Asymptotic Statistics-III

Theorem

(Multivariate CLT for iid case) *Let* $\mathbf{X}_i$ *be iid random p-vectors with mean* $\boldsymbol{\mu}$ *and and covariance matrix* $\boldsymbol{\Sigma}$. *Then*

$$\sqrt{n}\,(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

- By the Cramer-Wold device, this can be proved by finding the limit distribution of the sequences of real variables

$$\mathbf{c}^T \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{X}_i - \boldsymbol{\mu}) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\mathbf{c}^T \mathbf{X}_i - \mathbf{c}^T \boldsymbol{\mu}).$$

The multivariate central limit theorem

- Because the random variables $\mathbf{c}^T\mathbf{X}_i - \mathbf{c}^T\boldsymbol{\mu}$ are iid with zero mean and variance $\mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c}$, this sequence is $AN(0, \mathbf{c}^T\boldsymbol{\Sigma}\mathbf{c})$ by Theorem **??**.
- This is exactly the distribution of $\mathbf{c}^T\mathbf{X}$ if $\mathbf{X}$ possesses an $N_p(\mathbf{0}, \boldsymbol{\Sigma})$.

The multivariate central limit theorem

Example

Suppose that $X_1, \ldots, X_n$ is a random sample from the Poisson distribution with mean θ. Let Z_n be the proportions of zero observed, i.e., $Z_n = 1/n \sum_{i=1}^n I_{\{X_j=0\}}$. Let us find the joint asymptotic distribution of $(\bar{X}_n, Z_n)$

- Note that $E(X_1) = \theta$, $EI_{\{X_1=0\}} = e^{-\theta}$, $\text{var}(X_1) = \theta$, $\text{var}(I_{\{X_1=0\}}) = e^{-\theta}(1 - e^{-\theta})$, and $EX_1 I_{\{X_1=0\}} = 0$.
- So, $\text{cov}(X_1, I_{\{X_1=0\}}) = -\theta e^{-\theta}$.
- $\sqrt{n}\left((\bar{X}_n, Z_n) - (\theta, e^{-\theta})\right) \xrightarrow{d} N_2(\mathbf{0}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} \theta & -\theta e^{-\theta} \\ -\theta e^{-\theta} & e^{-\theta}(1 - e^{-\theta}) \end{pmatrix}.$$

Exercises

- Consider two sequences of random variables X_n and Y_n. If $(X_n - EX_n)/\sqrt{\text{var}X_n} \xrightarrow{d} X$ and $\text{corr}(X_n, Y_n) \to 1$, then $(Y_n - EY_n)/\sqrt{\text{var}Y_n} \xrightarrow{d} X$.
- Let $X_1, X_2, \ldots$ be iid double exponential (Laplace) random variables with density, $f(x) = (2\tau)^{-1}\exp\{-|x|/\tau\}$, where τ is a positive parameter that represents the mean deviation, i.e., $\tau = E|X|$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y}_n = n^{-1}\sum_{i=1}^n |X_i|$.
  - (a) Find the joint asymptotic distribution of $\bar{X}_n$ and $\bar{Y}_n$.
  - (b) Find the asymptotic distribution of $(\bar{Y}_n - \tau)/\bar{X}_n$.

- (a) Let $Y_i = |X_i|$. Then (X_i, Y_i) are iid with $E(X_i, Y_i) = (0, \tau)$. Since $EX^2 = 2\tau^2$, we have $\text{var}X_i = 2\tau^2$ and $\text{var}Y_i = \tau^2$. We also have $\text{cov}(X_i, Y_i) = 0$. Therefore, from the multivariate CLT,

$$\sqrt{n}(\bar{X}_n, (\bar{Y}_n - \tau)) \xrightarrow{d} N_2\left(\mathbf{0}, \begin{pmatrix} 2\tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix}\right).$$

  (b) From CMT with $g(x, y) = y/x$, continuous except on the line $x = 0$, we have
  $(Y_n - \tau)/X_n = \sqrt{n}(Y_n - \tau)/(\sqrt{n}X_n) \xrightarrow{d} V/U$, where U and V are independent normal random variables with zero means and $2\tau^2$ and τ^2 respectively. This has a Cauchy distribution with median zero and scale parameter $1/\sqrt{2}$, independent of τ. Of course, $(Y_n - \tau)/(X_n/\sqrt{2})$ has a standard Cauchy distribution.

