Statistics and Probability Letters 184 (2022) 109387

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Long term behavior of incomplete and time varying product ratings

Piotr Kokoszka [*], Deepak Singh, Haonan Wang

Colorado State University, United States of America

ARTICLE INFO

Article history:
Received 29 October 2021
Received in revised form 13 January 2022
Accepted 22 January 2022
Available online 4 February 2022

MSC:
60F15
60G50

Keywords:
Central limit theorem
Incomplete non-iid observations
Product ratings
Strong law of large numbers

ABSTRACT

Customer feedback is widely used to choose a product among various competing products. Such feedback is most commonly available to consumers via *average* 0–5 star ratings. These ratings are based only on opinions of purchasers who decided to rate a product and reflect a long term average of those available responses. We develop the SLLN and the CLT applicable to this realistic situation. In particular, we establish a relationship between the true and the reported long term ratings and study the impact of the probability of leaving a rating.

1. Introduction

Online product reviews have a huge impact on purchase decisions and hence the revenue of manufacturers and sellers. Consequently, their random structure has been extensively studied in the fields of marketing and operations research. We cannot attempt a comprehensive review in this note that is dedicated to the study of limit problems of probability that naturally arise in this setting. We develop limit results that account for the fact that many, or even most, purchasers may not leave a review. The question we seek to answer, is how the reported average ratings are affected by this consumer behavior.

Before formulating the framework we consider and our results, we refer the reader seeking more background to Linden et al. (2003) who study recommendation algorithms for online sellers, McAuley and Leskovec (2013) who propose a statistical model that combines numeric rating and review text, Subbian et al. (2016) who propose recommendations and ratings prediction algorithms and Besbes and Scarsini (2018) who formulate a model accounting for a limited feedback from purchasers and study how average ratings impact purchase decisions. These are just a few examples of contributions in the fields of market and operations research. Our objective is to add a different dimension to such research by establishing relevant limit theorems.

The setting of this paper is as follows. Consider $U_1, \ldots, U_n$ to be a sequence of independent Bernoulli random variables; $U_i = 1$ means that the ith customer leaves a review and $U_i = 0$ means the ith customer does not leave a review. If $U_i = 1$, the customer's rating, X_i, is observed. The iid assumption means that customers do not interact with each other to decide

[*] Corresponding author.
E-mail address: piotr.kokoszka@colostate.edu (P. Kokoszka).

https://doi.org/10.1016/j.spl.2022.109387
0167-7152/© 2022 Elsevier B.V. All rights reserved.

whether to leave a review or not, which is a reasonable assumption for online purchases. The average rating after n purchases is

$$W_n = \frac{X_1U_1 + X_2U_2 + \cdots + X_nU_n}{U_1 + U_2 + \cdots + U_n}. \tag{1.1}$$

In practical scenarios, n can be of the order of thousands or millions, so an asymptotic theory as $n \to \infty$ is suitable. In most rating systems, X_i is identically distributed on $\{1, 2, \ldots, K\}$. If the ratings do not change over time and $P(X_i = k) = r_k$, then the "true" rating is $R = EX_i = \sum_{k=1}^{K} kr_k$. We want to understand the convergence of the available, "incomplete" average W_n to R in situations when the ratings change over time. In our theory, we do not assume that the range of the X_i is $\{1, 2, \ldots, K\}$, we formulate general assumptions.

Ratios of random variables have been studied over several decades. Useful results have been obtained by Marsaglia (1965) who addressed several problems related to the ratio of two normal and two uniform random variables. Hinkley (1969) computed the exact distribution of the ratio of two correlated normal random variables and compared it with an approximation. For other results on the product and ratio of variables, we refer to Lomnicki (1967), Bohrnstedt and Marwell (1978) and Tang and Gupta (1984), and references therein. Most closely related to our work are the results of Novak and Utev (1990) who studied the asymptotic of the first two moments of the ratio

$$Z_n = \frac{\xi_1 + \xi_2 + \cdots + \xi_n}{\eta_1 + \eta_2 + \cdots + \eta_n}. \tag{1.2}$$

We now state their main result.

