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

May 28, 2019

1 Reviewer 1: General Comments

The general comments from reviewer 1 are presented below, and are discussed at some length. We attempt to address all the key issues raised there, and to highlight how we are responding to those comments in our revisions. We thank the reviewer very much for such a careful and helpful review of our paper.

1.1 Comment 1

> General comments Eaton et al. (2019) present what seems to be a new way to compute confidence intervals around grain-size distributions that is based on the binomial approach. Encouraging the routine computation of confidence intervals around sampled grain-size distributions is a worthwhile undertaking and often a monitoring requirement for detecting change in rivers beds over time or space. The study by Eaton et al.sets out to provide such a tool. However, the authors do not succeed in making their tool easily accessible: in fact as presented, their approach remains a black box to most potential users. The manuscript does not provide more than general statistical background information and no step-by-step explanations are given on how a potential user could apply the authors. approach to his/her field data. The reader is not much the wiser even after downloading the supplemental material which contains computer code but still no instructions on how to apply the computations. For a user whose basic work tool is spreadsheet computation, the study by Eaton et al. (2019) provides no help for computing confidence intervals.

1.2 reply by authors

This is very useful feedback for us. Our intention is indeed to provide a user-friendly tool that implements binomial statistical theory to calculate confidence bands about grain size distributions to prevent type 1 statistical errors. The revised manuscript now provides an overview of the confidence interval calculation procedure, and then lays out the precise statistical basis for the calculations for different kinds of data (i.e. raw observations and binned data). We have also written a new pair of functions to perform sample to sample comparisons to determine whether sample grain sizes for a percentile of interest are statistically different.

We have re-written the introduction and statistical basis sections of the paper, and we have added an overview section to better explain

1. how the binomial distribution can be applied to both raw data comprising n measurements of b-axis diameters and also to the typical binned data collected in the field; and
2. how the binomial theory can be used to generate confidence intervals about an estimate of a given grain size percentile.

The process is summarized in the new overview section, which describes how to estimate the percentile confidence interval (a term we introduce and use throughout the revised paper), and how to map that onto the sample cumulative frequency distribution to estimate the associated grain size confidence interval. The distinction between these two things is at the root of much of the confusion generated by our original manuscript.

In addition, we have written an appendix to the paper that describes how to use the simpler normal approximation to the binomial distribution to calculate the confidence interval, as well as a spreadsheet implementation of that approach.

We have also created an appendix containing reference tables of the percentile confidence interval bounds for a range of percentiles of interest (i.e. D_{10}, D_{15}, $D_{20} \ldots D_{90}$), sample size (n), and acceptable confidence limit (α).

1.3 Comment 2

> Computation of confidence bands around grain-size distributions without assuming an underlying distribution type is not a new idea. Fripp and Diplas (1993) presented a binomial approach to compute the relation between sample size and error around individual percentiles. The study by Church and Rice (1996) applied a bootstrap approach to a large pebble count of 3500 particles and computed error bands around various percentiles of the grain size distribution. The grain-size distributions did not fit a particular distribution type, but the bootstrap confidence limits were reasonably close to those computed assuming an underlying skewed log-normal distribution. Petrie and Diplas (2002) cautioned that ...the binomial distribution considers only two possibilities for each particle sampled: (1) the particle is within a specific size class (e.g.,smaller than a certain size) or (2) the particle is not within the specified size class. The binomial distribution is then inadequate to use for representing entire size distributions. To overcome this limitation and to compute confidence bands around the cumulative frequency distribution from a pebble count with data binned into size classes while considering distribution characteristics of the distribution, Petrie and Diplas (2000) developed a multinomial approach.

1.4 reply by authors

This is also very useful information for us, and we have read the papers with interest. The work by Diplas and colleagues is particularly relevant and strengthens our paper. The analysis by Fripp and Diplas (1993) is now used as a jumping off point for our analysis: we have re-written our manuscript to use that paper as the basis from which we start, we describe that approach in the appendix, and we have implemented a version of it in a spreadsheet that accompanies this paper.

The paper by Rice and Church (1996) was the inspiration for the re-sampling analysis that we presented in our original paper. However, we have clearly not done justice to the analysis presented therein, so we have expanded that section.

The work by Petrie and Diplas with multinomial theory is primarily focused on determining the sample size required for a given level of accuracy for estimating the shape and relative position of the cumulative grain size distribution, using binned data. Our approach and intent is different: we develop our statistical theory using individual measurements of b-axis diameters, and we develop confidence bounds to be plotted when comparing distributions to avoid type 1 and 2 statistical errors. In this context, the binomial approach is most appropriate. Our implementation of binomial theory is based on the interpretation that a measured stone is either (a) greater than a percentile of interest for the population, or (b) less than or equal to the percentile of interest, with no reference to or limitation imposed by having binned data. In this context, the estimation of j percentiles involves the execution of j independent binomial experiments with assumed probabilities corresponding to the percentile of interest. To test the difference between our approach and the traditional binned data, we use the scheme described in our paper to directly compare the distributions based on all measurements, and the binned data.

While current practice in the field is still to collect binned data, the automated techniques for grain size analysis that are standard practice in most experimental laboratories, and which are being increasingly deployed in the field promise to deliver much more data than can be collected manually, and will obviate the need for binned data. Our methodology is best leveraged in that context, using the automated data analysis approach possible using languages like R and Python. Therefore, our differentiation between binned data and the underlying b-axis diameter measurements is not simply a technical one, it is based on our perceptions of the future data types that will be commonly used.

1.5 Comment 3

> While the study presented by Eaton et al. (2019) is successful in raising awareness that the n=100 sample size is too low to attain reasonable accuracy for pebble counts in most gravel beds and that sample sizes of 400 or 500 particles are required to enable statistical evaluations about sameness or difference, the study does not succeed in presenting its computational approach in an easy to understand way. Providing

> computer code in R-language is not helpful for most users, hence the authors computations cannot be repeated or applied by users who are not expert statisticians but are seeking to determine confidence limits around their sampled grain-size distributions.The authors display the confidence bands that they drew with their binomial approach around grain-size distributions sampled in other studies (Kondolf, 1992; Bunte et al.2009, Bunte and Abt, 2011) and go on to discuss whether the now-drawn confidence bands warrant the interpretations made in the original studies. In the final sections of the study, the authors show general relations between sampling error, as computed with their binomial approach, and sample size as well as distribution sorting.

1.6 reply by authors

We are very grateful for the feedback about the relative difficulty in understanding our approach, and about the need for addition means of implementing our tools for estimating the confidence bands. We have responded to the first point by re-writing the section of the paper presenting the method, and to the second by developing reference materials in two appendices, as well as a spreadsheet implementing the normal approximation to our solution, as described by Fripp and Diplas (1993).

2 Recommendations for improving the paper

The reviewer made several helpful suggestions for improving the paper, listed below:

> Reference prior work and build on it Eaton et al. (2019) should discuss prior studies that likewise compute errors around percentiles without assuming an underlying distribution type and explain the improvements and advantages offered in the study presented. What reason is therefor a user to select the authors approach if the authors do not explain WHY their approach constitutes an improvement?

We have improved the links between our paper and the previous work. We also re-iterate in the revised paper that our main purpose is to produce a user-friendly introduction to the basic method for estimating confidence bounds using binomial theory. We point out that our approach is statistically conventional, has precedents in the literature, and is consistent with empirical analyses. We also more strongly articulate our key message – that all grain size curves ought to be plotted with confidence intervals, particularly when two distributions are being compared.

> Provide explanations and instructions In order for readers to apply the binomial approach to their own data, the authors need to provide a step-by step explanation on how to use their approach rather than referring to a book on statistics, pointing to a website, and offering computer code in R-language. Offering a reader access to computer code is a courtesy, but not a substitute for a step-by step explanation, especially not for a very hands-on and applied topic of monitoring bed-material changes.

With this particular comment in mind, we have re-written the manuscript and generated various reference materials.

> Comparison of results to those from prior work: How do percentile errors computed from the authors binomial approach compare to percentile errors computed from other approaches? Apart from a similarity of sampling errors around the D50 and D84that the authors computed from their binomial as well as a bootstrap approach for asymmetrical grain-size distribution (the authors flume experiment), the authors do not show how their binomial approach to computing confidence bands relates to confidence bands computed from other approaches. The authors should apply their binomial approach together with the approaches suggested by Fripp and Diplas (1993),Petrie and Diplas (2000), and Rice and Church (1996) as well as simply to sample-size equations for an error around the mean to a few pebble-count distributions that differ in their sorting and skewness (esp. the extent of a fine tail) and then assess difference sand similarities between results.

In our revised paper, we make the links to the cited literature clear, and we replicate the approach described by Rice and Church, and then compare it to the binomial methodology we describe.

> Explain whether or how confidence intervals computed from the binomial approach are affected by sorting and skewness of a sampled grain-size distribution While the authors show that confidence bands increase in width with a distribution's sorting co-efficient, the authors do not explain how exactly sorting (and skewness) of a sampled grain-size distribution (e.g., a tail of fines) flow into the computation of confidence

> intervals based on the binomial approach. The binomial approach introduced by Fripp and Diplas (1993) does not seem to involve sorting or skewness of the sampled distribution, suggesting that confidence intervals from a binomial approach are similar for all percentiles within a sampled grain-size distribution with a known sample size and number of size classes.

The revised text and several new figures address this point.

> Have a user in mind and offer a procedure that is reasonably easy to be applied by the user The authors provide a study that is of interest to users who are involved in relations of sample size to error. However, the study is geared towards a statistically expert audience rather than the needs of non-expert potential users. If the authors' work is to be applied for monitoring purposes by staff from environmental agencies or consulting and by those whose main interest is not statistical but who need to apply such relations, then the authors need to provide detailed explanation and instruction.A spreadsheet implementation of their computations of a percentile error would be considerably more helpful than code in R-language.

We have developed additional resources that address this point, and we are particularly thankful for this feedback, since our main purpose is to make it easy for people to use our approach.

> Editing suggestions Figures provided by the authors are generally fine, but considering that the study discusses plotted details of whether or not confidence bands overlap,a larger figure size would be helpful. It would also be helpful to place the figures below their first mention in the text, not simply at the top of the page with a mention some-where below on the page. With respect to writing style and typos (etc.), the manuscript is well written and clean

We have re-worked many of our figures, but will leave it to the editorial staff to properly place the figures in the final version of the manuscript.

3 Reviewer 1: 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.

> p.2, l. 15: ". . .but the largest source of uncertainty in many cases is likely to be sampling variability, which is a function of sample size." How do the authors know that sampling variability (do they mean statistical uncertainty due to a poorly sorted channel bed?) rather than methodological differences (e.g., measurements of particle sizes, spatial heterogeneity, differences in the sampled channel width or leaving poorly accessible stream locations unsampled) is the most likely factor causing uncertainty? The comparative study by Bunte et al. (2009) showed that differences in sampling outcomes due to methodological variability can be huge.

In order to avoid confusion, we have rewritten the sentence to read "but the largest source of uncertainty in many cases is likely to be associated with sample size, particularly for standard pebble counts of about 100 stones."

> p. 2, line 21: ". . . We then use this approach to demonstrate that the higher percentiles, such as D84, are subject to substantial uncertainty for typically used sample sizes, and that. . ." 1) Given this statement, it is odd then that the confidence bands drawn by the authors around the size-distributions from two streams sampled by Bunte et al. (2009) and another stream sampled by Bunte and Abt (2001) are all narrower for the D84 than for the D75 and the D95. 2) That statement is not backed by results from other studies: Rice and Church (1996) have shown for a very large pebble count that uncertainty was lowest for the D75 and D84 sizes, followed by the D50 and D95 sizes, and highest for the smallest percentiles. Green (1993) corroborated this finding; on average, the D73 could be determined with the least uncertainty. Similarly, Bunte and Abt (2001a) found in their field study that uncertainty was lowest for the D50 and the D75, slightly higher for the D84, D95 and D25, and percentiles lower than D25 were subject to the highest uncertainties.

This section of the paper has been re-written, and this sentence is no longer included. The underlying issue that that the standard graphs use lognormal axes. As a result, the uncertainty expressed in mm for the D84 is in fact larger than it is for the D50, even when the uncertainty expressed in phi units is smaller. The point is not an essential one in any case, and is no longer relevant, given the revisions we have made.

> p. 3, l. 3: ". . .since we preserve each measurement rather than grouping them into size classes, the data can be treated as a binomial experiment, . . ." Does that mean that the binomial computations is not applicable to field data binned in 0.5 phi units which results from measuring particle size using a 0.5-phi template?

We have hopefully addressed this question more clearly in the new section presenting an overview of the method we use and in the revised section where we discuss how to apply binomial theory to binned data. (in any case this section containing this sentence has been rewritten to improve clarity).

> In Eq. 1, Pr and p are not defined

This equation is now introduced (and defined) in the overview section to improve clarity. It is used first in an example of the standard coin toss binomial experiment, and then in the directly analogous problem of estimating the bed surface D_{50}.

> p. 4, line 10-19: The description of the methodology is too vague. To allow a reader to replicate the computations, authors need to provide step-by-step guidance. Reference to websites and other studies is not sufficient for a paper that would like to introduce a new approach to computing confidence bands.

The new overview section and the re-written statistical basis section hopefully address this point. We have also adopted the term *percentile confidence interval* and *grain size confidence interval* throughout the text to more clearly explain how binomial theory can be used address the uncertainty associated with sampling (i.e. the percentile confidence interval), and how the shape of the cumulative frequency curve determines the uncertainty for a given grain size percentile estimate (i.e. the grain size confidence interval). In the overview section, we use a new figure to explain the relation between percentile and grain size confidence intervals.

> p. 4, line 21: That statement comes out of the blue . . .what areas? What tails?? Fig. 2 does not provide much help either.