CLT: existence of a variance is not necessary

Definition

A function $g : \mathbb{R} \to \mathbb{R}$ is called slowly varying at ∞ if, for every $t > 0$, $\lim_{x\to\infty} g(tx)/g(x) = 1$.

Examples: $\log x$, $x/(1+x)$, and indeed any function with a finite limit as $x \to \infty$; x or e^{-x} are not slowly varying.

Theorem

Let $X_1, X_2, \dots$ be iid from a CDF F on $\mathbb{R}$. Let $v(x) = \int_{-x}^{x} y^2 dF(y)$. Then, there exist constants $\{a_n\}, \{b_n\}$ such that

$$\frac{\sum_{i=1}^n X_i - a_n}{b_n} \xrightarrow{d} N(0,1),$$

if and only if $v(x)$ is slowly varying at ∞.

If F has a finite second moment, $v(x)$ is slowly varying at ∞.

CLT: existence of a variance is not necessary

Example

- Suppose $X_1, X_2, \ldots$ are iid from a t-distribution with 2 degrees of freedom ($t(2)$) that has a finite mean but not a finite variance.
- The density is given by $f(y) = c/(2+y^2)^{\frac{3}{2}}$ for some positive c.
- by a direct integration, for some other constant k,

$$v(x) = k\sqrt{\frac{1}{2+x^2}}\left[x - \sqrt{2+x^2}\,\mathrm{arcsinh}(x/\sqrt{2})\right].$$

- on using the fact that $\mathrm{arcsinh}(x) = \log(2x) + O(x^{-2})$ as $x \to \infty$, we get, for any $t > 0$, $\frac{v(tx)}{v(x)} \to 1$.
- the partial sums $\sum_{i=1}^n X_i$ converge to a normal distribution
- The centering can be taken to be zero for the centered t-distribution; it can be shown that the normalizing required is $b_n = \sqrt{n \log n}$

CLT for the independent not necessarily iid case

Theorem

(Lindeberg-Feller) *Suppose X_n is a sequence of independent variables with means μ_n and variances $\sigma_n^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for any $\epsilon > 0$*

$$\frac{1}{s_n^2} \sum_{j=1}^{n} \int_{|x-\mu_j|>\epsilon s_n} (x-\mu_j)^2 dF_j(x) \to 0, \tag{1}$$

where F_i is the CDF of X_i, then

$$\frac{\sum_{i=1}^{n} (X_i - \mu_i)}{s_n} \xrightarrow{d} N(0,1).$$

The condition (1) is called Lindeberg-Feller condition.

Example

Let $X_1, X_2 \ldots,$ be independent variables such that X_j has the uniform distribution on $[-j, j], j = 1, 2, \ldots$. Let us verify the conditions of the theorem are satisfied.

- Note that $EX_j = 0$ and $\sigma_j^2 = \frac{1}{2j}\int_{-j}^{j} x^2 dx = j^2/3$ for all j.
- $$s_n^2 = \sum_{j=1}^{n} \sigma_j^2 = \frac{1}{3}\sum_{j=1}^{n} j^2 = \frac{n(n+1)(2n+1)}{18}.$$
- For any $\epsilon > 0$, $n < \epsilon s_n$ for sufficiently large n, since $\lim_n n/s_n = 0$.
- Because $|X_j| \leq j \leq n$, when n is sufficiently large,
$$E(X_j^2 I_{\{|X_j| > \epsilon s_n\}}) = 0.$$
- Consequently, $\lim_{n\to\infty} \sum_{j=1}^{n} E(X_j^2 I_{\{|X_j| > \epsilon s_n\}}) < \infty$. Considering $s_n \to \infty$, Lindeberg's condition holds.

CLT for the independent not necessarily iid case

It is hard to verify the Lindeberg-Feller condition.

