Estimating confidence intervals for gravel bed surface grain size distributions: Reply to reviewer comments

Brett C. Eaton[1], R. Dan Moore[1], and Lucy G. MacKenzie[1]

[1]Geography, The University of British Columbia, 1984 West Mall, Vancouver, BC, Canada

July 19, 2019

1 Reviewer 2: General Comments

The second round of comments from Reviewer 2 are presented below, and are discussed at some length. We attempt to address all the key issues, and highlight how we have responded to them.

> P2 L24 I would have to agree with Kristin Buntes original review about this line in the text. It is very difficult to say what is the largest source of uncertainty in measured grain size distributions and all of the other things that she lists could be just as important as sample size. For example, not choosing an appropriate representative sample location in a reach (with wide spatial variations in grain sizes) could lead to larger errors than not sampling enough grains in the right location. I appreciate the authors changes to the text but I think that this sentence is still misleading and ignores many other error sources. I would suggest simply changing the sentence to ...is ONE of the largest sources of uncertaintIES... to address this.

We have made the recommended change.

> Section 2.1.3 on approximate solution. I think it would be good to note here, beyond the title of this section, that this relies on assuming equal area tails. Doesnt this mean that the more the grain size distribution deviates from the equal area tail assumption, the larger the errors in approximate solution? If so, a note of caution is really needed for potential users of this approximation.

Appologies for the confusing wording: the distribution in question is the binomial distribution which describes percentile uncertainty. No assumptions about the distribution of the grain size is made. Therefore the equal area tail approach can be applied to all grain size distributions regardless of shape. We have changed the subsection heading to make this less confusing. The purpose of the method described in Section 2.1.3 is to produce symmetrical confidence intervals by interpolating from the binomial distribution. So the heading now says that.

> Section 3.1 I think that in step 3 of the bootstrap approach, you need to clarify that you are calculating the desired percentile value from the bootstrap samples, rather than just the samples as stated here. The original data used in the bootstrapping approach were also called samples so this step may be confusing as to which exact samples are being used here (original or bootstrapped data). Also shouldnt the subscripts in the equations for this step be xk rather than x and yl rather than y? I would recommend the same changes are made in Section 3.2, which has similar steps, equations, and use of the word sample in step 3.

We have re-written these to sections, as suggested. The re-written text should make it clearer why the subscripts are written as they are.

> Figure 5. I am somewhat confused as to why the 95% confidence interval bounds seem to be different between Figure 5a and 5b. If the confidence bounds are calculated from the true population of 3411 measurements, then shouldnt they be identical between the two figures because they are independent of the sample size of the randomly chosen grains? If the confidence bounds are not based on the entire population of rocks, can you please explain this more in the text because as currently written you state

along the with 95% grain size confidence interval bracketing the true grain size population. Also, why do the confidence intervals not extend though the entirety of the distribution (the tails of the distributions are missing confidence bounds)?

We have added text to the mansucript explaining both of these results. The computed confidence intervals use the population grain size distribution, but sets the sample size to 100 and 400, respectively. That way they are directly comparable to the samples of 100 and 400 stones taken from the population, and are different from the confidence intervals that would be produced using $n = 3411$ (that confidence interval is every narrow). We also observe that researchers almost never make use of percentiles outside the range D5 to D95, and therefore we end our confidence intervals there (this also happens to be the limit of applicability for the normal approximation to the binomial distribution used by Fripp and Diplas, for a sample size of 100, and is thus a good practical limit to use). We also adopt the same convention for all subsequent plots showing the confidence intervals on a graph (a change from the last version), and state this convention in the manuscript at this point.

Figure 6. Please see my comment in Figure 5, why do the confidence bounds not extend to all grain size percentiles and seem to end at about 5 and 95 percentiles? I think there is also an error in labeling this figure (three of the symbols/lines are labeled with 100).

The labelling error has been corrected. And the confidence interval issue is addressed by text added in response to the previous comment.

Figure 7. Although I generally agree with what is stated in the text about this figure, I am not clear how the comparisons between the three different distributions are being made here. The differences between pools, bars and riffles are not being individually discussed in terms of confidence intervals (paragraph on page 14) and you just state that the samples are or are not different. Are you somehow pooling all samples in this comparison, analogous to an ANOVA test and if so, please specify this? If not, please specify which of the distributions you mean are actually statistically different in this discussion rather than just saying samples. In Figure 7, it seems to me that most grain size percentiles could be different for the bars than for the pools or riffles, but this is not discussed until the paragraph on page 15. If bars are different from pools and riffles, then I do not understand what all of the various statistical comparisons of samples mean in the paragraph on page 14; that you state many of samples grain size percentiles are not different on page 14 but on page 15, you state that bars are different from the other two locations. I am also unclear as to why significance at 99% is now being used as the level for significance when throughout the rest of the paper, 95% was used. Why is a stricter significance level now being included?

