BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part I

Introduction

Objectives

Categorical data and common distributions

Statistical methods for one proportion

- Maximum likelihood estimation
- Inference (Hypothesis testing, Confidence interval)
    - Wald
    - Score
    - Likelihood Ratio
- Small sample scenario

Agresti (2002), Section 1.2, 1.3, 1.4

Categorical Data

A categorical variable has a measurement scale consisting of a set of categories

- Nominal - unordered categories
  e.g., choice of transport (walk, car, bike, bus)
- Ordinal - ordered categories
  e.g.., disease severity (mild, moderate, severe)

Types of Variables in Statistical Analysis

Response variable

- Outcome variable, dependent variable, Y variable
- Values are dependent on explanatory variables

Explanatory variable

- Predictor, covariate, independent variable, X variable
- Values are set at predetermined levels

In linear regression analysis, response variables are continuous (e.g., blood pressure).

In categorical data analysis, response variables are categorical (e.g., diabetic/non-diabetic, disease severity).

Explanatory variables can be both.

Context is important! The context of the study and the relevant questions of interest are important in specifying what kind of variable we will analyze.

For example,

Did you have a side effect with the high-intensity drugs?

- Yes or No
- This is a binary nominal categorical variable

How severe is the side effect?

- Low, Moderate, or Severe
- This is an ordinal categorical variable

Binomial Distribution

The binomial distribution (and its generalization, the multinomial distribution) plays a similar role to that of the normal distribution for continuous response.

- n Bernoulli trials: two possible outcomes for each trial (success, failure)
- π is the probability of success for a given trial
- n trials are independent and identical
- Y denotes the number of successes out of the n trials

Under these assumptions, Y has the *binomial distribution* with index n and parameter π.

Facts about the Binomial Distribution

Probability function
$\Pr(Y = y) = \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n$, where $y! = y(y-1)(y-2)\ldots 1$ with $0! = 1$.

Mean $E(Y) = n\pi$

Variance $V(Y) = n\pi(1-\pi)$

The binomial distribution is symmetric when $\pi = 0.5$.

For fixed π, it becomes more bell-shaped as n increases. For this normal approximation to work, n should large enough, i.e., $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

Check out examples : *R_eg_1.1.html*

Multinomial Distribution

- n trials: each trial has c possible outcomes. (Binomial distribution is the special case with $c = 2$.)
- $\{\pi_1, \pi_2, \ldots, \pi_c\}$, is the vector of probability for outcomes in each category, where $\sum_i \pi_i = 1$.
- n trials are independent and identically distributed
- N_1 = number of trials having outcomes in category 1,
  N_2 = number of trials having outcomes in category 2,
  $\ldots$,
  where $\sum_i N_i = n$.

$\{N_1, N_2, \ldots, N_c\}$ follows a multinomial distribution with probability function

$$P(n_1, n_2, \ldots, n_c) = \frac{n!}{n_1! n_2! \cdots n_c!} \pi_1^{N_1} \pi_2^{N_2} \cdots \pi_c^{N_c}$$

Facts about Multinomial Distribution

Multinomial is a multivariate distribution, i.e., more than one dimension.

Marginal distribution of the count in each category is Binomial with mean $n\pi_i$ and variance $n\pi_i(1-\pi_i)$.

Counts in different categories are negatively correlated with $cov(N_i, N_j) = -n\pi_i\pi_j$.

Poisson Distribution

We observe the counts of events within a set unit of time, area, volume, length etc.

Let Y be the number of occurrences in a given interval with the average number λ.

- events occur randomly in time or space
- outcomes in disjoint periods or regions are independent

Y follows a Poisson distribution with probability function

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!},$$

where $y = 0, 1, 2, \ldots$

Facts about Poisson Distribution

Mean $E(Y) = \lambda$

Variance $V(Y) = \lambda$

When λ increases, the Poisson approach normal distribution.

The Poisson is also limiting case of the binomial. For $X \sim$ Binomial (n, π) , when n is big and π in the way that $n\pi \to \lambda$, $X \sim$ Poisson (λ) as the limit.

