Statistics and Probability Letters 156 (2020) 108623

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

A hypothesis-testing perspective on the G-normal distribution theory

Shige Peng [a], Quan Zhou [b,*]

[a] Institute of Mathematics, Shandong University, Jinan 250100, China
[b] Department of Statistics, Texas A&M University, College Station 77843, USA

ARTICLE INFO

Article history:
Received 15 July 2019
Received in revised form 8 September 2019
Accepted 9 September 2019
Available online 18 September 2019

Keywords:
Heteroskedasticity
Nonlinear heat equation
p-hacking
Sublinear expectation
Tail capacity

ABSTRACT

The G-normal distribution was introduced by Peng (2007) as the limiting distribution in the central limit theorem for sublinear expectation spaces. Equivalently, it can be interpreted as the solution to a stochastic control problem where we have a sequence of random variables, whose variances can be chosen based on all past information. In this note we study the tail behavior of the G-normal distribution through analyzing a nonlinear heat equation. Asymptotic results are provided so that the tail "probabilities" can be easily evaluated with high accuracy. This study also has a significant impact on the hypothesis testing theory for heteroscedastic data; we show that even if the data are generated under the null hypothesis, it is possible to cheat and attain statistical significance by sequentially manipulating the error variances of the observations.

1. Introduction

The primary goal of this note is to study the asymptotic tail behavior of the G-normal distribution, providing a key result to the theory of sublinear expectation spaces developed by Peng (2007, 2008). To statisticians, our result can be interpreted from a hypothesis-testing perspective. Suppose for heteroscedastic observations $X_1, X_2, \ldots$, one wants to conduct a statistical test regarding their common mean. Then by manipulating their variances, the experimenter is able to reject the null hypothesis with probability greater than the nominal significance level when the data are actually generated under the null. This can be seen as a new type of "cheating with the data", which in spirit is similar to the well-known "p-hacking" phenomenon[1] (Head et al., 2015).

As suggested by its name, G-normal distribution plays a central role in the sublinear expectation theory as normal distribution does in the classical probability theory. Indeed, it is the limiting distribution in the generalized "central limit theorem" for sublinear expectation spaces. A more detailed review of the G-normal distribution (and sublinear expectation spaces) will be given in Section 2. As noted in Fang et al. (2017), to characterize the tail behavior of the G-normal distribution, equivalently we can consider the following stochastic control problem (see also Theorem 1.)

Problem 1. Let $\epsilon_1, \epsilon_2, \ldots$ be a sequence of i.i.d. random variables such that $\mathsf{E}(\epsilon_i) = 0$, $\mathsf{E}(\epsilon_i^2) = 1$ and $\mathsf{E}(|\epsilon_i|^3) < \infty$, defined on some filtered probability space $(\Omega, \mathcal{F}, \mathsf{P}, \{\mathcal{F}_i\}_{i=0}^{\infty})$ where $\{\mathcal{F}_i\}_{i=0}^{\infty}$ is the natural filtration generated by $\{\epsilon_i\}_{i=1}^{\infty}$, i.e. $\mathcal{F}_i = \sigma(\epsilon_1, \ldots, \epsilon_i)$. Let $\Sigma(\underline{\sigma}, \overline{\sigma})$ be the collection of all predictable sequences with respect to $\{\mathcal{F}_i\}_{i=0}^{\infty}$ that always take

* Corresponding author.
E-mail addresses: peng@sdu.edu.cn (S. Peng), quan@stat.tamu.edu (Q. Zhou).

[1] "p-hacking" refers to the phenomenon that researchers may try out different data analysis methods until they obtain a *p*-value small enough.

https://doi.org/10.1016/j.spl.2019.108623
0167-7152/

value in $[\underline{\sigma}, \overline{\sigma}]$ where $\underline{\sigma}, \overline{\sigma}$ are given constants ($0 \le \underline{\sigma} \le \overline{\sigma} < \infty$.) For any $\{\sigma_i\}_{i=1}^n \in \Sigma(\underline{\sigma}, \overline{\sigma})$, define $X_i = \sigma_i \epsilon_i$ and $\bar{X}_n = (X_1 + \cdots + X_n)/n$. The problem is to compute the following two functions and find the sequences $\{\sigma_i\}_{i=1}^\infty$ that attain the corresponding supremums,

$$\begin{aligned} p_1(c; \underline{\sigma}, \overline{\sigma}) &:= \lim_{n\to\infty} \sup_{\{\sigma_i\}\in\Sigma(\underline{\sigma},\overline{\sigma})} \mathsf{E}[\mathbb{1}(\sqrt{n}\bar{X}_n > c)], \\ p_2(c; \underline{\sigma}, \overline{\sigma}) &:= \lim_{n\to\infty} \sup_{\{\sigma_i\}\in\Sigma(\underline{\sigma},\overline{\sigma})} \mathsf{E}[\mathbb{1}(\sqrt{n}|\bar{X}_n| > c)], \end{aligned} \tag{1}$$

where $c \in [0, \infty)$ and $\mathbb{1}$ denotes the indicator function.