A simpler theorem

Theorem

(Liapounov) *Suppose X_n is a sequence of independent variables with means μ_n and variances $\sigma_n^2 < \infty$. Let $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for some $\delta > 0$*

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^{n} E|X_j - \mu_j|^{2+\delta} \to 0 \tag{2}$$

as $n \to \infty$, then

$$\frac{\sum_{i=1}^{n} (X_i - \mu_i)}{s_n} \xrightarrow{d} N(0, 1).$$

Liapounov CLT

- $s_n \to \infty$, $\sup_{j \geq 1} E|X_j - \mu_j|^{2+\delta} < \infty$ and $n^{-1}s_n$ is bounded
- In practice, work with $\delta = 1$ or 2.
- If X_i is uniformly bounded and $s_n \to \infty$, the condition is immediately satisfied with $\delta = 1$.

Liapounov CLT

Example

Let $X_1, X_2, \ldots$ be independent random variables. Suppose that X_i has the binomial distribution $\mathrm{BIN}(p_i, 1), i = 1, 2, \ldots$.

- For each i, $EX_i = p_i$ and $E|X_i - EX_i|^3 = (1-p_i)^3 p_i + p_i^3(1-p_i) \leq 2p_i(1-p_i)$.
- $\sum_{i=1}^n E|X_i - EX_i|^3 \leq 2s_n^2 = 2\sum_{i=1}^n E|X_i - EX_i|^2 = 2\sum_{i=1}^n p_i(1-p_i)$.
- Liapounov's condition (2) holds with $\delta = 1$ if $s_n \to \infty$.
- For example, if $p_i = 1/i$ or $M_1 \leq p_i \leq M_2$ with two constants belong to $(0, 1)$, $s_n \to \infty$ holds.
- Accordingly, by Liapounov's theorem, $\frac{\sum_{i=1}^n (X_i - p_i)}{s_n} \xrightarrow{d} N(0, 1)$.

CLT for double array and triangular array

Double array:

$$X_{11} \quad \text{with distribution } F_1$$
$$X_{21}, X_{22} \quad \text{independent, each with distribution } F_2$$
$$\dots$$
$$X_{n1}, \dots X_{nn} \quad \text{independent, each with distribution } F_n$$

Triangular array:

$$X_{11} \quad \text{with distribution } F_1$$
$$X_{21}, X_{22} \quad \text{independent, with distribution } F_{21}, F_{22}$$
$$\dots$$
$$X_{n1}, \dots X_{nn} \quad \text{independent, with distributions } F_{n1}, \dots, F_{nn}.$$

CLT for double array

Theorem

Let the X_{ij} be distributed as a double array. Then

$$P\left(\frac{\sqrt{n}(\bar{X}_n - \mu_n)}{\sigma_n} \leq x\right) \to \Phi(x)$$

as $n \to \infty$ for any sequence F_n with mean μ_n and variance σ_n^2 for which

$$E_n|X_{n1} - \mu_n|^3/\sigma_n^3 = o(\sqrt{n}).$$

Here E_n denotes the expectation under F_n.

Eg, $\mathrm{Bin}(p_n, n)$, where the success probability depends on n.

Theorem

Let the X_{ij} be distributed as a triangular array and let $E(X_{ij}) = \mu_{ij}$, $var(X_{ij}) = \sigma_{ij}^2 < \infty$, and $s_n^2 = \sum_{j=1}^n \sigma_{nj}^2$. Then,

$$\frac{\sum_{j=1}^n (X_{nj} - \mu_{nj})}{s_n} \xrightarrow{d} N(0,1),$$

provided that

$$\frac{1}{s_n^{2+\delta}} \sum_{j=1}^n E|X_{nj} - \mu_{nj}|^{2+\delta} \to 0$$

Hajek-Sidak CLT

Theorem

(Hajek-Sidak) *Suppose* $X_1, X_2, \ldots$ *are iid random variables with mean* μ *and variance* $\sigma^2 < \infty$. *Let* $c_n = (c_{n1}, c_{n2}, \ldots, c_{nn})$ *be a vector of constants such that*

$$\max_{1\leq i\leq n} \frac{c_{ni}^2}{\sum_{j=1}^n c_{nj}^2} \to 0 \tag{3}$$

as $n \to \infty$. *Then*

$$\frac{\sum_{i=1}^n c_{ni}(X_i - \mu)}{\sigma\sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{d} N(0,1).$$

Hajek-Sidak CLT

- The condition (3) is to ensure that no coefficient dominates the vector c_n, and is referred as Hajek-Sidak condition.
- For example, if $c_n = (1, 0, \ldots, 0)$, then the condition would fail and so would the theorem.