The sentence now reads "One disadvantage of the exact solution described above is that the areas under the tails of the binomial distribution differ". The Figure shows the binomial distribution, so the link between the figure and the text is now more explicit. We have also modified the figure caption and legend labels to explicitly identify the distribution tails.

> p. 5, line 1-5: Again, step-by-step instructions are needed to allow a reader to replicate the authors' approach.

We have tried to address this confusion by creating the overview section that precedes the admittedly rather dense description of the statistical basis for our approach. The precise mathematical approach is laid out in the code behind our functions in the `GSDtools` package (note: we have changed the name of the R package to reflect its more general nature since the addition of two hypothesis testing tools); the underlying calculations which are described in the text can be viewed mathematically by installing the package and then typing `WolmanCI` at the command line prompt. We have also included the source code for the functions in the online archive of code and data associated with this paper. The purpose of publishing an R package is to make our exact code and methodology available for both scrutiny and practical use. We have implemented the simpler normal approximation used by Fripp and Diplas (1993) in a spreadsheet version, and we have described the basis for this approximation in a new appendix to the paper. Hopefully, these additions will help potential users replicate our approach.

Also, we have included step-by-step instructions for the two new functions we have created to test hypotheses about differences between two samples.

> p. 6, line 5-6 ". . .Based on the overlap in 5 confidence intervals for the eight samples, the distributions do not appear to be statistically different (see Fig. 3). . .. 1) Confidence bands plotted by the authors for their stream table sediment overlap for samples 2 and 3, but not for samples 1 and 4 (Fig. 3, panel A). 2) With respect to their multinomial approach, Petrie and Diplas (2000) stated that error bands are identical for all particle- size distributions as long as the value for alpha (e.g., 0.05) and the number of sampled size classes remain the same. For the authors' 8 samples from the stream table sediment surface, I assume that the same number of size classes were collected in each of the 8 samples and that the same alpha value was applied to all computed confidence bands. If the statement by Petrie and Diplas (2000) was true for the error bands conducted by the authors, then why do the error bands plotted in Fig. 3

> differ between samples? 3) The authors use as basis for their analyses a sand-rich sand-gravel mixture with a D50 near 1.5 mm. The lengths of b-axes appear to have been determined to a precision of two decimals (e.g., 0.53 mm). It is difficult to imagine how a pebble count was performed and particle sizes were measured on sediment this small.

This section has been completely re-written, and the text and figure referred to has been removed. In summary though, the data collected were not binned into size classes, individual grain diameters were recorded; the error bands referred to by Petrie and Diplas are percentile confidence intervals, not grain size confidence intervals (an issue we explain in our new overview section); and the measurements were made from a digital photographs of the bed taken 15 cm above the bed with a pixel resolution of about 50 microns. Obviously this introduces the possibility of grains being partially hidden in the photo, but this effect is far less pronounced in laboratory sediments because, due to scaling issues, sediment finer than the field equivalent of 10 mm grains are not included in the bulk mixture (i.e. there are relatively few ‘fine’ grains that can fill in pores and obscure the larger grains the way they can in the field). In addition, the purpose of these data is simply to represent a known population of grains from which to draw samples, not to actually represent the bed surface GSD of the experiment accurately.

> p. 7, Fig. 4: 1) While the box of box and whisker plots typically shows the quartiles, there is less standardization of what the whiskers represent. Please indicate what the whiskers in this plot represent. It can’t be the overall spread because “outliers” are plotted as dots. Please define. 2) What parameter is plotted on the y-axis? Please clarify. 3) It would have been useful to show the 95

We have abandoned this figure, and instead used a different approach to test the binomial predictions against bootstrap error estimates for a much wider range of percentiles. The new figure plot the predicted and bootstrap errors on a typical grain size distribution curve, and we evaluate their goodness of fit using s 1:1 model (i.e. a model of perfect agreement) and the Nash Sutcliffe goodness of fit statistic (which is basically the same as an R^2 value, where 1 equals a perfect model). The completely re-written section on confidence interval testing now engages with previous approaches more explicitly and is more extensive. Note that we replicated the entire confidence interval testing using a different population of grain sizes defined by 1,000,000 observations drawn from a log normal distribution with virtually the same results.

> p. 6, line 9-19. The authors state that they found a close match between the confidence bands computed from the binomial approach and a bootstrap approach (Fig. 4) for an unskewed grain-size distribution (i.e., their stream table sediment). The comparison plot by Petrie and Diplas (2000) for a pebble count from the Mamquam River shows that the confidence bands computed with the approach by Fripp and Diplas (1993) are between n 0.02 and 0.06 phi-units higher than those from the bootstrap approach computed by Rice and Church (1996). Is the binomial approach by Fripp and Diplas (1993) similar or different to the authors’ binomial approach? Does a binomial approach yield wider confidence bands than a bootstrap approach?

We address all of these points in revisions to the introduction (where we talk about the Fripp and Diplas approach), and in the confidence interval testing section. We write in the revised paper “The advantage of a bootstrap approach is that is replicates the act of sampling, and therefore does not introduce any additional assumptions or approximations. The accuracy of the bootstrap approach is limited only by the number of samples collected, and the degree to which the individual estimates of a given percentile reproduce the distribution that would be produced by an infinite number of samples.” The differences observed by Petrie and Diplas are presumably due to their use of the normal approximation of the binomial distribution.

> p. 7, lines 11-19: In the authors’ reassessment of particle-size distributions from Kondolf (1997) and from Bunte et al. (2009), the authors need to clearly state to what percentage confidence the plotted confidence bands refer? I assume they are 95% confidence bands. Please clarify.

Figure captions all now clearly indicate that the polygons represent 95% confidence intervals.

> p. 8, Fig. 6: The study by Eaton et al. (2019) has drawn confidence limits around grain- size distributions from three Rocky Mountain gravel-bed streams sampled by Bunte et al. (2009) and Bunte and Abt (2001). 1) Based on visual examination of the error bands plotted in Fig. 6, I’d say that for Willow Creek, the error bands for riffles and pools are different except for the narrow range between 20 and 50 mm within which they cross. 2) The plotted confidence intervals for Willow Creek and the St. Vrain are jagged around the sampled distribution and seem to widen notably for the flatter sections of the cumulative size distribution but neck down for the steeper sections. The authors offer no explanation for this phenomenon.

The observed changes in the width of the grain size confidence interval do indeed correlate with the shape of the cumulative frequency curve. This effect is due to the mapping of the percentile confidence interval onto the grain size confidence interval. We have added a new figure and an overview section to better explain this point. Comparing samples to determine whether a given percentile of interest is different or whether the samples can be considered different as a whole can only be approximately done using a visual interpretation of the confidence intervals. We have developed two new functions to rigorously compare samples; these functions (and the step-by-step instructions for how to conduct the analysis) are presented in the statistical basis section; they are also used in the reanalysis section; and they are included in the online demonstration of how to use the GSDtools package.

> p. 10, line 9-14: The authors write: “Our method for estimating uncertainty requires only the cumulative distribution and the number of measurements used to construct the distribution. Therefore, confidence intervals can be constructed and plotted for virtually all existing surface grain size distributions (provided that the number of stones that were measured is known, which is almost always the case),. . .” If computation of the width of the confidence interval for any percentile of interest re- quires only knowledge of the sampled distribution and sample size n, and if the computation is conducted for each percentile individually, then how does the spread or sorting of the sampled distribution influence the computed confidence interval? Please CLARIFY!

This is explained in the overview section, and relates to the difference between the percentile confidence interval and the grain size confidence interval.

> p. 11, Fig. 9 and p. 12, Fig. 10: 1) The units in which the error is computed needs to be clearly stated. Somewhere down in the text the reader gets a hint that the error pertains to a percentage error in mm units. 2) The findings that percentile errors decrease with sample size and with the distribution sorting is in and of itself nothing new. What is new here is that the error is computed from the authors’ binomial approach (assuming an underlying log-normal distribution for Fig. 10). To allow a reader to see whether there is a difference between errors computed from the authors’ binomial approach and other approaches (e.g., Fripp and Diplas (1993) or simply errors around a mean), the computed relations between errors and n should be compared to errors computed with other approaches. 3) For comparison with other studies that compute percentile errors in terms of absolute +- error in phi-units it would be helpful if the error-n relations in Fig. 9 had a second y-axis with error in terms of the absolute +- error in phi-units. 4) It would be useful if the relation of error to n was also provided for the error around the D16.

The intention of this section is to provide the user with some guidance related to sample size required to reach a specified level of precision. As should be now clear in the revised paper, the grain size confidence interval cannot actually be estimated until the sample is collected. As a result, we have compared our results to those from others in the confidence interval testing section. This section has been edited to better emphasize that the analysis is only meant to guide sample size estimation, but does not obviate the necessity of calculating the grain size confidence intervals once the sample has been collected. With respect to the units, Eq 3 is now written so as to make it clear that we are calculating a normalized difference, which is by definition dimensionless.

> p. 12, line 8: The authors state that for a given n and sorting, errors are largest for steep gravel fans and bar top surfaces and smaller for typical gravel beds with a sorting near 1. That is a useful comment. It would be even more useful to elaborate a little bit here on what kind of sorting values to expect for different morphological or sedimentary channel units and hence what a user needs to expect in terms of the error - sample size relation.

We agree with this comment, which is what motivated us to model the effect of grain size distribution spread on uncertainty using log normal grain size distribution (the following section). Unfortunately, our data do not support even finer resolution of the issue on a sedimentary unit by unit basis.

> p. 14, line 12-13: I am afraid that the authors’ time estimates refer to dry deposits of mainly mid-sized gravels. The time requirements for a 500-particle pebble count in- creases to about 5 hours when sampling in poorly wadeable conditions, in the presence of abundant algae and large woody debris, under overhanging bushes, and with particles being next to irretrievable from the bed because they are tightly wedged within neighboring particles or small particles placed in tiny pockets between large clasts. The necessity for a large sample size remains, but users and their funding agencies need to commit to realistic time requirement.

We have incorporated the reviewer's time estimate for more arduous samples in a sentence that reads " In less ideal conditions or when working alone, it may take upwards of 5 hours to collect a 500 stone sample, but as we have demonstrated, the uncertainty of the data increases quickly as sample size declines (see Figs. 10 and 11), which may make the extra effort worthwhile in many situations."

> Typos etc. p. 2, l.5: The value should be 22.6 (=2ˆ0.5*16), not 22.7. p. 3 L. 5. . .compute the quantiles of the (Fig. 1). Something is off in that sentence. p. 4, Footnote: The access date is in the future.

We have fixed all of this smaller issues.

4 Reviewer 2: General Comments

The comments provided by Reviewer 2 are presented below, along with our responses. Many of the points have been addressed in our reply to Reviewer 1 above, but these comments were equally helpful in re-shaping the paper, particularly in those instances when Reviewer 2 has identified the same points raised by Reviewer 1.

4.1 Comment 1

> The submitted paper focuses on estimating uncertainties in measured grain size distributions using statistical analysis of grain size data from experiments, field measurements and synthetic data. I think that the authors make an important main point, which is that uncertainties in grain size distributions should be reported especially when used to assess grain size changes over time or in space. Although I am supportive of the overall goals, topics, and messages of this manuscript, I think that there are many details missing from the methods. This makes it difficult to evaluate how this calculation is actually applied, the assumptions involved, and finally how it compares to previously published studies on uncertainties in grain sizes. I suggest adding these details such that your paper can be understood by a broader audience.

4.2 reply by authors

To address these concerns, we have re-written much of the paper and generated additional figures that we hope better describe how our approach actually works. The revised paper also includes an expanded results section that clarifies the links to previous work, as well as reference appendices providing supporting information. We also now provide a spreadsheet that implements the normal approximation to our technique (as described by Fripp and Diplas, 1993) to estimating percentile confidence limits. Finally, we added two functions for explicitly comparing two samples to determine whether differences in the grain size estimate for a given percentile are actually significant. We appreciate all of the suggestions that are made in this review, and we are confident that the revised version will reach a broader audience.

4.3 Comment 2

> I would really like to see a more detailed review of what previous studies have done to quantify uncertainties in the D50 and other percentiles of the grain size distributions. Do approaches without an assumed grain size distribution exist? If so, what is wrong with these approaches that motivates this current study? I'm a bit confused because in the introduction you state that there is no easy way to estimate the required sample size. In the abstract you also write that you propose a simple approach to estimate sample size, but this also relies on assuming a log-normal distribution as in previous studies highlighted on p 2 lines 8-9. What is the difference between your approach that assumes a log normal distribution to estimate sample size and other log normal approaches? It is not entirely clear to me in reading the introduction what is new in this study compared to previous approaches. A more in depth review of previous approaches and a statement of how this new approach is different would really help.

4.4 reply by authors

We have extended our discussion of previous approaches by re-writing the paper to leverage the previous work by Diplas and colleagues as the starting point, and we describe in more detail how we replicated the bootstrap approach of Rice and Church to estimate the uncertainty of samples with various sizes drawn from our population of 3411 b-axis measurements. Basically, we believe that our approach is entirely consistent with that proposed by Fripp and

Diplas (1993), and replicates the empirical results presented by Rice and Church (1996). The main issue that we try to address in this paper is not that previous methods are flawed, but rather that we as a community have failed to use those approaches to quantify sampling uncertainty (despite the precedents in the literature). As a result, there are published results that are clearly not statistically defensible, and it is our impression that many people continue to collect relatively small samples with limited appreciation of what that means in terms of uncertainty.

In our revisions, we will also emphasize that we think are the main contributions of this paper, which are:

- to describe clearly how surface sampling can be described as a binomial experiment, analogous to a traditional coin toss experiment;
- to present a simple set of tools based on binomial theory with which anybody can easily calculate the grain size confidence interval about any sample percentile that will contain the population percentile size;
- to demonstrate the importance of considering uncertainty when comparing samples of the bed surface, or when making calculations based on those samples; and finally
- to make some assumptions about distribution shape so that we can provide some general guidance on the sample size required to reach a desired level of sampling precision.

