Introdução à teoria assintótica para os estimadores

Prof. Caio Azevedo

Introdução

- Vimos que sob a modelagem probabilística não são feitas suposições acerca da distribuição das variáveis aleatórias que representam o processo de seleção das amostras $(Y_1, ..., Y_n)^\top$ (veja aqui e aqui, por exemplo).
- Além disso, os estimadores para as variâncias de alguns estimadores são não viesados (somente) assintoticamente.
- Em geral, quando não se considera formas paramétricas (distribuições de probabilidade, p.e.) para as variáveis aleatórias de interesse, tem-se que utilizar resultados assintóticos e/ou numéricos, na construção das ferramentas inferenciais.

Introdução

- Mesmo que seja possível calcular quantidades como: esperança e variância (EQM) de forma exatas, as distribuições dos estimadores, utilizadas para se fazer inferência (intervalos de confiança, testes de hipótese e obtenção do tamanho da amostra), são (as) assintóticas (eventualmente aproximações numéricas podem ser utilizadas).
- Objetivo: estudar a obtenção dos resultados assintóticos, para alguns estimadores.

Resultados centrais

- Dois resultados centrais em amostragem são:
  - Convergência em distribuições de estimadores (normalidade assintótica).
  - Consistência: convergência em probabilidade.
- Condições "sine quibus non" (ou seja, "sem as quais não pode ser", para a validade dos resultados assintóticos) são os tamanhos amostral e populacional (conglomerados) serem suficientemente grandes, além de certas características que os estimadores e os dados populacionais devem ter.
- Uma breve revisão sobre convergência estocástica pode ser encontrada aqui.
- Cap. 5, 6 e 7 do livro do Prof. Barry James.

Convergências

- Seja $X_i \sim F_i$, i =1,2,..., uma sequência de v.a.'s com fda's (F_i, fda: função de distribuição acumulada). Então, dizemos que $\{X_i\}_{i \geq 1}$ (sequência), converge em distribuição para $X \sim F$ (F - fda), se $\lim_{i \to \infty} F_i = F$, denotando-se por:

$$X_i \xrightarrow[i \to \infty]{D} X$$

Convergências

- Dizemos que a mesma sequência converge em probabilidade para $a, a \in \Re$, se para algum $\epsilon > 0$,

$$\lim_{i \to \infty} P\left(|X_i - a| > \epsilon\right) = 0,$$

denotando-se por

$$X_i \xrightarrow[i \to \infty]{P} a$$

Comentários

- Convergência em distribuição - Teorema central do limite.

  Essencialmente três tipos:

  - Variáveis aleatórias IID/iid (independente e identicamente distribuídas).
  - Variáveis aleatórias independentes mas não identicamente distribuídas.
  - Variáveis que apresentam algum tipo de dependência (p.e., do tipo martingal).

- Convergência em probabilidade - Desigualdades (Chebyshev, Markov etc)

Estrutura estocástica

- Concentrar-nos-emos no plano AAS_s (os resultados podem ser adaptados para outros planos que dependem desse, p.e., AE_2, AC_2).
- Sob AAS_s as variáveis resposta $(Y_1, Y_2, ...)$ em geral são dependentes e eventualmente identicamente distribuídas (veja aqui).
- Nesse caso precisamos utilizar TCL's que exigem condições mais gerais, como a condição de Lindeberg-Hájek.
- Para o(s) plano(s) (do tipo AAS_c) podemos usar TCL's mais simples (variáveis IID), como aqui e aqui.

Estrutura estocástica

- Uma outra diferença em relação à modelagem clássica, reside no fato de que as condições são verificadas em relação as observações da população $(y_1, y_2, ..., y_N)'$ e não das variáveis aleatórias (amostrais) $(Y_1, Y_2, ..., Y_n)'$.
- O ponto anterior tem, em parte, relação com o fato de que estamos lidando com populações finitas $(N < \infty)$, diferentemente da modelagem clássica $(N \equiv \infty)$.