If $\underline{\sigma} = \overline{\sigma} = \sigma$, the observations $X_1, X_2, \ldots$ are i.i.d. and thus by the classical central limit theorem, we have $p_2(c) = 2\Phi(-c/\sigma) = 2p_1(c)$ where Φ denotes the distribution function of the standard normal distribution. When $\underline{\sigma} < \overline{\sigma}$, the functions p_1 and p_2 are called tail capacities of the G-normal distribution, where "capacity" can be understood as a generalization of probability. The characterization of p_1 and p_2 is vital to the understanding of G-normal distribution. To evaluate p_1 and p_2, we need solve a nonlinear heat equation, which is studied in Section 3. It turns out that p_1 admits a closed-form expression but p_2 does not. The main technical result of this paper is an asymptotic approximation for p_2, which is highly accurate and very easy to compute.

Now we explain how Problem 1 relates to hypothesis testing. Suppose we observe $X_1, X_2, \ldots$, which are generated by the model given in Problem 1 and consider the null hypothesis $H_0 : \mathsf{E}(X_i) = 0$ for every i. When $\underline{\sigma}$ is slightly smaller than $\overline{\sigma}$, both the heteroscedasticity (i.e. the fact that $\mathrm{Var}(X_i)$ is not a constant) and the dependence structure of the observations could be very difficult to detect; if $\{X_i\}_{i=1}^n$ is treated as an i.i.d. sample, the null hypothesis can be tested using the t-statistic,

$$T_n(X) = \frac{\sqrt{n}\bar{X}_n}{\sqrt{s_n^2}}, \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2. \tag{2}$$

For sufficiently large n, $T_n(X)$ can be treated as a standard normal variable and the probability of s_n^2 being greater than $\overline{\sigma}^2$ quickly decreases to zero. Hence, for a one-sided test with level α, the null hypothesis would be rejected if $\sqrt{n}\bar{X}_n > \overline{\sigma}\Phi^{-1}(1-\alpha)$. (One can also use the $(1-\alpha)$% quantile of the t-distribution here and our theory will apply equally.) Imagine that an experimenter is able to choose any $\{\sigma_i\}_{i=1}^n$ from the set $\Sigma(\underline{\sigma}, \overline{\sigma})$ (defined in Problem 1) and wants to maximize the probability of the event $\{\sqrt{n}\bar{X}_n > \overline{\sigma}\Phi^{-1}(1-\alpha)\}$. Then, as will be shown in Section 4, the asymptotically optimal strategy is to simply choose either $\sigma_i = \underline{\sigma}$ or $\sigma_i = \overline{\sigma}$ depending on whether $X_1 + \cdots + X_{i-1}$ is greater than $\sqrt{n}\overline{\sigma}\Phi^{-1}(1-\alpha)$. Further, $p_1(\overline{\sigma}\Phi^{-1}(1-\alpha))$ is always strictly greater than α given that $\overline{\sigma} > \underline{\sigma}$. A similar analysis can be conducted for the two-sided test as well. Simulation studies with unknown $\underline{\sigma}, \overline{\sigma}$ will be provided in Section 4.

We point out that in many applications, it is possible for the experimenter to affect the error variances. For example, consider an economist planning to survey individuals of different ages to study whether some variable has an effect on personal income. The errors are heteroscedastic because the income of older people tends to have a larger variance. Whether the economist deliberately surveys more younger (or older) people seems unimportant since age is included in the regression model as a confounding variable. But the result of this paper implies that this is not true if the economist decides who to survey next (in terms of age) based on previous observations.

2. G-normal distribution and Peng's central limit theorem

The sublinear expectation theory was motivated by capturing the model uncertainty in real-world markets (Artzner et al., 1999; Chen and Epstein, 2002) and has found applications in economics, mathematical finance and statistics (Epstein and Ji, 2014; Peng et al., 2018; Lin et al., 2016). Concepts such as "distribution" and "independence" are redefined for a sublinear expectation space. But to make the present note easier to understand, we will present all the results using the language of classical probability theory, except the use of the terms "G-normal distribution" and "tail capacity".

The central limit theorem for sublinear expectation spaces, first developed by Peng (2008) (see also Peng, 2019), has been presented in various forms. It can be interpreted as a generalization of the classical central limit theorem to controlled stochastic processes (Rokhlin, 2015; Fang et al., 2017), which we summarize in the following theorem.[2]

Theorem 1. *Let $\{\epsilon_i\}_{i=1}^\infty$, $\{\sigma_i\}_{i=1}^n$, $\{X_i\}_{i=1}^n$ and $\Sigma(\underline{\sigma}, \overline{\sigma})$ be as given in Problem 1. Then for any Lipschitz function φ,*

$$\lim_{n\to\infty} \sup_{\{\sigma_i\}\in\Sigma(\underline{\sigma},\overline{\sigma})} \mathsf{E}\left[\varphi(\sqrt{n}\bar{X}_n)\right] = u(1, 0; \varphi), \tag{3}$$

where $\{u(t, x;\ \varphi)$: $(t, x) \in [0, \infty) \times \mathbb{R}\}$ is the unique viscosity solution to the Cauchy problem,

$$u_t = \frac{1}{2}\left(\overline{\sigma}^2(u_{xx})^+ - \underline{\sigma}^2(u_{xx})^-\right), \qquad u(0, x) = \varphi(x). \tag{4}$$

In the above expression, $u_t = \partial u/\partial t$, $u_{xx} = \partial^2 u/\partial x^2$, and the superscripts $+$ and $-$ denote the positive and negative parts respectively.

[2] We choose to present Rokhlin's (2015) version of central limit theorem (see also Fang et al., 2017) since it can be easily understood without knowledge about sublinear expectation spaces. In essence, it is an immediate corollary of Peng's central limit theorem.