Revision of the manuscript

"Correcting exponentiality test for binned earthquake magnitudes"

by A. Stallone and I. Spassiani

Dear Editor,

Please find attached the revised version of the manuscript "Correcting exponentiality test for binned earthquake magnitudes" by Angela Stallone and Ilaria Spassiani, submitted for publication in *Seismica*. The manuscript has been revised in response to the reviewers' comments, which we greatly appreciated and carefully considered.
The main changes are:

- Correction of Figure 1 (CDF plot);
- Update of the test of Mc as a function of catalog size (Table 1), now based on 50 independent synthetic catalogs instead of one per catalog size;
- Extension of the same test to the case of truncated exponential dithering (Table 3);
- Refinement of the test on rejection probability as a function of sample size using independent random seeds for each bin width-catalog size combination (Figures 3 and 4);
- Relocation of remarks to the Supplementary Material.

Below, you can find our response to each of the reviewers' points.

Sincerely,
Angela Stallone and Ilaria Spassiani

REVIEWER #1

1. Line 1 (abstract)

REVIEWER - The abstract is well written and informative; however, I suggest reshaping it. In current version, is too focused on an intuitive result: replacing the usual uniform dithering of magnitudes within discrete bins with an exponential dithering recovers continuous magnitude exponentiality.
I think the important take-home messages should be:
1) Homogenized earthquakes catalogs should not be presented with uniformly (but exponentially) distributed uncertainty digit. This is the case of the HORUS catalog which is available in two versions. One of them lists magnitudes with uniformly distributed second fractional digit.
2) When implemented, the proposed correction decreases the value of Mc. Provide an order of magnitude of this effect in the abstract.
3) The Lilliefors test is just an example of application, authors' result is universal. Stress this in the abstract. Indeed, the overestimation effect is not test dependent.

AUTHORS - Thank you very much for this helpful comment. We have revised the abstract following the reviewer's suggestions, thus stressing the consequences of dithering discretized magnitudes with uniform noise, quantifying the impact on Mc and generalizing the result (which is indeed not specifically related to

the Lilliefors test, even if this is the test commonly implemented for evaluating the exponentiality of earthquake magnitudes). While we also agree with the reviewer about point 1, our study focuses on the statistical testing of discretized magnitudes rather than on catalog construction or reporting practices. For this reason, we prefer not to address this point.

2. Line 16

REVIEWER - I propose the following:
Earthquakes magnitude tend to follow a continuous exponential distribution
Indeed, the GR law is not obeyed uniformly; in physical terms, it is a (weak) asymptotic dynamic attractor across a limited range of magnitudes where self-similarity holds. Outside this range, the GR is not expected to be the theoretical distribution.

AUTHORS - We agree with the reviewer that the exponential behavior of earthquake magnitudes is not expected to be obeyed uniformly, but over a restricted magnitude range where self-similarity applies. We have revised both the technical abstract and the non-technical summary to make this point clearer. We did not write “tend to follow” because our study, as any study based on testing magnitude exponentiality, assumes exponentiality as the null hypothesis. In statistics, a null hypothesis is not something that “tends” to hold, but something that is assumed and tested.

3. Line 24

REVIEWER - General comment: the paper has been written using remarks, limiting cases as in journals of applied mathematics. Authors shoud consider most of the readers of Seismica are geophysicists and they are interested, more than in demonstrations and mathematical analyses, in the implications of their investigations.

AUTHORS - We have moved the remarks to the Supplementary Material.

4. Line 25

REVIEWER - "complementary cumulative" or simply "the f-m distribution".

AUTHORS - The reviewer is right, we modified accordingly.

5. Line 27

REVIEWER - I disagree. I copy and past my comment above: the GR law is not obeyed uniformly; in physical terms, it is a (weak) asymptotic dynamic attractor across a limited range of magnitudes where self-similarity holds. Outside this range, the GR is not expected to be the theoretical distribution.
****
Additional general comment for the authors on the general use of the Lilliefors test
This is the physical reason why I do not think the Lilliefors test should be used to estimate Mc.
There is also a mathematical motivation: the Lilliefors test is nothing more than a KS test with adjusted parameters (estimated directly from the data); however, in the GR law are not well defined (the mean and variance of power laws with scaling exponent smaller than 1 are not determined and they are dominanted by the occurrence of the largest events)
Finally, a statistical reason: the power of KS and Lilliefors tests is weaker than others, e.g., Anderson-Darling, CSN test.
****
I would write as follows: ... over a wide range of magnitudes

AUTHORS - As for the first point raised by the reviewer, we have revised the text accordingly. About the second point, we understand the reviewer's concern, but the Lilliefors test is commonly adopted for testing magnitude exponentiality. Our study does not aim to propose alternative/better tests, but only to show that exponential dithering prior to the Lilliefors test should be preferred to uniform dithering.

