STA4030: Categorical Data Analysis
Preliminaries: Part II

Instructor: Bojun Lu

School of Science and Engineering
CUHK(SZ)

September 10, 2020

Agenda

1. 1.3 Likelihood and Maximum-likelihood Estimation
2. 1.4 Large Sample Inference

1.3 Likelihood and Maximum-likelihood Estimation

1.3.1 Likelihood functions

The distributions we have seen depend on parameters, many of which (e.g. π, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_2)$, μ, etc.) are unknown. Much of this course will involve making inferences about these unknown parameters. Our principal tool for doing so is likelihood.

Take a probability distribution (a PMF or PDF) $p(y)$. This depends on some unknown parameter(s), θ. So let's make that dependence explicit by writing $p(y) = p(y; \theta)$.

e.g. $Y \sim Po(\mu)$, hence $p(y) = p(y; \mu) = \frac{e^{-\mu}\mu^y}{y!}$

e.g. $Y \sim Bern(\pi)$, hence $p(y) = p(y; \pi) = \pi^y(1-\pi)^{1-y}$

1.3 Likelihood and Maximum-likelihood Estimation

If we plug an observed value into $p(y;\theta)$, we end up with a function of the unknown parameter(s) θ only

e.g. $Y \sim Po(\mu)$, we observe value of 3 hence

$$p(3) = p(3;\mu) = \frac{e^{-\mu}\mu^3}{6} := L(\mu)$$

e.g. $Y \sim Bern(\pi)$, we observe value of 0 hence

$$p(0) = p(0;\pi) = (1-\pi) := L(\pi)$$

Thus $L(\theta)$, called the likelihood function, is the result of plugging in an observed value into distribution function $p(y;\theta)$. To make the influence of the observed data y on the likelihood function explicit, we can use the notation $L(\theta) := L(\theta;y)$.

1.3 Likelihood and Maximum-likelihood Estimation

Of course, in practice we don't just observe one single value - we usually have an independent sample of size n, e.g. $y = (y_1, \ldots, y_n)$. In that case, the overall likelihood is a product of the individual likelihoods

$$L(\theta; y) = L(\theta; y_1) \times \cdots \times L(\theta; y_n) = \prod_{i=1}^{n} L(\theta; y_i) = \prod_{i=1}^{n} p(y_i; \theta).$$

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence

$$L(\mu; y) = e^{-n\mu} \frac{\mu^{\sum_{i=1}^{n} y_i}}{y_1! \cdots y_n!}.$$

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence

$$L(\pi; y) = \pi^{\sum_{i=1}^{n} y_i} (1 - \pi)^{n - \sum_{i=1}^{n} y_i}.$$

1.3 Likelihood and Maximum-likelihood Estimation

1.3.2 Loglikelihood function

For computational reasons, we will usually work with the loglikelihood function $l(\theta; y)$, which is just the natural logarithm of the likelihood function

$$l(\theta; y) := \log(L(\theta; y)) = \log(\prod_{i=1}^{n} L(\theta; y_i)) = \sum_{i=1}^{n} l(\theta; y_i)$$

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence
$l(\mu; y) = -n\mu + \sum_{i=1}^{n} y_i \log(\mu) - \sum_{i=1}^{n} \log(y!)$.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence
$l(\pi; y) = \sum_{i=1}^{n} y_i \log(\pi) + (n - \sum_{i=1}^{n} y_i) \log(1 - \pi)$.

1.3 Likelihood and Maximum-likelihood Estimation

1.3.3 Maximum likelihood estimation

As the sample size grows, two things happen for "nice" loglikelihood functions:

- they become more and more peaked around a maximum value
- their shape becomes more and more quadratic

Clearly this maximum value is important. We know how to find it: differentiate $l(\theta; x)$ with respect to θ; equate this to zero and solve. What results is a numerical value for θ which maximizes the loglikelihood and therefore the likelihood. Intuitively, it seems a good guess for what the true value of θ might be.

We usually label this number $\hat{\theta}$. It is called the **maximum likelihood estimate (MLE) of θ.**

1.3 Likelihood and Maximum-likelihood Estimation

e.g. $Y \sim Po(\mu)$, we observe $y = (y_1, \ldots, y_n)$ hence $l(\mu; y) = -n\mu + \sum_{i=1}^n y_i \log(\mu) - \sum_{i=1}^n \log(y!)$.

- Thus $\frac{\partial l(\mu;y)}{\partial \mu} = -n + \frac{1}{\mu}\sum_{i=1}^n y_i$.
- MLE $\hat{\mu}$ satisfies $0 = -n + \frac{1}{\hat{\mu}}\sum_{i=1}^n y_i$.
- Solving, we find that $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$, the sample mean.

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$ hence $l(\pi; y) = \sum_{i=1}^n y_i \log(\pi) + (n - \sum_{i=1}^n y_i)\log(1-\pi)$.

