36-705 11/05/2020

YouTube 19/09/2016

Convergence Theory

- Previous review:

$X_1, \dots, X_n \sim P$

CLT:- $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \xrightarrow{d} N(0,1)$

$\sqrt{n}(\bar{X}-\mu) \xrightarrow{d} N(0,\sigma^2)$

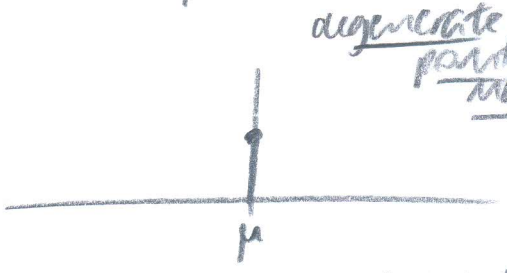
Informally $\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$ (not formal but prac.)

(*) original distn arbitrary, but limiting distn is approx normal

(*) Remember you want lim distn of rescaled

LLN (via convergence taxonomy)

$\bar{X} \xrightarrow{p} \mu \Rightarrow \bar{X} \xrightarrow{d} \delta_\mu$



- limiting distn (but not interesting)

Theorem 17 (Multivariate CLT)

- $\underline{X}_1, \dots, \underline{X}_n \sim P$ IID r. vectors

$\underline{X}_i \in \mathbb{R}^k$, $\underline{\mu} \in \mathbb{R}^k$, $\underline{\Sigma} \in \mathbb{R}^{k \times k}$

$\underline{X}_i = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{pmatrix}$ $\underline{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$ $\underline{\Sigma}$

(*) Intuit: Take measure multiple characteristics of a person (e.g. height, weight, ...)

(*) Elements of vector are not independent; independence is across vectors (which are observations) (people in this example)

(i.e. exact.)

$\underline{\Sigma} = \begin{pmatrix} \mathrm{Var}(X_1) & \dots & \mathrm{Cov}(X_1, X_k) \\ \mathrm{Cov}(X_2, X_1) & & \vdots \\ \vdots & & \\ \mathrm{Cov}(X_k, X_1) & & \mathrm{Var}(X_k) \end{pmatrix}$

- $\underline{\bar{X}} = \begin{pmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_k \end{pmatrix}$ $\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ji}$

Then $\sqrt{n}(\underline{\bar{X}} - \underline{\mu}) \xrightarrow{d} N(\underline{0}, \underline{\Sigma})$

5. Delta method (Propagation of errors)

Q: What if I want distri of $g(\bar{X})$ for some smooth function $g(\cdot)$?

- Maybe → continuous mapping theorem?
- will not help
- We want a rescaled version of continuous mapping

(Scalars)

- $g(\bar{X})$
- Consider:- $$\frac{\sqrt{n}(g(\bar{X}) - g(\mu))}{\sigma |g'(\mu)|} \xrightarrow{d} N(0,1)$$
- Assume $g(\cdot)$ is differentiable; $g'(\mu) \neq 0$
- Informally: $$g(\bar{X}) \approx N\left(g(\mu), (g'(\mu))^2 \frac{\sigma^2}{n}\right) \qquad \bar{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Allows any limit distri of any smooth function of av.

LW: When you take a function of $\bar{X}$, we have

i) mean becomes $g(\mu)$

ii) variance multiplies by derivative

- Example 19
- $X_1, \ldots, X_n$ IID with finite mean μ, finite variance σ^2
- We know $\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} N(0,1)$

(WAI): Check you can square the rot. intuitively

- Consider $W_n = e^{\bar{X}_n}$ (non linear fn)

$g(s) = e^s \quad g'(s) = e^s$

- LW: immediately know limiting distri:-

$$\sqrt{n}(W_n - e^{\mu}) \xrightarrow{d} N(0, \sigma^2 (e^{\mu})^2) \text{ or}$$

