DATA 606: Statistical Methods in Data Science

—— Introduction of categorical data analysis

Wenjun Jiang

Department of Mathematics & Statistics
The University of Calgary

Lecture 7

UNIVERSITY OF CALGARY

General intro

Definition 1 (Categorical variable)

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

Example 1

1. Political philosophy: liberal, moderate or conservative.
2. Diagnoses regarding some cancer: normal, benign or malignant.

General intro

Definition 1 (Categorical variable)

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

Example 1

1. Political philosophy: liberal, moderate or conservative.
2. Diagnoses regarding some cancer: normal, benign or malignant.

A categorical variable

- Could be either response or explanatory variable.
- Could be binary, nominal or ordinal scale.
- Could be discrete or continuous.
- Could be qualitative or quantitative.

Distributions

Three key distributions for categorical data: **binomial, multinomial and Poisson**.

Distributions

Three key distributions for categorical data: **binomial, multinomial and Poisson**.

1. Binomial distribution (π, n)
   - Fixed number of observations, e.g. n.
   - Observations $(y_i, i = 1, 2, \ldots, n)$ are binary, e.g. $y_i = 1$ or $y_i = 0$.
   - Fixed probability, e.g. $\mathbf{P}(Y_i = 1) = \pi$.
   - Observations are independent.

   Let $Y = \sum_{i=1}^{n} Y_i$, then Y is said to follow the binomial distribution, denoted as $\text{bin}(n, \pi)$.

Binomial distribution

The statistical properties of a binomial distribution

- The probability mass function

$$\mathbf{P}(Y = y) = \binom{n}{k} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n.$$

- The mean and variance

$$\mathbf{E}[Y] = n\pi, \quad \mathrm{Var}(Y) = n\pi(1-\pi).$$

Multinomial distribution

2. Multinomial distribution $(\pi_1, \dots, \pi_c)$
   - Fixed number of observations/trials, e.g. n.
   - c categories of outcomes $(1, 2, \dots, c)$.
   - Each kind of outcome appears with fixed probability, e.g. $\pi_1, \dots, \pi_c$.

Let $N_1, \dots, N_c$ count the number of appearance of each kind of outcome, then $(N_1, \dots, N_c)$ is said to follow the multinomial distribution.

- The probability mass function $(n_1 + n_2 + \cdots + n_c = n)$

$$\mathbf{P}(N_1 = n_1, \dots, N_c = n_c) = \frac{n!}{n_1! \cdots n_c!} \pi_1^{n_1} \cdots \pi_c^{n_c},$$

- Statistical properties

$$\mathbf{E}[N_i] = n\pi_i, \quad \mathrm{Var}(N_i) = n\pi_i(1 - \pi_i), \quad \mathrm{Cov}(N_i, N_j) = -n\pi_i\pi_j.$$

Poisson distribution

3. Poisson distribution (μ) Rare events?
   - Count data do not result from a fixed number of observations/trials.
   - There is no upper bound for the number of appearance of the outcome.

- The probability mass function

$$\mathbf{P}(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2 \ldots .$$

- Statistical properties

$$\mathbf{E}[Y] = \mu, \quad \mathrm{Var}(Y) = \mu.$$

Overdispersion

Definition 2 (Overdispersion)

count observations often exhibit variability exceeding that predicted by the preset distribution.

Overdispersion

Definition 2 (Overdispersion)

count observations often exhibit variability exceeding that predicted by the preset distribution.

An example:

- We assume each day there is a fixed probability for the tornado to occur.
- The probability actually is changing w.r.t other factors, such as temperature, moisture, whether it is rainy or windy, etc..
- Suppose Y is a random variable which is Poisson distributed conditional on μ. This μ is not fixed in reality.
- Using conditional mean and variance formulas

$$\mathbf{E}[Y] = \mathbf{E}[\mathbf{E}[Y|\mu]] = \mathbf{E}[\mu],$$
$$\mathrm{Var}(Y) = \mathbf{E}[\mathrm{Var}(Y|\mu)] + \mathbf{E}[\mathrm{Var}(Y|\mu)] = \mathbf{E}[\mu] + \mathbf{E}[\mathrm{Var}(Y \mid \mu)].$$

An alternative

4. Negative binomial (π, k)
   - Each time binary outcome, success or failure.
   - Each time, fixed probability for success, e.g. π.
   - Y is the number of failures before k successes occur.

