Central Limit Theorem

Gerardo Palafox

December 8, 2020

Abstract

In this work, the central limit theorem is presented, along some applications of the same.

1 Introduction

First, a definition concerning convergence of random variables is given. With this, the central limit theorem (CLT) can be stated. The proof is omitted but can be seen in the work of Casella and Berger [2002]. In Section 2, an application of the CLT to Markov chains is explained. Then, an application of the central limit theorem to defining and detecting hierarchical community structures in networks is given in Section 3. Finally, in Section 4, a CLT for an SIR epidemics in a configuration model is discussed.

1.1 Basic theory

Definition 1. *A sequence of random variables $X_1, X_2, \ldots,$ converges in distribution to a random variable X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \tag{1}$$

at all points x where $F_X(x)$ is continuous.

Theorem 1 (Central limit theorem [Casella and Berger, 2002])**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $0 < \operatorname{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cumulative distribution function of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy; \tag{2}$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to a standard normal random variable.

2 Markov Chains

A stochastic process is a sequence of random variables $\{X_i\}_{i\in I}$. A Markov chain is a stochastic process that takes on a finite or countable number of possible values, called states, and such that

$$\mathbb{P}\left\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\right\} = \mathbb{P}\left\{X_{n+1} = j | X_n = i\right\}. \tag{3}$$

Markov chains have several applications, for example, modeling of epidemic processes, such as the one described in Section 4. More about the basic theory of Markov chains can be found in the texts of Ross [2000], Feller [1964], Lawler [2006]. For a state i, let f_i denote the probability that, starting at state i, the process will ever reenter state i. If $f_i = 1$, state i is said to be *recurrent*, and if $f_i < 1$, the state is said to be *transient* . If state i is recurrent, starting in state i, the process will enter state i infinitely often. Let k be a recurrent state in a finite Markov chain. Let N_n denote the number of passages up to time n of the system through state k. Then, N_n is normally distributed as $n \to \infty$ [Feller, 1964]. For example, consider a Markov chain with states $\{0, 1\}$ and probabilities $\mathbb{P}\{X_{n+1} = j | X_n = i\} = P_{ij}$ given by the entries of the matrix

$$P = \begin{bmatrix} .3 & .7 \\ .6 & .4 \end{bmatrix}. \tag{4}$$

Then the number N_n of passages up to time n of the system through the state 0 is normally distributed as $n \to \infty$. This Markov chain was simulated[1] in R [R Core Team, 2020] on a Jupyter notebook [Kluyver et al., 2016]. The number N_n for $n = 10,000$ was computed a thousand times. The histogram of these results is shown in Figure 1.

[1] The code of this simulation, as well as this report, can be found in the Github repository https://github.com/palafox794/AppliedProbabilityModels/tree/master/Assignment14

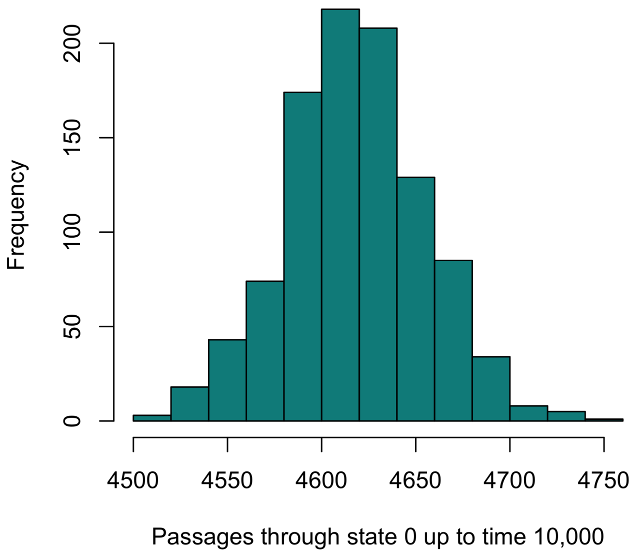

Figure 1: Histogram of N_n for large n.

3 Hierarchical community structure in networks

Schaub and Peel [2020] studied hierarchical community structures in networks. In particular, their work addresses how to define hierarchies of communities, how to determine if such hierarchical structure exists in a network, and how to detect such structures efficiently. To define similarities between nodes, some terminology is introduced. Groups of nodes r and s are called *stochastically equivalent* if any node in group r has the same probability Ω_{rs} of linking to any node in group s. Then, the stochastic block model (SBM) is used to represent the community structure of a network. The SBM defines the probability of a link existing between two nodes depending on their community assignment, with a group indicator binary matrix H, where $H_{ir} = 1$ if node i is assigned to group r and $H_{ir} = 0$ otherwise. Denoting by $H_{i\cdot}$ the ith row of H, A the adjacency matrix of the network, and Ω the affinity matrix, the probability of nodes i and j being linked is given by

$$\mathbb{P}\{A_{ij} = 1\} = H_{i\cdot}\Omega H_{j\cdot}^{\top}. \tag{5}$$

The SBM provides a parametric probability distribution over adjacency matrices. The expected adjacency matrix of this distribution can be calculated from the affinity matrix Ω and group indicator matrix H as