Estrutura geral (no caso da amostragem probabilística)

- Considere uma sequência de populações $\{\mathcal{U}_{\nu}\}_{\nu \geq 1}$, de tal forma que $N_{\nu+1} > N_{\nu}$, $\nu \geq 1$.
- Da população $\mathcal{U}_{\nu}$, uma amostra s_{ν} de tamanho n_{ν} $(n_{\nu+1} > n_{\nu})$ é selecionada segundo o PA - AAS_s.
- Associada à população $\mathcal{U}_{\nu}$ temos:
  - $(y_{1\nu}, ..., y_{n_{\nu}\nu})'$(população) e $(Y_{1\nu}, ..., Y_{n_{\nu}\nu})'$ (amostra)
  - μ_{ν}, s^2_{ν}, $\widehat{\mu}_{\nu}$ e $\widehat{s}^2_{\nu}$ (definidos aqui).
- Anteriormente, demonstrou-se que $\mathcal{E}(\widehat{\mu}_{\nu}) = \mu$ e $\mathcal{V}(\widehat{\mu}_{\nu}) = (1 - f_{\nu})\dfrac{s^2_{\nu}}{n_{\nu}}$, em que $f_{\nu} = \dfrac{n_{\nu}}{N_{\nu}}$, $\nu \geq 1$.

TCL para variáveis IID (AAS_c)

- Suponha que $n_\nu \to \infty$ e $N_\nu - n_\nu \to \infty$, quando $\nu \to \infty$. Considere também que a sequência $\{y_{i\nu}\}_{i\nu}$, é tal que: $\mu_\nu < \infty$ e $s^2_\nu < \infty$, então:

$$\frac{\widehat{\mu}_\nu - \mu_\nu}{\sqrt{s^2_\nu / n_\nu}} \xrightarrow[\nu \to \infty]{D} N(0,1) \tag{1}$$

- Além disso, se $\{y_{i\nu}\}_{i\nu}$ satisfaz a condição $\frac{s^2_\nu}{n_\nu} \xrightarrow[\nu \to \infty]{} 0$, então:

$$\widehat{\mu}_\nu - \mu_\nu \xrightarrow[\nu \to \infty]{P} 0 \tag{2}$$

Cont.

- Na prática, não conhecemos s_{ν}^2, então usamos $\widehat{s}_{\nu}^2$, no lugar. Com efeito, se a sequência $\left\{\frac{(y_{i\nu} - \mu_{\nu})^2}{s_{\nu}^2}\right\}_{i\nu}$, é tal que $\frac{s_{\nu}^2}{n_{\nu}} \xrightarrow[\nu \to \infty]{} 0$, então:

$$\frac{\widehat{s}_{\nu}^2}{s_{\nu}^2} \xrightarrow[\nu \to \infty]{P} 1 \tag{3}$$

- Sob a validade dos resultados (1) e (3), temos que:

$$\frac{\widehat{\mu}_{\nu} - \mu_{\nu}}{\sqrt{\widehat{s}_{\nu}^2 / n_{\nu}}} \xrightarrow[\nu \to \infty]{D} N(0, 1)$$

Cont.

- Raciocínio(s) similar(es) pode(m) ser aplicado(s) para os estimadores (aqui e aqui, respectivamente) $\widehat{\tau}_{\nu}$ e $\widehat{p}_{\nu}$ (e para os respectivos estimadores sob AE_1 e AC_1).
- Para os estimadores razão e regressão, veja Bolfarine & Bussab (2005) e Scott & Wu (1981).