Theorem 1.1. *Let $\{\xi_i, \eta_i, i \geq 1\}$ be a pairs of i.i.d. random variables such that $P(\eta_1 \geq 0) = 1$ and $0 < E(\eta_1) < \infty$. Further, let $a = E(\xi_1)$, $b = E(\eta_1)$, $\overline{\xi_i} = \xi_i - \eta_i(a/b)$, and consider Z_n defined by (1.2).*

In order to prove $E(Z_n) \to a/b$, it is necessary and sufficient to show that for some $m \geq 1$

$$E(|\xi_1|/(\eta_1 + \eta_2 + \cdots + \eta_m)) < \infty. \tag{1.3}$$

If (1.3) holds and $E(|\xi_1|\eta_1^2) + E(\eta_1^4) < \infty$, then as $n \to \infty$,

$$|E(Z_n) - a/b + n^{-1}b^{-2}E(\overline{\xi_1}\eta_1)| \leq O(n^{-2});$$

and if $E(\xi_1^2(1 + \eta_1) + \xi_1^2(\eta_1 + \eta_2 + \cdots + \eta_m)^{-2}) < \infty$, then as $n \to \infty$

$$|E(Z_n - a/b)^2 - n^{-1}b^{-2}E(\overline{\xi_1^2})| \leq O(n^{-2}).$$

Assuming that the central limit theorem holds, Novak (1997) and Novak (2000) established Berry–Esseen type inequalities. In this paper, we establish conditions for the SLLN and the CLT to hold for the ratio W_n in (1.1). In our setting, the assumption that the ξ_i and the η_i have the same distribution is violated. We study the behavior of the average rating under the assumption that the actual ratings may exhibit a general, nonlinear trend. We also allow the probability of submitting a rating to change over time. In Section 2, we state our main results whose proofs are given in Section 3.

2. Main results

Recall that the X_i in (1.1) represent ratings. The following assumption is designed to accommodate ratings that may evolve over time.

Assumption 2.1. The X_i have the form

$$X_i = g(i) + Y_i,$$

where the Y_i are i.i.d. random variables with mean zero and $|Y_n| \leq M$, a.s., and g is a bounded function. The U_n, $n \geq 1$, are independent Bernoulli random variables with $P(U_i = 1) = p > 0$. The sequences $\{Y_i\}$ and $\{U_i\}$ are independent.

Note that the expected rating of the ith customer is $EX_i = g(i)$. The true average rating is then $\bar{g}_n = n^{-1}\sum_{i=1}^{n} g(i)$. The question we seek to answer is under what assumptions the recorded average rating W_n given by (1.1) is a good approximation to the unobservable $\bar{g}_n$. Our first result states that the distance between W_n and $\bar{g}_n$ converges to zero with probability 1. In particular, if $\bar{g}_n$ has a limit, then W_n converges a.s. to this limit. We note that the probability p of submitting a rating can be arbitrarily small.

Theorem 2.1. *Under Assumption 2.1,*

$$\left| W_n - \frac{1}{n}\sum_{i=1}^{n} g(i) \right| \to 0$$

with probability 1.

We next state the corresponding CLT. Recall a sequence $a(n)$ is Cesàro summable if $\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} a(i)$ exists.

Theorem 2.2. *If Assumption 2.1 holds and the sequence $g^2(n)$ is Cesàro summable, then*

$$\sqrt{n}\left(W_n - \frac{1}{n}\sum_{i=1}^{n} g(i)\right) \xrightarrow{d} N\left(0, \frac{(1-p)\Phi_1 + \Phi_2}{p}\right), \tag{2.1}$$

where $\Phi_1 = \lim_{n\to\infty} \dfrac{1}{n}\sum_{i=1}^{n} g^2(i)$ and $\Phi_2 = EY_1^2$.

