

# 1 Introduction: Distributions and inference for categorical data

## 1.1 Categorical response data

- ◆ A categorical variable has a measurement scale consisting of a set of categories. Examples include
  - political philosophy: liberal/moderate/conservative
  - diagnoses regarding breast cancer: normal/benign/suspicious/malignant
- ◆ Categorical data arises from social and biomedical sciences but appear in every field (example?)
- ◆ Response-explanatory variable distinction
  - Response (dependent) variable
  - Explanatory (independent) variable
  - Regression context; We focus on categorical response variables
- ◆ Nominal-ordinal scale distinction
  - Nominal variable: no natural ordering. E.g. religious affiliation, mode of transportation
  - Ordinal variables: natural ordering of categories but distances b/w categories are unknown. E.g. social class, patient condition
  - Interval variables: numerical distances between any two values, e.g. blood pressure level, annual income
  - The classification of a variable depends on the way the variable is measured: e.g. education level can be ordinal or interval.
- ◆ Continuous-discrete variable
  - The number of values they can take, if large then it's continuous, if a few then it's discrete.
- ◆ We deal with "Discretely measured responses" which arise as:
  - Nominal variables
  - Ordinal variables
  - Discrete interval variables with few values
  - Continuous variables grouped into a small number of categories
- ◆ We basically learn about regression models but not for continuous response variables with normal distribution but for discrete/categorical response variables having binomial, multinomial, or Poisson distributions. Mainly,

- Logistic regression models: for a binary response with a binomial distribution (generalizes to a multcategory response with a multinomial distribution)
- Loglinear models: for count data with a Poisson distribution.
- ◆ The course covers:
  - Descriptive and inferential methods for univariate and bivariate categorical data (Ch. 1-3)
  - GLM for categorical responses (Ch. 4)
  - Logistic regression models and multinomial extensions(Ch. 5-7)
  - Loglinear regression models (Ch. 8)
  - Model buildings (Ch. 9)
  - And more if time permits

## 1.2 Distributions for categorical data

### ◆ *Binomial distribution*

- A fixed number  $n$  of binary observations
- Let  $y_1, \dots, y_n$  denote responses for  $n$  independent and identical trials such that  $P(Y_i=1)=\pi$  and  $P(Y_i=0)=1-\pi$ . The total number of successes  $Y=\sum(Y_i)\sim\text{bin}(n,\pi)$ .
- The probability mass function (PMF)
- Mean and variance
- Convergence in distribution to a normal distribution as  $n$  increases.
- Sampling binary outcomes WOR from a finite populations  $\rightarrow$  Hypergeometric distribution

### ◆ *Multinomial distribution*

- Each of  $n$  independent, identical trials can have outcome in any of  $c$  categories.
- Let  $y_{ij}=1$ (trial  $i$  has outcome  $j$ )
- $\mathbf{y}_i=(y_{i1}, \dots, y_{ic})$  represents a multinomial trial ( $\sum_j y_{ij} = 1$ , making  $y_{ic}$  redundant)
- Let  $n_j=\sum_i y_{ij} = \#$  of trials having outcome  $j$
- The counts  $(n_1, \dots, n_c)$  have the multinomial distribution
- Let  $\pi_j=P(Y_{ij}=1)$
- The PMF=
- Mean vector and covariance matrix =
- Binomial distribution is a special case of multinomial distribution with  $c=2$
- The marginal distribution of each  $n_j$  is binomial

### ◆ *Poisson distribution*

- No fixed upper limit  $n$  for some count data e.g. # of deaths from automobile accidents on motorways in Italy.
- Still, it needs to be integer and nonnegative.

- Poisson is a simplest such distribution
- $Y \sim \text{Poisson}(\mu)$
- The PMF, the mean and variance (sample counts vary more when their mean is higher)
- Unimodal
- Asymptotic normality as  $\mu$  increases
- Used for counts of events that occur randomly over time or space when outcomes in disjoint periods or regions are independent
- Also,  $\text{bin}(n, \pi) \sim \text{Poisson}(n\pi)$  if  $n$  is large and  $\pi$  is small. E.g.  $\text{bin}(5e7, 2e-6) \sim \text{Poisson}(100)$
- ◆ **Over-dispersion**
  - Count exhibit variability exceeding that predicted by binomial or Poisson models
  - Variation in individual 'success' probability causes such overdispersion (example?)
  - Suppose  $Y|\mu \sim (E(Y|\mu), \text{var}(Y|\mu))$
  - Unconditionally,
  - $E(Y) = E[E(Y|\mu)]$ ,
  - $\text{var}(Y) = E[\text{var}(Y|\mu)] + \text{var}[E(Y|\mu)]$
  - Let  $E(\mu) = \theta$ . If  $Y|\mu \sim \text{Poisson}(\mu)$ ,
  - $E(Y) = E(\mu) = \theta$
  - $\text{var}(Y) = E(\mu) + \text{var}(\mu) = \theta + \text{var}(\mu) > \theta$  : overdispersion
  - The negative binomial for count data permits  $\text{var} > \text{mean}$
- ◆ *Analyses assuming binomial/multinomial distributions, as well as those assuming Poisson distribution, can become invalid too due to overdispersion.*
  - E.g. The true distribution is a mixture of different binomial distributions or  $\pi$  itself is a random variable
- ◆ *Connection between Poisson and multinomial distributions*
  - Let  $(y_1, y_2, y_3)$  =# of people who die in {automobile, airplane, railway} accidents.
  - A Poisson model:  $Y_i \sim \text{Poisson}(\mu_i)$ , independent.
  - The join pmf for  $\{Y_i\}$  is the product of Poisson pmfs.
  - The total  $n = \sum_i Y_i \sim \text{Poisson}(\sum_i \mu_i)$  is random, not fixed.
  - If we assume the Poisson model but condition on  $n$ ,  $\{Y_i\}$  no longer have Poisson since they all need to  $\leq n$ , nor independent.
  - For  $c$  independent Poisson variates with  $E(Y_i) = \mu_i$ , the conditional distribution of  $(Y_1, \dots, Y_c) | n \sim \text{multinomial}(n, \{\pi_i\})$  derive?

