The Introduction of Asymptotic Statistics

Jianbin Tan

School of Mathematics
Sun Yat-sen University

Online Seminar

The Introduction of Asymptotic Statistics

1. What does statistics do?
    - Point Estimation
    - Hypothesis Test
    - Interval Estimation
2. Stochastic Convergence
    - Convergence in Probability
    - Convergence in Distribution

We assume that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, which exists to ensure the conceptually feasibility of the law of the observed data:

$$(X_1, ..., X_n) \sim P^n,$$

where $X_i \in \mathcal{X}$.

What can we do if we possess the raw data: $X_1, ..., X_n$?

Let $\theta \in \Theta$ be any "parameter" that you are interesting in:

1. Point estimation: Construct a suitable estimator of θ: $\hat{\theta}$.
2. Hypothesis test: If we wonder "$H_0 : \theta \in \Theta_1$" achieve or not, we construct a powerful test statistics: $\hat{\phi}$ and reject H_0 if $\hat{\phi} \in R_\alpha$, where α is the upper bound of $\mathbb{P}\{\hat{\phi} \in R_\alpha | H_0 = T\}$.
3. Interval estimation: Construct a random interval: $[\widehat{L}_\alpha, \widehat{U}_\alpha]$, which is as "short" as possible to satisfies:

$$\mathbb{P}\{\theta \in [\widehat{L}_\alpha, \widehat{U}_\alpha]\} \geq 1 - \alpha.$$

A fundamental problem of point estimation is to make comparisons among all the point estimators of given parameters.

Example

If $X_1, ..., X_n \overset{iid}{\sim}$ Pois(λ), note that:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

are both unbiased estimators of λ. Which is better?

For the estimators for given parameter, we can define a criterion for the "better" statistics, e.g. we can compare two estimators of θ by:

$$R_{\hat{\theta}}(\theta) := \mathbb{E}(\hat{\theta} - \theta)^2.$$

Rao-Blackwell Theorem indicated that given a sufficient statistics of θ: $\hat{\phi}$, and any other point estimator of θ: $\hat{\theta}$,

$$\hat{\theta}_1 := \mathbb{E}(\hat{\theta}|\hat{\phi})$$

is never worse than $\hat{\theta}$, i.e. $R_{\hat{\theta}_1}(\theta) \leq R_{\hat{\theta}}(\theta)$.

If $\hat{\phi}$ is also the complete statistic of θ and $\hat{\theta}$ is unbiased, then $\hat{\theta}_1$ is the minimum variance unbiased estimator (MVUE) by Lehmann-Scheffé Theorem.

While sufficiency principle is remarkable, the tricky part for this procedure is:

$$\mathbb{E}(\hat{\theta}|\hat{\phi}),$$

which is hard to calculate, or even has no closed form.

To overcome this, a more powerful method: Maximum likeli -hood estimation (MLE), is used in the point estimation procedure for the parameter θ s.t. $\theta \leftrightarrow P^n$:

$$\hat{\theta} = \text{argmax}_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\log p_\theta(X_i)$$

.

Lemma

$\theta_0 = \mathit{argmax}_{\theta\in\Theta}\mathbb{E}\log p_\theta(X)$, *where* $X \sim p_{\theta_0}$.

Proof.

Note that:

$$\mathbb{E}\log p_\theta(X) = -\mathbb{E}\log\frac{p_{\theta_0}(X)}{p_\theta(X)} + C = -\mathrm{KL}(p_{\theta_0}||p_\theta) + C,$$

which maximum is achieved iff $\theta = \theta_0$. □

The logic of MLE is straight-forward, which just replaces the law P for expectation in $\mathbb{E}\log p_\theta(X)$ by the empirical distribution of data.

Apart form this aspect, likelihood principle also relates to the sufficiency principle. If $\hat{\phi}$ is the sufficient statistics of θ, we have:

$$p_\theta(x_{1:n}) = f\,(\theta, \hat{\phi})\; h(x_{1:n})$$

by Fisher-Neyman factorization Theorem, which means that

$$\begin{aligned}\hat{\theta} &= \mathrm{argmax}_{\theta\in\Theta} \log f\,(\theta, \hat{\phi}) \\ &:= \mathcal{H}(\hat{\phi}).\end{aligned}$$

While $\mathcal{H}$ may not have closed form, it can get effective approxi -mation by many optimization methods.