This last point involves making assumptions about the underlying distribution (i.e. we assume a log normal grain size curve), but that is simply to generate synthetic data with which to model the effect of sample size and the spread of the distribution on the precision of a percentile estimate. We will make it clear that any distribution form could have been used, but that we chose a log-normal distribution because (1) it is the simplest to describe (i.e. it can be described by a mean and standard deviation), (2) it has been used previously by others, and (3) many gravel beds are approximately log-normal. We more clearly emphasize our central message in our revisions, and de-emphasize the point about sample size.

4.5 Comment 4

The reviewer made several comments about our calculations that we would like to address:

> In section 2.1, how is equation (1) used? Please provide a step wise explanation nohow someone would perform these calculations and what information is needed. Right now it is somewhat difficult to understand how equation (2) is actually solved. Although I appreciate the inclusion of the R code that is part of this paper, a simple explanation of your detailed methodology is really needed in the main text to properly evaluate your methods. What are successes, please define. I am also somewhat confused about the definition of p, earlier you state it is the percentile of a distribution but on P 4 L6 is it called a probability.

We have completely re-written the statistical basis, including an overview section that walks the user through the idea of a binomial experiment, the probabilities of a particular outcome (and the relation of those probabilities to the grain size percentiles for the population being sampled), and the relation between percentile confidence intervals and grain size confidence intervals.

> In section 2.2, please also provide more details on this approach, one brief sentence on interpolation really does not make this calculation clear.

We have re-written the entire section to improve clarity.

> Section 3 and Figure 4 How many times did you create a sample with 100 grains to make these distributions in Figure 4? It seems like the results could really vary with the number of 100 grain samples? Also, some explanation of the boxplots is needed to evaluate the results. What are the horizontal lines at the top and bottom ends of the distributions? This information is needed to validate that the two predictions actually provide similar results. Can you provide the actual numeric values of the 99%confidence interval bounds for the two methods in the figures to enable quantitative comparisons?

We have re-written the entire section with these comments in mind. We repeat the kind of bootstrap error estimates presented by Rice and Church (1996), and make a more extensive comparison of the binomial predictions and the bootstrap estimates. We ended up taking 5000 samples from the population to ensure that the distributions of estimates stabilized. In addition, the entire analysis was repeated using samples from a synthetic log normal population of 1,000,000 observations; the re-analysis yielded nearly identical results.

5 Reviewer 1: Specific comments

The reviewer also provides a list of specific comments that will improve the paper, listed below, followed by our response to them.

> P 1 L 21-22 For facies mapping, my understanding of the Buffington approach is not that it is meant to be purely qualitative as implied here. They have visual classification of patches that are then verified by numerous pebble counts on the patches. So their approach likely provides a more accurate representation of the grain size distribution because they use many pebble counts in a single reach.

This is a good point, and we now refer to semi-qualitative methodologies to avoid the issue.

> P3 L5 Missing word(s) here.

This text has been deleted during the revisions.

> P4 L 12-16 Please state if this text is for a specific sample (e.g. the data shown in Figure 1), right now it seems to be written as if it applies to all grain size measurements but I don't think that is actually the case?

It is in fact true for the percentile confidence interval for all samples, but not the grain size confidence interval (which depends on the shape of the cumulative frequency distribution for the sample in question). We have made extensive edits to existing sections and we have added an overview section that addresses this point explicitly.

> P4 L 15-16 Please explain what you mean by 19 times out of 20. I'm not clear why these exact numbers are chosen instead of a percent of trials. It is also not clear how this percent of trails was calculated or how the range of 159-180 was determined.

This section of the paper has been re-written and is augmented by the new overview section that now better explains the how the bounds to the percentile confidence intervals are determined.

> P4 L 21-23 Stating that the area under the tails differs is pretty vague. Do you mean tails of the distribution? How are the tails of the distribution defined? Please state why these different areas are problematic. Similarly, upper and lower limits of what exactly? What do you mean by a one-sided interval and how does this relate to your calculations? I can guess what you mean but the lack of language specificity here makes your text somewhat difficult to follow.

This section has been re-written to improve clarity, and is also augmented by Appendix A, which describes how the confidence interval bounds are determined using the more familiar normal approximation to the binomial distribution. We have also added text to the caption of the figure that explicitly references the distribution tails. We also define one-tailed distributions (though admittedly it remains a technical, statistical definition).

> Figure 3 More details are needed as to how the grain size data were collected, through a random sample or grid count? Were the samples in different locations on stream table and using the same or different operators? It is a little difficult to see the confidence bounds in this figure to assess overlap of various distributions, not sure though how you can easily address this problem.

This figure has been deleted during the re-write of the paper. The main point is that we have a population of 3411 measurements that we can use to replicate the bootstrap error calculations performed by Rice and Church (1996). Since the time and space distribution of the sub-samples used to generate this population is never referred to in the rest of the paper, we chose to delete the figure and simplify the text. Where this population is first introduced, we provide a bit more information about the sampling, as requested. The sentence in the paper reads "the population shown is defined by 3411 measurements of bed surface b-axis diameters at randomly selected locations in the wetted channel of a laboratory experiment performed by the authors."

> P 7 L 8 typo here

Fixed

> Figure 5. I appreciate this reanalysis but I don't think that you can say that the distributions are statistically similar or different without a similar confidence bound on the bulk sample data. Previous studies have demonstrated that bulk samples also have considerable uncertainty depending on the size of the actual bulk sample and the portion of the sample that is occupied by the largest grain sizes. So the bulk sample is also not free from uncertainties and this needs to be acknowledged.

This is a fair point. Given that we have added new sections and figures to the paper, and that we have extended our comparison of our method to previous methods, we chose to remove this figure and the associated analysis.

> P 8 L 3-5 The statement that fine sediment would be deposited preferentially in the pool rather than in the run/riffle during the waning limb of the preceding hydrograph needs some references to support it.

We have added references to some of the seminal work on this topic.

> P 12 L 6-7 Please explain why you are assuming the standard deviation of the distribution is related to logD84-logD50.

To make the paper clearer and to improve the comparability of the field data and the results of the log normal simulations, we now use a sorting index ($\phi_{84} - \phi_{16}$) to quantify the spread of the distribution. This is, we think, a clearer way of conveying what we did without introducing unnecessary complications.

> P 12 L 10-12 I do not entirely why you are simulating log-normal samples with this given range of D50 values and SDlog values? How were these distributions simulated by defining D50 and SDlog beforehand? Figure 10 does not seem to be referenced or explained anywhere in the text.

We have added edits at various points in this section to make a few points related to this comment. We point out that the purpose of this section of the paper is simply to provide some guidance to choosing an appropriate sample size, and that this is a secondary objective of the paper (the primary objective being the articulation of the importance and relative ease of generating confidence intervals about bed surface grain size distributions). We also now clearly state that we approach this problem first using a set of field data to estimate the grain size uncertainty associated with different sample sizes, and second by using log normal distributions to quantify the effect of data spread, indexed by standard deviation. We generated the log normal distributions using the `rnorm` function in R (e.g. `GSD = 2^rnorm(n = 352, mean = 5.6, sd = 1.3)`).

> P 13 L. 14-22 More details are needed as to how you estimated that this grain size is entrained at a certain shear stress and discharge. Did you use Shields equation? What critical Shields stress did you assume? How did you then translate this shear stress into a discharge beyond using a stage-discharge relation; did you have a measured channel bed slope and are you assuming stage is equivalent to the average flow depth in a reach? What is the basis of the assumption that D50 becomes fully mobile at twice the shear stress needed to initiate D50 movement? Some rational and supporting references are needed to support this argument. I am also a little confused about this uncertainty in grain size because all of these sizes (46, 55, 64 mm) are essentially in the same half-phi bin. I may be mistaken but if you have binned your data into half phi intervals for this analysis, wouldn't you expect a similar, although likely smaller, level of uncertainty in the D50 anyway? This uncertainty would occur because you are deter- mining the measured stream bed D50 value (55 mm) by interpolation between the two percentiles straddling the 50th percentile value, and these two bounding percentiles correspond to grain size bins 45 and 64 mm. But you do not actually have any grain size resolution finer than half phi bin size. So when you calculate a median grain size of 55 mm, you are interpolating this grain size to a finer resolution than you actually have data. Doesn't this already seem to imply that your uncertainty in D50 might be somewhere within a half phi bin size when you only have binned data, depending of course on how the actual grain sizes are distributed within that half phi bin?

We now explain how we determined the entrainment threshold (visual observation of painted tracers, confirmed to occur at a dimensionless shear stress of about 0.045). The other details of the methodology to estimate shear stress are described in the referenced papers. We have added a reference supporting full mobility at twice the entrainment threshold. The issue of interpolation using binned data, and the accuracy of that kind of data relative to individual measurements of b axis diameters is now addressed in the overview section and in the re-written statistical basis section. In particular, our new Fig. 3 demonstrates that the differences between binned data and interpolations from cumulative data are small compared to the sampling confidence interval, which means that, in practice, binned data can be treated as if they were not binned.

> P 15 L 12-13 Although I certainly agree that having more than 100 sampled particles would be better for uncertainties in most studies, these time estimates assume a team of people performing pebble counts. Having conducted a very large number of pebble counts on my own, these can take much longer than 20 minutes. The time also really depends as to whether you are binning grain sizes or measuring individual b axes. Finally, setting up and finding grains on a grid also adds to the pebble count time, so I would argue that this 20 minute estimate is a minimum.

We have added a note to this section of the paper that does acknowledge the difficulties of collecting large samples in arduous conditions.

~~Estimating confidence intervals for gravel bed surface~~ Percentile-based grain size ~~distributions~~distribution analysis tools (GSDtools) – estimating confidence limits and hypothesis tests for comparing two samples

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

Correspondence: Brett Eaton (brett.eaton@ubc.ca)

Abstract. Most studies of gravel bed rivers present at least one bed surface grain size distribution, but there is almost never any information provided about the uncertainty of the percentile estimates. We present a simple method for estimating the grain size confidence intervals about ~~the grain size~~ sample percentiles derived from standard Wolman or pebble count samples of bed surface texture. ~~Our approach uses binomial probability theory to generate confidence intervals for all grain sizes in the distribution. We~~ The width of a grain size confidence interval depends on the confidence level selected by the user (e.g., $\alpha = 0.05$ for a 95% confidence interval), the number of stones sampled to generate the cumulative frequency distribution, and the shape of the frequency distribution itself. For a 95% confidence interval, the true grain size of the underlying population will fall within the confidence interval for the sample 95% of the time. The method uses binomial theory to calculate a percentile confidence interval for each percentile of interest, then maps that confidence interval onto the cumulative frequency distribution of the sample in order to calculate the more useful grain size confidence interval. The validity of this approach is confirmed by comparing the predictions using binomial theory with estimates of the grain size confidence interval based on repeated sampling from a known population. We also developed a two-sample test of the equality of a given grain size percentile (e.g., D_{50}), which can be used to compare different sites, sampling methods or operators. The test can be applied with either individual or binned grain size data. These analyses are implemented in the freely available `GSDtools` package, written in the R language. A solution using the normal approximation to the binomial distribution is implemented in a spreadsheet. Applying our approach to various samples of grain size distributions in the field, we find that the standard sample size of 100 observations is ~~associated with errors~~ typically associated with uncertainty estimates ranging from about ±15% to ±30%, which may be unacceptably large for many applications. In comparison, a sample of 500 stones produces ~~an uncertainty~~ uncertainty estimates ranging from about ±9% to ±18%. In order to help workers develop appropriate sampling approaches that produce the desired level of precision, we present simple equations that approximate the proportional uncertainty associated with the ~~median size and the 84th percentile~~ 50^{th} and 84^{th} percentiles of the distribution as a function of ~~the~~ sample size and ~~the standard deviation of the distribution, assuming that the underlying distribution is log-normal. However, the~~ sorting coefficient; the true uncertainty of any sample depends on the shape of the sample distribution, and can only be accurately estimated once the sample has

been collected~~, so these simple equations complement – but do not replace – the basic uncertainty analysis using binomial probability theory~~.

1 Introduction

A common task in geomorphology is to estimate one or more percentiles of a particle size distribution, denoted ~~D_p~~D_P, where D represents the particle diameter (mm) and the subscript ~~p~~ P indicates the percentile of interest. Such estimates are typically used in calculations of flow resistance, sediment transport, and channel stability; they are also used to track changes in bed condition over time, and to compare one site to another. In fluvial geomorphology, commonly used percentiles include D_{50} (which is the median) and D_{84}. In practice, sampling uncertainty for the estimated grain sizes is almost never considered during data analysis and interpretation. This paper presents a simple approach based on binomial theory for calculating grain size confidence intervals, and for testing whether or not the grain size percentiles from two samples are statistically different.