- The probability mass function

$$\mathbf{P}(Y = y) = \binom{y + k - 1}{y} (1 - \pi)^y \pi^k, \quad y = 0, 1, 2 \ldots .$$

- Statistical properties

$$\mathbf{E}[Y] = \frac{\pi k}{1 - \pi}, \quad \mathrm{Var}(Y) = \frac{\pi k}{(1 - \pi)^2}$$

Likelihood function

Definition 2 (Likelihood function)

A likelihood function is the probability of the observed data.

Likelihood function

Definition 2 (Likelihood function)

A likelihood function is the probability of the observed data.

Example 3

Suppose Y_1, Y_2, Y_3 all follow Poisson distribution with parameter μ, then given $Y_1 = 4, Y_2 = 2, Y_3 = 1$, the likelihood function is

$$L(\mu) = \frac{e^{-\mu}\mu^4}{4!} \cdot \frac{e^{-\mu}\mu^2}{2!} \cdot \frac{e^{-\mu}\mu}{1!} = \frac{e^{-3\mu}\mu^7}{48}$$

Maximum likelihood estimation

Maximum likelihood estimation (MLE)

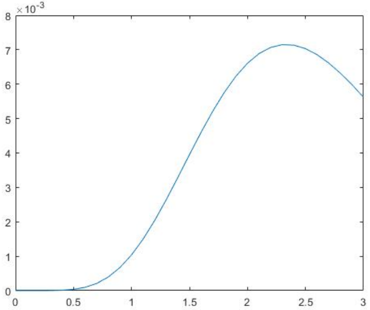
Find the parameters which could maximize the likelihood function.

Maximum likelihood estimation

Maximum likelihood estimation (MLE)

Find the parameters which could maximize the likelihood function.

Example 4 (Cont.)



$$L'(\mu) = -\frac{e^{-3\mu}\mu^7}{16} + \frac{7e^{-3\mu}\mu^6}{48} = 0 \Longrightarrow \hat{\mu} = \frac{7}{3}.$$

Maximum likelihood function

Note that maximizing the likelihood function is **equivalent to** maximizing its log version:

$$\max \ L(\mu) \iff \max \ l(\mu) = \log(L(\mu)).$$

It is easier to calculate based on the log-likelihood function.

Example 5 (Cont.)

$$\log(L(\mu)) = \log(e^{-3\mu}) + \log(\mu^7) - \log(48) = -3\mu + 7\log(\mu) - \log(48),$$

$$\frac{d\log(L(\mu))}{d\mu} = -3 + \frac{7}{\mu} = 0 \Longrightarrow \hat{\mu} = \frac{7}{3}.$$

MLE for binomial distribution

Suppose Y follows binomial distribution (n, π) where π is unknown, and we observe $Y = y$, then

$$L(\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y},$$

$$l(\pi) = \log L(\pi) = \log \binom{n}{y} + y \log \pi + (n-y) \log(1-\pi).$$

MLE for binomial distribution

Suppose Y follows binomial distribution (n, π) where π is unknown, and we observe $Y = y$, then

$$L(\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y},$$

$$l(\pi) = \log L(\pi) = \log \binom{n}{y} + y \log \pi + (n-y) \log(1-\pi).$$

Then

$$\frac{dl(\pi)}{d\pi} = \frac{y}{\pi} - \frac{n-y}{1-\pi} = 0 \Longrightarrow \hat{\pi} = \frac{y}{n}.$$

More about MLE

In the previous examples, the estimator $\hat{\mu}$ and $\hat{\pi}$ are both **random**.

Question: without parametric form, can you tell the variance of the MLE estimator?

Answer:

- *Information matrix*: $\nu(\mu) = -\mathbf{E}\left(l''(\mu)\right) = -\mathbf{E}\left(\frac{d^2 l(\mu)}{d\mu^2}\right)$.
- The asymptotic variance of $\hat{\mu}$ is $\frac{1}{\nu(\mu)}$.

Variance for binomial MLE

In binomial example, the log-likelihood function is (neglecting the constant term)

$$l(\pi) = Y \log \pi + (n - Y) \log(1 - \pi).$$

Its second-order derivative is

$$l''(\pi) = -\frac{Y}{\pi^2} - \frac{n - Y}{(1 - \pi)^2}.$$

Then the information matrix of π is

$$\nu(\pi) = -\mathbf{E}(l''(\pi)) = \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1 - \pi)^2} = \frac{n}{\pi(1 - \pi)}.$$

Therefore, the asymptotic variance is $\frac{\pi(1-\pi)}{n}$[1].

