STAT 5107: Discrete Data Analytics

Yuanyuan LIN

Department of Statistics

The Chinese University of Hong Kong

Chapter 1. Distribution and Inference for Categorical Data

1.1.1 Definition

Categorical variable: A variable has a measurement scale consisting of a set of categories.

Examples:

1. x_1 = Grade received in a class
   Five categories: A, B, C, D, E
2. x_2 = Social class
   Three categories: upper, middle, lower
3. x_3 = Gender of a patient
   Two categories: male, female
4. x_4 = Mode of transportation to work
   Five categories: automobile, bicycle, bus, subway, walk
5. x_5 = political philosophy
   Three categories: liberal, moderate, conservative

1.1 Categorical Response Data

Categorical data are by no means restricted to the social and biomedical sciences. They frequently occur in

- behavioral sciences (e.g., type of mental illness with categories schizophrenia, depression, neurosis)
- epidemiology and public health
- education (e.g., whether a student response to an exam question is correct or incorrect)
- marketing (e.g., consumer preference among the three leading brands of a product)

1.1 Categorical Response Data

- **Response-Explanatory variable distinction**: Statistical analysis distinguish *response* (or *dependent*) variable and *explanatory* (or *independent*) variables. This course focuses on methods for categorical response variable. As in ordinary regression modelling, explanatory variables can be any type.

- **Discrete-Continuous variable distinction**: Variables are classified as *discrete* or *continuous*, according to whether the number of value they can take is countable. Actual measurement of all variables occurs in a discrete manner, in practice, distinguishes between variables that take few values and variable that take lots of values.

- **Discrete-Continuous variable distinction**:
  For instance: statisticians often treat discrete interval variables having a large number of values (such as test scores) as continuous, using them in methods for continuous responses.

1.1.2 Data set

A data set of categorical variables consists of frequency counts for the categories.

e.g. Observations of X_1 in a class with $N = 50$ students:

Grade received	A	B	C	D	E
Frequency counts	15	25	7	2	1

1.1.3 Classifying categorical variables

Nominal variables: variables having categories without a natural ordering.

e.g. x_3 – Gender of a patient

x_4 – Mode of transportation to work

For a nominal variable, the order of listing the categories is irrelevant.

Ordinal variables: variables having ordered categories.

e.g. x_1 – Grade received in a class

x_2 – Social economic status

Ordinal variables have ordered categories, but distances between categories are unknown.

Interval variables: variables having numerical distances between any two values

e.g. blood pressure level, annual income

The levels of categorical variables depend on the amount of information they include:

nominal variables -> ordinal variables ->interval variables
(lowest level) (highest level)

1.1 Categorical Response Data

This course deals with certain types of discretely measured responses:

(1) binary variables

(2) nominal variables

(3) ordinal variables

(4) discrete interval variables having relatively few values

(5) continuous variables groups into a small number of categories

1.2.1 Binomial distribution

Let $y_1, y_2, \cdots, y_n$ denote responses for n independent and identical trials such that $p(y_i = 1) = \pi$ and $p(y_i = 0) = 1 - \pi$. Then, $y = \sum_{i=1}^{n} y_i$ has the binomial distribution $B(n, \pi)$:

$$p(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, 2, \cdots, n.$$

Mean:

$$\mu = E(y) = n\pi$$

Variance:

$$\sigma^2 = Var(y) = n\pi(1 - \pi)$$

For a fixed π, the distribution converges to normality as $n \to \infty$.

