Chapter 1 Introduction

Zhenghui Feng

Outline

1. 1.1 Categorical Response Data
2. 1.2 Probability Distributions for Categorical Data
3. 1.3 Statistical Inference for A Proportion
4. 1.4 More on Statistical Inference for Discrete Data
5. Homework

Outline

1. 1.1 Categorical Response Data
2. 1.2 Probability Distributions for Categorical Data
3. 1.3 Statistical Inference for A Proportion
4. 1.4 More on Statistical Inference for Discrete Data
5. Homework

Outline

1. 1.1 Categorical Response Data
2. 1.2 Probability Distributions for Categorical Data
3. 1.3 Statistical Inference for A Proportion
4. 1.4 More on Statistical Inference for Discrete Data
5. Homework

1.2.1 Binomial Distribution

Bernoulli Distribution.

A DRV Y is called a Bernoulli(π) ($0 < \pi < 1$) random variable if its PMF

$$f_Y(y) = \begin{cases} \pi & \text{if } y = 1, \\ 1 - \pi & \text{if } y = 0. \end{cases}$$

Binomial Distribution.

A DRV Y is called a Binomial(n, π) ($n \geq 0$ and $0 < \pi < 1$) if its PMF is

$$f_Y(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \; x = 0, 1, \cdots, n$$

Remark: A person throws a coin n times independently. Each time the head has probability π and the tail has probability $q = 1 - \pi$. The number of heads is a random variable following Binomial(n, π) distribution.

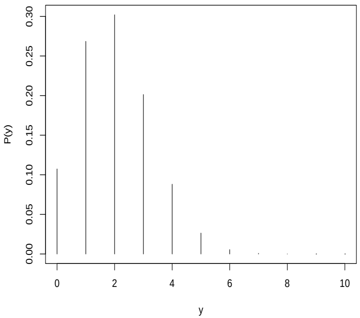
- n Bernoulli trials, two possible outcomes for each (success, failure)
- $\pi = P(\text{success})$, $1 - \pi = P(\text{failure})$ for each trial
- Y = number of successes out of n trials
- Trials are independent

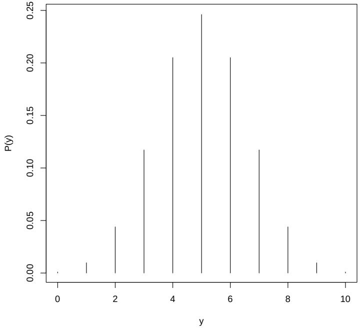
Then Y has binomial distribution, with the probability density function (pdf)

$$P(y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y},\ y = 0, 1, 2, ..., n.$$

Example (Quiz)

Suppose a quiz has 10 multiple-choice questions, with five possible answers for each. A student who is completely unprepared randomly guesses the answer for each question. The probability of a correct response is 0.20 for a given question.





$\pi = 0.2$ $\pi = 0.5$

Figure: Binomial PMF

Note

- $E(Y) = \mu = n\pi$, $\mathrm{Var}(Y) = n\pi(1-\pi)$, $\sigma = \sqrt{n\pi(1-\pi)}$
- $\hat{\pi} = \frac{Y}{n}$=proportion of success, $E(\hat{\pi}) = \pi$, $\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}$
- When n is large, the distribution of Y can be approximated by a normal distribution with $\mu = n\pi$, $\sigma = \sqrt{n\pi(1-\pi)}$
- The approximation has a guideline that $\mathbf{n}\pi$, $\mathbf{n}(\mathbf{1}-\pi)$, should be ≥ 5
- When each trial has $n > 2$ possible outcomes, numbers of outcomes in various categories have *multinomial distribution*.

1.2.2 Multinomial Distribution

Multinomial Distribution.

Let c denote the number of outcome categories. Their probabilities are denoted by $\{\pi_1, \pi_2, \ldots, \pi_c\}$, where $\sum_j \pi_j = 1$. For n independent observations, the multinomial probability that n_1 fall in category 1, n_2 fall in category 2, ..., n_c fall in category c, where $\sum_j n_j = n$, equals

$$P(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n2} \cdots \pi_c^{n_c}.$$

Note

- The multinomial is a multivariate distribution
- The marginal distribution of the count in any particular category is binomial. For category j, the count n_j has mean $n\pi_j$ and standard deviation $\sqrt{n\pi_j(1 - \pi_j)}$

Outline

1 1.1 Categorical Response Data

2 1.2 Probability Distributions for Categorical Data

3 1.3 Statistical Inference for A Proportion

4 1.4 More on Statistical Inference for Discrete Data

5 Homework

1.3.1 Likelihood Function and Maximum Likelihood Estimation

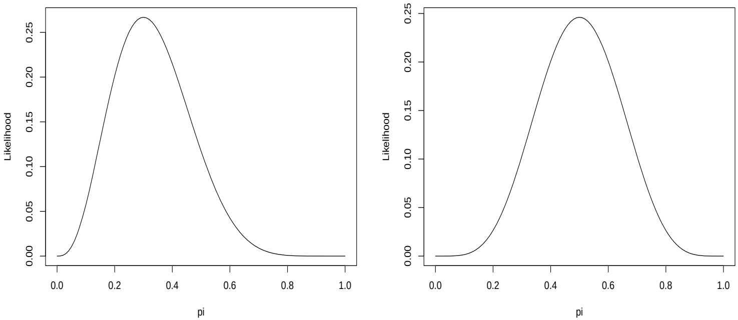
Likelihood Function

The likelihood function is the probability of the observed data, expressed as a function of the parameter value.