Various methods for measuring bed surface sediment texture have been reviewed by previous researchers (**???**). While some approaches have focused on using ~~qualitative~~ semi-qualitative approaches such as facies mapping (e.g. **?**), or visual estimation procedures (e.g. **?**), the most common means of characterizing the texture of a gravel bed surface is still the cumulative frequency analysis of some version of the pebble count (**????**). Pebble counts are sometimes completed by using a random walk approach, wherein the operator walks along the bed of the river, sampling those stones that are under the toe of each boot and recording the b-axis diameter. In other cases, a regular grid is superimposed upon the sedimentological unit to be sampled, and the b-axis diameter of all the particles under each vertex is measured. In still other cases, computer-based photographic analysis identifies the b-axis of all particles in an image of the bed surface. Data are typically reported as cumulative grain size distributions for 0.5ϕ size intervals (e.g., 8 - 11.3 mm, 11.3 to 16 mm, 16 - ~~22.7 mm, 22.7~~ 22.6 mm, 22.6 - 32 mm, and so on), from which the grain sizes corresponding to various percentiles are extracted. ~~Attempts to characterize the uncertainty of this approach have focused on estimating the uncertainty of D_{50}, and have typically assumed that the underlying distribution is log normal (???). Attempts to characterize the uncertainty associated with other percentiles besides the median have relied on statistical analysis of extensive field data sets (????), and do not provide an easy means of calculating the sample size required to achieve a given confidence level.~~

Operator error and the technique used to randomly select bed particles have frequently been identified as important sources of uncertainty in bed surface samples (**????**), but the largest source of uncertainty in many cases is likely to be ~~sampling variability, which is a function of sample size~~associated with sample size, particularly for standard pebble counts of about 100 stones. Unfortunately, the magnitude of the confidence interval bounding an estimated grain size is seldom calculated and/or reported, and the implications of this uncertainty are – we believe – generally under-appreciated. To address this issue, we believe that it should become standard practice to calculate and graphically present the confidence intervals about surface grain size distributions.

For the most part, attempts to characterize the uncertainty of pebble counts have focused on estimating the uncertainty of D_{50}, and have typically assumed that the underlying distribution is log normal (**???**); when used to determine the number of measurements required to reach a given level of sample precision, these approaches also require that the standard deviation of the underlying distribution be known, beforehand.

Attempts to characterize the uncertainty associated with other percentiles besides the median have relied on empirical analysis of extensive field data sets (**????**), which cannot be easily applied to pebble counts from other gravel bed rivers having a different population of grain sizes. Perhaps because of the complexity involved in extending the grain size confidence intervals about the median to the rest of the distribution, researchers almost never present confidence intervals on cumulative frequency distribution plots, or constrain comparisons of one distribution to another by any estimate of statistical significance. While others have recognized the limitations of relatively small sample sizes (**????**), it still seems to be standard practice to rely on surface samples of about 100 observations.

? do present a means of generating confidence intervals bounding a grain size distribution. They present a method for determining the minimum sample size required to achieve a desired level of sample precision using the normal approximation to the binomial distribution, wherein uncertainty is expressed in terms of the percentile being estimated (i.e., they estimate the percentile confidence interval), but not in terms of actual grain sizes (i.e., the grain size confidence interval). **?** demonstrate that the percentile confidence interval predicted by **?** is similar to the empirical estimates produced by **?**, who repeatedly sub-sampled a known population of grain size measurements in order to quantify the confidence interval; they also recommend plotting the confidence intervals on the standard cumulative distribution plots as an easy way of visualizing the implications of sampling uncertainty. It is worth noting that the primary focus of the previous analyses has been directed toward determining the sample size necessary to achieve a given level of sample precision; it has not been adapted to the analysis and interpretation of surface distribution samples, once they have been collected.

A number of studies have compared grain size distributions for two or more samples to assess differences among sites, sampling methods or operators (**??????**). A simple approach would be to construct confidence intervals for the two estimates. If the confidence intervals do not overlap, one can conclude that the estimates are significantly different at the confidence level used to compute the intervals (e.g., 95%); and if a percentile estimate from one sample falls within the confidence interval for the other sample, then one cannot reject the null hypothesis that the percentile values are the same. However, the conclusion is ambiguous when the confidence intervals overlap but do not include both estimates; even for populations with significantly different percentile values, it is possible for the confidence intervals to overlap. Therefore, there is a need for a method to allow two-sample hypothesis tests of the equality of percentile values.

The objective of this note is ~~introduce a~~ to introduce robust, distribution-free ~~approach to computing confidence intervals~~ approaches to (a) computing percentile confidence intervals and then mapping them onto a given cumulative frequency distribution from a standard pebble count in order to estimate the grain size confidence interval for the sample, and (b) conducting two-sample hypothesis tests of the equality of grain size percentile values. The approaches can be applied not only in cases in which individual grain diameters are measured, but also to the common situation in which grain diameters are recorded within phi-based classes, so long as the number of stones sampled to derive the cumulative distribution is also known.

The primary purpose of this work is to guide the analysis and interpretation of the grain size samples. While grain size confidence intervals are most applicable when comparing two samples to ascertain whether or not they are statistically different, we also demonstrate how knowledge of grain size uncertainty could be applied in a management context, where flood return period is linked to channel instability (for example). As we demonstrate in the paper, percentile uncertainty is distribution-free, and can be estimated using standard look-up tables similar to those used for ~~percentile estimates. We then use this approach to demonstrate that the higher percentiles, such as~~ t-tests, or using the normal approximation to the binomial distribution referred to by **?** (see Appendix A). Translating percentile confidence intervals to grain size confidence intervals requires information about the grain size distribution, but is essentially a mapping exercise, not a statistical one. We implement both the estimation of a percentile confidence interval and the mapping of it onto a grain size confidence interval using: (1) a spreadsheet that we provide which uses the normal approximation to the binomial distribution, described by (**?**); and (2) an R package called `GSDtools` that we have written for this purpose that uses the statistical approach described in this paper. A demonstration is available online at `https://bceaton.github.io/GSDtools_demo_2019.nb.html`, which provides instructions for installing and using the `GSDtools` package; the demonstration is also included in the data repository associated with this paper. Finally, we use both existing data sets and the results from a Monte Carlo simulation to develop recommendations regarding the sample sizes required to achieve a pre-determined precision for estimates of the D_{50} and the D_{84}~~, are subject to substantial uncertainty for typically used sample sizes, and that this uncertainty translates into significant uncertainty in estimates of sediment entrainment thresholds. We then provide recommendations regarding sample sizes for estimating particle size percentiles~~.

2 ~~Statistical basis~~Calculating confidence intervals

2.1 Overview

The key to our approach is that the estimation of any grain size ~~quantile D_p~~ percentile can be treated as a binomial experiment~~during which~~, much like predicting the outcome of a coin-flipping experiment. For example, we could toss a coin 100 times and count the number of times the coin lands head-side up. For each toss (of a fair coin, at least), the probability (p) of obtaining a head is 0.50. The number of times that we get heads during repeated experiments comprising 100 coin tosses will vary about a mean value of 50, following the binomial distribution (see Fig. **??**).

The probability of getting a specific number of heads (B_k) can be computed from the binomial distribution:

$$B_k(k,n,p) = p^k(1-p)^{n-k}\frac{n!}{k!(n-k)!} \tag{1}$$

for which k is the number successes (in this case, the number of heads) observed during n trials for which the probability of success is p. The probabilities of obtaining between 40 and 60 heads calculated using Eq. **??** are shown in Fig. **??**. The sum of all the probabilities shown in the figure is 0.96, which represents the coverage probability, P_c, associated with the interval from 40 to 60 successes.

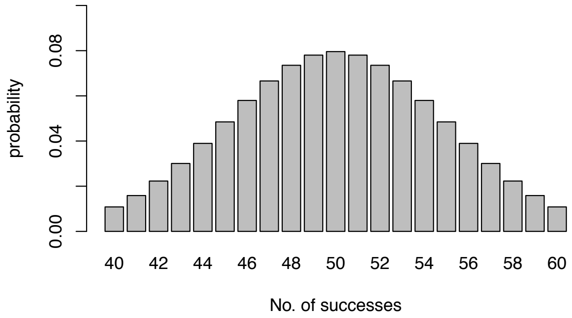

Figure 1. Binomial probability distribution for obtaining between 40 and 60 successes in 100 trials when the probability of success is 0.5. The probabilities for each outcome are calculated using Eq. **??**.

We can apply this approach to a bed surface grain size sample. Imagine that we are sampling a population of surface sediment sizes like that shown in Fig. **??**a, for which the true median grain size of the population (D_{50}) is known (the population shown is defined by 3411 measurements of bed surface b-axis ~~diameter of n particles is measured, some of which will be smaller than the true value of D_p for the population of grains on the bed,~~ diameters at randomly selected locations in the wetted channel of a laboratory experiment performed by the authors, and has a median surface size of 1.7 mm). We know that half of the surface grains are smaller than the D_{50}, so for each stone that we select, the probability of it being smaller than the D_{50} is 0.50. If we measure 100 stones and compare them to the D_{50}, then binomial sampling theory tells us that the probability of selecting exactly 50 stones that are less than D_{50} is just 0.08, but that the probability of selecting between 40 and ~~some of which will be larger.For repeated samples~~ 60 stones less than D_{50} is 0.96 (see Fig. **??**).

Figure **??**b shows a random sample of 100 stones taken from the population ~~, the number of measured stones that will be~~ shown in Fig. **??**a. Each circle represents a measured b-axis diameter, and all 100 measurements are plotted as a cumulative frequency distribution; the median surface size of the sample, d_{50}, is 1.5 mm. There are clear differences between the distribution of the sample and the underlying population, which is to be expected.

The first step in calculating a grain size confidence interval that is likely to contain the true median value of the population is to choose a confidence level; in this example, we set the confidence level to 0.96, corresponding to the coverage probability shown in Fig. **??**. As a result, the true value of the D_{50} will fall between the sample d_{40} and the sample d_{60} 96% of the time. This represents the *percentile confidence interval* (see Fig. **??**c), and it does not depend on the shape of the grain size distribution. For reference, a set of percentile confidence interval calculations are presented in Appendix B.

Once a confidence level has been chosen and the percentile confidence interval has been identified, a *grain size confidence interval* can be estimated by mapping the percentile confidence interval onto the sampled grain size distribution, as indicated

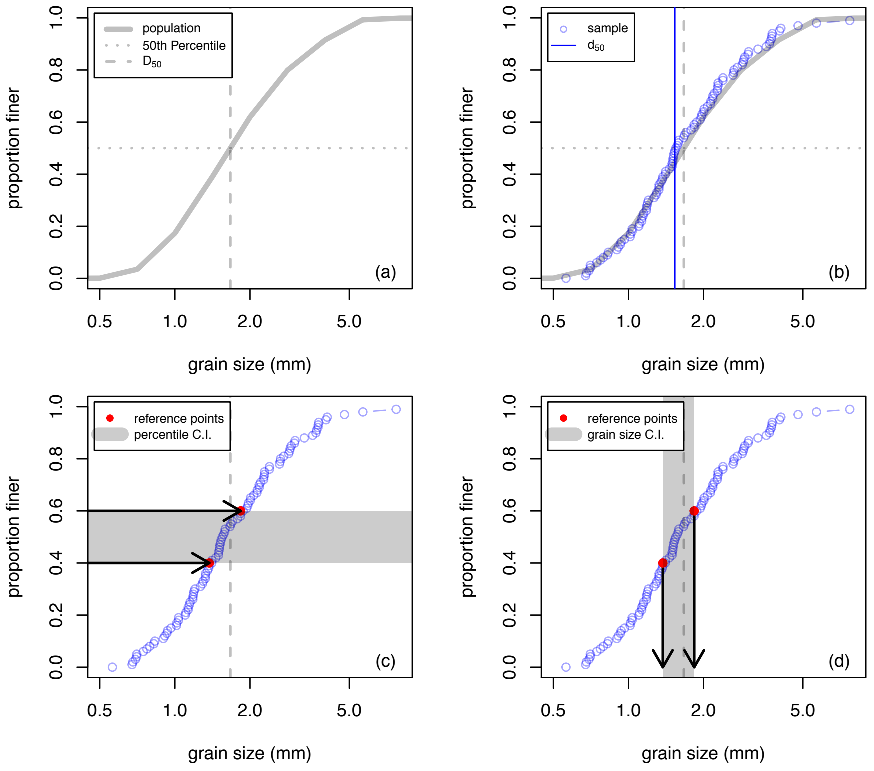

Figure 2. Defining the relation between the percentile confidence interval and the grain size confidence interval for a sampled d_{50} value. (a) Begin with the known distribution for the population being sampled, with a vertical line indicating the true D_{50}. (b) Derive a sample distribution from 100 measurements from the population shown in (a) (note that the sample d_{50} and the population D_{50} are different). (c) Use binomial theory to estimate the percentile confidence interval that contains the population D_{50}. (d) Map the percentile confidence interval onto the sample cumulative frequency distribution to estimate the grain size confidence interval around the sample estimate, d_{50} (note that the confidence interval does indeed contain the true D_{50} for the population).

graphically in Fig. **??**d. Unlike the percentile confidence interval, the grain size confidence interval depends on the shape of the cumulative frequency distribution, and can only be calculated once the sample has been collected.