We have re-written this section to avoid the confusion identified above. We now begin by pointing out that the bars are significantly different from the other two units (pool and run/riffle) for both streams. We then go on to compare the pools to the run/riffle units using the statistical comparision tool. We use both the 95% and the 99% confidence levels merely as an example of the sorts of comparisions that researchers might wish to make. There is no single confidence level that we recommend, and the purpose for making the comparision will determine the appropriate confidence level that a researcher might choose.

Figures 8 and 9. Why is the Bonferroni correction only applied when comparing these data and not any of the other comparisons in the paper, and what is m in these cases (each percentile compared?)?

We have added text to the manuscript which indicates that the Bonferroni correction is appropriate when comparing two distributions based on several statistical metrics, but not when comparing individual metrics, as was the case for the previous comparisons. If we wish to say this population is different from that population, then we need to use several metrics and apply the correction; if we wish to compare the D84 from this population to the D84 for another population, no correction should be applied.

P20 L9-19 Thank you for answering some of my questions about these data but I have a few remaining concerns. For example, how was visual observation of tracer movement conducted? I assume this was not during the flow that actually moved the tracers and that you simply observed that tracers were in different locations before and after a hydrograph and that the assumed critical discharge was the peak discharge value of that hydrograph? If so, can you please specify this because visual observations of tracer stone movement are usually impossible during sediment transport events and is somewhat misleading.

We are gratified that you appreciate how difficult it is to actually observe when tracers begin to move, but in this particular case, that is exactly what we did. A crew of researchers was in the field for the entire duration of the snowmelt season. Because the stream has a snowmelt generated hydrograph, the increase in flows takes place gradually over about one week. The crew was tasked with measuring discharge using a moving boat ADCP, and between discharge estimates would visit each tracer line to observe whether or not tracers had begun to move. Since the tracers were originally deployed in straight lines across the channel, were painted very bright colours, and since the stream was relatively shallow and clear up to about a discharge of about 4 m3/s. it was easy to determine when they began to move.

1.1 Comment 1

2 Reviewer 3: General Comments

The general comments from Reviewer 3 (Dr. R. Hodge) are presented below, and are discussed at some length. We attempt to address all the key issues, and highlight how we have responded to them.

2.1 Comment 1

One comment I had was about the assumptions made in the model. It is assumed that the probability of sampling grains larger than D50 is 0.5, with the statement that half the surface grains are smaller than D50. Both are not necessarily true. D50 is the median grain size sampled using whatever sampling technique is applied. However, if the grains are sampled on a grid, then more than half of the surface grains (by number) could be smaller than D50. The discrepancy is because larger grains are more likely to be sampled than smaller ones as they take up more space and so there is a higher probability of a grid node or foot landing on them. In the case of equal numbers of large and small grains, the larger grains would occupy more than half the surface area and be more likely to be sampled. I dont think that this sampling bias affects your analysis, but it needs more careful wording in the section around page 5, line 5. (See the Bunte and Abt 2001 technical report, section 4.3, for converting between distributions collected using different sampling techniques, e.g. grid to area.)

2.2 reply by authors

This is an important point that we overlooked. The wording identifed by the reviewer is innacurate, and we have changed it. Furthermore, we have added text that this approach only applies to grid-based samples or Wolman based samples, not to other kinds of samples, as the reviewer points out. For the standard Wolman (and the equivalent grid samples), the probabily of picking up a particle of a certain size depends on the relative area of the bed covered by particles of that size, which is at the heart of the statistical argument for this paper.

2.3 Comment 2

From a quick look through the response to reviewers, it looks like the authors have worked on incorporating previous literature. There were some places where this could have been developed a bit further. For example, demonstrating the range of different recommendations that are currently in the literature. Could you have also compared the results of your bootstrapping with the findings of Rice and Church, rather than just saying that you used the same method?