However, there are not always exact optimal theory for procedure like MLE, then asymptotic optimality theory may help.

For example, Cramér-Rao bound is stated that for any unbias -ed estimator $\hat{\theta}$:

$$\mathsf{Var}(\hat{\theta}) \geq I_n(\theta)^{-1},$$

where Fisher Information $I_n(\theta) := \mathsf{Var}(\frac{\mathsf{d}\log p_\theta(X_{1:n})}{\mathsf{d}\theta})$ for broad class of density functions. CR bound implies that a point estimator $\hat{\theta}_n$ may be feasible if

$$I_n(\theta)^{-1} \asymp \mathsf{Var}(\hat{\theta}_n).$$

Somehow, there exists several problems in MLE:

1. We need to specify a parametric law of data, which is unrealistic for some statistical problems.
2. The optimization procedure is often computationally intrac -table.

Therefore, some more complicate methods, e.g. Machine Learning, Deep Learning, are proposed to adapt the new situation, which optimality is evaluated asymptotically or even non-asympto -tically (like MVUE) in the framework of learning theory. The technique of asymptotic statistics plays an important role among them.

Problems of hypothesis test seems to be similar to the point estimation for classification, which main point are both to detach which class within the given choices that the real situation belongs to.

However, the soul of hypothesis test is to reject the rejection, not to accept. Unlike mathematics, if we want to prove something by data, we should try our best to deny it, and "accept" it if you can't.

The procedure of "negation of negation" gives a more confident "positive" answer for null assumption H_0, while the decision of rejection is a "pity".

Therefore, for a given testing statistics $\hat{\phi}$, the key part is to choose a reject domain R_α s.t. $\mathbb{P}\{\hat{\phi} \in R_\alpha | H_0 = T\} \leq \alpha$, i.e. we want to control the probability of rejecting the H_0.

Example

$X_1, ..., X_n \overset{iid}{\sim}$ N(μ, σ^2), we want to test $H_0 : \mu = \mu_1$. Note that:

$$\mathbb{P}\left(\frac{|\hat{\mu}_n - \mu_1|}{\sigma/\sqrt{n}} > t | H_0 = T\right) \leq 2\,\text{exp}(-\frac{t^2}{2}),$$

for given α and σ^2, we can calculate a suitable t for construction of rejection domain.

What if the distribution of X is unknown but the second moment exists. By Central Limit Theorem, we have:

$$\frac{\hat{\mu}_n - \mu_1}{\sigma/\sqrt{n}} \to_d \mathsf{N}(0,1),$$

then $\mathbb{P}\left(\frac{|\hat{\mu}_n - \mu_1|}{\sigma/\sqrt{n}} > t | H_0 = T\right) \to \int_{|x|>t} \mathsf{N}(x;0,1)\mathrm{d}x$, which implies that if n is large, then we can use this approximation to propose a test.

Similarly, for a given testing problem, we can construct many testing procedures, which emphasizes the necessity for evaluation of the optimal test for a given problem.

The concept: misclassification rate, could be used:

$$\sum_{i=1}^{M} \mathbb{P}\left(\hat{\phi} \text{ rejects } H_i | H_i = T\right).$$

If $M = 2$, since we control the error $\mathbb{P}\left(\hat{\phi} \text{ rejects } H_0 | H_0 = T\right)$, we can use the power of test:

$$\mathrm{W}(\hat{\phi}) := 1 - \mathbb{P}\left(\hat{\phi} \text{ rejects } H_1 | H_1 = T\right)$$

for evaluating different tests.