The approach demonstrated above for the median size can be applied to all other grain size percentiles by varying the probability p in Eq. **??**, accordingly. For example, the probability of picking up a stone smaller than the true ~~value of D_p will vary about a mean value $n \cdot p$, just as the number of heads observed during n tosses of~~ D_{84} of a ~~fair coin will vary about a mean value of $0.5n$. The~~ population is 0.84, while the probability of picking up a stone smaller than the true D_{16} is just 0.16. If we define P to be the percentile of interest for the population being sampled, then the probability of selecting a stone smaller than that percentile is $p = P/100$, meaning that there is a direct correspondence between the grain size percentile and the probability of encountering a grain smaller than that percentile. As we show in the next section, the binomial distribution can be used to

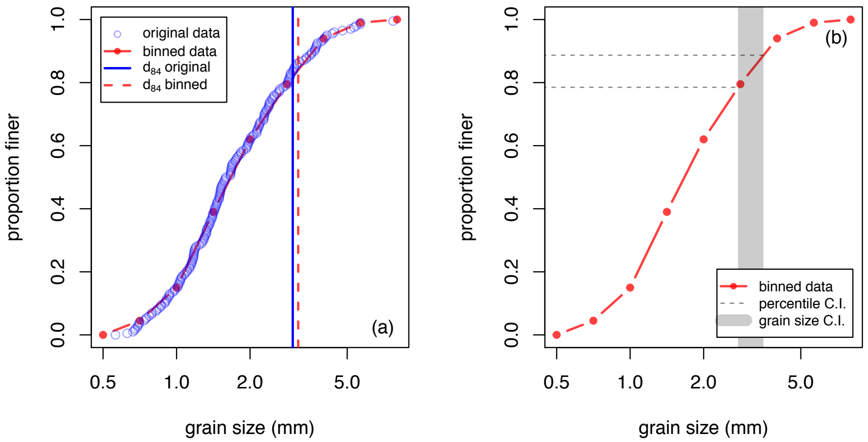

Figure 3. A grain size distribution from a stream table experiment based on a sample size of 200 observations. ~~Blue~~ In Panel (a), blue circles indicate individual grain size measurements ($d_{(i)}$), and the red line is the cumulative frequency distribution for binned data using the standard 0.5 ϕ bins. ~~Dashed lines indicate~~ In Panel (b), the ~~interpolation procedure for translating the estimated~~ interpolated upper and lower percentile confidence ~~intervals~~ bounds for the binned data are shown as ~~percentiles (i.e., the~~ horizontal lines~~) into~~ , and the ~~corresponding~~ associated 95% grain size ~~quantiles (i.e.,~~ confidence interval containing the ~~vertical lines) that bound the estimate of the~~ true D_{84} ~~(represented as black solid lines)~~for the population is shown in grey.

derive grain size confidence intervals for any estimate of ~~D_p~~ d_P for a sample that can be expected to contain the true value of ~~D_p~~ D_P for the entire population.

2.1.1 Statistical basis

In order to illustrate our approach for estimating confidence intervals in detail, we will use ~~grain size data from a recent laboratory experiment, comprising~~ a sample of 200 measurements of b-axis diameters ~~; since we preserve each measurement rather than grouping them into size classes, the data can be treated as a binomial experiment, analogous to flipping a coin, wherein each measurement represents the outcome of a single coin flip~~from our laboratory population of 3411 observations. These data are sorted in rank order and then used to compute the quantiles of the ~~(Fig. ??). The difference in granularity between the raw data~~ sample distribution. The difference between the cumulative distribution of raw data (based on 200 measurements of b-axis diameters) and the standard ~~binned data is illustrated on the figure by adding a cumulative frequency curve based on binned data using the standard 0.5ϕ size classes.~~

~~A variety of approaches has been proposed in the statistical literature for estimating quantiles from a sample (?). The differences among methods are greatest for smaller sample sizes, and decrease as~~ 0.5ϕ binned data (which is typical for most field samples) is illustrated in Figure **??**. While the calculated d_{84} value for the binned data shown in Fig. **??**a is not identical to that from the original data, the difference is small compared with the grain size confidence interval associated with a sample size of 200, shown in Fig. **??**b. We first develop a method to apply to samples comprising n ~~increases.The first step~~

~~in all approaches is to sort the measured values from lowest to highest and use these to define order statistics $d_{(i)}$ such that $d_{(1)} \le d_{(2)} \le ... \le d_{(n)}$, where, for example, $d_{(1)}$ is the minimum value of d_i~~individual measurements of grain diameter, and then describe an approximation that can be applied to the more commonly encountered 0.5ϕ binned cumulative grain size distributions.

2.2 ~~Exact solution for a confidence interval~~

2.1.1 Exact solution for a confidence interval

Suppose we wish to compute a ~~specific quantile, say D_p~~confidence interval containing the population percentile, D_P, from our sample of ~~sediment particles. The probabilities of drawing a specific number of particles, k, that are smaller than D_p (i. e., $d_{(k)} < D_p$ and $d_{(k+1)} > D_p$) can be computed from the binomial distribution :~~

$$\underline{Pr(k,n,p) = p^k(1-p)^{n-k}\frac{n!}{k!(n-k)!}}$$

200 b-axis diameter measurements. The first step is to generate order statistics, $d_{(i)}$, by sorting the measurements into rank order from lowest to highest (such that $d_{(1)} \le d_{(2)} \le ... \le d_{(n)}$). Figure ??a plots $d_{(i)}$ against the ratio $(i-1)/n$, which is a direct representation of the proportion of the distribution that is finer than that grain size.

To define a confidence interval, we first specify the confidence level, usually expressed as $100\cdot(1-\alpha)$%. For 95% confidence, $\alpha = 0.05$. Following ?, we then find lower and upper values of the order statistics ($d_{(l)}$ and $d_{(u)}$, respectively) that determine the percentile confidence interval, such that the coverage probability ~~(P_c)~~ is as close as possible to $1-\alpha$, but no smaller. Note that, in our example of 100 coin tosses from the previous section, we made a calculation by setting $l = 40$ and $u = 60$, which gave us a coverage probability of 96%. Coverage probability is defined as:

$$P_c = \sum_{k=0}^{u-1} B_k(\underline{u-1}k,n,p) - \sum_{k=0}^{l-1} B_k(\underline{l-1}k,n,p) \qquad (2)$$

where ~~$B(j,n,p)$ is the cumulative distribution function for j "successes "~~ B_k is the binomial probability distribution for k successes in n trials for probability p~~.~~, defined in Eq. ??. The goal, then, is to find integer values l and u that satisfy the condition that $P_c \ge 1-\alpha$, with the additional condition that l and u be approximately symmetric about the expected value of k (i.e., $n \cdot p$). The lower ~~and upper confidence limits are then given by~~ grain size confidence bound for the estimate of D_P is then mapped to grain size measurement $d_{(l)}$ and upper bound is mapped to $d_{(u)}$. Obviously, this approach cannot be applied to the binned data usually collected in the field, but is intended for the the increasingly common automated, image-base techniques that retain individual grain size measurements.

We have created an R function (`QuantBD`) that determines the upper and lower confidence ~~limits~~bounds, and returns the coverage probability, which is included in the ~~supplementary material for this paper~~`GSDtools` package. Our function is based on a script published online by W. Huber [1], which follows the approach described in ?. For $n = 200$, $p = 0.84$ and $\alpha = 0.05$

[1] https://stats.stackexchange.com/q/284970~~!~~, last accessed on 19 September, ~~2019~~2018

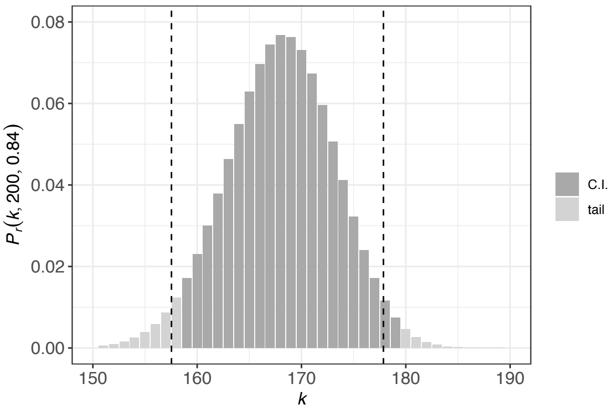

Figure 4. Binomial distribution values for $n = 200$ and $p = 0.84$, displaying the range of k values included in the coverage probability. The dark grey bars indicate which order statistics are included in the 95% confidence interval, and light grey indicates ~~order statistics~~ the tails of the distribution that lie outside the interval. The vertical dashed lines indicate confidence limits computed by an approximate approach that places equal area under the two tails outside the confidence interval.

(i.e., 95% confidence level), $l = 159$ and $u = 180$, with a coverage probability (0.953) that is only slightly greater than the desired value of 0.95. This implies that the number of particles in a sample ~~of200~~ of 200 measurements that would be smaller than the true D_{84} should range from 159 particles to 180 particles, ~~19 times out of 20.~~ 95% of the time. This in turn implies that the true D_{84} could correspond to sample estimates ranging from the 80^{th} percentile (i.e., $159/200$) to the 90^{th} percentile (i.e., $180/200$). We can translate the percentile confidence bounds into corresponding grain size ~~values~~ confidence bounds using our ranked grain size measurements: the lower bound of 159 corresponds to a measurement of ~~2.7~~ 2.8 mm, and the upper bound corresponds to a measurement of ~~3.7~~ 3.6 mm.

2.2 ~~Approximate solution for equal-area tails~~

2.1.1 Approximate solution for equal-area tails

One disadvantage of the exact solution described above is that the areas under the tails ~~differ, as evident from~~ of the binomial distribution differ (Fig. **??**), such that the expected value is not located in the center of the confidence interval. **?** described an alternative approach based on interpolation for finding lower or upper limits for one-sided intervals ~~.~~ (i.e., confidence intervals pertaining to a one-tailed hypothesis test). This approach can be applied to find two-sided intervals by finding one-sided intervals, each with a confidence level of $1-\alpha/2$~~.~~, which results in a confidence interval that is symmetric about the expected value (see the dashed lines in Fig. **??**). By interpolating between the integer values of k, we can find real numbers for which the binomial distribution has values of $\alpha/2$ and $1-\alpha/2$, which we refer to as l_e and u_e. The corresponding grain sizes can be found by interpolating between measured diameters whose ranked order brackets the real numbers l_e and u_e.

The values of l_e and u_e are indicated on Fig. **??** by dashed vertical lines. As can be seen, the values of l and u generated using the equal tail approximation are shifted to the left of those found by the exact approach~~. Consequently, the approximate confidence limits are also shifted to the left of the exact approach~~, resulting in a symmetrical confidence interval. The corresponding grain sizes representing the confidence interval are 2.7 mm and ~~3.6~~ 3.4 mm, which are similar to the exact solution presented above.

2.2 ~~Approximate solution for binned data~~

2.1.1 Approximate solution for binned data

We have adapted the approximate solution described above to allow estimation of confidence limits for binned data, which is accomplished by our R function called `WolmanCI` ~~. We~~ in the `GSDtools` package. Just as before, we use the equal area approximation of the binomial distribution to compute upper and lower limits ~~of k, and then~~ (l_e and u_e), but then we transform these ordinal values into percentiles by normalizing by the number of observations. Using our sample data, the ordinal confidence bounds $l_e = 157.03$ and $u_e = 177.36$ thus become the ~~percentiles 79% and 89%~~percentile confidence bounds d_{79} and d_{89}, respectively.

~~To estimate the confidence limit in terms of grain sizes~~Next, we simply interpolate from the ~~empirical~~ binned cumulative frequency distribution ~~based on the classed sediment diameters~~ to find the corresponding ~~quantiles~~grain sizes that define the grain size confidence interval. Note that the linear interpolation is applied to $log_2(d)$, and that the interpolated values are then transformed to diameters in mm.

This interpolation procedure is represented graphically ~~on~~ in Fig. **??**~~. The dashed~~ b; the horizontal lines represent ~~percentile values of~~ the percentile confidence interval (defined by l_e/n and u_e/n), while the ~~solid horizontal line represents the percentile of interest (i. e., $p = 0.84$).~~ grey box indicates the associated grain size confidence interval. Our binned sample data yield a grain size confidence interval for the D_{84} that ~~ranges from 2.7~~ range from 2.8 mm to 3.5 mm.

~~Clearly, the~~ The binomial probability approach ~~requires that the sample distribution be known in order~~ uses the sample cumulative frequency distribution to calculate the ~~confidence intervals in units of length. While this is problematic when attempting~~ grain size confidence interval. This makes it difficult to predict the statistical power ~~associated with a given~~ of sample size, n, ~~before actually~~ prior to collecting the sample~~, it is possible to use~~. However, the approach can be applied to any previously collected distribution~~to calculate and plot confidence intervals of the bed surface grain sizes~~, provided the number of observations used to generate the distribution is known. ~~The approach can also be used to estimate the confidence intervals about any previously published grain size distribution, and~~

3 Two-sample hypothesis tests

3.1 When individual grain diameters are available

Suppose we have two samples for which individual grain diameters have been measured (e.g., two sites, two operators, two sampling methods). The values in the two samples are denoted as X_i, (where i ranges from 1 to ~~assess whether or not a given set of distributions is statistically different or not~~ n_x) and Y_j ($j = 1$ to n_y) where n_x and n_y are the number of grains in each sample. In this case, one can use a resampling method (specifically the bootstrap) to develop a hypothesis test. A straightforward approach is based on the percentile bootstrap (**?**), and involves the following steps:

1. Take a random sample of n_x diameters, with replacement, from the set of values of X_i. This bootstrap sample is denoted as x_k, $k = 1$ to n_x.

4 ~~Confidence interval testing~~

~~The approximate method presented in the preceding section can easily be tested numerically by sub-sampling a large population of observations, determining the distribution of resulting percentile size estimates produced by the sub-samples,~~

2. Take a random sample of n_y diameters, with replacement, from the set of values of Y_j. This bootstrap sample is denoted as y_l, $l = 1$ to n_y.

3. Determine the desired percentile value from each sample, $(d_P)_x$ and ~~comparing it to the confidence interval based on binomial theory.We have eight samples of about 400 observations each from a stream table experiment. Based on the overlap in confidence intervals for the eight samples~~ $(d_P)_y$, and compute the difference: $\Delta d_P = (d_P)_x - (d_P)_y$.

4. Repeat steps 1 to 3 n_r times (e.g., n_r = 1000), each time storing the value of Δd_P.

5. Determine a confidence interval for Δd_P by computing the quantiles corresponding to $\alpha/2$ and $1 - \alpha/2$, where α is the desired significance level for the test (e.g., α = 0.05).

6. If the confidence interval determined in step 5 does not overlap 0, then one can reject the null hypothesis that the sampled populations have the same value of D_P.

This analysis is implemented with the function `CompareRAWs` in the `GSDtools` package. The required inputs are two vectors listing the measured b axis diameters for each sample.

3.1 When only binned data are available and sample size is known

For situations in which only the cumulative frequency distribution is available, an approach similar to parametric bootstrapping can be applied, which employs the inverse transform approach (see Chapter 7 in **?**) to convert a set of random uniform numbers in the interval (0, 1) to a random sample of grain diameters by interpolating from the binned cumulative frequency distribution, similar to the procedure described above for determining confidence intervals for binned data.