1.2.2 Multinomial distribution

For $i = 1, \cdots, n$, $j = 1, \cdots, c$, let

$$y_{ij} = \begin{cases} 1, & \text{if trial } i \text{ has an outcome in category } j, \\ 0, & \text{otherwise,} \end{cases}$$

$\mathbf{y}_i = (y_{i1}, y_{i2}, \cdots, y_{ic})$ represents a multinomial trial, with $\sum_{j=1}^{c} y_{ij} = 1$. Let $n_j = \sum_{i=1}^{n} y_{ij}$ denote the number of trials having outcome in category j, $\pi_j = P(y_{ij} = 1)$, then the counts $(n_1, n_2, \cdots, n_c)$ have the multinomial distribution:

$$p(n_1, n_2, \cdots, n_{c-1}) = \left(\frac{n!}{n_1! n_2! \cdots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}$$

1.2 Distribution for Categorial Data

The marginal distribution of each n_j is binomial. If we let the j-th cell as success and lump the remaining cells into a single cell as failure, we then have $n_j \sim B(n, \pi_j)$, and

Mean:

$$\mu_j = E(n_j) = n\pi_j,$$

Variance:

$$Var(n_j) = n\pi_j(1 - \pi_j),$$

Covariance:

$$Cov(n_j, n_h) = -n\pi_j\pi_h.$$

1.2 Distribution for Categorial Data

Example: $c = 5$,

y	1	2	3	4	5
p	π_1	π_2	π_3	π_4	π_5

3 — (0,0,1,0,0) , 2 — (0,1,0,0,0)

5 — (0,0,0,0,1) , 3 — (0,0,1,0,0)

1 — (1,0,0,0,0) , 4 — (0,0,0,1,0)

$\vdots$

Repeat n multinomial trials ($\sum_{j=1}^{5} n_j = n, \ \sum_{j=1}^{5} \pi_j = 1$):

$$P(n_1, n_2, n_3, n_4) = \left(\frac{n!}{n_1!\, n_2!\, n_3!\, n_4!\, n_5!} \right) \pi_1^{n_1} \pi_2^{n_2} \pi_3^{n_3} \pi_4^{n_4} \pi_5^{n_5}.$$

- A special case:

$$z_j = \begin{cases} 1, & \text{if } y_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}$$

Then, we have

z_j	1	0
$P(z_j = 1)$	π_j	1 -π_j

Repeat n Bernoulli trials, and let n_j be the number of $(z_j = 1)$,

$$P(n_j) = \frac{n!}{n_j!(n-n_j)!}\pi_j^{n_j}(1-\pi_j)^{n-n_j} = \binom{n}{n_j}\pi_j^{n_j}(1-\pi_j)^{n-n_j}.$$

1.2.3 Poisson distribution

The Poisson distribution is used for describing the counts of events that occur randomly over time or space, when outcomes in disjoint periods or regions are independent.

The poisson distribution Pois(μ):

$$p(y) = \frac{e^{-\mu}\mu^y}{y!}, \quad y = 0, 1, 2, \cdots$$

Mean: $E(y) = \mu$,
Variance: $Var(y) = \mu$.

The distribution converges to normality as $\mu \to \infty$.
It is an approximation of the binomial when n is large and π is small, with $\mu = n\pi$.

1.2.4 Negative Binomial distribution

Duality between Binomial and Negative Binomial:

- Binomial:

n — Number of Bernoulli trials (fix)
y — Number of successes among n Bernoulli trials (random)

$$P(y=y)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y},\quad y=0,1,\cdots,n$$

- Negative Binomial:

r — Number of successes (fix)
y — Number of Bernoulli trials until r successes (random)

$$P(y=y)=\binom{y-1}{r-1}\pi^{r}(1-\pi)^{y-r},\quad y=r,r+1,\cdots$$

- Geometric distribution:

When $r = 1$ (a special case of Negative Binomial),
y — Number of Bernoulli trials until the first success (random)

$$P(y = y) = \pi(1 - \pi)^{y-1}, \; y = 1, 2, \cdots$$

1.3.1 Likelihood function and maximum likelihood estimation

The part of a likelihood function involving the parameters is called the kernel. Since the maximization of the likelihood is with respect to the parameters, the rest is irrelevant.

The binomial log-likelihood is

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

Differentiating with respect to π and equating it to 0 yields

$$\frac{\partial L(\pi)}{\partial \pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi} = \frac{y-n\pi}{\pi(1-\pi)} = 0.$$

So, $\hat{\pi} = y/n$, the sample proportion of successes for the n trials.