The following corollary to Theorems 2.1 and 2.2 applies to the scenario when the ratings do not change over time.

Corollary 2.1. *If $g(i) = \mu$, for all $i \geq 1$, then $\lim_{n\to\infty} W_n \overset{a.s.}{\to} \mu$, and*

$$\sqrt{n}\,(W_n - \mu) \xrightarrow{d} N\left(0, \frac{(1-p)\mu^2 + EY_1^2}{p}\right).$$

We now consider an example when the ratings decline to zero. It illustrates what insights can be gained from Theorems 2.1 and 2.2.

Example 2.1. Suppose for constants $C > 0, \alpha > 0$,

$$g(i) = Ci^{-\alpha}, \ \ i \geq 1.$$

Then $\frac{1}{n}\sum_{i=1}^{n} g(i) \to 0$ and, by Theorem 2.1, $W_n \overset{a.s.}{\to} 0$. Regarding the CLT, $\frac{1}{n}\sum_{i=1}^{n} g^2(i) \to 0$, for any $\alpha > 0$, so $\Phi_1 = 0$ and by Theorem 2.2,

$$\sqrt{n}W_n \xrightarrow{d} N\left(0, \frac{1}{p}EY_1^2\right),$$

provided $n^{-1/2}\sum_{i=1}^{n} g(i) \to 0$. We see that if the true ratings tend to zero sufficiently fast, $\alpha > 1/2$, then the asymptotic distribution of the observed average ratings is the same as if the true ratings were all equal to zero. Clearly, the zero rating can be replaced by any fixed value in the above argument.

We now consider the case when customers leave reviews with a non-constant probability. Theorems 2.3 and 2.4 generalize Theorems 2.1 and 2.2, respectively. It is however useful to first prove the theorems with $p_i = p$ to see the central idea of the proof clearly.

Theorem 2.3. *Suppose $P(U_i = 1) = p_i > 0$ in Assumption 2.1. Then*

$$\left| W_n - \frac{\sum_{i=1}^{n} g(i)p_i}{\sum_{i=1}^{n} p_i} \right| \to 0$$

with probability 1.

Theorem 2.4. *Suppose $P(U_i = 1) = p_i > 0$ in Assumption 2.1 and the limits*

$$\ell_1 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} p_i(1-p_i)g^2(i), \quad \ell_2 = E(Y_1^2)\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} p_i > 0$$

exist. Then

$$\frac{\sum_{j=1}^{n} p_j}{\sqrt{n}}\left(W_n - \frac{\sum_{i=1}^{n} g(i)p_i}{\sum_{j=1}^{n} p_j}\right) \xrightarrow{d} N(0, \ell_1 + \ell_2). \tag{2.2}$$

Theorem 2.4 provides, for example, the following insight. If the ratings are constant, $g(i) = \mu$, then $W_n - \mu$ is of the order $\sqrt{n}/\sum_{j=1}^{n} p_j$. For the true rating μ to be recovered from incomplete observations, the average sampling rate $n^{-1}\sum_{j=1}^{n} p_j$ must be much greater that $n^{-1/2}$. In particular, if the p_j decay like a power function, $p_j \sim j^{-\kappa}$, for some $\kappa > 0$, the κ cannot be too large. The precise requirement is $\kappa < 1/2$.

3. Proofs of the results of Section 2

For ease of reference, we begin by listing several known results. The first two are the Khintchine–Kolmogorov convergence theorem and Lindeberg's CLT, see e.g. Kallenberg (1997).

