STA4010 Causal Inference Lecture 1 - 2023-2024 Term 1

Lecture 1: Statistics Review

Lecturer: Yunyi Zhang

Suggested Reading: Agresti and Kateri 2021.

1 Estimator, confidence interval, and hypothesis testing

Suppose we have observations $X_1, \cdots, X_n \in \mathbf{R}^d$, an estimator is defined to be $\widehat{\mu} = h(X_1, \cdots, X_n)$, here $h(\cdot)$ is a known function, and there is no extra unknown parameters. In other words, estimator should be able to derive after we collect the data.

Example 1 Common estimator includes the sample mean and the sample variance

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

An $1-\alpha-$confidence interval for parameter θ is defined to be an interval $[\widehat{a}, \widehat{b}]$ such that $\widehat{a}$ and $\widehat{b}$ are estimators, and we have

$$\lim_{n\to\infty} Prob\left(\theta \in [\widehat{a},\ \widehat{b}]\right) = 1-\alpha,$$

here $1-\alpha$ is called the confidence level or coverage probability. Normally we choose $1-\alpha = 0.95$.

Example 2 Suppose the data $X_1, \cdots, X_n$ satisfy normal distribution $N(\mu, \sigma^2)$, then

$$\boxed{\frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \text{ has } t_{n-1} \text{ distribution.}}$$

Therefore, we choose the corresponding 2.5% and 97.5% quantile $c_{2.5\%}$ and $c_{97.5\%}$, then

$$Prob\left(c_{2.5\%} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \leq c_{97.5\%}\right) = 95\%$$
$$\Rightarrow Prob\left(\overline{X}_n - c_{97.5\%} \times \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X}_n - c_{2.5\%} \times \frac{S}{\sqrt{n}}\right),$$

so the lower bound estimator is $\overline{X}_n - c_{97.5\%} \times \frac{S}{\sqrt{n}}$ while the upper bound estimator is $\overline{X}_n - c_{2.5\%} \times \frac{S}{\sqrt{n}}$.

Example 3 Suppose the data $X_1, \cdots, X_n$ satisfy $Binom(p)$, and we want to estimate p. From central limit theorem and Slutsky's theorem,

$$\boxed{\frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \to_d N(0,1).}$$

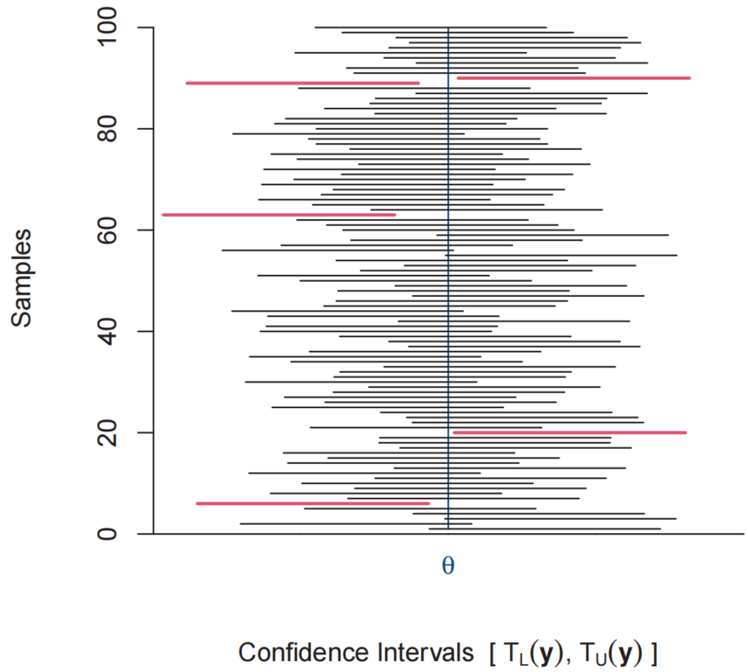

Figure 1: When a data set changes, the endpoints of the confidence interval change, but the underlying parameter θ is fixed.

Therefore, choose standard normal quantile (search the quantile table) $z_{2.5\%}$ and $z_{97.5\%}$, we have

$$\lim_{n\to\infty} Prob\left(z_{2.5\%} \le \frac{\sqrt{n}(\overline{X}_n - p)}{\sqrt{\overline{X}_n(1-\overline{X}_n)}} \le z_{97.5\%}\right) = 95\%,$$

so the 95% confidence interval for p is

$$\left[\overline{X}_n - z_{97.5\%} \times \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}, \overline{X}_n - z_{2.5\%} \times \frac{\sqrt{\overline{X}_n(1-\overline{X}_n)}}{\sqrt{n}}\right]$$

Now suppose we want to test the hypothesis about the parameter θ. Suppose we have two sets H_0 and hypothesis H_1, the test is formulated to be

null hypothesis $\theta \in H_0$ vs the alternative $\theta \in H_1$.

We need to construct a test statistics $\widehat{T}$ for θ. After that, we need to determine a rejection region R_α, which is related to the confidence level $1-\alpha$. We reject H_0(in other words, we believe that $\theta \in H_1$) if $\widehat{T} \in R_\alpha$.

Define the $P-$value to be

$$\widehat{p} = \inf\{\alpha \in [0,1] : T \in R_\alpha\}.$$

In the hypothesis testing literature, we have two types of errors: type-1 error and the type-II error, which are summarized below:

		Null hypothesis is	
		True	False
Decision based on $\widehat{T}$	Fail to reject	Correct	Type-II error
	Reject	Type-I error	Correct

Table 1: Illustration of different types of errors.