1.3.2 Test about a binomial parameter

Consider $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$

1. The Score test statistic is

$$z_s = \frac{\widehat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}}$$

2. The Wald test statistic is

$$z_w = \frac{\widehat{\pi} - \pi_0}{\sqrt{\widehat{\pi}(1 - \widehat{\pi})/n}}$$

One refers z_s and z_w to the standard normal table to obtain one- or two-sided P-values.

3. The likelihood ratio test statistic is

$$
\begin{aligned}
& G^2 = -2[L(\pi_0) - L(\widehat{\pi})] = 2[L(\hat{\pi}) - L(\pi_0)] \\
&= 2\left[\log\{\hat{\pi}^y(1-\hat{\pi})^{n-y}\} - \log\{\pi_0^y(1-\pi_0)^{n-y}\}\right] \\
&= 2\left[y\log(\hat{\pi}) + (n-y)\log(1-\hat{\pi}) - y\log(\pi_0) - (n-y)\log(1-\pi_0)\right] \\
&= 2\left[y\log\frac{\widehat{\pi}}{\pi_0} + (n-y)\log\frac{1-\widehat{\pi}}{1-\pi_0}\right].
\end{aligned}
$$

The statistic has a limiting null χ_1^2 distribution, as $n \to \infty$.

1.3.3 Confidence intervals for a binomial parameter

1. The Score confidence interval of π_0 is $|z_s| < z_{\alpha/2}$, or

$$\widehat{\pi} \pm z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)/n}$$

2. The Wald confidence interval of π_0 is $|z_w| < z_{\alpha/2}$, or

$$\widehat{\pi} \pm z_{\alpha/2}\sqrt{\widehat{\pi}(1-\widehat{\pi})/n}$$

3. The likelihood-ratio-based confidence interval is

$$[\pi_0 : -2[L(\pi_0) - L(\widehat{\pi})] < \chi^2_{1,\alpha}] \text{ or } [\pi_0 : G^2 \leq \chi^2_{1,\alpha}].$$

A rough guideline for a large-sample test and confidence intervals is $n\pi > 5$ and $n(1-\pi) > 5$.

1.4.1 Likelihood function and maximum likelihood estimation

The multinomial log-likelihood function is

$$L(\pi) = \sum_j n_j \log \pi_j,$$

where

$$\sum_j \pi_j = 1, \quad \sum_j n_j = n.$$

From $\partial L(\pi)/\partial\pi = 0 \rightarrow \widehat{\pi}_j = n_j/n$.

Find maximum likelihood estimation (MLE) of π_j:

$$L(\pi) = \sum_{j=1}^{c} n_j \log \pi_j =$$

$$n_1 \log \pi_1 + n_2 \log \pi_2 + \cdots + n_{c-1} \log \pi_{c-1} + n_c \log(1 - \pi_1 - \cdots - \pi_{c-1})$$

$$\frac{\partial L(\pi)}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{1 - \pi_1 - \cdots - \pi_{c-1}} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c} = 0$$

$$\Rightarrow n_j \pi_c = \pi_j n_c \Rightarrow (\sum_{j=1}^{c} n_j)\pi_c = (\sum_{j=1}^{c} \pi_j) n_c$$

$n\pi_c = 1 n_c \Rightarrow \hat{\pi}_c = \frac{n_c}{n},$ and $\hat{\pi}_j = \frac{n_j}{n}, \ \ j = 1, \cdots, c.$

1.4.2 Hypothesis testing

1. Pearson statistic for testing a specified multinomial.

Case A: cell probabilities are completely specified by H_0.

$H_0 : \pi_j = \pi_{j0},\ j = 1, \cdots, c$, where $\sum_j \pi_{j0} = 1$.

If H_0 is true, the expected frequencies $E_j = n\pi_{j0},\ j = 1, \cdots, c$.
The test statistic (Pearson's χ^2) is:

$$X^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^{c} \frac{(n_j - n\pi_{j0})^2}{n\pi_{j0}},$$

where O_j is the observed cell frequency, and E_j is the expected cell frequency under H_0. When n is large enough ($E_j \geq 1$ for all j and no more than 20% of E_j are less than 5; combine cells if necessary), X^2 is distributed as chi-square with $df = c - 1$.