Theorem 3.1. *Let $\{Z_{i,n}, n \geq 1, 1 \leq i \leq n\}$ be a sequence of independent random variables with zero mean. If $\lim_{n\to\infty}\sum_{i=1}^{n} EZ_{i,n}^2 < \infty$, then $\sum_{i=1}^{n} Z_{i,n}$ converges almost surely.*

Theorem 3.2. *Let $\{Z_{i,n}, n \geq 1, 1 \leq i \leq n\}$ be a sequence of independent random variables with mean zero and finite variance. Consider $T_n = \frac{1}{n}\sum_{i=1}^n Z_{i,n}$ and $s_n^2 = \frac{1}{n}\sum_{i=1}^n V(Z_{i,n})$. If 1. $s_n^2 \to s^2$, 2. for every $\epsilon > 0$, $\frac{1}{n}\sum_{i=1}^n E[Z_{i,n}^2 I(|Z_{i,n}| \geq \epsilon\sqrt{n})] \to 0$, then $\sqrt{n}T_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_{i,n} \xrightarrow{d} N(0, s^2)$.*

We will also use the Kronecker lemma, see e.g. Rohatgi and Saleh (2015).

Lemma 3.1. *If $\sum_{n=1}^\infty x_n$ converges to a finite limit and $\{b_n\}$ is an increasing sequence diverging to infinity, then $b_n^{-1}\sum_{k=1}^n b_k x_k \to 0$.*

Proof of Theorem 2.1. Set

$$Z_{i,n} = X_iU_i - \frac{E(X_iU_i)}{\sum_{j=1}^n E(U_j)}\sum_{j=1}^n U_j = X_iU_i - \frac{g(i)}{n}\sum_{j=1}^n U_j. \tag{3.1}$$

Observe that $EZ_{i,n} = 0$ and for a constant B_1,

$$|Z_{i,n}| \leq |X_i||U_i| + |g(i)|\left|\frac{\sum_{j=1}^n U_j}{n}\right| \leq B_1 \ \ a.s.$$

Set $\Gamma_{i,n} = i^{-1}Z_{i,n}$. The sequence $\{\Gamma_{i,n}\}$ is also a bounded sequence of mean zero random variables satisfying

$$\lim_{n\to\infty}\sum_{i=1}^n E(\Gamma_{i,n}^2) = \lim_{n\to\infty}\sum_{i=1}^n \frac{1}{i^2}E(Z_{i,n}^2) \leq K\sum_{i=1}^\infty \frac{1}{i^2} < \infty,$$

where K is a constant. By Theorem 3.1, the series $\sum_{i=1}^n \Gamma_{i,n}$ converges almost surely. Using the Kronecker lemma, given as Lemma 3.1, we get

$$\frac{1}{n}\sum_{i=1}^n Z_{i,n} = \frac{1}{n}\sum_{i=1}^n i\Gamma_{i,n} \overset{a.s.}{\to} 0. \tag{3.2}$$

Next, observe that

$$W_n - \frac{1}{n}\sum_{i=1}^n g(i) = \frac{\sum_{i=1}^n X_iU_i - \frac{1}{n}\sum_{i=1}^n g(i)\sum_{j=1}^n U_j}{\sum_{j=1}^n U_j} = \frac{\frac{1}{n}\sum_{i=1}^n Z_{i,n}}{\frac{1}{n}\sum_{j=1}^n U_j}.$$

Since, by the SLLN, $\frac{1}{n}\sum_{j=1}^n U_j \overset{a.s.}{\to} p$, (3.2) gives

$$W_n - \frac{1}{n}\sum_{i=1}^n g(i) \to 0,$$

with probability 1. This completes the proof.

Proof of Theorem 2.2. Consider the sequence $\{Z_{i,n}, n \geq 1, 1 \leq i \leq n\}$ defined in (3.1). We begin by verifying the assumptions of Theorem 3.2. We have verified that $|Z_{i,n}| \leq B_1$ a.s., where B_1 is a constant, so $V(Z_{i,n})$ is finite. Consider next the decomposition

$$Z_{i,n}^2 = X_i^2U_i^2 + g^2(i)\left(\frac{1}{n}\sum_{j=1}^n U_j\right)^2 - 2g(i)X_iU_i\left(\frac{1}{n}\sum_{j=1}^n U_j\right)$$

that gives

$$E(Z_{i,n}^2) = E[X_i^2U_i^2] + g^2(i)E\left(\frac{1}{n}\sum_{j=1}^n U_j\right)^2 - 2g(i)E\left[X_iU_i\left(\frac{1}{n}\sum_{j=1}^n U_j\right)\right].$$

