Report on the paper "Number of distinct values in a large sample with dependent observations from an infinite discrete distribution"

The paper under review constructs a sequence (X_n) with a certain dependence structure and with values in a the set $\{1, 2, \ldots\}$. The dependence structure is governed by fractional Gaussian noise with Hurst paramter $H \geq 1/2$. The object of interest is the number of distinct values up to time n, i.e. $T_n := |\{X_1, \ldots, X_n\}|$. There are two results: Theorem 1 shows that $T_n \to \infty$ almost surely and Theorem 2 says that $\mathbb{E}T_2^{(H)} < \mathbb{E}T_2^{(1/2)}$ where the dependence on the Hurst parameter is denoted in the superscript.

In my opinion, the object studied here is interesting. The results appear to be new. However, I think they are very weak (in particular Theorem 2 is only obtained for $n = 2$). On the technical level, the paper ignores some more recent works, which might lead to improvements of the results. Furthermore, the presentation of the paper needs a thorough revision. I believe that the paper could be published in *Siberian Electronic Mathematical Reports*.

Here are some remarks to be taken into consideration:

- The title and intro talks about 'dependent observations'. This suggests a generality that is not maintained. One should find a more specific title referring to 'generated by fractional Gaussian noise' or 'Gaussian copula' or anything similar.
- abstract: 'mathematical expectation' is an uncommond word in English.
- intro: Recall the results from [1] and [2] in the introduction
- Sort the list of references alphabetically
- The notation $T^{(i)}$ and $T^{(d)}$ are (a) never used and (b) misleading as i and d are typically indicies.
- Sec 2.1: First sentence. It would be shorter and easier to understand if one only gives the definition without 'corresponding to...' and without the parenthesis.
- I do not understand at which place in the paper the assumption $H \geq 1/2$ is used. This definitely has to be made clear!
- The notation $\Phi_{0,1}$ is overly complicated. Φ suffices.
- Remark 1, Reference to [3]. Note that [3] is a book. Give a more concrete reference.
- Remark 1, Reference to [5]: This is certainly *not* the standard reference for LRD.

- Remark 2 seems like a digression. I'd suggest to either write more so that one can understand the context or delete this.

- After Theorem 1: Delete "From"

- After Lem 1: Mention $c < d$.

- Lem 1: This seems like a technical result. Why does it appear in the "Theorems" section?

- I believe that (3) for $c = -\infty$ is Slepian's lemma. This should be cited and referenced correctly.

- In Section 2.3, the same symbol n is used both for the number of Monte Carlo replications and for the sample size.

- Text below Figure 2: Again the digression on word generation. This might be interesting if explained correctly.

- I do not immediately understand (6). It think it is not "clearly" as mentioned before.

- There are results similar to Proposition 1 in the paper: Wenbo V. Li and Qi-Man Shao, A normal comparison inequality and its applications, Probability Theory and Related Fields 122 (2002), 494–508. There are other papers referenced by that paper and other papers that reference that paper. I think this more modern literature should be included. And it should be checked if that helps to extend the results.

- The notation in Prop. 1 is inconsistent: ξ_j vs. ε_j.

- In (12) to (13), the application of Jensen is quite quick. Insert a step.

- Lemma 3 is just Slepian's lemma. Reference correctly.

- I cannot follow the proof of Lemma 1. The word 'immediately' is certainly inappropriate.

- The list of references looks inconsistent. In [6] it says "In-formational"?