6. Line 47

REVIEWER - The key physical issue for small magnitudes in AI-enhanced catalog comes from the the bias introduced by crustal attenuation patterns variability which dominate source terms at local and regional scales. Essentially, below magnitude 2, the dominant contribution to the final (local) magnitude estimation comes from wave amplitude attenuation.

AUTHORS - We understand the reviewer's comment, but we did not add this to the revised text, as in this context we refer to the issue with small magnitudes in general, and not specifically related to AI-enhanced catalogs.

7. Line 98

REVIEWER - I checked the calculations in the Supplementary Materials; they are correct.
Perhaps, it may be useful for the readers to clarify the structure of the modified Lilliefors tests for power laws.

AUTHORS - We are not completely sure to have understood the reviewer's comment about clarifying the structure of the modified Lilliefors test. We guess he is referring to clarifying that it was initially a normality test, later modified to apply also to exponential distributions. We briefly included such specifications in the Introduction.

8. Line 121

REVIEWER - I suggest using the following LaTeX representation:
\mathbb{1}_{(a,b)}

AUTHORS - We used this LaTeX representation at the beginning, but the LaTeX syntax \mathbb requires the LaTeX package amssymb, and adding this package to the template returned an error. We then used the LaTeX syntax \mathbbm, which requires adding \usepackage{bbm}. Now the notation is the proper mathematical one for indicator functions, and we get no error. However, in the guidelines the authors are asked not to add packages to the *Seismica* template. We will solve this problem with the publication team.

9. Line 149

REVIEWER - Memory is not a property of distributions; I understand the meaning of this sentence; but I think it should be written more clearly

AUTHORS - We have revised the text accordingly.

10. Figure 1

REVIEWER - This figure has some issues:
-1) Using a black line for the exponential distribution is not optimal for visualization: use a red one.
0) Missing a), b) for each subplot.
More serious comments:

1) on the left: pdf is not normalized
2) on the right: CDF must go from 0 to 1 and, also, it is not clear which is the starting magnitude for binning and the meaning of the systematic deviation of the distribution upwards as \Delta m increases.It seems a spurious effect due to the arbitrary choice of the starting point used for the approximation of the true distribution.

AUTHORS - We have edited Fig. 1 according to the reviewer's suggestions (red colour for the exponential distribution and a), b) for the subplots).
Regarding the PDF, the staircase distribution is normalized by construction (the areas of the piecewise-constant rectangles of width Δm sum to 1). We have clarified this in the figure caption.
Regarding the CDF, we thank the reviewer for spotting this. Indeed, while the analytical CDF was evaluated at the right edge of each bin, the plotted x-axis values were wrongly associated with the left edges. This has now been corrected. In addition, the exponential CDF is now correctly starting from the minimum magnitude (1.0), with no shift, so that all CDFs now range from 0 to 1. As discussed in the text, shrinking the bin width Δm forces the staircase to follow the exponential distribution more tightly, thus

reducing the local approximation error, whereas decreasing the number of bins N affects the bin heights and worsens the global fit.

11. Figure 2

REVIEWER N = 50 is the standard binning condition (\Delta m = 0.1) --> the effect us small
PDF is not normalized and CDF must be plotted from 0 to 1.
Comments about labelling and colors as in Figure 1

AUTHORS - We have edited Fig. 2 according to the reviewer's suggestions

12. Table 1

REVIEWER - Motivate why \alpha = 0.1. Do the same in the Supplementary Materials

AUTHORS - The choice of alpha = 0.1 is quite common when using the LIlliefors test for estimating the magnitude of completeness. As stated in Herrmann & Marzocchi (2021): "Choosing $\alpha = 0.1$ is conservative in a statistical sense
(Clauset et al., 2009); less conservative choices $\alpha < 0{:}1$ increase the probability to not reject models that have only a very small chance to follow an exponential distribution.". Moreover, this is the default choice in Herrmann's code (https://zenodo.org/records/4162497) that we use to get the results shown in Table 1. We have added this clarification in the text.

References

- *Herrmann, M., & Marzocchi, W. (2021). Inconsistencies and lurking pitfalls in the magnitude–frequency distribution of high‐resolution earthquake catalogs. Seismological Society of America, 92(2A), 909-922.*
- *Clauset, A., Shalizi, C. R., and Newman, M. E. Power-law distributions in empirical data. SIAM review, 51(4):661–703, 2009.*

13. Table 1

REVIEWER - This trend is suspicious.
How many simulations did the auhtors perform?
Can you add a second fractional digit?
Compare your results with the correct method introduced here.
Finally, in the caption, more information about the simulation should be added.
Finally, please, do not set Mmin = = Mc.
Simulate a catalog as in real conditions with known Mc AND events below Mc (Mmin < Mc).