[1]The MLE estimator $\hat{\pi} = \frac{y}{n}$, its variance is also $\frac{\pi(1-\pi)}{n}$.

More about MLE

Asymptotic property of MLE estimator: when the number of observations n is large, the estimator is more close to normal distributed.

More about MLE

Asymptotic property of MLE estimator: when the number of observations n is large, the estimator is more close to normal distributed.

In other words, if β is the parameter to be estimated, then when $n \to \infty$,

$$\hat{\beta} \sim \text{Normal}(\mathbf{E}[\hat{\beta}], \sigma(\hat{\beta})),$$

where $\hat{\beta}$ is the MLE estimator.

Several tests

Motivation: in daily life, we always encounter hypothesis tests, such that

$$\begin{aligned} \mathcal{H}_0: &\quad \beta = \beta_0 \\ \mathcal{H}_1: &\quad \beta \neq \beta_0. \end{aligned}$$

β is some parameter in the model we apply to the practical problem.

Several tests

Motivation: in daily life, we always encounter hypothesis tests, such that

$$\begin{aligned} \mathcal{H}_0 : &\quad \beta = \beta_0 \\ \mathcal{H}_1 : &\quad \beta \neq \beta_0. \end{aligned}$$

β is some parameter in the model we apply to the practical problem.

How to test whether $\mathcal{H}_0$ is acceptable or not?
We use MLE estimator and the following three constructed tests.

Wald statistic

- With MLE, we could obtain $\hat{\beta}$, as well as its information $\nu(\hat{\beta}) - \mathbf{E}[\frac{\partial^2 l(\beta)}{\partial \beta^2}|_{\beta=\hat{\beta}}]$.
- Under the NULL hypothesis $\mathcal{H}_0$, the following statistic is proved to asymptotically follow standard normal distribution

$$Z = \frac{\hat{\beta} - \beta_0}{\sigma(\hat{\beta})}$$

where

$$\sigma^2(\hat{\beta}) = \frac{1}{\nu(\hat{\beta})}.$$

- As $Z \sim \text{Normal}(0, 1)$, if Z is too small or too large, then the null hypothesis should be rejected.

Likelihood ratio test

- You have a vector of parameters to estimate: $\boldsymbol{\beta} = (\beta_0, \beta_1)$.
- You are given a hypothesis: $\mathcal{H}_0 : \beta_0 = 0$ and want to test if this hypothesis is acceptable.
- You apply MLE to the data and obtain the following parameters:

$$\mathcal{H}_0 \text{ is assumed : } \quad (0, \tilde{\beta}_1),$$

$$\mathcal{H}_0 \text{ is NOT assumed : } \quad (\hat{\beta}_0, \hat{\beta}_1).$$

- You get two likelihoods:

$$L_0 = L(\boldsymbol{x}; 0, \tilde{\beta}_1), \quad L_1 = L(\boldsymbol{x}; \hat{\beta}_0, \hat{\beta}_1).$$

Likelihood ratio test

- Apparently, $L_1 \geq L_0$. Furthermore, the following statistic

$$-2\log\Delta = -2\log\frac{L_0}{L_1} = -2(l_0 - l_1)$$

  follows χ^2_q distribution, where q is equal to the difference between the dimensions of the two different parameter spaces[2].

- We hope L_0 is not far away from L_1, therefore, if $-2\log\Delta$ is too large, then $\mathcal{H}_0$ should be rejected.

[2] In this case, the df is equal to 1.

Score test

- You are given the hypothesis: $\mathcal{H}_0 : \beta = \beta_0$
- The first-order derivative of the log-likelihood function evaluated at β_0:

$$u(\beta_0) = \frac{\partial l(\beta)}{\partial \beta}\Big|_{\beta=\beta_0}.$$

- The expected second-order derivative of the log-likelihood function evaluated at β_0:

$$\nu(\beta_0) = -\mathbf{E}\Big[\frac{\partial^2 l(\beta)}{\partial \beta^2}\Big|_{\beta=\beta_0}\Big].$$

- The score statistic is $\frac{u(\beta_0)}{\sqrt{\nu(\beta_0)}}$[3] and it is proved to approximately follow standard normal distribution.

[3] Understand it as slope/curvature.