Example

$H_0 : \theta = \theta_1$ and $H_1 : \theta = \theta_2$ ($\theta \leftrightarrow P^n$), *Neyman-Pearson Lemma shows that: If any $\hat{\phi}$ s.t.*

$$\mathbb{P}\left(\hat{\phi}_0 \textit{ rejects } H_0 | H_0 = T\right) \geq \mathbb{P}\left(\hat{\phi} \textit{ rejects } H_0 | H_0 = T\right),$$

then

$$\mathbb{P}\left(\hat{\phi} \textit{ rejects } H_1 | H_1 = T\right) \geq \mathbb{P}\left(\hat{\phi}_0 \textit{ rejects } H_1 | H_1 = T\right),$$

where $\hat{\phi}_0 = \frac{p_{\theta_1}(X_{1:n})}{p_{\theta_2}(X_{1:n})}$ and it rejects H_0 if $\hat{\phi} < K$.

For many other problems, it may be hard to compare them since the power function may not be easy to compute, we might need to compare approximations to their power functions.

Furthermore, we often need to construct a testing statistics with well defined asymptotic distribution, while the asymptotic distribution could be computed by some permutation procedures.

These non-parametric method indicates that the asymptotic distribution is not necessary to be known accurately, but the existence of it should be guaranteed, which is the main job for asymptotic statistics.

Point estimator is not always the unique thing we care about for the true parameters, the boundary is also of interest to us.

Given a confident level α, we need construct a random cover -ing of true parameter s.t. the probability of covering the "oracle" is larger than $1 - \alpha$.

Example

$X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. *Note that:*

$$\mathbb{P}\left(\frac{|\hat{\mu}_n - \mu|}{\sigma/\sqrt{n}} > t\right) = \int_{|x|>t} N(x; 0, 1)\, dx,$$

for given α *and* σ^2, *we claim that*

$$\mathbb{P}\left(\mu \in \left[\hat{\mu}_n - \frac{t_\alpha \sigma}{\sqrt{n}}, \hat{\mu}_n + \frac{t_\alpha \sigma}{\sqrt{n}}\right]\right) = 1 - \alpha.$$

The length of random confident interval (CI) should be as short as possible. (If the distribution of X is unknown, this CI is valid if n is large enough)

The construction of CI are mostly based on the sample distribution of statistics. If we don't know the specific form of data, the techniques of asymptotic statistics are used to give an approximation of sample distribution.

The development of Monte Carlo is often employed to the situation that the asymptotic distribution only exists but has no closed form, e.g. Bootstrap.

Also, in Bayesian framework, a covering of the posterior distribution of the given parameters is another way to construct the random interval (Credible interval).

Definition

$X_n \to_p X$: *For any* $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ *as* $n \to \infty$.

Indeed, this kind of convergence could be metrization. Define

$$d(X, Y) := \mathbb{E}\frac{|X - Y|}{|X - Y| + 1},$$

then $d(X_n, X) \to 0 \Leftrightarrow X_n \to_p X$, which implies that the conver -gence in probability is similar to other convergences in metric space.

We mark that $X_n \to_p 0$: $X_n = o_p(1)$ and $X_n = o_p(Y_n)$ iff $\frac{X_n}{Y_n} = o_p(1)$.

Definition

$X_n = O_p(1)$, *i.e. is said to be* ***bounded in probability*** *if:* $\forall \delta > 0$, $\exists$ $M > 0$ *s.t.* $\mathbb{P}(|X_n| > M) \leq \delta$ *for all* n.

Example

If $X_1, ..., X_n \sim N(\mu, \sigma^2)$, *then*

$$\mathbb{P}\left(|\hat{\mu}_n - \mu| > \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \Rightarrow \hat{\mu}_n - \mu = o_p(1),$$

$$\mathbb{P}\left(\frac{|\hat{\mu}_n - \mu|}{\sigma/\sqrt{n}} > t\right) = \int_{|x|>t} N(x; 0, 1)\, dx$$

$$\Rightarrow \hat{\mu}_n - \mu = O_p(\frac{1}{\sqrt{n}}).$$

Theorem

Continuous Mapping Theorem: If g is continuous, and $X_n \rightarrow_p X$, then

$$g\left(X_n\right) \rightarrow_p g\left(X\right).$$

Property

If $X_n \rightarrow_p X$, then $X_n = O_p(1)$;

$o_p(1) + o_p(1) = o_p(1)$, $o_p(1) + O_p(1) = O_p(1)$;

If $X_n = o_p(1)$, $R(h) = o(|h|^q)$, then $R(X_n) = o_p(|X_n|^q)$;

If $X_n = O_p(1)$, $R(h) = O(|h|^q)$, then $R(X_n) = O_p(|X_n|^q)$.