$$\mathbb{E}[A] = H\Omega H^{\top}. \tag{6}$$

Nodes in the network $\mathbb{E}[A]$ which are in the same group are caled *structurally equivalent.* A partition of an adjacency matrix A such that every node in a group r has the same number of links to nodes in a group s is called an *equitable partition.* A *stochastic equitable partition* is an equitable partition in expectation. Partitions that are equitable only between different groups are called *externally equitable partitions* (EEP). A *stochasitc externally equitable partition* (sEEP) is a partition that is externally equitable in expectation. A hierarchical partition is a *valid hierarchy* if at each level the partition is a sEEP and is not degenerate. In detecting hierarchies via spectral methods, authors approximate the true affinity matrix via an estimated affinity matrix $\hat{\Omega}$. To measure how well a partition of $\hat{\Omega}$ approximates an EEP of Ω, the authors use the central limit theorem to conclude the spectral properties of $\hat{\Omega}$ will closely approximate the true Ω, since for large n the entry $\hat{\Omega}_{ij}$ will be approximated by a normal $\mathrm{N}(\mu_{ij}, \sigma_{ij})$ random variable, and $\mu_{ij} = \Omega_{ij}$, $\sigma_{ij}^2 = \Omega_{ij}(1 - \Omega_{ij})/n_i n_j$, where n_i, n_j are the number of nodes in group i and j respectively.

4 Central limit theorems for SIR epidemics on random graphs

Ball [2018] develops central limit theorems for a stochastic susceptible - infectious - recovered epidemic defined on a configuration model [Newman, 2018] random graph. Graphs where degrees of individuals are deterministic (Molloy-Reed) and where degrees are i.i.d. (Newman-Strogatz-Watts) are considered. A population of n individuals, labeled $1, 2, \ldots, n$, is considered. Let $T_i^{(n)}$ be the total number of degree-i susceptibles infected by the epidemic, and $T^{(n)} = \sum_{i=0}^{d_{\max}} T_i^{(n)}$ the final size of the epidemic, where $d_{\max}$ is the maximum degree among all nodes. Upon a suitable standardization, it is proven that the final size of the epidemic converges in distribution to a normal random variable. Let $a_i^{(n)}$ be the number of degree i initial infectious, and $a^{(n)} = \sum_i a_i^{(n)}$ the total number of initial infectious. Let $\epsilon^{(n)} = n^{-1}a^{(n)}$, $\epsilon_i^{(n)} = n^{-1}a_i^{(n)}$, $\epsilon := \lim_{n\to\infty} \epsilon^{(n)}$, and let ϵ_i be such that $\lim_{n\to\infty} \sqrt{n}(\epsilon_i^{(n)} - \epsilon_i) = 0$. If the degree distribution is a random variable D, let $\mathbb{P}\left\{D = i\right\} = p_i$ and $\mu_D = \mathbb{E}\left[D\right]$. Let p_I be the probability that the neighbor of an infectious individual is contacted, and $q_I = 1 - p_I$ the probability that an infectious fails to contact a given neighbor. Let $f_{D_\epsilon}(s) = \sum_{i=0}^{d_{\max}} (p_i - \epsilon_i)s^i$. Define $z \in [0, 1)$ as the unique solution of

$$z - q_I = \mu_D^{-1} p_I f_{D_\epsilon}(z), \tag{7}$$

and

$$\rho = 1 - \epsilon - f_{D_\epsilon}(z). \tag{8}$$

It is proven [Ball, 2018] that $\sqrt{n}(n^{-1}T^{(n)} - \rho)$ converges in distribution to a normal random variable.

References

F. Ball. Central limit theorems for SIR epidemics and percolation on configuration model random graphs. 2018. URL https://arxiv.org/abs/1812.03105v1.

G. Casella and R. L. Berger. *Statistical inference*. Thomson Learning, 2nd edition, 2002. ISBN 978-0-534-24312-8.

W. Feller. *An introduction to probability theory and its applications. Vol I*. John Wiley and Sons, Inc., 2nd edition, 1964.

T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, et al. Jupyter notebooks—a publishing format for reproducible computational workflows. In *Positioning and Power in Academic Publishing: Players, Agents and Agendas: Proceedings of the 20th International Conference on Electronic Publishing*, page 87. IOS Press, 2016.

G. F. Lawler. *Introduction to Stochastic Processes*. Taylor and Francis/CRC Press, 2nd edition, 2006.

M. Newman. *Networks*. Oct 2018. ISBN 978-0-19-880509-0. doi: 10.1093/oso/9780198805090.001.0001. URL https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198805090.001.0001/oso-9780198805090.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

S. M. Ross. *Introduction to probability models*. Harcourt/Academic Press, 7th edition, 2000. ISBN 978-0-12-598475-1.

M. T. Schaub and L. Peel. Hierarchical community structure in networks. 2020. URL https://arxiv.org/abs/2009.07196.