Since a large value of the overall discrepancy indicates a disagreement between the data and the hypothesis, reject H_0 if $X^2 \geq \chi^2_{c-1,\alpha}$, where $\chi^2_{c-1,\alpha}$ is the upper α probability point of the chi-square distribution with $df = c - 1$.

- Example 1.1:

H_0: $\pi_{10} = 1/7, \pi_{20} = 1/7, \pi_{30} = 2/7, \pi_{40} = 3/7$

Cell j	1	2	3	4	Total
O_j = frequency	12	13	20	25	70
E_j = nπ_{j0}	10	10	20	30	70

$X^2 = \sum_{j=1}^{c} \frac{(O_j - E_j)^2}{E_j} = \frac{(12-10)^2}{10} + \cdots + \frac{(25-30)^2}{30} = 2.133,$

$\chi^2_{3,0.05} = 7.814$, do not reject H_0.

Case B : cell probabilities are not completely specified by H_0

e.g. $H_0 : \pi_1 = \pi_2,\ \pi_3 = \pi_4$

$H_0 : \pi_1 + \pi_2 = \pi_3$

Test statistic (Pearson's χ^2):

$$X^2 = \sum_{j=1}^{c} \frac{(O_j - \widehat{E}_j)^2}{\widehat{E}_j} = \sum_{j=1}^{c} \frac{(n_j - n\widehat{\pi}_j)^2}{n\widehat{\pi}_j},$$

where $\widehat{\pi}_j$ is the estimate of π_j under H_0. When n is large enough ($n\widehat{\pi}_j \geq$ 1 for all j and no more than 20% of $n\widehat{\pi}_j$ are less than 5), X^2 is distributed as chi-square with degree of freedom:
$df =$ No. of cells$-1-$No. of independent parameters estimated.

- Find the MLE of π_j in Case B

Under $H_0 : \pi_1 = \pi_2,\ \pi_3 = \pi_4$,

$$L(\pi) = n_1 \log \pi_1 + n_2 \log \pi_1 + n_3 \log \pi_3 + n_4 \log \pi_3$$
$$= (n_1 + n_2) \log \pi_1 + (n_3 + n_4) \log \pi_3$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 2\pi_1 + 2\pi_3 = 1 \Rightarrow \pi_1 + \pi_3 = \tfrac{1}{2} \ \Rightarrow\ \pi_3 = \tfrac{1}{2} - \pi_1$$

So, $\frac{\partial \pi_3}{\partial \pi_1} = -1 \ \Rightarrow\ \frac{\partial L(\pi)}{\partial \pi_1} = \frac{n_1+n_2}{\pi_1} - \frac{n_3+n_4}{\pi_3} = 0 \ \Rightarrow$

$$(n_1+n_2)\pi_3 = (n_3+n_4)\pi_1 \ \Rightarrow\ (n_1+n_2+n_3+n_4)\pi_3 = (n_3+n_4)(\pi_1+\pi_3)$$
$$\Rightarrow\ n\pi_3 = (n_3 + n_4)/2 \ \Rightarrow$$

$$\hat{\pi}_3 = \frac{n_3 + n_4}{2n} = \hat{\pi}_4, \qquad \hat{\pi}_1 = \frac{n_1 + n_2}{2n} = \hat{\pi}_2.$$

- Example 1.2:

$H_0 : \pi_1 = \pi_2, \; \pi_3 = \pi_4$

Cell j	1	2	3	4	Total
O_j = frequency	12	13	20	25	70
$\hat{E}_j = n\hat{\pi}_{j0}$	12.5	12.5	22.5	22.5	70

Under H_0, the MLEs are:

$$\hat{\pi}_1 = \hat{\pi}_2 = \frac{n_1+n_2}{2n} = \frac{25}{140} = \frac{5}{28}, \qquad \hat{\pi}_3 = \hat{\pi}_4 = \frac{n_3+n_4}{2n} = \frac{45}{140} = \frac{9}{28}.$$

We then have

$$X^2 = \sum_{j=1}^{4} \frac{(n_j - n\widehat{\pi}_j)^2}{n\widehat{\pi}_j} = \frac{(12-12.5)^2}{12.5} + \cdots + \frac{(25-22.5)^2}{22.5} = 0.596,$$

$$df = 4 - 1 - 1 = 2,$$

$\chi^2_{2,0.05} = 5.991 > 0.596$, do not reject H_0.