An illustrative graph

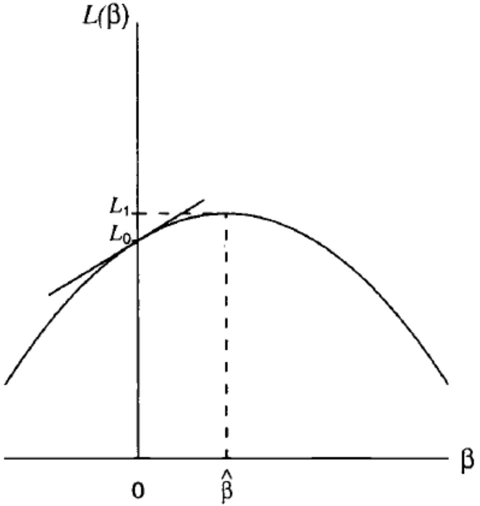


Figure 1.1 Log-likelihood function and information used in three tests of H_0: $\beta = 0$.

Tests about a binomial distribution

In a binomial distribution, the parameter to be estimated is π.

The MLE estimator for π is $\hat{\pi} = \frac{y}{n}$, its statistical characteristics are

$$\mathbf{E}[\hat{\pi}] = \pi, \quad \mathrm{Var}(\hat{\pi}) = \frac{\pi(n - \pi)}{n}.$$

Now we have our hypothesis:

$$\mathcal{H}_0 : \quad \pi = \pi_0,$$

$$\mathcal{H}_1 : \quad \pi \neq \pi_0.$$

Tests about a binomial distribution

- The Wald statistics

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

Tests about a binomial distribution

- The Wald statistics

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

- The slope and curvature

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \quad \nu(\beta_0) = \frac{n}{\pi_0(1-\pi_0)}.$$

The score statistic simplifies to

$$z_S = \frac{u(\pi_0)}{\sqrt{\nu(\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Tests about a binomial distribution

- Under $\mathcal{H}_0$,

$$l_0 = \log L_0 = y \log \pi_0 + (n - y) \log(1 - \pi_0).$$

- Without $\mathcal{H}_0$,

$$l_1 = \log L_1 = y \log \hat{\pi} + (n - y) \log(1 - \hat{\pi}).$$

- The likelihood-ratio test statistic is

$$-2(l_0 - l_1) = 2 \left[y \log \frac{\hat{\pi}}{\pi_0} + (n - y) \log \frac{1 - \hat{\pi}}{1 - \pi_0} \right].$$

  As here the difference between the dimensions of parameter spaces is 1. Therefore

$$-2(l_0 - l_1) \sim \chi_1^2.$$

CI for binomial distribution

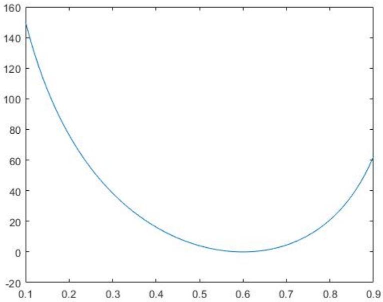
We could again utilize the aforementioned three constructed tests to obtain the *confidence intervals* for π.

- *Wald statistic* $|z_W| \leq z_{1-\alpha/2}$:

$$\left[\hat{\pi} - z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}},\ \hat{\pi} + z_{1-\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right].$$

CI for binomial distribution

- *Likelihood ratio test*



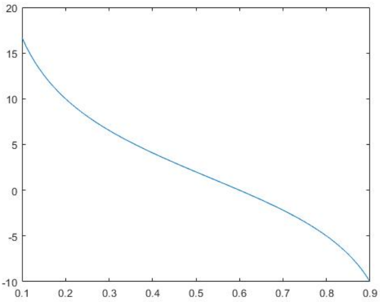
$$-2(l_0 - l_1) \leq \chi_1^2(0.95) \implies [\pi_0(1),\ \pi_0(2)],$$

where $\pi_0(1)$ and $\pi_0(2)$ are the roots of

$$2\left[y \log \frac{\hat{\pi}}{\pi_0} + (n-y)\log \frac{1-\hat{\pi}}{1-\pi_0}\right] = \chi_1^2(0.95).$$

CI for binomial distribution

- *Score statistic*



$$|z_S| \le z_{1-\alpha/2} \implies [\pi_0(1),\ \pi_0(2)],$$

where $\pi_0(1)$ and $\pi_0(2)$ are the roots of

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

An example

Example 6 (The proportion of vegetarians)

The instructor questioned the students in one class whether he or she was a vegetarian. Of $n = 25$ students, $y = 0$ answered "yes". Give a 95% confidence interval for the proportion of vegetarians over the whole student population on campus.