Definition

Let $F_X(x) := \mathbb{P}(X \leq x)$. $X_n \to_d X$ *iff*

$$F_{X_n}(x) \to F_X(x),$$

for all continuous point x *of* F_X*.*

This kind of convergence is defined in the distributions of random variables, but not random variables themselves, which means that

$$X_n + Y_n \to_d X + Y$$

may not hold if $X_n \to_d X$ and $Y_n \to_d Y$.

We also mark $X_n \to_d X$: $F_{X_n} \to_w F_X$, which is called the weak convergence of the distributions (compared with other kind of convergence, total variation).

Property

If $X_n \to_p X$, then $X_n \to_d X$. If $X = c$ a.s., $X_n \to_p X$ iff $X_n \to_d X$.

There are essential difference between convergence in law and convergence in probability. While the convergence in probability is defined on the sequence of random elements from the same latent probability space $(\Omega, \Sigma, \mathbb{P})$, which means that the dependence of random elements could effect this kind of convergence.

Theorem

Skorokhod's Representation Theorem: If $X_n \to_d X$, then exists Y_n, Y defined from the same probability space s.t. $F_{X_n} =_d F_{Y_n}$, $F_X =_d F_Y$ and $Y_n \to Y$ a.s..

The trick is frequently used if the quantities only depends on the distributions of random variables.

Property

If $X_n \to_d X$, g is continuous, then $g(X_n) \to_d g(X)$.

If $X_n \to_d X$ and $|X_n| < Y$, $\mathbb{E}|Y| < \infty$, then $\mathbb{E}X_n \to \mathbb{E}X$.

Theorem

Portmanteau: $X_n \to_d X$ *iff one of the following achieves:*

(1) $\forall$ *bound Lipschitz function* f, $\mathbb{E}f(X_n) \to \mathbb{E}f(X)$.

(2) liminf$_n \mathbb{P}(X_n \in G) \geq \mathbb{P}(X \in G)$, *for* $\forall$ *open* G.

(3) limsup$_n \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$, *for* $\forall$ *closed* F.

(4) lim$_n \mathbb{P}(X_n \in B) = \mathbb{P}(X \in B)$, $\forall$ B *s.t.* $\mathbb{P}(X \in \partial B) = 0$.

Indeed, this theorem quantified the meaning of “weakness” (e.g. the testing class of functions, the testing collection of subsets).

Theorem

Prohorov's Theorem: If $X_n \to_d X$, then $X_n = O_p(1)$. Conversely, if $X_n = O_p(1)$, then $\exists$ X_{n_k} and X s.t.

$$X_{n_k} \to_d X.$$

We also said that $X_n = O_p(1)$ is tight. Tightness of a given set of random elements $\mathcal{A}$ related to the topological property: relative compactness, i.e. every subsequence of $\mathcal{A}$ exists a convergent subsubsequence.

The evaluation of relative compactness of a sequence is the prerequisite of the convergence of it, e.g. $o_p(O_p(1)) = o_p(1)$.

Theorem

Slutsky Theorem: *If* $X_n \to_d X$, $Y_n \to_d c$, *then*

(1) $X_n + Y_n \to_d X + c$;

(2) $X_n Y_n \to_d cX$;

(2) $X_n / Y_n \to_d X/c$, *if* $c \neq 0$.

The proofs of them depends on this property:

$$(X_n, Y_n) \to_d (X, c),$$

but $(X_n, Y_n) \to_d (X, Y)$ may not hold in general if $Y_n \to_d Y$.

Theorem

Polya's Theorem: If $X_n \to_d X$ and F_X is continuous, then

$$||F_{X_n} - F_X||_\infty := \sup_t |F_{X_n}(t) - F_X(t)| \to 0.$$

If $X \sim \mathrm{N}(0,1)$, we will revisit this property in Berry-Esseen Bound, which claim that:

$$\sup_t |\mathbb{P}(\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sigma} \leq t) - \mathbb{P}(X \leq t)| \precsim \frac{1}{\sqrt{n}}.$$