X and Y are independent Poisson random variables with parameters λ_X and λ_Y , then $X + Y$ has a Poisson distribution with parameter $\lambda_X + \lambda_Y$.

The conditional distribution of X given $X + Y = n$ is Binomial$\left(n, \pi = \frac{\lambda_X}{\lambda_X + \lambda_Y}\right)$.

Statistical Methods for One Proportion

According to Wikipedia, 15% of people in the US diagnosed with lung cancer survive five years after the diagnosis.

Data from the Surveillance, Epidemiology and End Results (SEER) Program: 269,136 people are diagnosed with lung cancer between 2000 and 2006. Of which 45,483 survived after 5 years.

Practical questions

Estimate π

Test if the 5-year survival rate is 15% (determine if a particular value for a parameter is plausible)

Confidence interval for π (determine the range of plausible values for a parameter)

Likelihood Function

Assume we have a distribution $f(x)$ with unknown parameter θ. Since the probability distribution depends on θ, we can make this dependence explicit by writing $f(x)$ as $f(x;\theta)$.

Once a value of X has been observed, we can plug this observed value x into $f(x;\theta)$ and obtain a function of p only.

This probability of the observed data, expressed as a function of the parameter θ, is called the *likelihood function*.

We will write the likelihood function as $L(\theta;x)$ or just $L(\theta)$.

$L(\theta; x)$ vs $f(x; \theta)$

Algebraically, the likelihood $L(\theta; x)$ is just the same as the distribution $f(x; \theta)$, but its meaning is quite different because it is regarded as a function of θ rather than a function of x.

Consequently, a graph of the likelihood usually looks very different from a graph of the probability distribution.

For discrete random variables, a graph of the probability distribution $f(x; \theta)$ has spikes at specific values of x, whereas a graph of the likelihood $L(\theta; x)$ is a continuous curve (e.g. a line) over the parameter space, the domain of possible values for θ.

Maximum Likelihood Estimation

$L(\theta; x)$ summarizes the evidence about θ contained in the event $X = x$. $L(\theta; x)$ is high for values of θ that make $X = x$ likely, and small for values of θ that make $X = x$ unlikely.

Maximum Likelihood Estimation estimates the unknown parameter θ by the value for which the likelihood function $L(\theta; x)$ is largest. The value obtained in such way is denoted by $\hat{\theta}$ and called the *maximum-likelihood estimate* (MLE) of θ.

Besides intuitive, MLEs have good large-sample properties.

- Their distributions are approximately normal
- Their large-sample standard errors are the smallest.

Example on Binomial Distribution ($n = 10$ trials)

Observe $y = 2$ successes.

$$\begin{aligned} & L(\pi; y = 2) \\ = & \frac{10!}{2!8!}\pi^2(1-\pi)^8 \\ = & 45\pi^2(1-\pi)^8 \end{aligned}$$

$L(0.1; y = 2) = 0.194$
$L(0.2; y = 2) = 0.302$
$L(0.3; y = 2) = 0.233$
$L(0.4; y = 2) = 0.121$
$L(0.5; y = 2) = 0.044$
$L(0.6; y = 2) = 0.011$

Observe $y = 4$ successes.

$$\begin{aligned} & L(\pi; y = 4) \\ = & \frac{10!}{4!6!}\pi^4(1-\pi)^6 \\ = & 210\pi(1-\pi)^9 \end{aligned}$$

$L(0.1; y = 4) = 0.011$
$L(0.2; y = 4) = 0.088$
$L(0.3; y = 4) = 0.200$
$L(0.4; y = 4) = 0.251$
$L(0.5; y = 4) = 0.205$
$L(0.6; y = 4) = 0.111$

Check out examples : *R_eg_1.2.html*

The $L(\pi; y)$ is maximized at sample proportion $p = \frac{y}{n}$.

Mean $E(\frac{Y}{n}) = \pi$

Variance $V(\frac{Y}{n}) = \pi(1 - \pi)/n$

Sample proportion is an unbiased estimator of π (on average do not overestimate or underestimate the true value).