### 1.3 Statistical inference for categorical data

- ◆ We use maximum likelihood method for parameter estimation.

- ◆ *An MLE has desirable properties like*
  - Asymptotic normality
  - Asymptotic consistency
  - Asymptotic efficiency
- ◆ Given the data under a certain probability model, the “likelihood function” is the probability of those data treated as a function of the unknown parameter.
- ◆  $\text{MLE} = \arg \max (\text{likelihood}) = \arg \max (\log \text{likelihood})$
- ◆ Some review of the maximum likelihood theory

### 1.3.1 Likelihood function and ML estimate for binomial parameter

- ◆ ML estimate for the success probability in the binomial model and its asymptotic (and exact) variance

### 1.3.2 Wald/Likelihood Ratio/Score Test Triad

- ◆ Significance test  $H_0: \beta = \beta_0$  exploiting the asymptotic normality of MLE.
- ◆ “Wald statistic”:
  - Compute  $z = \frac{\hat{\beta} - \beta}{SE} \sim N(0,1)$  approximately. Use  $z$  for one- or two-sided p-values. For the two sided alternative,  $z^2 \sim \chi^2(1)$  under the null. Multivariate extension is  $W = (\hat{\beta} - \beta_0) \text{cov}(\hat{\beta})^{-1} (\hat{\beta} - \beta_0) \sim \chi^2(\text{rank}(\text{cov}(\hat{\beta})))$
- ◆ “Likelihood ratio”
  - Compute (1) the maximum over the possible parameter values under  $H_0$
  - (2) the maximum over the larger set of parameter values permitting  $H_0$  or an alternative  $H_1$  to be true; Call their ratio  $\Lambda = (2)/(1)$ . One can show  $-2 \log \Lambda = -2 \log(l_0 / l_1) = -2(L_0 - L_1) \sim \chi^2(\dim(H_1 \cup H_0) - \dim(H_0))$
- ◆ “score test” :
  - compute the score function evaluated at  $\beta_0$ .
- ◆ *As  $n$  increases, all three have certain asymptotic equivalences. For small to moderate sample sizes, the LR test is usually more reliable.*

### 1.3.3 Constructing confidence intervals(CI)s

- ◆ *Equivalence to testing:*
  - a 95% CI for  $\beta$  is the set of  $\beta_0$  for which the test of  $H_0: \beta = \beta_0$  has a p-value exceeding 0.05.
- ◆ We usually use
  - $z_a$  : 100(1-a)percentile of  $N(0,1)$  distribution.

- $\chi_{df}^2(a)$  : 100(1-a)percentile of chi-squared distribution with d.f. df.
- ◆ The “Wald CI”
- ◆ The “LR CI”
- ◆ *If the two are significantly different, asymptotic normality may not be holding up well. (sample size too small) What to do?*
  - Exact small-sample distribution
  - Higher-order asymptotic
- ◆ *For classical linear regression with normal response, all three provide identical results.*

## 1.4 Statistical inference for binomial parameters

### 1.4.1 Tests for a binomial parameter

- ◆ Consider  $H_0: \pi = \pi_0$ .
- ◆ The Wald statistic
- ◆ The normal form of the score statistic
  - Score  $u(\pi_0)$  and information  $i(\pi_0)$
- ◆ The LR test statistic
  - $-2(L_0 - L_1) = ??? \sim \chi^2(1)$

### 1.4.2 Confidence intervals for a binomial parameter

- ◆ The Wald CI : poor performance. Especially near 0 and 1. Can be improved by adding  $.5 z_{\alpha/2}^2$  observations of each type to the sample.
- ◆ The score CI : complicated but performs better. (similar to the modifications above)
- ◆ The LR CI: also complicated.
- ◆ *Vegetarian example: of 25 students, none were vegetarians. ( $y=0$ ) What's the 95% CI for the proportion of the vegetarians?*
  - Wald CI gives (0,0)
  - Score CI gives (0.0, 0.133)
  - LR CI gives (0.0, 0.074)
  - When  $\pi \sim 0$ , the sampling distribution of the estimator is highly skewed to the right. Worth considering alternative methods not requiring asymptotic approximations.

### 1.4.3 Statistical inference for multinomial parameters

- ◆ Estimation of multinomial parameters

- ◆ Pearson statistic for testing a specified multinomial:

For  $H_0: \pi_j = \pi_{j0}$  for  $j=1, \dots, c$ , compute the expected frequencies  $\mu_j = n\pi_{j0}$  and use the statistic

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j} \sim \chi^2(c-1) \text{ approximately. It's called "Pearson chi-squared statistic"}$$

- ◆ Example: testing Mendel's theories.

- For cross of pea plants of pure yellow strain with plants of pure green strains, he predicted that second-generation hybrid seed would be 75% yellow, 25% green, and 25% green. One experiment shows out of  $n=8,023$  seeds,  $n_1=6,022$  were yellow and  $n_2=2,001$  were green. The Pearson chi-squared statistic and the P-value are...
- Fisher summarized Mendel's data as a single chi-squared stat whose value is 42 when it follows  $\chi^2(84)$ . The P-value is .99996. Too perfect??