- Thus $\frac{\partial l(\pi;y)}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^n y_i - \frac{1}{1-\pi}(n - \sum_{i=1}^n y_i)$.
- MLE $\hat{\pi}$ satisfies $0 = \frac{1}{\hat{\pi}}\sum_{i=1}^n y_i - \frac{1}{1-\hat{\pi}}(n - \sum_{i=1}^n y_i)$.
- Solving, we find $\hat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i$, the sample mean.

Note that for finding MLEs, only the part of $l(\theta; y)$ involving θ is relevant. This part is called the **kernel.**

1.3 Likelihood and Maximum-likelihood Estimation

e.g. $Y \sim Bern(\pi)$, we observe $y = (y_1, \ldots, y_n)$. We have shown $l(\pi; y) = \sum_{i=1}^{n} y_i \log(\pi) + (n - \sum_{i=1}^{n} y_i) \log(1 - \pi)$ and $\hat{\pi} = y/n$.

- $\frac{\partial^2 l(\pi;y)}{\partial \pi^2} = -\frac{\sum_{i=1}^{n} y_i}{\pi^2} - \frac{n - \sum_{i=1}^{n} y_i}{(1-\pi)^2}$.
- Observations $y_1, \ldots, y_n$ only appear in the sum $\sum_{i=1}^{n} y_i$. The random version of this sum, $\sum_{i=1}^{n} Y_i$, follows the $B(n, \pi)$ distribution. Hence $E(\sum_{i=1}^{n} Y_i) = n\pi$.
- $-E\left(\frac{\partial^2 l(\pi; Y=(Y_1,\ldots,Y_n))}{\partial \pi^2}\right) = \frac{n\pi}{\pi^2} + \frac{n - n\pi}{(1-\pi)^2}$
- We find $I(\pi) = \frac{n}{\pi(1-\pi)}$ and $\hat{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{n})$, or $\frac{\hat{\pi} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0, 1)$ for n large.

1.4 Large Sample Inference

- The asymptotic properties of maximum likelihood estimators provide ways for us to make large sample inference on the parameters of discrete distributions.
- Next we would consider three significance tests of a null hypothesis,

$$H_0 : \theta = \theta_0 \quad v.s. H_1 : \theta \neq \theta_0.$$

  - ♯ 1.4.1 Wald test and CI
  - ♯ 1.4.2 Score test and CI
  - ♯ 1.4.3 Likelihood Ratio test and CI

1.4 Large Sample Inference

1.4.1 Wald test and CI

- The asymptotic variance of the MLE $\hat{\theta}$ derived from the Fisher Information $I(\theta)$ is a function of θ, the unknown parameter. If we plug in the unrestricted MLE $\hat{\theta}$, we obtain an estimated variance/standard error of $\hat{\theta}$. Let $\iota(\hat{\theta})$ be the Fisher Information evaluated at $\hat{\theta}$. Then the statistic,

$$Z := (\hat{\theta} - \theta_0)/SE, \;\; \text{SE} = 1/\sqrt{\iota(\hat{\theta})},$$

  has an approximate standard normal distribution when $\theta = \theta_0$. Alternatively, the statistic Z^2 has an approximate chi-squared distribution with df $= 1$ (df: degree of freedom), under $\theta = \theta_0$.

- This kind of statistic which uses the non-null estimated standard error, is called a **Wald statistic**.

1.4 Large Sample Inference

e.g. We have a sample of n IID Bernoulli random variables with probability of success π (equivalently, we observe a binomially distributed random variable with parameters n and π).

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

The Wald test statistic $z = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$ can be used to obtain one- or two-sided P-values.

The related $100(1-\alpha)\%$ confidence interval for π is given by $|z| < z_{\alpha/2}$, or

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

1.4 Large Sample Inference

1.4.2 Score test and CI

- The **score function** $u(\theta)$ is the first derivative of the loglikelihood, i.e.,

$$u(\theta) := \frac{\partial l(\theta; y)}{\partial \theta}.$$

- Evaluated at the MLE $\hat{\theta}$, the score function is zero.
- Evaluated at the value of θ_0, the score function tends to be larger in absolute value if $\hat{\theta}$ is far from θ_0.
- Hence, generally speaking, the larger the absolute value of $u(\theta_0)$, the less the data supports the null hypothesis H_0.

1.4 Large Sample Inference

- The test statistic,

$$Z := u(\theta_0)/\sqrt{\iota(\theta_0)},$$

  has an approximate standard normal distribution.
- Alternatively, the statistic Z^2 has an approximate chi-squared distribution with df $= 1$.
- Note: the score statistic Z (or Z^2) uses the null *SE* and does not require the computation of $\hat{\theta}$, the MLE.

1.4 Large Sample Inference

e.g. Again consider a sample of n IID Bernoulli random variables with probability of success π. We wish to test $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$.