An example

Example 6 (The proportion of vegetarians)

The instructor questioned the students in one class whether he or she was a vegetarian. Of $n = 25$ students, $y = 0$ answered "yes". Give a 95% confidence interval for the proportion of vegetarians over the whole student population on campus.

- Since $y = 0$, the MLE estimate $\hat{\pi} = \frac{0}{25} = 0$. With the *Wald method*, the 95% CI for π is

$$\left[\hat{\pi} - 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n},\ \hat{\pi} + 1.96\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right]$$

which is $[0, 0]$.

An example

- With *Score statistic*, we first need to solve the equation

$$|\hat{\pi} - \pi| = 1.96\sqrt{\pi(1-\pi)/n},$$

which yields $\pi(1) = 0$ and $\pi(2) = 0.133$. Therefore the CI based on score statistics is $[0, 0.133]$

An example

- With *Score statistic*, we first need to solve the equation

$$|\hat{\pi} - \pi| = 1.96\sqrt{\pi(1-\pi)/n},$$

which yields $\pi(1) = 0$ and $\pi(2) = 0.133$. Therefore the CI based on score statistics is $[0, 0.133]$

- With *likelihood ratio test*, as we know the likelihood function is

$$l(\pi) = \pi^2(1-\pi)^{25} = (1-\pi)^{25},$$

The likelihood-ratio-based 95% CI is given by

$$-2(l_0 - l_1) = -2(l(\pi) - l(\hat{\pi})) = -50\log(1-\pi) \leq \chi_1^2(0.95) = 3.84$$

which gives $\pi \in [0, 0.074]$.

MLE for multinomial distribution

In a multinomial distribution, the parameters to be estimated are $\pi_1, \ldots, \pi_c$.

- We have observations $\pi_1, \ldots, \pi_c$ such that $n = n_1 + \cdots + n_c$.
- The likelihood function is $L(\pi_1, \ldots, \pi_c) = \pi_1^{n_1} \cdots \pi_c^{n_c}$, where the log-likelihood function is

$$l(\pi_1, \ldots, \pi_c) = \log L(\pi_1, \ldots, \pi_c) = \sum_{i=1}^{c} n_i \log \pi_i.$$

- Note that $\pi_1 + \cdots + \pi_c = 1$, therefore only first $c - 1$ parameters need to be estimated and $\pi_c = 1 - \pi_1 - \cdots - \pi_{c-1}$.

MLE for multinomial distribution

- Write $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_c)$, for $j \in \{1, 2, \ldots, c-1\}$, we have

$$\frac{\partial \log \pi_i}{\partial \pi_j} = \begin{cases} 0, & i \neq j \text{ and } i \neq c, \\ \frac{1}{\pi_j}, & i = j, \\ -\frac{1}{\pi_c}, & i = c. \end{cases}$$

and

$$\frac{\partial l(\boldsymbol{\pi})}{\partial \pi_j} = \frac{n_j}{\pi_j} - \frac{n_c}{\pi_c}.$$

Hence $\frac{\hat{\pi}_j}{\hat{\pi}_c} = \frac{n_j}{n_c}$. Recall $\sum_{i=1}^{c} \hat{\pi}_i = 1$, therefore $\hat{\pi}_j = n_j/n$.

Pearson Chi-square test

Now we are faced with another hypothesis test problem

$$\mathcal{H}_0 : \quad \pi_1 = a_1, \ldots, \pi_c = a_c.$$

With observations $n_1, \ldots, n_c$, how to check whether $\mathcal{H}_0$ is acceptable?

Pearson Chi-square test

Now we are faced with another hypothesis test problem

$$\mathcal{H}_0: \quad \pi_1 = a_1, \ldots, \pi_c = a_c.$$

With observations $n_1, \ldots, n_c$, how to check whether $\mathcal{H}_0$ is acceptable?

Let $n = n_1 + \cdots + n_c$ and $\mu_i = n \times a_i$. Pearson proposed the following test statistic

$$X^2 = \sum_{i=1}^{c} \frac{(n_i - \mu_i)^2}{\mu_i}$$

If the null hypothesis is wrong, then at least one term of the above sum is large.

Pearson Chi-square test

For large samples, the above test statistic

$$X^2 \sim \chi^2_{c-1}.$$

Hence, if $\underline{X^2 > \chi^2_{c-1}(0.95)}$, the hypothesis $\mathcal{H}_0$ could be rejected.