Heuristic derivation

- consider $g(\bar{X})$
- use Taylor series approx, expand $g(\bar{x})$ about μ (not including all terms → heuristic)

$$g(\bar{X}) \approx g(\mu) + (\bar{x}-\mu) g'(\mu)$$

$$\sqrt{n}(g(\bar{X}) - g(\mu)) \approx \sqrt{n}(\bar{x}-\mu) g'(\mu)$$

$\sqrt{n}(\bar{x}-\mu) \xrightarrow{d} N(0, \sigma^2)$ via CLT

- $g'(\mu)$ is constant, so multiply by constant $(g'(\mu))^2$ in variance
- $\text{var}(aX) = a^2 \text{var}(X)$

Theorem 20 (multivariate delta method)

- $\sqrt{n}(\underline{\bar{X}} - \underline{\mu}) \xrightarrow{d} N(\underline{0}, \underline{\Sigma})$ (CLT) $\quad \underline{\bar{X}} \in \mathbb{R}^k$
- suppose $g : \mathbb{R}^k \to \mathbb{R}$ (scalar fn) (also one for vector fn)
- take gradient of $g(y)$ wrt y:-

$$\nabla_y g(y) = \begin{pmatrix} \frac{\partial g}{\partial y_1} \\ \vdots \\ \frac{\partial g}{\partial y_k} \end{pmatrix}$$

- let $\underline{\nabla}_\mu$ denote $\nabla_y g(y)\big|_{y=\mu}$ (i.e. gradient of g wrt y evaluated at μ)
- Then:

$$\sqrt{n}(g(\underline{\bar{X}}) - g(\underline{\mu})) \xrightarrow{d} N(0, \underline{\nabla}_\mu^T \underline{\Sigma} \underline{\nabla}_\mu)$$

- **example 21** $\quad \underline{X}_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix}$, have $\underline{X}_1, \underline{X}_2, \underline{X}_3, \ldots \underline{X}_n$
- i.e. n IID random vectors with mean (μ_1, μ_2) and variance $\underline{\Sigma}$

- Denote:

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i} \quad \bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$$

- Then $\sqrt{n}\begin{pmatrix}\bar{X}_1 - \mu_1 \\ \bar{X}_2 - \mu_2\end{pmatrix} \xrightarrow{d} N(\underline{0}, \underline{\Sigma})$ $\quad \sqrt{n}(\bar{\underline{X}} - \underline{\mu}) \xrightarrow{d} N(0, \underline{\Sigma})$

- define a function (product) $g(s_1, s_2) = s_1 s_2$

(Q) How do I get the limiting distri of $\bar{X}_1\bar{X}_2$? (complicated, no longer sum)

$$\nabla_{\underline{s}} g(\underline{s}) = \begin{pmatrix} s_2 \\ s_1 \end{pmatrix}$$

$$\nabla_{\underline{\mu}}^T \underline{\Sigma} \nabla_{\underline{\mu}} = (\mu_2 \ \mu_1)\begin{pmatrix}\sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22}\end{pmatrix}\begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} = \mu_2^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22}$$

$$\therefore \sqrt{n}(\bar{X}_1\bar{X}_2 - \mu_1\mu_2) \xrightarrow{d} N(0, \mu^2\sigma_{11} + 2\mu_1\mu_2\sigma_{12} + \mu_1^2\sigma_{22})$$

---

informally $g(\bar{X}) \overset{\cdot}{\sim} N\left(g(\mu), \frac{1}{n}\nabla_{\underline{\mu}}^T \underline{\Sigma} \nabla_{\underline{\mu}}\right)$

LW: Have probabilistic tools to do statistics/ML

- use tools to statistical inference (part II) - extracting insight from data

LW: In real-world, stats is opposite to prob, in sense that:-

Probability: we have a distri

Stats: - given data, figure out what generated data

36705 11/05/2020

YouTube 19/09/16

- Statistical inference (Lecture Notes 5)

- LW: I give you data, you have to do things e.g. tell me about DGP.

(*) Difference between statistics and probability.