The score function is
$u(\pi) := \frac{\partial l(\pi;y)}{\partial \pi} = \frac{1}{\pi}\sum_{i=1}^{n} y_i - \frac{1}{1-\pi}(n - \sum_{i=1}^{n} y_i)$.

Thus the score test statistic is

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

The related $100(1-\alpha)\%$ confidence interval for π is given by all possible values for π_0 for which $|z| < z_{\alpha/2}$.

1.4 Large Sample Inference

1.4.3 Likelihood Ratio test and CI

- The Likelihood Ration (LR) test takes two maximizations of the likelihood function: one maximum over the possible parameter values under the null hypothesis H_0; the other is the maximum over the larger set of possible parameter values under H_0 or H_1, the alternate hypothesis.
- Let l_0 be the maximized likelihood under H_0. Let l_1 be the maximized likelihood under $H_0 \cup H_1$. The ratio

$$\Lambda := \frac{l_0}{l_1} \leq 1.$$

- It is known that the LR test statistic, $-2\log(\Lambda)$, has a chi-squared distribution in the limit as $n \to \infty$. The df is the difference between the dimensions of the parameter spaces under $H_0 \cup H_1$ and H_0.

1.4 Large Sample Inference

e.g. Again consider a sample of n IID Bernoulli random variables with probability of success π. We wish to test $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$.

Recall $L(\pi; y) = \pi^{\sum_{i=1}^n y_i}(1-\pi)^{n-\sum_{i=1}^n y_i}$. Under H_0, π can only take one possible value, π_0. Hence, under H_0, the maximum (and only) value $L(\pi; y)$ can take is $\ell_0 = L(\pi_0; y)$.

Alternatively, under $H_0 \cup H_1$, π can take any possible value in $[0, 1]$. We have already shown that $L(\pi; y)$ is maximized at $\pi = \hat{\pi} = \sum_{i=1}^n y_i/n$. Hence $\ell_1 = L(\hat{\pi}; y)$.

1.4 Large Sample Inference

The LR test statistic is given by

$$
\begin{aligned}
-2\log(\ell_0/\ell_1) &= 2(\log(\ell_1) - \log(\ell_0)) \\
&= 2\log(\hat{\pi})\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log(1-\hat{\pi}) \\
&\quad - 2\log(\pi_0)\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log(1-\pi_0) \\
&= 2\log\left(\frac{\hat{\pi}}{\pi_0}\right)\sum_{i=1}^{n} y_i + 2(n - \sum_{i=1}^{n} y_i)\log\left(\frac{1-\hat{\pi}}{1-\pi_0}\right)
\end{aligned}
$$

No unknown parameters occur under H_0 but one occurs under $H_0 \cup H_1$. Thus the LR test statistic will have the χ^2_1 distribution.

1.4 Large Sample Inference

1.4.4 Comparing the tests

- The three tests are asymptotically equivalent, which means that, in the limit, their (squared for Wald and Score tests) test statistics will follow a chi-squared distribution with the same df, if H_0 is true.

- If H_0 is not true, the test statistics may take very different values. But in such a situation, usually the test statistics will be large and so H_0 will be rejected nevertheless.

1.4 Large Sample Inference

- The Wald test uses $\hat{\theta}$ and the curvature of likelihood at $\hat{\theta}$. The Score test depends on the slope and curvature of likelihood at θ_0. The LR test uses the values of likelihood at $\hat{\theta}$ and θ_0.
- The Wald test is the most commonly used, because it is simplest. However, the other two are increasingly available in software.
- For small to moderate sample sizes, the LR and Score tests are usually more reliable than the Wald test.
- All three tests rely on “large" sample sizes. A rule of thumb for testing binomial parameter π is $n\pi \geq 5$ and $n(1-\pi) \geq 5$.

1.4 Large Sample Inference

1.4.5 Example: Eyesight of students and staff

We randomly selected 100 CUHK Statistics students. 53 of these wear glasses. We wish to test whether the (binomial) proportion of CUHK stats students who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals,

- Wald CI: (0.432, 0.627)
- Score CI: (0.433, 0.625)
- LR CI: (0.432, 0.626)

1.4 Large Sample Inference

1.4.5 Example: Eyesight of students and staff

We randomly selected 10 CUHK Statistics staff. 6 of these wear glasses. We wish to test whether the (binomial) proportion of CUHK stats staff who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals,

- Wald CI: (0.296, 0.904)
- Score CI: (0.313, 0.832)
- LR CI: (0.300, 0.854)

1.4 Large Sample Inference

1.4.5 Example: Eyesight of students and staff

We randomly selected 10 CUHK Fine Arts students. 1 of these wears glasses. We wish to test whether the (binomial) proportion of CUHK Fine Arts students who wear glasses is equal to 0.5 or not. The triumvirate of tests yields the following confidence intervals,

- Wald CI: (0.086, 0.286)
- Score CI: (0.018, 0.404)
- LR CI: (0.006, 0.372)