2.4 reply by the authors

We believe that replicating the analysis used by Rice and Church is more fundamental that simply comparing our results to their empirical results. Furthermore, there is no statistical basis for making such direct comparisions in the first place. It would be like using the distribution of heights and weights for some country's Olympic swimming team to describe the distribution of heights and weights for another country's Olymic rowing team: the distributions might be similar (we are talking about Olympic atheletes), but we would expect there to be differences based on the particulars of the sport in question. So too would be expect differences in the shape of the grain size distribution curves for different rivers, based on the flow regime, sediment supply regime and local bedrock lithology. Furthermore, we are presenting a statistically based argument that overcomes the need to make this kind of flawed comparision. Therefore, we replicated the analysis, demonstrated that it is appropriate and useful, and showed that it is consistent with the predictions of binomial theory.

3 Reviewer 3: Specific comments

The reviewer also provides a list of specific comments that improved the paper. Those comments are quoted below, along with our responses to them. The comments are linked to page number/line number.

> 1/15: Make it clear that the spreadsheet is available with this paper?

We have added text making this explicit.

> 2/19: The problem with image based analysis is that you dont know whether you are seeing the b-axis, which could introduce a different bias into the data.

We have added text pointing out this potential bias for photographic-based methods.

> 3/3: Can you demonstrate how different the results from the different empirical analyses are, e.g. in terms of % error in D50?

How different the results are depend in large part on the population being sampled. The main point is that there is no statistical basis for applying the empirical results from one site to other sites. The text has been modified to point this out.

> 3/11: It might be helpful to summarise what Fripp and Diplas presented, e.g. the sample size suggested for a given level of precision.

We have added a summary statement to the manuscript.

> 3/17: Not clear who they is referring to.

Text added to clarify this.

> 3/28: This issue of overlapping intervals that dont include both estimates will not be unique to analysing grain size data. Why cant we use methods that have been developed by other disciplines to address it?

The methods we develop are based on existing, more widely applied statistical theory. In that sense, it is a general solution. The main problem (and the one we are trying to address with our GSDtools package) is that not many geomorphologists are using these existing, standard methods for their analyses.

> 4/13: Move bracket to before 1993.

Done.

> 5/16: It took me a couple of reads to get my head round this; it might be useful to add an additional statement (as you do later on) along the lines of In the case that 60 stones are smaller than D50, then d60 = D50.

We have added text the manuscript, along the lines suggested above.

> 7/3: After a clear overview, I wasnt sure where these subsections (2.1.1 onwards) were going. Can you add a bit more signposting? In particular, I wasnt sure what the aim of this paragraph was.

We have added a sentence indicating that the next section goes into the statistical basis for the method described in overview section, but in a more rigorous way.

> 7/10: It seemed a bit odd to be referring to a confidence interval, when you hadnt yet addressed how it was calculated.

We agree that it is a bit awkward to introduce the confidence interval here in this way. However it is discussed in the previous section, albiet without the statistical basis laid out. And, this discussion was added to emphasize the relatively small effect (relative to the size of the sample taken) of grouping data into phi size classes, an issue raised during the first round of reviews.

> 7/13: I needed a bit more help with comparing the two different approaches in 2.1.2 and 2.1.3. The differences seem to be that 2.1.2 gives asymmetric intervals and is mapped to specific grain size measurements. 2.1.3 seems to be symmetric, and allows interpolation between grain size measurements. When would you use the different approaches? Could you have a version that was asymmetric, but allowed interpolation?

We have modified the subsection headings to help distinguish these two approaches. The first approach is the exact solution, and strictly applies binomial theory. The key problem with the exact solution is that it produces asymmetrical confidence intervals. The second approach is an approximation based on interpolation of the binomial distribution that does produce symmetrical confidence intervals. Given the desirability of symmetrical confidence intervals, we use the approximation. The exact solution presented first because it is the basis for the approximation.

> 11/17: Change to measurements.

Done.

> 14/Fig.7: How have the polygons been calculated, i.e. which percentiles have confidence intervals been calculated for? Also, add something to the caption to explain why 22.6 mm is highlighted.

We have added text to the figure caption addressinb both these points.

> 14/16: Some of this explanation was a bit confusing, because the earlier analysis only refers to comparisons between two GSDs, but there are three in Fig. 7. Are all the significance values for a three-way comparison? I was surprised that the text didnt report more differences between the bars and the other two units as there seems to be almost no overlap in the figure.

This entire section has been substantially re-written. It was unclear, as written, and now we make it clear that only two units are being compared at a time (pools vs run/riffle units). We now begin the section by pointing out that, for both streams, the bars are statistically different from the other two units.

> 18/9: In this line and the next, clarify that you are referring to the mean . It might be helpful to give the range as well.

This has been done.

> 18/12: Change to wide range of.

This has been done.