Hajek-Sidak CLT

Example

(**Simplest linear regression**) Consider the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i's are iid with mean 0 and variance σ^2 but are not necessarily normally distributed. The least squares estimate of β_1 based on n observations is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}.$$

Hajek-Sidak CLT

- $\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^n \varepsilon_i c_{ni} / \sum_{j=1}^n c_{nj}^2$, where $c_{ni} = x_i - \bar{x}_n$.
- By the Hajek-Sidak's Theorem

$$\sqrt{\sum_{j=1}^n c_{nj}^2}\frac{\widehat{\beta}_1 - \beta_1}{\sigma} = \frac{\sum_{i=1}^n \varepsilon_i c_{ni}}{\sigma\sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{d} N(0,1),$$

  provided

$$\frac{\max_{1\le i\le n}(x_i - \bar{x}_n)^2}{\sum_{j=1}^n (x_j - \bar{x}_n)^2} \to 0$$

  as $n \to \infty$.
- Under some conditions on the design variables

Lindeberg-Feller multivariate CLT

Theorem

(Lindeberg-Feller multivariate) *Suppose* $\mathbf{X}_i$ *is a sequence of independent vectors with means* $\boldsymbol{\mu}_i$*, covariances* $\boldsymbol{\Sigma}_i$ *and distribution function* F_i*. Suppose that* $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\Sigma}_i \to \boldsymbol{\Sigma}$ *as* $n \to \infty$*, and that for any* $\epsilon > 0$

$$\frac{1}{n}\sum_{j=1}^{n}\int_{||\mathbf{x}-\boldsymbol{\mu}_j||>\epsilon\sqrt{n}} ||\mathbf{x}-\boldsymbol{\mu}_j||^2 dF_j(\mathbf{x}) \to 0,$$

then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu}_i) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Example

(multiple regression) In the linear regression problem, we observe a vector $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ for a fixed or random matrix $\mathbf{X}$ of full rank, and an error vector ε with iid components with mean zero and variance σ^2. The least squares estimator of β is $\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. This estimator is unbiased and has covariance matrix $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. If the error vector ε is normally distributed, then $\widehat{\beta}$ is exactly normally distributed. Under reasonable conditions on the design matrix, $\widehat{\beta}$ is asymptotically normally distributed for a large range of error distributions.

Lindeberg-Feller multivariate CLT

Here we fix p and let n tend to infinity. This follows from the representation

$$(\mathbf{X}^T\mathbf{X})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{n} \mathbf{a}_{ni}\varepsilon_i,$$

where $\mathbf{a}_{n1}, \ldots, \mathbf{a}_{nn}$ are the columns of the $(p \times n)$ matrix $(\mathbf{X}^T\mathbf{X})^{-1/2}\mathbf{X}^T =: \mathbf{A}$.

- This sequence is asymptotically normal if the vectors $\mathbf{a}_{n1}\varepsilon_1, \ldots, \mathbf{a}_{nn}\varepsilon_n$ satisfy the Lindeberg conditions.
- The norming matrix $(\mathbf{X}^T\mathbf{X})^{1/2}$ has been chosen to ensure that the vectors in the display have covariance matrix $\sigma^2\mathbf{I}_p$ for every n.
- The remaining condition is

$$\sum_{i=1}^{n} ||\mathbf{a}_{ni}||^2 E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}|||\varepsilon_i|>\epsilon\}} \to 0.$$

Lindeberg-Feller multivariate CLT

- Because $\sum ||\mathbf{a}_{ni}||^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^T) = p$, it suffices that $\max_i E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}|||\varepsilon_i|>\epsilon\}} \to 0$
- The expectation $E\varepsilon_i^2 I_{\{||\mathbf{a}_{ni}|||\varepsilon_i|>\epsilon\}}$ can be bounded $\epsilon^{-k}E|\varepsilon_i|^{k+2}||\mathbf{a}_{ni}||^k$
- a set of sufficient conditions is

$$\sum_{i=1}^n ||\mathbf{a}_{ni}||^k \to 0; \quad E|\varepsilon_1|^k < \infty, \; k > 2.$$

the number of terms present in a partial sum is a random variable. Precisely, $\{N(t)\}, t \geq 0$, is a family of (nonnegative) integer-valued random variables, and we want to approximate the distribution of $T_{N(t)}$