2. The likelihood-ratio test (LRT)

$X_1, \cdots, X_n$ are sampled from $f(x|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \in R^k$.

$H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta - \Theta_0$ ($\Theta_0 \subset \Theta$)

Likelihood function: $l(\boldsymbol{\theta}) = \prod_i f(x_i|\theta)$,
Log-likelihood function: $L(\boldsymbol{\theta}) = \log l(\boldsymbol{\theta})$
Likelihood-ratio test statistic:

$$G^2 = -2\log\wedge = -2\log\frac{l(\widehat{\theta_0})}{l(\widehat{\theta})} = -2[L(\widehat{\theta_0}) - L(\widehat{\theta})],$$

where

$\hat{\theta}_0$ = MLE of $\boldsymbol{\theta}$ under Θ_0,
$\hat{\theta}$ = MLE of $\boldsymbol{\theta}$ under Θ.

Under H_0, we have:

$$G^2 = -2\log \wedge \to \chi^2_r,$$

r = number of parameters estimated under $H_0 \cup H_1$ $-$ number of parameters estimated under H_0.

Reject H_0 if $G^2 \geq \chi^2_{r,\alpha}$.

Note: Likelihood-ratio test statistic G^2 and Pearson's χ^2 test statistic X^2 are asymptotically equivalent.

The likelihood-ratio test statistic in Case A:

$H_0 : \pi_j = \pi_{j0},\ j = 1, \cdots, c,$ where $\sum_j \pi_{j0} = 1.$

The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\pi_{j0})^{n_j}}{\prod_j (n_j/n)^{n_j}} = \prod_j \left(\frac{n\pi_{j0}}{n_j} \right)^{n_j}.$$

The likelihood-ratio test statistic is:

$$G^2 = -2\log\Lambda = 2\sum_j n_j \log\frac{n_j}{n\pi_{j0}} = 2\sum_j O_j \log\frac{O_j}{\widehat{E}_j} \sim \chi^2_{c-1},$$

where

O_j = observed cell frequency
$\widehat{E}_j$ = estimated expected cell frequency under H_0.

The likelihood-ratio test statistic in Case B:

e.g. $H_0 : \pi_1 = \pi_2,\ \pi_3 = \pi_4$
$H_0 : \pi_1 + \pi_2 = \pi_3$

The ratio of the likelihoods equals

$$\Lambda = \frac{\prod_j (\hat{\pi}_j)^{n_j}}{\prod_j (n_j/n)^{n_j}} = \prod_j \left(\frac{n\hat{\pi}_j}{n_j} \right)^{n_j}.$$

The likelihood-ratio test statistic is

$$G^2 = -2\log\Lambda = 2\sum_j n_j \log\frac{n_j}{n\hat{\pi}_j} = 2\sum O_j \log\frac{O_j}{\hat{E}_j},$$

where $\widehat{E}_j = n\hat{\pi}_j$.

- Example 1.3:

Cell j	1	2	3	4	Total
O_j = frequency	12	13	20	25	70
$\widehat{E}_j = n\widehat{\pi_{j0}}$	12.5	12.5	22.5	22.5	70

$H_0 : \pi_1 = \pi_2, \ \pi_3 = \pi_4$

$H_0 \cup H_1$: a multinomial model

Under H_0, the MLEs of $\pi_1 = \pi_2$ and $\pi_3 = \pi_4$ are

$$\widehat{\pi}_1 = \widehat{\pi}_2 = \frac{O_1+O_2}{2n} = \frac{23}{140} = \frac{5}{28},$$
$$\widehat{\pi}_3 = \widehat{\pi}_4 = \frac{O_3+O_4}{2n} = \frac{45}{140} = \frac{9}{28}.$$

$$G^2 = 2\sum O_j \log \frac{O_j}{\widehat{E}_j}$$

$$= 2[12 \times \log(\frac{12}{12.5}) + 13 \times \log(\frac{13}{12.5}) + 20 \times \log(\frac{20}{22.5}) + 25 \times \log(\frac{25}{22.5})]$$