Observe that

$$E\left(\frac{1}{n}\sum_{j=1}^n U_j\right)^2 = \frac{1}{n^2}[np(1-p) + n^2p^2] \to p^2$$

and

$$E\left[X_iU_i\left(\frac{1}{n}\sum_{j=1}^n U_j\right)\right] = \frac{1}{n}EX_i\left\{EU_i^2 + \sum_{j\neq i} EU_iEU_j\right\} \to p^2g(i).$$

Therefore, using $X_i = g(i) + Y_i$, for each i,

$$E(Z_{i,n}^2) \to pEX_i^2 + p^2g^2(i) - 2p^2g^2(i) = p(1-p)g^2(i) + pEY_1^2.$$

It follows that, in the notation of Theorem 3.2,

$$s_n^2 := \frac{1}{n}\sum_{i=1}^n E(Z_{i,n}^2) \to p(1-p)\Phi_1 + p\Phi_2 =: s^2. \tag{3.3}$$

Since $|Z_{i,n}| \leq B_1$, the Lindeberg condition is also satisfied. Using Theorem 3.2, we thus conclude that

$$n^{-1/2}\sum_{i=1}^n Z_{i,n} \xrightarrow{d} N(0,\ p(1-p)\Phi_1 + p\Phi_2).$$

Observing that

$$n^{-1/2}\sum_{i=1}^n Z_{i,n} = \frac{\sum_{j=1}^n U_j}{\sqrt{n}}\left(\frac{\sum_{i=1}^n X_iU_i}{\sum_{j=1}^n U_j} - \frac{1}{n}\sum_{i=1}^n g(i)\right)$$

we obtain

$$p\sqrt{n}\left(W_n - \frac{1}{n}\sum_{i=1}^n g(i)\right) \xrightarrow{d} N(0, p(1-p)\Phi_1 + p\Phi_2),$$

completing the proof.

In the remaining two proofs, we use the fact that by the SLLN,

$$\frac{1}{n}\sum_{j=1}^n (U_j - p_j) \xrightarrow{a.s.} 0.$$

Proof of Theorem 2.3. The proof follows by setting

$$Z_{i,n} = X_iU_i - \frac{E(X_iU_i)}{\sum_{j=1}^n E(U_j)}\sum_{j=1}^n U_j = X_iU_i - \frac{g(i)p_i}{\sum_{j=1}^n p_j}\sum_{j=1}^n U_j. \tag{3.4}$$

and using similar steps as in the proof of Theorem 2.1.

Proof of Theorem 2.4. Let $\{Z_{i,n}, n \geq 1, 1 \leq i \leq n\}$ be the sequence random variables defined in (3.4). The idea of the proof is similar to the proof of Theorem 2.2 with a few technical differences. Observe that

$$Z_{i,n}^2 = X_i^2U_i^2 + A_{i,n}^2\left(\sum_{j=1}^n U_j\right)^2 - 2A_{i,n}X_iU_i\left(\sum_{j=1}^n U_j\right), \quad A_{i,n} = \frac{g(i)p_i}{\sum_{j=1}^n p_j}.$$

This gives

$$E(Z_{i,n}^2) = E(X_i^2U_i^2) + A_{i,n}^2E\left(\sum_{j=1}^n U_j\right)^2 - 2A_{i,n}E\left\{X_iU_i\left(\sum_{j=1}^n U_j\right)\right\}.$$