- Probability → start with a distn P
- Statistics → start with a model (statistical)

(*) - i.e. what are the possible distributions (aka distn family) that could have generated the data

---

1. Statistical models

- A statistical model $\mathcal{P}$ is a set of probability distributions (or densities)
- LW: All we know is that the 'true' distribution is within that set of distributions
- LW: Introduce artificial distinction between nonparametric and parametric models

- Parametric model

LW: Set of distn that can be indexed by a finite no. params

$$\mathcal{P} = \left\{ p(x;\theta) : \theta \in \Theta \right\} \qquad \Theta \subset \mathbb{R}^d \qquad \Theta\text{-param space}$$

- set of Normal densities :- $\left\{ p(x;\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-(x-\mu)^2/(2\sigma^2)\right) \right\}$

$$\underline{\theta} : (\mu, \sigma^2) \qquad \Theta = \mathbb{R} \times \mathbb{R}^+$$

(param space)

LW: Highly restrictive, in general shouldn't use

LW: wean off parametric models late (start only starting point)

Nonparametric models

- A model (collection of prob distri) that cannot be specified by a finite dimensional param.

LW: should be infinitely-parametric 2 senses of non-parametric

example $\mathcal{P} = \{\text{all distri on } \mathbb{R}^d\}$ → non-param

example 2 $\mathcal{P} = \{p : \int (p''(x))^2 dx < \infty\}$ "set of all distri with densities (which have 2nd derivative and are square integrable (?)
- smoothness constraint on $\mathcal{P}$

(*) start with parametric models

LW: Given data from some distri ; and that (assumed) distri comes from some distri within parametric model
- Have to find parametes → can we estimate them from a finite sample?

LW: data reduction, sufficiency
- Ronald Fisher
- NO computes then
- Q: can I reduce data to a few nos → how much info loss
  (lots of emphasis on sufficiency in pre-compute age)
- sufficiency become less important with advent of computes
- But BIG DATA → sufficiency

Q: what is sufficiency (in context of a statistic)
- LW: A statistic is a function of the data (itself a random variable as a function of random variables)

(*) $\bar{X} - \mu$ is not a statistic, not a function of the data, but param μ

(*) key distinction between r.v.; parameters → these index the distri

(*) Assume parameters unknown.

Q: which statistics capture all info in the data → sufficient statistic

---

(*) Formal definition; intuition

- sufficient statistics

- "parametric model" ; θ - means family of densities indexed by the parameter (fixed) (freq.)

- Suppose $X_1, \ldots, X_n \sim p(x;\theta) = p_\theta(x)$

- $T(X_1, \ldots, X_n)$ is sufficient for θ if the condit. distri :- (*) This realisation of t (*) condition on $T=t$

$$p(x_1, \ldots, x_n | T=t, \theta) = p(x_1, \ldots, x_n | T=t)$$

(*) Intuition: - distri depends on θ

E.g. Normal:- $x \sim N(0,1)$ $\quad p(x) = \frac{1}{\sqrt{2\pi}} - \frac{(x-\mu)^2}{2}$

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2}}$$

- suppose I give you a statistic; and get you to estimate μ

- μs left

(*) amounts to condit. distri; conditioning on statistic $T=t$; then the cond. distri no longer dependent on parameters

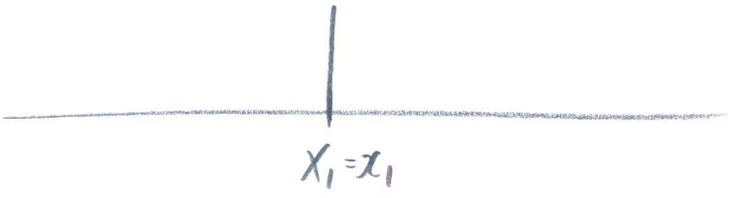
(*) Density no longer contains parameters conditioning on $T=t$

(*) Possible to take statistic as the data:-

$$T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$$

- what is distri of single r.v.; given I tell you what it is? (one obs.)

- Point mass at X (no free param.)



(*) The statistic equal to whole dataset is always sufficient
(data are sufficient)

(*) tautological; but tells us that there are good and bad statistics

(*) We want to build to minimal sufficient statistics(*)

- $\bar{X}_n$ better at data reduction than $(X_1, \dots, X_n)$

(*) Not enough to be sufficient

(*) sufficiency is only really useful for parametric models

(*) Next lecture