As an MLE, the sample proportion also enjoys good large-sample properties.

Large-sample inferential methods for π

Consider $H_0 : \pi = \pi_0$ for a fixed value π_0 and alternative H_1.

Under H_0, the **test statistic**

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

is approximately standard normal Z for large sample size n.

p-value the probability, computed assuming H_0 is true, of obtaining a sample test statistic as extreme or more extreme (either or both direction) than the observed value of the sample statistic.

If the p-value is small, we reject H_0 in favor of H_1; otherwise fail to reject H_0.

$H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$ (two sided)

- Compute test statistic z, or p-value $= 2 \times P(Z > |z|)$.
- Reject H_0 if $z > z_{\alpha/2}$ or $z < -z_{\alpha/2}$
- Reject H_0 if p-value $< \alpha$.

$H_0 : \pi \geq \pi_0$ vs. $H_1 : \pi < \pi_0$

- Compute test statistic z, or p-value $= P(Z < -z)$.
- Reject H_0 if $z < -z_{\alpha}$
- Reject H_0 if p-value $< \alpha$.

$H_0 : \pi \leq \pi_0$ vs. $H_1 : \pi > \pi_0$

- Compute test statistic z, or p-value $= P(Z > z)$.
- Reject H_0 if $z > z_{\alpha}$
- Reject H_0 if p-value $< \alpha$.

In the example,

$H_0 : \pi = 0.15$ vs. $H_1 : \pi \neq 0.15$

Sample proportion $p = \frac{45483}{269136} = 0.169$

Test statistic

$$z = \frac{0.169 - 0.15}{\sqrt{\frac{0.15 \times 0.85}{269136}}} = 27.6$$

p-value is $2 \times P(z > 27.6) < 10^{-10}$.

Since $z > 1.96$ (equivalently, $p < 0.05$), we reject the H_0.

Check out the example *R_eg_1.3.html*

Confidence Intervals for a Binomial Proportion

Consider a given α, the large-sample $100(1-\alpha)\%$ confidence interval (CI) for π is

$$\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right),$$

where $P(Z > z_{\alpha/2}) = \alpha/2$.

This is also called **Wald CI**, because we use the sample proportion p as an estimate of π in $\pi(1-\pi)/n$.

In the example, we have already shown that $p = 0.169$ for $n = 269136$ observations.

The 95% Wald CI is given by,

$$\left(0.169 - 1.96\sqrt{\frac{0.169 \times 0.831}{269136}}, 0.169 + 1.96\sqrt{\frac{0.169 \times 0.831}{269136}}\right).$$

Interpretation: we are 95% confident that the true five year survival is between 0.168 and 0.170.

Facts about Wald CI $\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}, p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right)$

Wald CI has poor performance when n is very small.

Wald CI can be out of range for $(0, 1)$.

Wald CI for π collapses if $p = 0$ or 1.

Actual coverage probability is much less than 0.95 when π is close to 0 or 1.

Wald CI relates to the test $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$ using the test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}},$$

where the denominator is estimated standard error evaluated at MLE. The corresponding CI is the set of π_0 values not rejected in this test, i.e, solving $|z| = z_{\alpha/2}$.

Score Test and Score CI

Score test and Score CI use the null standard error.

Score 95% CI relates to testing $H_0 : \pi = \pi_0$ vs. $H_1 : \pi \neq \pi_0$ ($\alpha = 0.05$) using the test statistic

$$z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}},$$

where the denominator is null standard error.

To obtain the two end points in Score 95% CI, solve the following quadratic equation for π_0:

$$\frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = z_{\alpha/2}.$$

Check out the example *R_eg_1.3.html*

Likelihood ratio test and CI

Let l_0 be the maximum likelihood function under the null, and l_1 be the maximum likelihood function under all possible parameter values.

Likelihood ratio test is the ratio of two MLs, defined by

$$2\ln(l_1/l_0).$$

The test statistic is always nonnegative.

Large value of the test statistic is strong evidence against H_0 in favor of H_1.

Approximately χ^2_1 distribution for large sample size.