Example

Binomial, $n = 10$, $y = 3$.

$$P(Y = 3) = \frac{10!}{3!7!}\pi^3(1-\pi)^7 = l(\pi).$$



$y = 3$ $\qquad\qquad$ $y = 5$

Figure: Likelihood Function for π

Maximum Likelihood Estimate

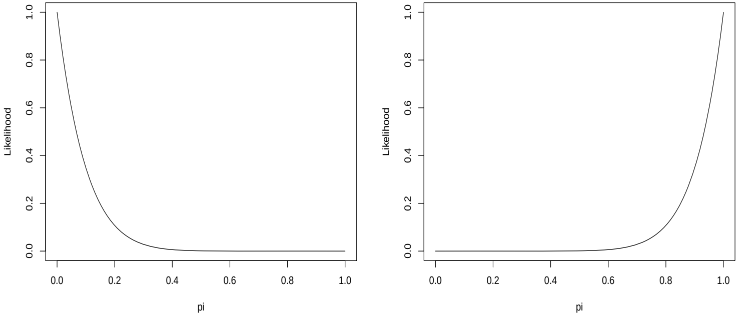
The maximum likelihood (ML) estimate is the parameter value at which the likelihood function takes its maximum.

Example

$n = 10$, $y = 3$.

$$l(\pi) = \frac{10!}{3!7!}\pi^3(1-\pi)^7.$$

is maximized at $\hat{\pi} = 0.3$.



$y = 0$ $\qquad\qquad$ $y = 10$

Figure: Likelihood Function for π

Note

- For binomial, $\hat{\pi} = \frac{y}{n}$ = proportion of successes.
- If $y_1, y_2, \ldots, y_n$ are independent from normal, ML estimate $\hat{\mu} = \bar{y}$.
- In ordinary regression Y ~normal, “least squares” estimates are ML.
- For large n for any distribution, ML estimates are optimal (no other estimator has smaller standard error)
- For large n, ML estimators have approximate normal sampling distributions (under weak conditions).

1.3.2 Significance Test About a Binomial Proportion

Significance Test

$H_0 : \pi = \pi_0 \qquad H_A : \pi \neq \pi_0$ (or 1-sided)
Test statistic is

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

has large-sample standard normal $N(0, 1)$ null distribution.

Question: How to do a hypothesis testing?

1.3.3 Example: Survey Results on Legalizing Abortion

Example

Do a majority, or minority, of adults in the United States believe that a pregnant woman should be able to obtain an abortion? Let π denote the proportion of the American adult population that responds “yes to the question, “Please tell me whether or not you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children. We test $H_0 : \pi = 0.50$ against the two-sided alternative hypothesis, $H_A : \pi \neq 0.50$.

1.3.4 Confidence Intervals for a Binomial Proportion

Let SE denote the estimated standard error of p. A large sample $100(1-\alpha)\%$ confidence interval for π has the formula

$$p \pm z_{\alpha/2} SE, \qquad \text{with } SE = \sqrt{p(1-p)/n},$$

where $z_{\alpha/2}$ denotes the standard normal percentile having right-tail probability equal to $\alpha/2$.

Example

For the attitudes about abortion example just discussed, $p = 0.448$ for $n = 893$ observations. The 95% confidence interval equals

$$0.448 \pm 1.96\sqrt{(0.448)(0.552)/893}$$

Note

- Unless π is close to 0.5, however, it doest not work well unless n is very large. It is especially poor when π is near 0 or 1.
- A better way: the CI contains all values π_0 for the null hypothesis that are not rejected: for given p and n, the π_0 values are the solution to the inequality

$$\frac{|p - \pi_0|}{\sqrt{\pi_0(1 - \pi_0)/n}} \leq 1.96$$

- A simple alternative approximation: add 2 to the number of successes and 2 to the number of failures (and thus 4 to n) and then use the ordinary formula with the estimated standard error. *Agresti-Coull confidence interval*