Using $X_i = g(i) + Y_i$, for each i, we get $E(X_i^2U_i^2) = p_i\{g^2(i) + E(Y_1)^2\}$. Observe that

$$E\left(\sum_{j=1}^n U_j\right)^2 = \sum_{j=1}^n p_j(1-p_j) + \left(\sum_{j=1}^n p_j\right)^2$$

and

$$E\left\{X_iU_i\left(\sum_{j=1}^n U_j\right)\right\} = p_ig(i) + p_ig(i)\sum_{j\neq i} p_j.$$

Some algebraic calculations give

$$E(Z_{i,n}^2) = p_i g^2(i) + p_i E(Y_1)^2 + A_{i,n}^2 \sum_{j=1}^{n} p_j(1-p_j) - p_i^2 g^2(i) + o(1).$$

Setting $G_2 = \sup_{i\geq 1} g^2(i)$, observe that

$$A_{i,n}^2 \sum_{j=1}^{n} p_j(1-p_j) = \frac{g^2(i)p_i^2}{\left(\sum_{j=1}^n p_j\right)^2} \sum_{j=1}^{n} p_j(1-p_j) \leq \frac{G_2}{\sum_{j=1}^n p_j} = \frac{1}{n}\frac{G_2}{\frac{1}{n}\sum_{j=1}^n p_j} = O\left(\frac{1}{n}\right).$$

This gives

$$s_n^2 := \frac{1}{n}\sum_{i=1}^{n} E(Z_{i,n}^2) \to \ell_1 + \ell_2 =: s^2.$$

Since $|Z_{i,n}| \leq B_1$, the Lindeberg condition is also satisfied. Using Theorem 3.2, we conclude that

$$n^{-1/2}\sum_{i=1}^{n} Z_{i,n} \xrightarrow{d} N(0, \ell_1 + \ell_2).$$

Similarly as in the proof of Theorem 2.4, it follows that

$$\frac{\sum_{j=1}^n p_j}{\sqrt{n}}\left(W_n - \frac{\sum_{i=1}^n g(i)p_i}{\sum_{j=1}^n p_j}\right) \to N(0, \ell_1 + \ell_2),$$

completing the proof.

Acknowledgements

Research partially supported by NSF grants DMS–1923142, DMS–1914882 and DMS–2123761.

References

Besbes, O., Scarsini, M., 2018. On information distortions in online ratings. Oper. Res. 66, 597–610.
Bohrnstedt, G.W., Marwell, G., 1978. The reliability of products of two random variables. Sociol. Methodol. 9, 254–273.
Hinkley, D.V., 1969. On the ratio of two correlated normal random variables. Biometrika 56, 635–639.
Kallenberg, O., 1997. Foundations of Modern Probability. Springer.
Linden, G., Smith, B., York, J., 2003. Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput. 7, 76–80.
Lomnicki, Z.A., 1967. On the distribution of products of random variables. J. Roy. Statist. Soc.: Ser. B (Methodol.) 29, 513–524.
Marsaglia, G., 1965. Ratios of normal variables and ratios of sums of uniform variables. J. Amer. Statist. Assoc. 60, 193–204.
McAuley, J., Leskovec, J., 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In: Proceedings Of The 7th ACM Conference On Recommender Systems. pp. 165–172.
Novak, S.Y., 1997. On the distribution of the ratio of sums of random variables. Theory Probab. Appl. 41, 479–503.
Novak, S.Y., 2000. On self-normalised sums. Math. Methods Stat. 9, 415–436.
Novak, S.Y., Utev, S.A., 1990. Asymptotics of the distribution of the ratio of sums of random variables. Sib. Math. J. 31, 781–788.
Rohatgi, V.K., Saleh, A.K., 2015. An Introduction to Probability and Statistics. John Wiley & Sons.
Subbian, K., Aggarwal, C., Hegde, K., 2016. Recommendations for streaming data. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. pp. 2185–2190.
Tang, J., Gupta, A.K., 1984. On the distribution of the product of independent beta random variables. Statist. Probab. Lett. 2, 165–168.