  $H_0 : \beta = \beta_0$ , the wald test statistic  $z$  is approximately standard normal. Therefore  $H_0$  is rejected if  $|z| > z_{\alpha/2}$ .

# Likelihood ratio test

**The likelihood ratio test (LRT)** is defined as

$$-2 \log \Lambda = -2 \log(l_0/l_1) = -2(L_0 - L_1)$$

where  $l_0$  and  $l_1$  are the maximized likelihood under  $H_0$  and  $H_0 \cup H_1$ . The null hypothesis is rejected if  $-2 \log \Lambda > \chi^2_{\alpha}(df)$  where  $df$  is the difference in the dimensions of the parameter spaces under  $H_0 \cup H_1$  and  $H_0$ .

## Score (a.k.a. Wilson) test

**Score test**, also called the *Wilson* or *Lagrange multiplier test*, is based on the slope and expected curvature of the log-likelihood function  $L(\beta)$  at the null value  $\beta_0$ . It utilizes the size of the score function

$$u(\beta) = \partial L(\beta) / \partial \beta$$

evaluated at  $\beta_0$ . The score test statistic is

$$\frac{[u(\beta_0)]^2}{I(\beta_0)} = \frac{[\partial L(\beta) / \partial \beta_0]^2}{-E[\partial^2 L(\beta) / \partial \beta_0^2]}$$

The null is rejected when the absolute test statistic is greater than  $\chi_1^2(\alpha)$ .



## Example: Estimating the proportion of Vegetarians

Students in a class were surveyed whether they are vegetarians. Of  $n = 25$  students,  $y = 0$  answered “yes”.

- ▶ With the Wald method, the 95% confidence interval for  $\pi$  (true proportion of vegetarians in the population) is

$$\hat{\pi} \pm 1.96\sqrt{\hat{\pi}(1 - \hat{\pi})/n}$$

which is

$$0 \pm 1.96\sqrt{(0 \times 1)/25}, \text{ or } (0, 0)$$

When a parameter falls near the boundary of the sample space, often sample estimates of standard errors are poor and the Wald method does not provide a sensible answer.

- ▶ With the score method, the interval equals  $(0, 0.133)$

## Comparison of the tests

- ▶ For small to moderate sample sizes, the likelihood-ratio and score tests are usually more reliable than the Wald test, having actual error rates closer to the nominal level.
- ▶ There are lots of different methods to compute CIs for a binomial proportion!

```
library(binom)
binom.confint(x=0, n=25)
```

##		method	x	n	mean	lower	upper
## 1	agresti-coull	0	25	0.00000000	-0.02439494	0.15758719	
## 2	asymptotic	0	25	0.00000000	0.00000000	0.00000000	
## 3	bayes	0	25	0.01923077	0.00000000	0.07323939	
## 4	cloglog	0	25	0.00000000	0.00000000	0.13718517	
## 5	exact	0	25	0.00000000	0.00000000	0.13718517	
## 6	logit	0	25	0.00000000	0.00000000	0.13718517	
## 7	probit	0	25	0.00000000	0.00000000	0.13718517	
## 8	profile	0	25	0.00000000	0.00000000	0.12291101	

# Bayesian inference for binomial parameters

Bayesian analyses incorporate “prior information” about parameters using

- ▶ prior subjective belief about a parameter, or
- ▶ prior knowledge from other studies, or
- ▶ very little knowledge (a “weakly informative” prior)

Prior distribution ( $g$ ) is combined with the likelihood ( $f$ ) to create a posterior ( $h$ ):

$$\begin{aligned}h(\theta|y) &= \frac{f(\mathbf{y}|\theta)g(\theta)}{f(\mathbf{y})} \\ &\propto f(\mathbf{y}|\theta)g(\theta)\end{aligned}$$

## Using Beta distributions for priors

If  $\pi \sim \text{beta}(\alpha_1, \alpha_2)$  (for  $\alpha_1 > 0$  and  $\alpha_2 > 0$ ) then  
 $g(\pi) \propto \pi^{\alpha_1-1}(1-\pi)^{\alpha_2-1}$ .

Beta is a *conjugate prior distribution* for a binomial parameter, implying that the posterior is also a beta distribution, specifically,  $h$  follows a  $\text{beta}(y + \alpha_1, n - y + \alpha_2)$ .

Shiny app for Bayesian inference of a Binomial.

## An exercise

1. Write down your prior belief about the probability that this coin will land heads.
2. Share it with the class
3. Use the prior probabilities to estimate a beta distribution.

```
library(MASS)
fitdistr(x, "beta", list(shape1=1, shape2=1))
```

4. Use [the app](#) to see how the posterior changes as we flip the coin.