In practice, Type-I and Type-II errors always are controversial; i.e., a large type-I error always results in a small Type-II error, and vice versa. What statistician can do is balance both. They assign an α (like 5%) and make sure Type-I error asymptotically equals α. Then, they aim to make Type-II error tend to 0 as $n \to \infty$. **Example 4** Suppose the observed data $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$, and we want to test $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$. To achieve the goal, we use the sample mean $\overline{X}_n$ as the estimator. Notice that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \text{ has } t_{n-1} \text{ distribution.}$$

Therefore, we have

$$Prob\left(c_{\alpha/2} \leq \frac{\sqrt{n}(\overline{X}_n - \mu)}{S} \leq c_{1-\alpha/2}\right) = 1 - \alpha.$$

In other words, we may choose the rejection region

$$R_\alpha = \{X_1, \cdots, X_n : |\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S}| > c_{1-\alpha/2}\},$$

and the corresponding $P-$value is

$$\widehat{p} = 1 - G\left(\frac{\sqrt{n}(\overline{X}_n - \mu)}{S}\right) + G\left(-\frac{\sqrt{n}(\overline{X}_n - \mu)}{S}\right),$$

here G is the cumulative distribution function of t_{n-1} distribution. By definition, the Type-II error of the test is given by

$$Prob\left(|\frac{\sqrt{n}(\overline{X}_n - \mu_0)}{S}| \leq c_{1-\alpha/2}\right)$$

under H_1. Based on this definition, you can calculate the corresponding error.

2 Maximum likelihood estimation

One of the important issues statisticians need to consider is how to establish a "good" estimator for a given parameter θ. This is hard to achieve in general. However, if the observations are (assumed to be) generated from a parametric model, then we have an ad-hoc way to find the good estimator, which is the well-known maximum likelihood estimation.

Suppose the observed data $X_1, \cdots, X_n$ are generated from the joint density/ probability mass function $f(x_1, \cdots, x_n, \theta)$. After observing the data, we can plug-in the data to the joint density and derive the so-called likelihood function

$$l(\theta) = f(X_1, \cdots, X_n, \theta).$$

The MLE $\widehat{\theta}$ is then defined to be the maximizer of $l(\theta)$.

Example 5 Suppose the data $X_1, \cdots, X_n$ are generated from the exponential distribution $f(x, \lambda) = \lambda \exp(-\lambda x)$, then the joint density of data satisfies

$$\log(f(x_1, \cdots, x_n, \lambda)) = n \log(\lambda) - \lambda \sum_{i=1}^{n} x_i,$$

by plugging-in data and taking derivative, one has

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} L(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = 0 \Rightarrow \widehat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i},$$

which is the MLE for the parameter λ.

The MLE has many good properties. Especially, the MLE's asymptotic distribution and its variance are very clear. Define the Fisher Information

$$\boxed{I(\theta) = \mathbf{E}\left(\frac{\partial}{\partial\theta} \log f(X_1, \cdots, X_n, \theta)\right)^2 = -\mathbf{E}\frac{\partial^2}{\partial\theta^2} \log f(X_1, \cdots, X_n, \theta),}$$

then the MLE's asymptotic distribution satisfies

$$\sqrt{I(\theta)}(\widehat{\theta} - \theta) \to_d N(0, 1)$$

Example 6 Further illustration The second-order derivative of the log-likelihood function is

$$\frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} L(\lambda) = -\frac{n}{\lambda^2}.$$

So the MLE's asymptotic distribution satisfies

$$\frac{\sqrt{n}}{\lambda}(\widehat{\lambda} - \lambda) \to_d N(0, 1).$$

Combine with other techniques such as the Slutsky's theorem, we can derive the pivot, which helps establish the CI/perform hypothesis testing.

3 Linear regression

Suppose we have a set of observations $(\mathbf{x}_i, y_i) \in \mathbf{R}^p \times \mathbf{R}$, where $\mathbf{x}_i$ are fixed and

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2).$$

Then $y_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$, and the log-likelihood function becomes

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 - \frac{n}{2} \log(\sigma^2).$$

By calculating gradients and setting them to 0, we have

$$\sum_{i=1}^n \mathbf{x}_i y_i = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}} \text{ and } \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})^2.$$

If we use matrix form, then we derive the frequently used least-square estimator

$$\widehat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top y.$$

Furthermore, we have

$$\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (X^\top X)^{-1}),$$

which supports statistical inference.

4 Logistic regression

The issue of linear regression is that it suggests $y_i \in \mathbf{R}$, which does not suit y with specific structures. Concerning this, we introduce the Logistic regression: suppose $(\mathbf{x}_i, y_i) \in \mathbf{R}^p \times \{0, 1\}$, where

$$\boxed{Prob(y_i = 1) = \frac{1}{1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})} \text{ and } Prob(y_i = 0) = \frac{\exp(-\mathbf{x}_i^\top \boldsymbol{\beta})}{1 + \exp(-\mathbf{x}_i \boldsymbol{\beta})}.}$$

In other words, $Prob(y_i = 1)$ is increasing with respect to the linear combination $\mathbf{x}_i^\top \boldsymbol{\beta}$. Similarly, the log-likelihood function is given by

$$\ell(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \log(1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})) - \sum_{i=1}^{n} (1 - y_i)\mathbf{x}_i^\top \boldsymbol{\beta},$$

and the coefficient is given by

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} \log(1 + \exp(-\mathbf{x}_i^\top \boldsymbol{\beta})) + \sum_{i=1}^{n} (1 - y_i)\mathbf{x}_i^\top \boldsymbol{\beta}.$$

However, in this class, we will not use logistic regression to perform classification. Instead, we are interested in "the propensity score"

$$\widehat{p}(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}^\top \widehat{\boldsymbol{\beta}})}$$

for a feature $\mathbf{x}$.

References

Agresti, Alan and Maria Kateri (2021). *Foundations of Statistics for Data Scientists: With R and Python*. Chapman and Hall/CRC.