Intro to Categorical Data Analysis

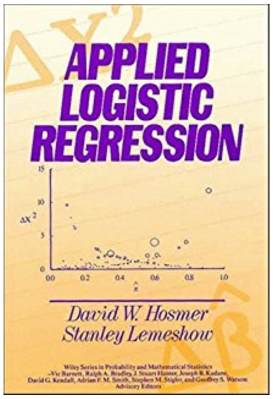
Nicholas Reich and Anna Liu, based on Agresti Ch 1

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.

History of the course

- Taught since mid-1980s at UMass-Amherst (PUBHLTH 743)
- Led to most cited statistics book in print (> 30,000 citations)



Focus of this course (different from the original)

- 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

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
  - What are different ways that a variable on education could be classified?

- 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 $y_1, y_2, \cdots, y_n$ 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 $bin(n, \pi)$. The probability mass function for the possible outcomes y for Y is

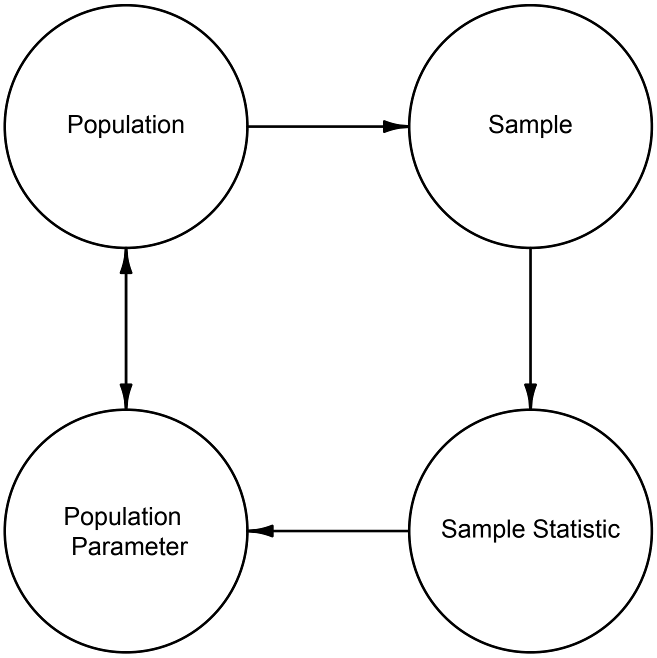
$$p(y) = \left(\begin{array}{c} n \\ y \end{array} \right) \pi^y (1-\pi)^{(n-y)}, y = 0, 1, ..., n$$

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

- The binomial distribution converges to normality as n increases, for fixed π, 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 $X_1, ..., X_n$, given the unknown parameter(s) β. Denoted as $l(\beta|X_1, ..., X_n)$ or simply $l(\beta)$.
- The MLE of β 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 $cov(\hat{\beta})$ denote the asymptotic convariance matrix of $\hat{\beta}$, where β is a multidimensional parameter.

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

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

- 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

$$L(\pi) = \log[\pi^y (1-\pi)^{(n-y)}]$$

$$= y\log(\pi) + (n-y)\log(1-\pi)$$

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

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

- The asympotic 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 π.

Wald test

Consider the hypothesis

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

The **Wald test** defines a test statistic

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

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.

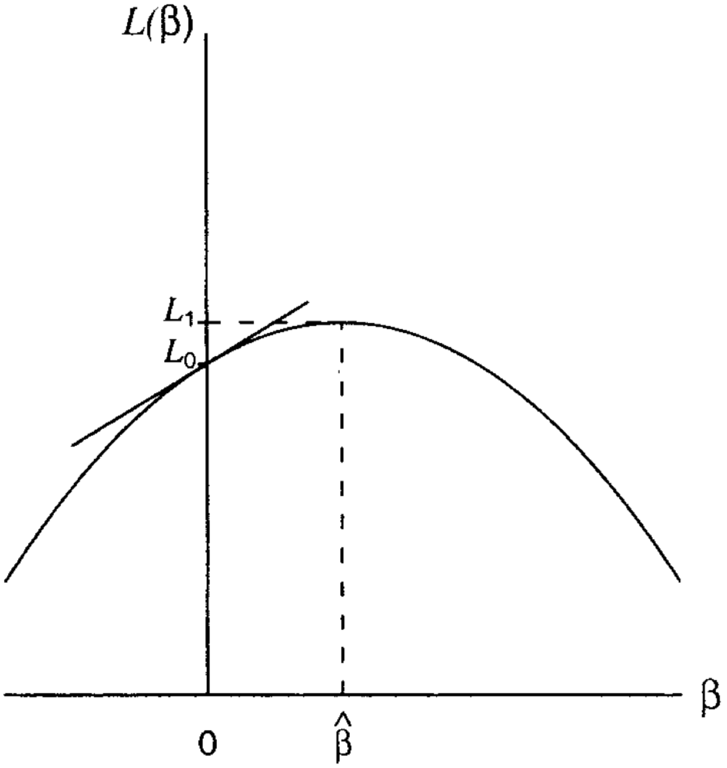


Figure 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

Figure 1: binomial likelihood

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 β_0. It utilizes the size of the score function

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

evaluated at β_0.

The score test statistic is

$$z = \frac{u(\beta_0)}{[I(\beta_0)]^{1/2}} = \frac{\hat{\pi} - \pi_0}{\pi_0(1-\pi_0)/n}$$

.

Example: Estimating the proportion of Vegetarians

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

- Using the Wald method, compute the 95% confidence interval for π (true proportion of vegetarians in the population):

- Using the Score method, compute the 95% confidence interval for π (true proportion of vegetarians in the population):

Warning about the Wald test

- 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.
- 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.

Comparison of the tests

There are lots of different methods to compute CIs for a binomial proportion!

```r
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
## 9            lrt 0 25 0.00000000  0.00000000 0.07398085
## 10     prop.test 0 25 0.00000000  0.00000000 0.16577301
## 11        wilson 0 25 0.00000000  0.00000000 0.13319225
```

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):

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

Using Beta distributions for priors

If $\pi \sim beta(\alpha_1, \alpha_2)$ (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 $beta(y + \alpha_1, n - y + \alpha_2)$.

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.

```r
library(MASS)
x <- c(
    ## enter probabilities here
    )
fitdistr(x, "beta", list(shape1=1,shape2=1))
```

4. Use the app to see how the posterior changes as we flip the coin.