Teorema (AAS_s)

- Suponha que $n_\nu \to \infty$ e $N_\nu - n_\nu \to \infty$, quando $\nu \to \infty$. Considere também que a sequência $\{y_{i\nu}\}_{i\nu}$, satisfaz a condição de Lindeberg-Hajek, ou seja

$$\lim_{\nu\to\infty} \sum_{T_\nu(\delta)} \frac{y_{i\nu} - \mu_\nu}{(N_\nu - 1)s_\nu^2} = 0,$$

para todo $\delta > 0$, onde $T_\nu(\delta)$ é o conjunto das unidades em $\mathcal{U}_\nu$, para os quais

$$\frac{|y_{i\nu} - \mu_\nu|}{\sqrt{(1-f_\nu)s_\nu^2}} > \sqrt{n_\delta}\delta$$

Teorema

- Assim, temos que:

$$\frac{\widehat{\mu}_{\nu} - \mu_{\nu}}{\sqrt{(1 - f_{\nu})s_{\nu}^2/n_{\nu}}} \xrightarrow[\nu \to \infty]{D} N(0,1) \tag{4}$$

- Na prática, não conhecemos s_{ν}^2, então usamos $\widehat{s}_{\nu}^2$, no lugar. Com efeito, se a sequência $\left\{\frac{(y_{i\nu} - \mu_{\nu})^2}{s_{\nu}^2}\right\}_{i\nu}$, é tal que $\frac{(1 - f_{\nu})s_{\nu}^2}{n_{\nu}} \xrightarrow[\nu \to \infty]{} 0$, então:

$$\frac{\widehat{s}_{\nu}^2}{s_{\nu}^2} \xrightarrow[\nu \to \infty]{P} 1 \tag{5}$$

Cont.

- Sob a validade dos resultados (4) e (5), temos que:

$$\frac{\widehat{\mu}_\nu - \mu_\nu}{\sqrt{(1-f_\nu)\widehat{s}_\nu^2/n_\nu}} \xrightarrow[\nu\to\infty]{D} N(0,1)$$

- Além disso, se $\{y_{i\nu}\}_{i\nu}$ satisfaz a condição $\dfrac{(1-f_\nu)\,s_\nu^2}{n_\nu} \xrightarrow[\nu\to\infty]{} 0$, então:

$$\widehat{\mu}_\nu - \mu_\nu \xrightarrow[\nu\to\infty]{P} 0 \qquad (6)$$

Teorema

- Assim, temos as condições que devem ser verificadas para que os principais resultados assintóticos sejam válidos.
- Em geral, tais condições são satisfeitas na prática, a não ser que os dados apresentem observações discrepantes (outliers).
- Podemos recorrer à métodos de reamostragem para verificar a validade dos resultados assintóticos e/ou como alternativa para obtenção da distribuição empírica do(s) estimador(es) de interesse.
- Para o caso AAS_s, as demonstrações específicas podem ser encontradas em Hajek (1960) e Rényi (1966).

Referências importantes

- Bolfarine, H., Bussab, W. O. (2005). Elementos de amostragem, primeira edição. Associação Brasileira de Estatística, Editora Buchler. Projeto Fisher.
- Chaudhuri, A., Stenger, H. (2020). Survey Sampling: Theory and Methods, second edition, CRC. (link)
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. Publication of the Hungarian Academy of Science, 5, 361-374.
- James, B. (2015). Probabilidade: um curso em nível intermediário, 4ª edição, IMPA.

Referências importantes

- Leite, J. G., Singer, J. M. (1990). Métodos assintóticos em estatística: fundamentos e aplicações. 9º SINAPE. (link)
- Mukhopadhyay, P. (2013). Theory and Methods of Survey Sampling, second edition. (link)
- Pfeffermann, D., Rao, C.R. (2009). Handbook of Statistics Sample Surveys: Design, Methods and Applications, North Holland. (link)
- Rényi, A., (1966): Wahrscheinlichkeitsrechnung: mit einem Anhang über Informationstheorie. 2. Aufl., Berlin (unpublished).

Referências importantes

- Sen, P. K., Singer, J. M. (2019) (1997). Large Sample Methods in Statistics (1994): An Introduction with Applications, CRC Press Revivals. (link)
- Sen, P. K., Singer, J. M., Lima, A. C. P. (2009). From Finite Sample to Asymptotic Methods in Statistics, Cambridge University Press. (link)
- Scott, A., Wu, C. F. (1981). Asymptotic distribution of ratio and regression. Journal of the American Statistical Association 76, 98-102. (link)