AUTHORS - In the original analysis, we generated one synthetic catalog and 100 random noise iterations for each catalog size. We have rerun the numerical experiment with 50 synthetic catalogs and 100 random noise iterations for each catalog size. Only two estimates of Mc slightly changed, thus unaffecting the implications of our test. Table 1 is updated accordingly in the revised paper, with a richer caption and the second digit, as suggested by the reviewer, which we thank for this comment.

As suggested by the reviewer, we have performed the same numerical test with the proposed (exponential) dithering. In this case, we did not use Herrmann's code, as it implements uniform dithering. We added the test's outcomes to the revised article (Table 3).

As for the reviewer's suggestion to simulate real catalogs (i.e., incomplete), we agree with the reviewer, but this is not the purpose of our work: here, we simulate complete catalogs with magnitudes that are exponentially distributed for M >= Mmin, and then binned. In this idealized case, we expect the test to recover Mc = Mmin. So, the question we start from is: does uniform dithering bias Mc estimates even for a perfectly complete catalog? We have made this point clearer in the revised article.

14. Table 2

REVIEWER - probability not rates
The same in the supplementary materials in the last lines of the attached simplified codes

AUTHORS - We agree with the reviewer. We have therefore replaced the term "rejection rate" with "rejection probability" throughout the text, figures and captions.

15. Line 213

REVIEWER - The power of a test is an intrinsic property; it does not depend on the data.
I suppose you mean: the larger the catalog and the wider the bins, the rejection rate increases because of the more abundant information and reduced data resolution respectively.

AUTHORS - We thank the reviewer for this clarification. The power of a statistical test is defined independently of data, but it depends on several factors, among them the sample size and the effect size (in our case, this is controlled by how well the staircase PDF approximates the exponential distribution). In particular, for a fixed significance level, the power increases with increasing sample size and/or stronger departures from the null hypothesis (e.g., De Muth, 2014, Eq. 8.1). To avoid any ambiguity, we have rephrased the sentence in terms of the observed rejection probability, which increases with catalog size and bin width, and which is consistent with the expected increase in test power for larger samples and larger deviations from the exponential distribution.

References

- *De Muth, J. E. (2014). Basic statistics and pharmaceutical statistical applications. CRC Press.*

16. Figure 3

REVIEWER - This value makes the rejection probability (again, it is not a rate - i.e., [probability] = pure number, [rate] = T^-1) saturate easily. Set \alpha = 0.05 as usual. It is not a crucial request. Just, I think this choice may me more reasonable and improve visualization.

AUTHORS - Usually, the significance level used when implementing the Lilliefors test for exponentiality test is 0.1 (see our reply at point 13). Also, while the choice of the significance level affects the rejection probabilities returned by the numerical test, this does not change the implications of our analysis: when dithering magnitudes with uniform noise, the rejection probability is substantially higher than the case of exponential dithering.

17. Line 250

REVIEWER - This sentence is wrong: the power is an intrinsic property of tests. It does not depend on data.
I understand what you mean. I suggested a possible way to rephrase this sentence above

AUTHORS - Please, see our reply to point 15.

18. Line 259

REVIEWER - I am wondering why.
I have not a clear back of the envelope explanation for this. My impression is that the slight increase with catalog size may be not significant.
Do the authors agree? Please, state clearly what you think.

AUTHORS - We agree with the reviewer that the increase in the rejection probability with the sample size is not statistically significant. We believe this behavior reflects numerical effects that are present at all sample sizes (e.g., floating-point rounding, ties, p-value approximation errors), but become more visible for very large catalogs. We have clarified this point in the revised text. In addition, we reviewed our numerical test and ensured that independent random seeds are used for each Δm–size combination. While this change slightly reduced the rejection probabilities at the largest sizes, the overall behavior and conclusions remain unchanged. Anyway, we have adopted this implementation to ensure maximum statistical robustness and updated the corresponding tables and figures accordingly.

19. Code

REVIEWER - I checked the codes.

A few comments:
1) In the License: THE SOFTWARE IS PROVIDED "AS IS" --> "... as IT is"
I think there are two issues in the mc_lilliefors_exp.py.
1) Line 26: return mags + noise - MinMag
adds noise then subtracts MinMag, which doesn't preserve the truncation point; you should dither relative to bin centers, not subtract MinMag. RIght?
return mags + noise
2) this is not an issue, but I think that, for future users, this point can be improved:
In the code, M_min is assumed = Mc. If users, without thinking, set M_min < Mc, the test output will become wrong because the test is applied to data, not to data - Mc as it should be.