Theorem

(Anscombe-Renyi) *Let X_i be iid with mean μ and a finite variance σ^2, and let $\{N_n\}$, be a sequence of (nonnegative) integer-valued random variables and $\{a_n\}$ a sequence of positive constants tending to ∞ such that $N_n/a_n \xrightarrow{p} c, 0 < c < \infty$, as $n \to \infty$. Then,*

$$\frac{T_{N_n} - N_n\mu}{\sigma\sqrt{N_n}} \xrightarrow{d} N(0,1) \quad as \quad n \to \infty.$$

CLT for a random number of summands

Example

(coupon collection problem) Consider a problem in which a person keeps purchasing boxes of cereals until she obtains a full set of some n coupons.

- The assumptions are that the boxes have an equal probability of containing any of the n coupons mutually independently.
- Suppose that the costs of buying the cereal boxes are iid with some mean μ and some variance σ^2.
- If it takes N_n boxes to obtain the complete set of all n coupons, then $N_n/(n \ln n) \xrightarrow{p} 1$ as $n \to \infty$.
- The total cost to the customer to obtain the complete set of coupons is $T_{N_n} = X_1 + \ldots + X_{N_n}$.
- $$\frac{T_{N_n} - N_n\mu}{\sigma\sqrt{n \ln n}} \xrightarrow{d} N(0,1).$$

CLT for a random number of summands

[On the asymptotic behavior of N_n].

- Let t_i be the boxes to collect the i-th coupon after $i-1$ coupons have been collected.
- the probability of collecting a new coupon given $i-1$ coupons is $p_i = (n-i+1)/n$.
- t_i has a geometric distribution with expectation $1/p_i$ and $N_n = \sum_{i=1}^{n} t_i$.
- By WLLN, we know

$$\frac{1}{n \ln n} N_n - \frac{1}{n \ln n} \sum_{i=1}^{n} p_i^{-1} \xrightarrow{p} 0$$

- $$\frac{1}{n \ln n} \sum_{i=1}^{n} p_i^{-1} = \frac{1}{n \ln n} \sum_{i=1}^{n} n \frac{1}{i} = \frac{1}{\ln n} \sum_{i=1}^{n} \frac{1}{i} =: \frac{1}{\ln n} H_n.$$
- $H_n = \ln n + \gamma + o(1)$; γ is Euler-constant
- $\frac{N_n}{n \ln n} \xrightarrow{p} 1.$

Exercises

- Suppose (X_i, Y_i), $i = 1, \ldots, n$ are iid bivariate normal samples with $E(X_1) = \mu_1$, $E(Y_1) = \mu_2$, $\text{var}(X_1) = \sigma_1^2$, $\text{var}(Y_1) = \sigma_2^2$, and $\text{corr}(X_1, Y_1) = \rho$. The standard test of the hypothesis $H_0 : \rho = 0$, or equivalently, $H_0 : X, Y$ are independent, rejects H_0 when the sample correlation r_n is sufficiently large in absolute value. Please find the asymptotic critical value.

- Suppose $X_i \overset{\text{indep}}{\sim} (\mu, \sigma_i^2)$, where $\sigma_i^2 = i\delta$. Find the asymptotic distribution of the best linear unbiased estimate of μ.

Homework

- Prove Theorem 1.3.9 or Theorem 1.3.10 (choose one of them);
- Suppose $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \mu^2)$, $\mu > 0$. Therefore, $\bar{X}_n$ and S_n are both reasonable estimates of μ. Find the limit of $P(|S_n - \mu| < |\bar{X}_n - \mu|)$;
- Consider n observations $\{(x_i, y_i)\}_{i=1}^n$ from the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, where ε_i's are iid with mean 0 and unknown variance $\sigma^2 < \infty$ (but are not necessarily normally distributed). Assume x_i is equally spaced in the design interval $[0, 1]$, say $x_i = \frac{i}{n}$. We are interested in testing the null hypothesis $H_0 : \beta_0 = 0$ versus $H_1 : \beta_0 \neq 0$. Please provide a proper test statistic based on the least squares estimate and find the critical value such that the asymptotic level of the test is α.