The approach involves the following steps:

1. Generate a set of n_x uniform random numbers, u_i, ~~the distributions do not appear to be statistically different (see Fig. ??). Therefore, the data have been pooled to form a single data set of~~ $i = 1$ to n_x. Transform these into a corresponding set of grain diameters x_i by using the cumulative frequency distribution for one sample.

2. Generate a set of n_y uniform random numbers, u_j, $j = 1$ to n_y. Transform these into a corresponding set of grain diameters y_j by using the cumulative frequency distribution for the second sample.

3. Determine the desired grain size percentile from each sample, $(d_P)_x$ and $(d_P)_y$, and compute the difference: $\Delta d_P = (d_P)_x - (d_P)_y$.

4. Repeat steps 1 to 3 n_r times (e.g., $n_r = 1000$), each time storing the value of Δd_P.

5. Determine a confidence interval for Δd_P by computing the quantiles corresponding to $\alpha/2$ and $1-\alpha/2$, where α is the desired significance level for the test (e.g., $\alpha = 0.05$).

6. If the confidence interval determined in step 5 does not overlap 0, then one can reject the null hypothesis that the sampled populations have the same value of D_P.

This analysis is implemented in the `CompareCFDs` function. It requires that the user provide the cumulative frequency distribution for each sample (as a data frame), as well as the number of measurement upon which each distribution is based.

4 Confidence interval testing

We can test whether or not our approach successfully predicts the uncertainty associated with a given sample size using our known population of 3411 ~~observations. For the purposes of our uncertainty analysis, we let these 3411 observations define the population of interest and then take repeated, random sub-samples~~ measurements from the lab. The effect of sample size on the spread of the data is demonstrated graphically in Fig. **??**. In Fig. **??**a, 25 random samples of 100 stones selected from the population are plotted, along the with 95% grain size confidence interval bracketing the true grain size population, calculated using our binomial approach. In Fig. **??**b, random samples of 400 stones are plotted, along the with corresponding confidence interval. A comparison of the two plots shows that sample size (i.e. 100 vs. 400 stones) has a strong effect on variability of the sampled distributions. It is also clear that the variability of the samples is well predicted by the binomial approach, since the sample data generally fall within the confidence interval for the population.

In order to more formally test the binomial approach, we collected 10,000 random samples (with replacement) ~~of 100 observationsfrom the larger data set. For each sub-sample, we generate the cumulative frequency distribution and then estimate the bed surface D_{16}, D_{50}, and D_{84}.~~ from our population of 3411 observations, calculated sample percentiles ranging from the d_5 to the d_{95} for each sample, and used the distribution of estimates to determine the grain size confidence interval. This resampling analysis was conducted twice; once for samples of 100 stones and then again for samples of 400 stones. This empirical approximation of the grain size confidence interval is the same technique used by **?**. The advantage of a resampling approach is that it replicates the act of sampling, and therefore does not introduce any additional assumptions

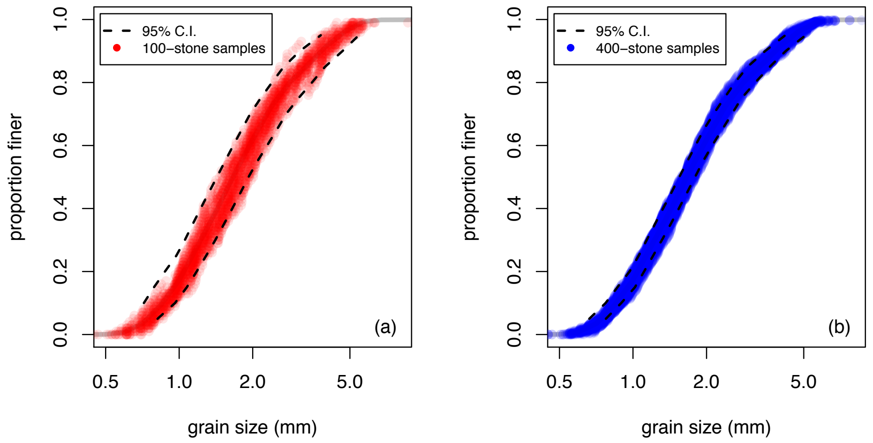

Figure 5. Effect of sample size on uncertainty. In Panel (a), 25 samples of 100 stones drawn from a known population are plotted, along with the 95% grain size confidence interval calculated for D_5 to D_{95} using the binomial method. In Panel (b) samples of 400 stones are plotted, along with the predicted grain size confidence interval.

or approximations. The accuracy of the resampling approach is limited only by the number of samples collected, and the degree to which the individual estimates of a given percentile reproduce the distribution that would be produced by an infinite number of samples. The only draw back of this approach is that the results are only strictly applicable to the population to which the resampling analysis has been applied (**?**). While it is an ideal way to assess the effect of sample size on variability for a known population, resampling confidence intervals cannot be calculated for individual samples drawn from an unknown grain size population.

~~The box-plots represent the distribution of estimates for the D_{16}, D_{50}, and D_{84} of the same bed surface, based on repeatedly selecting 100 measurements from the larger population of observations. The 99% confidence interval estimated using binomial theory is shown in red, the 50% confidence interval is shown in blue, and the ‘true’ percentile for the population is shown in black, for comparison.~~ In Fig. **??**, the resampling estimates of the 95% grain size confidence intervals for D_5 to D_{95} based on samples of 100 stones are plotted as red circles, and those based on samples of 400 stones are plotted as blue circles. For comparison, the confidence intervals predicted using our binomial approach are plotted using dashed lines. There is a close agreement between the resampling confidence intervals and the binomial confidence intervals, indicating that our implementation of binomial sampling theory captures the effects of sample size that we have numerically simulated using the resampling approach.

~~As seen in Fig. **??**, the spread of the estimates from the repeated sub-sampling of the data set is generally similar to the confidence intervals based on binomial theory; the predicted confidence interval containing 50% of the observations (shown in blue) corresponds approximately to the~~ We have also calculated the statistics of a 1:1 linear fit between the upper and lower bounds of the confidence intervals predicted by binomial theory and those calculated using the resampling approach for sample sizes of 100 and 400. For a sample size of 100 stones, the 1:1 fit had a Nash Sutcliffe model efficiency (NSE) of 0.998, a root

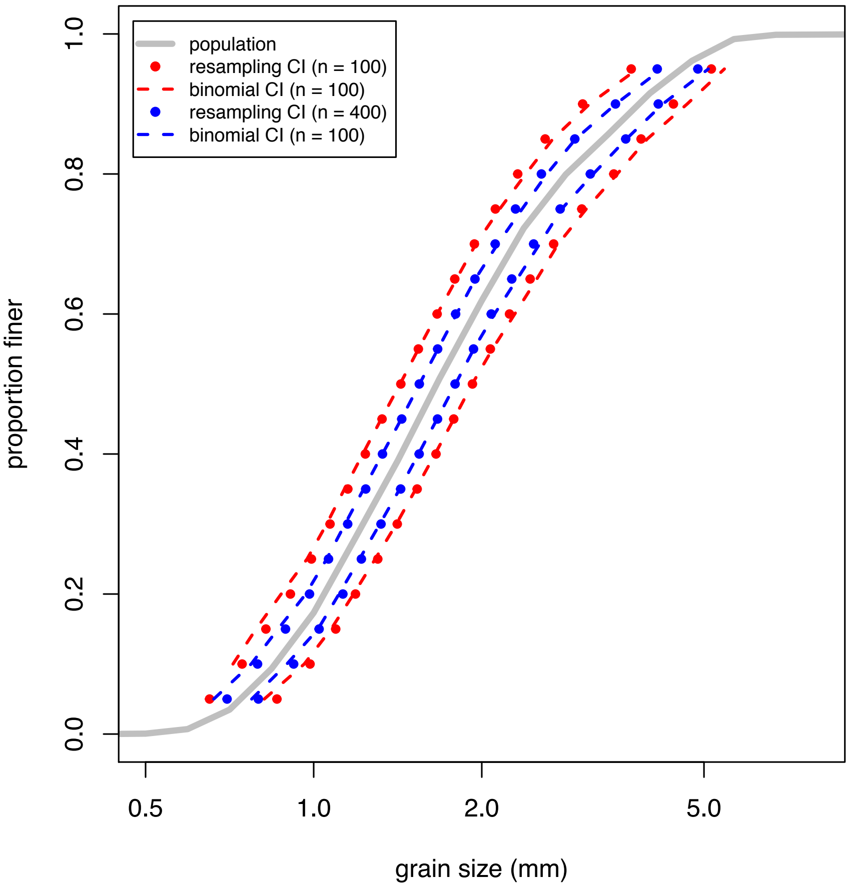

Figure 6. ~~All~~ Comparing calculated resampling grain size ~~distributions from a stream table experiment based on a sample size of about 400 observations~~confidence intervals to predicted intervals using the binomial approach. The ~~estimated~~ grain ~~sizes~~ size confidence intervals for samples of 100 stones are shown in red, ~~along with the 95% confidence intervals~~and those for samples of 400 stones are shown in blue.

mean standard error ($RMSE$) of 0.0353ϕ units, and a mean bias (MB) of -0.0035ϕ units. Since $NSE = 1$ indicates perfect model agreement (see **?**), and considering that MB is small relative to the $RMSE$, these fit parameters indicate a good 1:1 agreement between the resampling estimates and binomial predictions of the upper and lower ~~quartiles of the box plots, and the 95% confidence interval corresponds approximately to the overall spread of the numerical estimates. A more direct comparison shows that the calculated 50% confidence intervals contain 54% of the grain size estimates from the sub-samples,while the 95% confidence intervals contain 97% of the estimates.~~ confidence interval bounds. The results for a sample size of 400 stones were essentially the same ($NSE = 0.999$, $RMSE = 0.0262\phi$, and $MB = 7e-04\phi$).

In order to confirm that the size of the original population did not affect our comparison of the resampling and binomial confidence bounds estimates, we repeated the entire analysis using a simulated log-normal grain size distribution of 1,000,000 measurements. The graphical comparison of the binomial and resampling confidence intervals for the simulated distributions (not shown) was essentially the same as that shown in Fig. **??**, and the 1:1 model fit was similar to the fits reported above ($NSE = 0.998$, $RMSE = 0.043\phi$, and $MB = -0.0013\phi$).

The close match between the ~~confidence intervals calculated from~~ grain size confidence intervals predicted using binomial theory and ~~the distribution of percentiles based on sub-sampling~~ those estimated using the resampling analysis supports the validity of the proposed approach for computing confidence ~~limits about the cumulative grain size distribution. Since these confidence limits are straightforward to calculate, we argue that it should be standard practice to plot them on all grain size distribution graphs, particularly those that purport to show a difference between two distributions.~~ intervals.

5 Reassessing previous analyses

In order to demonstrate the importance of understanding the uncertainty, we have reanalyzed the results of ~~several~~ previous papers that have compared bed surface texture distributions, but which have not considered uncertainty associated with sampling variability. In ~~most~~ some cases, these re-analyses confirm the authors' interpretations, and strengthen them by highlighting which parts of the distributions are different and which are similar, thus allowing for a more nuanced understanding. In ~~some cases, however, the re-analyses~~ others, they demonstrate that the observed differences do not appear to be statistically significant, and suggest that the interpretations and explanations of those differences are not supported ~~but~~ by the authors' data. In either case, we believe that adding information about the grain size confidence intervals is a valuable step that should be included in every surface grain size distribution analysis.

The data published by **?** include pebble counts of about 400 stones for different channel units in two mountain streams (see Fig.~~-~~ **??**). Adding the ~~confidence bands~~ grain size confidence intervals to the distributions emphasizes the ~~advantages of taking larger sample sizes, since the confidence bands are narrower than those for a sample of only 100 stones (e. g., Fig. ??). It also emphasizes that the key difference for the bed texture in poolsand in runsor riffles is~~ differences and similarities between the distributions. Based on the data in Fig. **??**, it seems that clear differences in bed texture exist when comparing pools, runs, and riffles for the fraction of sediment less than about 22.6 mm; the distributions of sediment coarser than this are ~~not statistically different for either stream.This observation~~ quite similar. Using the `CompareCFDs` function to compare percentiles ranging from D_5 to D_{95} (in increments of 5), we found that the differences in the samples from Willow Creek for percentiles greater than D_{65} are significant for $\alpha = 0.05$, but not for $\alpha = 0.01$ (i.e., for a 99% confidence interval). For North St. Vrain Creek, there are significant differences at $\alpha = 0.05$ for percentiles finer than D_{20}, and for the D_{80} and D_{85}, though none of the differences for the coarser part of the distribution are significant for $\alpha = 0.01$.

The relative similarity of pool and run/riffle sediment textures for the coarser part of the distribution suggests that the most noticeable differences in bed surface texture are likely due to the deposition of finer bed-load sediment in pools on the waning limb of the previous flood hydrograph (as suggested by **???**), and that the bed surface texture of both kinds of mainstem units during flood events could generally be quite similar. The analysis also clearly demonstrates that size distributions of the exposed channel bars in these two streams are statistically different from both the pools and the runs/riffles. From these plots we can conclude that the bed roughness (which is typically indexed by the bed surface D_{50} or by sediment coarser than that) is similar

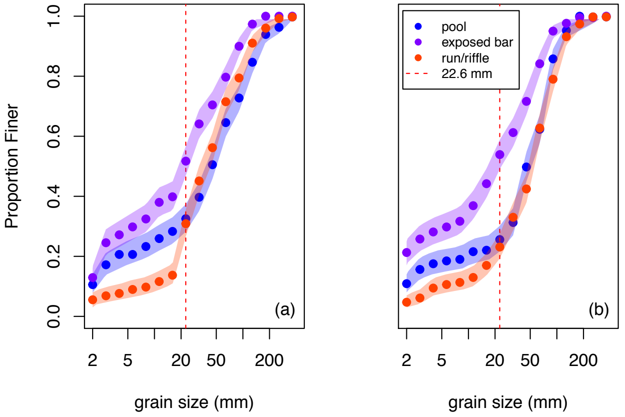

Figure 7. Comparing ~~the bulk surface sample and~~ pebble ~~count distributions, published by Kondolf (1997, their Fig~~counts from different channel units. ~~3).~~ Panel A ~~shows the traditional grain size distribution representation~~presents data reported by **?** for Willow Creek. Panel B ~~uses~~ presents data for North St. Vrain Creek. Shaded polygons represent the 95% confidence ~~band calculated for~~ intervals about the ~~pebble count to highlight where the distributions are statistically similar and where they are different~~sample distribution.