AUTHORS - The license text is automatically generated by GitHub, we have no control on that

1) This is required by the structure of the noise distribution. Uniform noise assumes a symmetric interval around the centered bin value, so bin centers are considered in this case. However, for the truncated exponential noise we assume support on [0, Δm] (see Sec. 1.3 in the main text), which requires considering lower bin edges. It follows that, in order to shift magnitudes to 0, we need to subtract MagnMin - Δm/2 for the case of uniform noise, and MagnMin for the case of truncated exponential noise.

2) Please, see our reply to point 13.

REVIEWER #2 - HANDLING EDITOR

1) Line 222, the q is with the minus "-" at the exponent?

AUTHORS - We thank the Editor for having spotted this error. We have corrected it.

2) Two different discretization conventions are used in the manuscript: in Section 1.2, magnitudes are discretized using centered bins, in Section 1.3, discretization is reformulated using left-edge bins. While both conventions are legitimate, switching between them has non-trivial implications for the support of the dithering noise, the definition of the shift applied before the goodness-of-fit test, and the interpretation of the resulting continuous variable. This change is only briefly mentioned in a remark, but it should be made fully explicit. I strongly recommend that the authors clearly summarize both conventions explicitly stating how the noise support and shifting operations differ between them.

AUTHORS - We thank the Editor for highlighting this point. We agree that switching between centered and left-edge bins has implications for the definition of the dithering noise and the shift before the

goodness-of-fit test. Bin centers are used in Section 1.2 to align with standard practice and with the standard magnitude class representation by Tinti&Mulargia (1987), while left-edge bins are used in Section 1.3 because they allow the truncated exponential dithering noise to be naturally defined over the full bin support (without breaking it into two parts). In the revised manuscript, we have summarized and clarified the differences between the two conventions. We have also stressed that the two conventions are mathematically equivalent and lead to identical conclusions.

3) In the numerical experiments, the authors generate multiple dithering realizations and assess goodness-of-fit by averaging the resulting p-values and comparing the mean to the significance level \alpha. While this approach may be reasonable as a heuristic stability measure, the mean of p-values is not a standard p-value combination method and does not, in general, preserve exact size under the null hypothesis. I recommend that the authors to briefly justify this choice (e.g., as a variance-reduction or robustness strategy).

AUTHORS - We agree with the Editor that the mean of p-values is not a standard p-value combination method and does not, in general, preserve exact test size under the null hypothesis. In our numerical experiments, however, the averaging of p-values is used as a robustness and variance-reduction strategy to mitigate the dependence of the goodness-of-fit result on a specific realization of the dithering noise. The goal is to ensure that the rejection behavior is not influenced by random fluctuations associated with a single noise realization. This procedure also follows the implementation adopted by Herrmann et al. (2021), where multiple noise realizations are generated and the resulting p-values are averaged to obtain a stable estimate. We have clarified this point in the revised manuscript.

4) The manuscript interprets rejection rates exceeding the nominal significance level \alpha as evidence of distributional mismatch. This interpretation is reasonable, but it should be noted more explicitly that the Lilliefors test for the exponential distribution relies on approximations, with very large sample sizes (e.g., n = 10^6), even minor numerical or structural deviations can become detectable. A short discussion clarifying whether the observed over-rejection is due to genuine distributional differences or to limitations of the test itself (e.g., by comparison with a KS test using a known parameter) would strengthen the argument.

AUTHORS - We agree with the Editor. In the revised manuscript, we have clarified that numerical precision effects could affect the observed rejection probability of the Lilliefors test. We note, however, that the interpretation of rejection probabilities exceeding alpha is based on the comparison between the two dithering cases rather than on the absolute values. When magnitudes are dithered using truncated exponential noise, the rejection probability remains indeed close to alpha across all catalog sizes, including very large ones. This is not happening with the uniform dithering, implying that numerical precision effects alone cannot explain the observed discrepancy.

5) The term “residual lifetime distribution” is used to describe the staircase density. While this is conceptually reasonable, a brief clarification of the terminology would help avoid confusion with the classical definition from reliability theory.

AUTHORS - We thank the Editor for pointing this out, as the terminology we used could indeed be misinterpreted . We have now explicitly clarified in the text that the term “residual lifetime distribution” is used to describe the process distribution of holding and jumping times, with the probability mass allocated across steps (i.e., a staircase density).