Likelihood ratio CI is the set of π_0 values not rejected in a Likelihood ratio test, i.e., $2\ln(l_1/l_0) = \chi^2_1(0.95)$.

[Example] Suppose observing 9 successes in 10 trials in a clinical trial to evaluate a new treatment, the MLE of the success rate is 0.9. $H_0 : \pi = 0.5$ vs. $H_0 : \pi \neq 0.5$.

Wald test
$z = \frac{0.9-0.5}{\sqrt{0.9\times 0.1/10}} = 4.22$, p-value $= 2.5 \times 10^{-5}$.
95 % CI: $(0.9 \pm 1.96\sqrt{0.9 \times 0.1/10})) = (0.71, 1.09)$.

Score test
$z = \frac{0.9-0.5}{\sqrt{0.5\times 0.5/10}} = 2.53$, p-value $= 0.011$.
Solve $\frac{0.9-\pi_0}{\sqrt{\pi_0(1-\pi_0)/10}} = 1.96 \rightarrow$ 95% CI: (0.60, 0.98).

Likelihood ratio test
$2\ln(l_1/l_0) = 2\ln\left(\frac{c\times 0.9^9 0.1}{c\times 0.5^9 0.5}\right) = 7.36$, p-value $= 0.007$.
Solve $2\ln\left(\frac{c\times 0.9^9 0.1}{c\times \pi_0^9 \pi_0}\right) = 3.84 \rightarrow$ 95% CI: (0.63, 0.99).

Small-Sample Binomial Inference

When $n\pi \geq 5$ and $n(1-\pi) \geq 5$ are not met, it is recommended to use binomial distribution to compute the exact p-value.

[Example] In $n = 10$ trial with $y = 9$ successes.

The p-value for testing $H_0 : \pi = 0.5$ versus $H_1 : \pi > 0.5$ is

$$\begin{aligned} \Pr(Y \geq 9) &= \frac{10!}{9!1!}(0.5)^9(0.5)^1 + \frac{10!}{10!0!}(0.5)^{10}(0.5)^0 \\ &= 0.0107. \end{aligned}$$

The p-value for testing $H_0 : \pi = 0.5$ versus $H_1 : \pi \neq 0.5$ is

$$\Pr(Y \geq 9) + \Pr(Y \leq 1) = 2\Pr(Y \geq 9) = 0.0214.$$

Generally, the two-sided p-value is
$\sum_y \{\Pr(Y = y) : \Pr(Y = y) \leq \Pr(Y = y_0)\}$.

Small-sample discrete inference is conservative

Although the exact test provides exact p-value, its type I error rate is conservative (smaller than the nominal level).

[Example] In the previous example for testing $H_0 : \pi = 0.5$ versus $H_1 : \pi > 0.5$, we have shown that the exact p-value equals $\Pr(Y \geq 9) = 0.0107$. Furthermore,

$$\begin{aligned} \Pr(Y \geq 8) &= \frac{10!}{8!2!}(0.5)^8(0.5)^2 + \Pr(Y \geq 9) \\ &= 0.0546. \end{aligned}$$

That is, the actual probability of rejecting H_0 is 0.0107 when $y = 9$ or $y = 10$ is observed. This probability is much less than 0.05. The exact test is too conservative.

One method is to use mid p-value: adding half the probability of the observed result to the probability of the more extreme results.

In the previous example, the mid p-value for $Y = 9$ is

$$\begin{aligned} \Pr(Y = 9)/2 + \Pr(Y > 9) &= \frac{10!}{9!1!}(0.5)^9(0.5)^1/2 + \Pr(Y = 10) \\ &= 0.006. \end{aligned}$$

Similarly, the mid p-value for

$Y = 8 \rightarrow \Pr(Y = 8)/2 + \Pr(Y > 8) = 0.033.$

$Y = 7 \rightarrow \Pr(Y = 7)/2 + \Pr(Y > 7) = 0.113.$

So at level 0.05, you would reject the null if $Y = 8, 9$, or 10. The probability of rejection is thus $\Pr(Y \geq 8) = 0.0546875$, that is, reject the null more often than the ordinary p-value.