~~Figure **??** plots data published by **?**, which were used to compare the bed surface grain size distribution estimated using a pebble count method, and from a truncated bulk sample of the bed surface. Re-plotting the analysis by **?** demonstrates that the coarse tail (i.e., $D_i > 22.6$ mm) of their bulk sample of the bed surface is statistically similar to the coarse end of the distribution for a pebble count, once the sediment finer than 4 mm is excluded from the analysis of the bulk sediment. Interestingly, the finer half of the two distributions appear to be statistically different. While **?** reached essentially the same conclusion, the use of confidence bands about the distributions highlights the statistical similarity of the coarse tail, and can be used to suggest that the transition occurs at a grain size of about 22.6 mm. Comparing pebble counts from different channel units. Panel A presents data reported by **?** for Willow Creek. Panel B presents data for North St. Vrain Creek.~~

for the mainstem units (i.e., pools, and runs/riffles), but that exposed bar surfaces in these two streams are systematically less rough. These kinds of inferences could have important implications for decisions about the spatial resolution of roughness estimates required to build 2D or 3D flow models; it is also possible to reach the same conclusions based on the original data plots in **?**, but the addition of confidence bands supports the robustness of the inference.

A more fundamental motivation for plotting the binomial confidence bands is illustrated in Fig. **??**, which compares the bed surface texture estimated by two different operators using the standard heel-to-toe technique to sample more than 400 stones from the same sedimentological unit. These data were published by **?** (see their Fig. 7). Based on their original representation of the two distributions (Fig. **??**~~, Panel A~~a), **?** concluded that

> "operators produced quite different sampling results . . . operator B sampled more fine particles and fewer cobbles . . . than operator A and produced thus a generally finer distribution."

However, once the ~~confidence bands~~ grain size confidence intervals are plotted (Fig.**??**~~, Panel B~~b), it is clear that the differences ~~do not appear~~ are not generally statistically significant. Using the `CompareCFDs` function to compare each percentile from

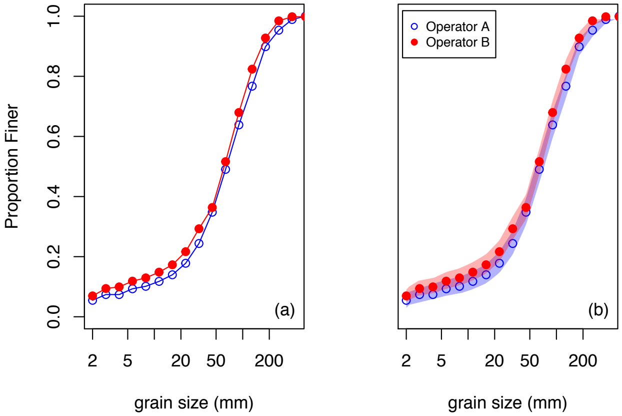

Figure 8. Comparing pebble counts of the same bed surface by different operators. The data plotted were published by **?**. Panel A shows the traditional grain size distribution representation. Panel B uses the 95% grain size confidence ~~band~~ intervals calculated for the pebble count to demonstrate that the two distributions are not statistically different.

D_5 to D_{95}, we found no statistically significant differences for any percentile at $\alpha = 0.01$; at $\alpha = 0.05$, only differences for the D_{80}, D_{85} and D_{95} are significant.When comparing distributions, it is common practice to apply the Bonferroni correction in which α is replaced by α/m, where m is the number of metrics being compared. Applying this correction, there is no statistical difference between the two samples for $\alpha = 0.05$. The value in considering sampling variability in the analysis is that it supports a more nuanced interpretation of differences in grain size distributions.

A similar analysis of the heel-to-toe sampling method and the sampling frame method advocated by **?** shows that the distributions produced by the two methods are not generally statistically different, either (Fig. **??**). The `CompareCFDs` function only found significant differences for grain size percentiles coarser than D_{70} for $\alpha = 0.05$, and between D_{75} and D_{90} for $\alpha = 0.01$. Once the Bonferroni correction is applied, none of the differences between the two samples would be considered significant at $\alpha = 0.05$.

In both cases, the uncertainty associated with sampling variability appears to be greater than ~~any~~ the difference between operators or between sampling methods, and thus one cannot claim these differences as evidence for statistically significant effects. It ~~may be~~ is likely the case that there are significant differences among operators or between sampling methods, but larger sample sizes would be required to reduce the magnitude of sampling variability in order to identify those differences.

Indeed, **?** found that operator errors were difficult to detect for small sample sizes (wherein the sampling uncertainties were comparatively large), but became evident as sample size increased, so the issue at hand is not whether there are important differences between operators, but whether the differences in Fig. **??** are statistically significant. Interestingly, **?** were able to detect operator differences at sample sizes of about 300 stones, whereas **?** did not detect statistical differences for samples of about 400 stones, indicating either that **?** had larger operator differences than did **?**, or smaller sample uncertainties due to the nature of the sediment size distribution.

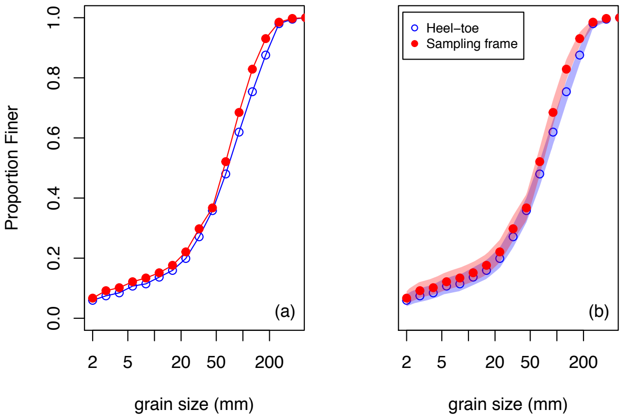

Figure 9. Comparing sampling methods for the same bed surface and operator. The data plotted were published by **?**, and were collected by operator B. Panel A shows the traditional grain size distribution representation. Panel B uses the 95% grain size confidence ~~band~~ intervals calculated for the pebble count to demonstrate that the two distributions do not appear to be statistically different.

6 Determining sample size

~~Our method for estimating uncertainty requires only the cumulative distribution and the number of measurements used to construct the distribution. Therefore,~~ As we demonstrated in the previous section, grain size confidence intervals can be constructed and plotted for virtually all existing surface grain size distributions (provided that the number of stones that were measured is known, which is almost always the case), and future sampling efforts need not be modified in any way in order to take advantage of our method. While the primary purpose of our paper is to demonstrate the importance of calculating grain size confidence intervals when analyzing grain size data, our method can also be adapted to predict the sample size required to achieve a desired level of sampling precision, prior to collecting the sample.

~~The actual uncertainty of an estimated grain size percentile cannot be predicted using our method~~ While the percentile confidence interval for any percentile of interest can be calculated based on the sample size, n, and the desired confidence level, α (see Appendix B, for example), it cannot be mapped onto the grain size confidence interval before the cumulative distribution has been generated. This problem is well recognized, and has been approached in the past by making various assumptions about the distribution shape (**????**), or using ~~computational approaches~~ empirical approximations (**????**), but in all cases it is still necessary to know something about the spread of the distribution – regardless of its assumed shape – in order to ~~predict the level of uncertainty associated with a given sample size~~ assess the implications of sample size for the precision of the resulting grain size estimates. It is perhaps the difficulty of predicting sample ~~uncertainty~~ precision that has led to the persistent use of the standard 100-stone sample. Here we provide a simple means of determining the appropriate sample size; first we use existing data to calculate the uncertainty of estimates for d_{50} and d_{84}; and then we use simulated log-normal grain size distributions to quantify the effect of the spread of the distribution on uncertainty.

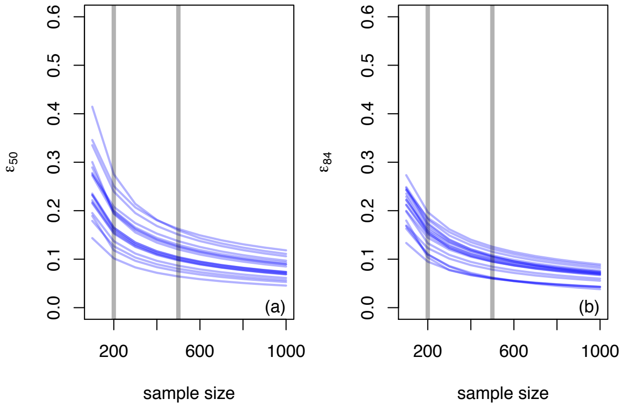

Figure 10. Estimated uncertainty for estimates of D_{50} (Panel A) and D_{84} (Panel B) are plotted against sample size. Curves were generated for ~~all the~~ bed surface samples ~~analysed in this paper (???), and for bed surface samples~~ collected by BGC Engineering and students from The University of British Columbia (unpublished data), and those published by **???**. Vertical lines highlight the range of uncertainties for sample sizes of 200 and 500 stones.

6.1 Uncertainty based on field data

Here, we demonstrate the effect of sample size on uncertainty. We begin by calculating the uncertainty of estimates for D_{50} and D_{84} for all the surface samples used in this paper, for eight samples collected by BGC Engineering from gravel bed channels in the Canadian Rocky Mountains, and for samples from two locations on Cheakamus River, British Columbia, collected by undergraduate students from the Department of Geography at The University of British Columbia. The number of stones actually measured to create these distributions is irrelevant, since it is the shape of the cumulative distribution that determines how the known percentile confidence interval maps onto the grain size confidence interval. Since these distributions come from a wide range of environments and have a range of distribution shapes, they are reasonable representation of the range grain size confidence intervals that could be associated with a given percentile confidence interval.

Uncertainty (ϵ) ~~is expressed as a proportion of the estimate,~~ in the grain size estimate is calculated as follows:

$$\epsilon_{P} = 0.5\left(\frac{D_{upper}-D_{lower}}{D_{est}}\frac{d_{upper}-d_{lower}}{d_{P}}\right) \tag{3}$$

where ~~D_{upper}~~ d_{upper} is the upper ~~95% confidence bound calculated for a given sample size , D_{lower}~~ bound of the grain size confidence interval, d_{lower} is the lower ~~confidence~~ bound, and ~~D_{est}~~ d_{P} is the estimated ~~size for~~ grain size of the percentile of interest. ~~For the sake of simplicity, we have assumed that uncertainty is symmetrically distributed about D_{est}, but this is not true for all distribution shapes. Therefore, we can be approximately 95% confident that the interval $D_{est}[1 \pm \epsilon]$ includes the true value of the percentile~~As a result, ϵ_{50} represents the half-width of the grain size confidence interval about the median grain size (normalized by d_{50}), and ϵ_{84} represents half-width of the normalized grain size confidence interval for the d_{84}.

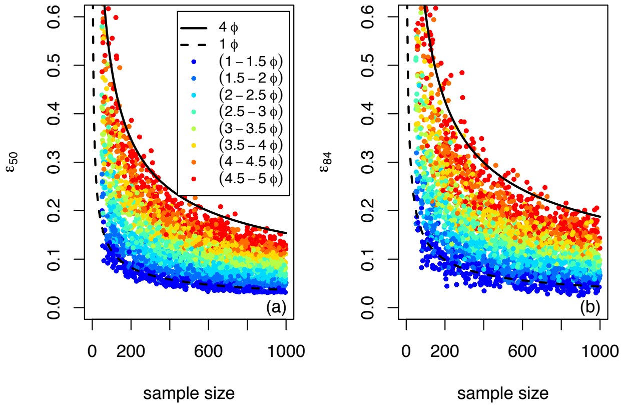

Figure 11. Estimated uncertainty for estimates of D_{50} (Panel A) and D_{84} (Panel B) are plotted against sample size for a simulated set of log normal surface distributions with a range of ~~standard deviations~~sorting indices. The markers are color-coded by ~~standard deviation~~si_{ϕ}. The bounding curves for ~~$SD_{log} = 0.5\ \phi$~~ $si_{\phi} = 1$ and ~~$SD_{log} = 2.0\ \phi$~~ $si_{\phi} = 5.0$ are shown for reference, calculated using Eq. (**??**) and Eq. (**??**).

Fig. **??** presents the ~~range of uncertainties~~ calculated values of ϵ_{50} and ϵ_{84} for various gravel bed surface samples, including those shown in Figs. (~~**??**), (**??**), (~~**??**), and (**??**). For a sample size of 100 stones, the uncertainties are relatively large, with a mean ~~uncertainty across all of the distributions of ±25% for D_{50} and of ±21% for D_{84}. The mean uncertainty drops to ±18% for D_{50} and ±15% for D_{84} for~~ ϵ_{50} value of 0.25 and a mean ϵ_{84} value of 0.21; for a sample of 200 stones, ~~and to ±11% (D_{50}) and ±9% (D_{84})~~ϵ_{50} drops to 0.18, and ϵ_{84} drops to 0.15, on average; and for ~~500 stones~~$n = 500$, $\epsilon_{50} = 0.11$, and $\epsilon_{84} = 0.09$. This analysis transforms the predictable, distribution-free contraction of the percentile confidence interval as sample size increases into the distribution-dependent contraction of the grain size confidence interval. Clearly there is a wide range cumulative frequency distribution shapes in our data set, resulting in a large differences in ϵ_{50} and ϵ_{84} for the same sample size (and therefore the same percentile confidence interval).