$= 0.597.$ (Close to $X^2 = 0.596$ in Example 1.2)

$$G^2 \sim \chi^2_r,$$

where r = Number of parameters estimated under $H_0 \bigcup H_1$
$-$ Number of parameters estimated under H_0

$\chi^2_{2,0.05} = 5.991$. Again, H_0 is not rejected.

1.5.1 Likelihood function and maximum likelihood estimation

The poisson log-likelihood function is:

$$L(\mu) = \log[e^{-\mu}\mu^{y}] = y\log(\mu) - \mu$$

From $\partial L(\mu)/\partial\mu = y/\mu - 1 = 0 \quad \Rightarrow \quad \widehat{\mu} = y$.

For a sample with size n: $y_1, \cdots, y_n$,

$$L(\mu) = \log[\prod_{i=1}^{n} e^{-\mu}\mu^{y_i}] = \sum_{i=1}^{n} y_i \log(\mu) - n\mu$$

From $\partial L(\mu)/\partial\mu = \sum_{i=1}^{n} y_i/\mu - n = 0 \quad \Rightarrow \quad \widehat{\mu} = \sum_{i=1}^{n} y_i/n = \bar{y}$.

The mean and standard error of $\widehat{\mu}$:

$$E(\widehat{\mu}) = \mu, \qquad Var(\widehat{\mu}) = \mu.$$

1.5.2 Hypothesis testing

Pearson's χ^2 statistic for testing a family of distribution:

H_0: The data come from a Poisson distribution

- Example 1.4:

X	0	1	2	3	4	5	6	≥ 7
O_j	22	53	58	39	20	5	2	1
$\hat{\pi}_j$	0.135	0.271	0.271	0.18	0.09	0.36	0.012	0.005
$\widehat{E}_j$	27	54.2	54.2	36	18	7.2	2.4	1.0

Total sample size = 200
O_j = Observed frequency
$\widehat{\pi}_j$ = Estimated probability under H_0
$\widehat{E}_j$ = Estimated expected frequency under H_0

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \qquad x = 0, 1, 2, \cdots$$

μ: the average number of occurrence per unit

Under $H_0,\ X \sim$ Poisson distribution.

To estimate π_j, we first find the MLE of μ:

$$\widehat{\mu} = \bar{X} = \frac{0 \times 22 + 1 \times 53 + \cdots + 7 \times 1}{200} = 2.05 \sim 2.0.$$

So,

$\widehat{\pi}_1 = P(X = 0) = \frac{e^{-2}2^0}{0!} = 0.135, \qquad \widehat{E_1} = 200 \times 0.135 = 27$

$\widehat{\pi}_2 = P(X = 1) = \frac{e^{-2}2^1}{1!} = 0.271, \qquad \widehat{E_2} = 200 \times 0.271 = 54.2$

$\vdots$

$\widehat{\pi}_8 = P(X \geq 7) = 1 - \sum_{j=1}^{7} \widehat{\pi}_j = 0.005, \qquad \widehat{E_8} = 200 \times 0.005 = 1.$

Since the sum of the expected frequency of the last two cells is smaller than 5, combine with the cell of $X = 5$, and thus $c = 6$.

$$X^2 = \sum_{j=1}^{6} \frac{(O_j - \widehat{E}_j)^2}{\widehat{E}_j} = 2.33.$$

$X^2 \sim \chi^2_{df}$,

df = Number of cells - 1 - Number of estimated parameters

= $6 - 1 - 1 = 4$,

$\chi^2_{4,0.05} = 9.49.$

So, H_0 is not rejected. That is, the Poisson model does not contradict the data.

(End of Chapter 1)