Outline

1 1.1 Categorical Response Data

2 1.2 Probability Distributions for Categorical Data

3 1.3 Statistical Inference for A Proportion

4 1.4 More on Statistical Inference for Discrete Data

5 Homework

1.4.1 Wald, Likelihood-Ratio, and Score Inference

Wald Test

Let β denote an arbitrary parameter. Consider a significance test of $H_0 : \beta = \beta_0$. Let SE denote the standard error of ML estimator $\hat{\beta}$, evaluated by substituting the ML estimator for the unknown parameter in the expression for the true standard error. When H_0 is true, the test statistic

$$z = (\hat{\beta} - \beta_0)/SE$$

has approximately a standard normal distribution. Equivalently, z^2 has approximately a chi-squared distribution with $df = 1$. *This type of statistic, which uses the standard error evaluated at the ML estimate, is called a Wald statistic.* The z or chi-squared test using this test statistic is called a *Wald test.*

Likelihood Ratio Test

Under $H_0 : \beta = \beta_0$, the likelihood ratio test statistic

$$-2\log(l_0/l_1)$$

has a large-sample chi-squared distribution with $df = 1$. l_0 is the likelihood function calculated at β_0, and l_1 is the likelihood function calculated at the ML estimate $\hat{\beta}$.

Score Test

The standard error are calculated under the assumption that the null hypothesis holds. E.g., $\sqrt{\pi_0(1-\pi_0)/n}$

1.4.2 Wald and Score Inference for Binomial Parameter

Wald Statistic for Binomial Parameter

$H_0 : \pi = \pi_0$, $H_A : \pi \neq \pi_0$, Wald statistic is

$$z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}}$$

Wald Confidence Interval (CI) for Binomial Parameter

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Example

For $n = 20$, $y = 5$, find the 95% Wald CI.

Score test, Score CI use null SE

Score 95% CI is the set of π_0 values for which p-value > 0.05 in testing

$$H_0 : \pi = \pi_0 \quad H_A : \pi \neq \pi_0$$

using

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Example

$y = 5$, $n = 20$, find the 95% Score CI.

Likelihood-ratio test

When $H_0 : \pi = 0.5$ is true, $l_0 = [10!/9!1!](0.5)^9(0.5)^1 = 0.00977$. The likelihood-ratio test compares this to the value of the likelihood function at the ML estimate of $p = 0.9$, which is $l_1 = [10!/9!1!](0.9)^9(0.1)^1 = 0.387$. The likelihood-ratio test statistic

$$-2\log(l_0/l_1) = 7.36$$

From the χ_1^2, the P-value is 0.007.

1.4.3 Small-Sample Binomial Inference

For small sample size, it is safer to use the binomial distribution directly (rather than a normal approximation) to calculate P-values. To illustrate, consider testing $H_0 : \pi = 0.50$ against $H_A : \pi > 0.50$ for the example of a clinical trial to evaluate a new treatment, when the number of successes $y = 9$ in $n = 10$ trials. The exact P-value, based on the right tail of the null binomial distribution with $\pi = 0.50$, is

$$P(Y \geq 9) = [10!/9!1!](0.50)^9(0.50)^1 + [10!/10!0!](0.50)^{10}(0.50)^0 = 0.011.$$

For the two sided alternative $H_A : \pi \neq 0.50$, the P-value is

$$P(Y \geq 9 \text{ or } Y \leq 1) = 2 \times P(Y \geq 9) = 0.021.$$

1.4.4 1.4.5 More about Small-Sample Inference

- With discrete probability distributions, small-sample inference using the ordinary P-value is *conservative*.This means that when H_0 is true, the P-value is ≤ 0.05 (thus leading to rejection of H_0 at the 0.05 sig. level) not exactly 5% of the time, but typically less than 5% of the time.
- Mid P-value: it adds only half the probability of the observed result to the probability of the more extreme results.

Example

$H_0 : \pi = 0.5$ v.s. $H_a : \pi > 0.5$ with $y = 9, n = 10$.The ordinary P-value is

$$\text{P-value} = P(9) + P(10) = 0.011.$$

The mid P-value is $\quad P(9)/2 + P(10) = 0.006.$

Outline

1. 1.1 Categorical Response Data
2. 1.2 Probability Distributions for Categorical Data
3. 1.3 Statistical Inference for A Proportion
4. 1.4 More on Statistical Inference for Discrete Data
5. Homework

Homework 1

1. Read Preface, Chapter 11 and Chapter 1 carefully.
2. Analysis the GDS5037 data.
   (1) Download the data from the NIH web in the Gene Expression Omnibus (GEO) database and read the file.
   (2) For all samples, there are 3 categories for patients' status: mild asthma (MMA), control, severe asthma (SA). Calculate the frequency and percentage of each category and plot a pie chart.
   (3) Plot the frequency bar chart for the 3 categories by patient's gender.
   (4) Classify patients into 3 groups according to patients' status: MMA, control, SA. Calculate the sample mean and variance of IDENTIFIER FAM174B in each group.
   (5) According to (4), conduct the hypothesis test that SA and control, the mean of two groups are equal, under 5% significant level.
3. Problems in textbook 1.2, 1.4, 1.5, 1.6, 1.8, 1.12.