6.2 Uncertainty for Log-normal distributions

~~We can also approach this problem by assuming that~~ In order to quantify the effect of distribution shape on the grain size confidence interval, we conducted a modelling analysis using simulated log-normal bed surface texture distributions ~~are approximately log-normal, but have varying degrees of gradation, indicated by a standard deviation expressed~~ that have a range of sorting index values. Here, sorting index (si_{ϕ}) is defined by the following equation.

$$si_{\phi} = \phi_{84} - \phi_{16} \tag{4}$$

The term ϕ_{84} refers to the 84th percentile grain size (in ϕ units~~(SD_{log})~~), and ϕ_{16} refers the 16th percentile. As a point of comparison, ~~if we estimate the SD_{log}~~ we estimated si_{ϕ} for the samples analyzed in the previous section~~by assuming that $SD_{log} = \log_2 D_{84} - \log_2 D_{50}$, then SD_{log} ranges from 0.8 to 1.8~~. For those samples, the sorting index ranges from 1.5ϕ to 5.6ϕ, with a median value of ~~1. For those samples, the~~ 2.5ϕ. The largest values of ~~SD_{log}~~ si_{ϕ} were associated with samples from channels on steep gravel bed fans and on bar top surfaces, while samples characterizing the bed of typical gravel bed streams had values close to the median value.

We ~~generated a relation between uncertainty and sample size by first simulating~~ simulated 3000 log-normal grain size distributions with D_{50} ranging from 22.6 mm to 90.5 mm, n ranging from ~~51 to 999~~ 50 to 1000 stones, and ~~SD_{log} ranging from 0.5~~ si_{ϕ} ranging from 1ϕ to ~~2~~ 5ϕ. ~~We then used~~ For each simulated sample, we calculated uncertainty for D_{50} and D_{84} using Eq. **??**. The calculated values of ϵ_{50} and ϵ_{84} are plotted in Fig. **??**. Using the data shown in the figure, we fit least-squares regression to fit models of the form

$$\ln(\epsilon_{P}) = a \cdot n + b \cdot \underline{SD_{log}} si_{\phi} + c \tag{5}$$

where a, b, and c are the estimated coefficients. The empirical model ~~describing the uncertainty of D_{50}~~ predicting ϵ_{50} has an adjusted R^2 value of 0.95, with the variable n explaining about ~~47~~43% of the total variance, and ~~SD_{log} explaining 47~~si_{ϕ} explaining 51% of the variance. The model for ~~D_{84}~~ ϵ_{84} has an adjusted R^2 value of ~~0.9~~ 0.91 with the variables n and ~~SD_{log} explaining the similar amounts~~ si_{ϕ} explaining similar proportions of the total variance ~~(46% and 45~~as they do in the ϵ_{50} model (41% and 50%, respectively).

After back-transforming from logarithms, the equation describing the ~~uncertainty in D_{50}~~ ϵ_{50} can be expressed as:

$$\epsilon_{50} = A \cdot n^{\underline{-0.498}-0.506} \tag{6}$$

where the coefficient A is given by:

$$A = \exp(\underline{-0.346}-0.171 + \underline{0.832 SD_{log}} 0.359 si_{\phi}) \tag{7}$$

The ~~equation for estimating uncertainty in D_{84}~~ equations for ϵ_{84} are:

$$\epsilon_{84} = B \cdot n^{-0.51} \tag{8}$$

where B is given by:

$$B = \exp(\underline{-0.1}0.021 + \underline{0.842 SD_{log}} 0.366 si_{\phi}) \tag{9}$$

Table **??** provides values of A and B for a range of ~~standard deviations.~~

sorting indices.

7 Practical implications of uncertainty

The implications of uncertainty can be important in a range of practical applications. ~~Here~~As an example, we translate ~~uncertainty in grain size percentiles into uncertainty in~~ grain size confidence intervals into confidence intervals for the critical discharge for significant morphologic change using data for Fishtrap Creek, a gravel bed stream in British Columbia that has been studied by the authors (**???**). The estimated bed surface D_{50} for Fishtrap Creek is about 55 mm, which we

Table 1. Coefficient values for estimating uncertainty in D_{50} and D_{84} as a function of ~~SD_{log} and sample size (n)~~ si_{ϕ} using Eqs. (**??**) and (**??**)

$Coef.$	~~0.75ϕ~~ 1.5ϕ	~~1.00ϕ~~ 2.00ϕ	~~1.25ϕ~~ 2.5ϕ	~~1.50ϕ~~ 3.0ϕ	~~1.75ϕ~~ 3.5ϕ	~~2.00ϕ~~ 4.00ϕ	4.50ϕ
A	~~0.278~~ 1.444	~~0.486~~ 1.728	~~0.694~~ 2.068	~~0.901~~ 2.474	~~1.109~~ 2.961	~~1.317~~ 3.543	4.240
B	~~0.531~~ 1.768	~~0.742~~ 2.123	~~0.952~~ 2.550	~~1.163~~ 3.062	~~1.374~~ 3.677	~~1.584~~ 4.415	5.302

estimate becomes entrained at a shear stress of 40 Pa, corresponding to a discharge of about 2.5 m^{3}s^{-1}(**?**); the threshold discharge is based on visual observation of tracer stone movement, and corresponds to a critical dimensionless shear stress of approximately 0.045. If we assume that significant channel change can be expected when D_{50} becomes fully mobile ~~(which occurs at about twice the entrainment threshold)~~(which occurs at about twice the entrainment threshold, according to **?**), then we would expect channel change to occur at a shear stress of 80 Pa, which corresponds to a critical discharge of 8.3 m^{3}s^{-1}, based on the stage-discharge relations published by **?**.

Since we used the standard technique of sampling 100 stones to estimate D_{50} and since the ~~standard deviation~~ sorting index of the bed surface ~~distribution is about 1.0ϕ~~is about 2.0ϕ, we can assume that the uncertainty will be about ~~±16~~±17%, based on Eqs. ~~(~~**??** and **??**~~)~~, which in turn suggests that we can expect the actual surface D_{50} to be as small as 46 mm or as large as 64 mm. This range of D_{50} values translates to shear stresses that produce full mobility that range from 67 Pa to ~~93~~ 94 Pa. This in turn translates to critical discharge values for morphologic change ranging from 5.9 m^{3}s^{-1}to ~~11.1~~ 11.2 m^{3}s^{-1}, which correspond to return periods of about 1.5 years and ~~7.2~~ 7.4 years, based on the flood frequency analysis presented in **?**. Specifying a critical discharge for morphologic change that lies somewhere between a flood that occurs virtually every year and one that occurs about once a decade, on average, is of little practical use, and highlights the cost of relatively imprecise sampling techniques.

If we had taken a sample of 500 stones, we could assert that the true value of D_{50} would likely fall between 51 mm and 59 mm, assuming an uncertainty of ±7%. The estimates of the critical discharge would range from 7.2 m^{3}s^{-1} to 9.5 m^{3}s^{-1}, which in turn correspond to return periods of 2 years and 4.1 years, respectively. This constrains the problem more tightly, and is of much more practical use for managing the potential geohazards associated with channel change.

Operationally, it takes about 20 minutes for a crew of two or three people to sample 100 stones from a typical dry bar in a gravel bed river, and a bit over an hour to sample 500 stones, so the effort required to sample the larger number of stones is often far from prohibitive. In less ideal conditions or when working alone, it may take upwards of 5 hours to collect a 500 stone sample, but as we have demonstrated, the uncertainty of the data increases quickly as sample size declines (see Figs. **??** and **??**), which may make the extra effort worthwhile in many situations. Furthermore, computer-based analyses using photographs of the channel bed may be able to identify virtually all of the particles on the bed surface, and generate even larger samples. The statistical ~~advantage~~ advantages of the potential increase in sample size are obvious, and justify further concerted development of these computer-based methods, in our opinion.

8 Conclusions

Based on the statistical approach presented in this paper, we developed a suite of functions in the R language that can be used to ~~estimate the uncertainty of any percentile in a cumulative grain size distribution~~ first calculate the percentile confidence interval and then translate that into the grain size confidence interval for typical pebble count samples (see the supplemental material for the source code). ~~The approach~~ We also provide a spreadsheet which uses the normal approximation to the binomial distribution to estimate the grain size confidence interval. The approach presented in this paper uses binomial theory to ~~generate uncertainty estimates for any~~ calculate the percentile confidence interval for any percentile of interest (e.g. $P = 50$ or $P = 84$), and then maps that confidence interval onto the cumulative grain size distribution based on pebble count data ~~, and~~ to estimate the grain size confidence interval. As a result, the approach requires only that the total number of stones used to generate the distribution is known ~~. Approaches were developed for cases~~ in order to generate grain size distribution plots that indicate visually the precision of the sample distribution (e.g. Fig. **??**). We have developed statistical approaches that can be used for samples in which individual grain sizes are known and for samples in which data are binned (e.g., into ϕ classes).

By estimating the ~~uncertainty~~ grain size confidence intervals for each percentile in the distribution, the ~~uncertainty~~ sample precision can be displayed graphically as a polygon surrounding the distribution estimates. When comparing two different distributions, this means of displaying grain size distribution data highlights which distributions appear statistically different, and which do not.

Our analysis of various samples collected in the field demonstrates that the ~~uncertainty~~ grain size confidence interval depends on the shape of the distribution, with more widely graded sediments having ~~higher uncertainty~~ wider grain size confidence intervals than narrowly graded ones. Our analysis also suggests that typical gravel bed river channels have a similar gradation, and that the typical uncertainty of the D_{50} varies from ±25% for a sample size of 100 observations to about ±11% for 500 observations.

When designing a bed sampling program, it is useful to estimate the precision of the sampling strategy and to select the sample size accordingly; to do so, we must first assume something about the spread of the data (assuming a log-normal distribution), and then verify the uncertainty after collecting the samples. Simple equations for predicting uncertainty (as a percent of the estimate) are presented here to help workers select the appropriate sample size for the intended purpose of the data.

Appendix A: Normal approximation

While it is difficult to determine the percentile confidence interval using Eq. **??** without using a scripting approach similar to the one we implement in the `GSDtools` package, we can approximate the percentile confidence interval analytically, and use the approximating equations in spreadsheet calculations. As **?** point out, the percentile of interest (P) can be approximated by a normally distributed variable with a standard deviation calculated as follows:

$$\sqrt{np(1-p)}n$$

(A1)

The term n refers to the number of stones being measured, and p refers to the probability of a single stone being finer than the grain size for a percentile of interest, D_P (recall from above that $p = P/100$, such that $p = 0.84$ for D_{84}). The standard deviation for $n = 100$ and $P = 84$ would be 3.7 . That means that the true D_{84} would be expected to fall between sampled $d_{80.3}$ and $d_{87.7}$ for a sample of 100 observations approximately 68% of the time, and would fall outside that range 32% of the time.

More generally, we can use the normal approximation to calculate the percentile confidence interval for any chosen confidence level (α). We simply need to find the appropriate value of the z statistic for the chosen values of α and n, and calculate the percentile confidence interval using the following confidence bounds:

$$P_{upper} = P + \sigma z \tag{A2}$$

$$P_{lower} = P - \sigma z \tag{A3}$$

The use of a normal distribution to approximate the binomial distribution is generally assumed to be valid for p values in the range $\frac{5}{n} \leq p \leq 1 - \frac{5}{n}$, although some have recommended the more stringent range of $\frac{20}{n} \leq p \leq 1 - \frac{20}{n}$ (e.g. **?**). For a sample size of 100 stones, the limits correspond to 5^{th} and 95^{th} percentiles of the distribution.

For ease of reference, Table **??** presents σ values for P ranging from 10 (i.e., the D_{10}) to 90 (D_{90}) and for n ranging from 50 observations to 3200 observations. For $\alpha = 0.10$, $z = 1.64$; for a $\alpha = 0.05$, $z = 1.96$; and for $\alpha = 0.01$, $z = 2.58$. The table can be used to estimate the approximate percentile confidence intervals for common values of α, P and n. However, the user will have to manually translate the percentile confidence intervals into grain size confidence intervals using the cumulative frequency distribution for their sample.

A spreadsheet (see supplemental material) implementing these calculations has also been developed. That spreadsheet maps the percentile confidence interval onto the user's grain size distribution sample in order to estimate the grain size confidence interval.

Table A1. Percentile standard deviations for various sample sizes (n) and percentiles (D_p)

n	D_{10}	D_{16}	D_{25}	D_{50}	D_{75}	D_{84}	D_{90}
50	4.2	5.2	6.1	7.1	6.1	5.2	4.2
100	3.0	3.7	4.3	5.0	4.3	3.7	3.0
200	2.1	2.6	3.1	3.5	3.1	2.6	2.1
400	1.5	1.8	2.2	2.5	2.2	1.8	1.5
800	1.1	1.3	1.5	1.8	1.5	1.3	1.1
1600	0.8	0.9	1.1	1.2	1.1	0.9	0.7
3200	0.5	0.6	0.8	0.9	0.8	0.6	0.5

Appendix B: Binomial distribution reference tables

This appendix presents reference tables for the percentile confidence interval calculations described above. The tables present calculations for a range of percentiles (P) and sample sizes (n). The calculations presented were made using the `GSDtools` package, hosted on Brett Eaton's GitHub page. It is freely accessible to download. You can also find a demonstration showing how to install and use the package at `https://bceaton.github.io/GSDtools_demo_2019.nb.html`. The source code for the package can be found in the online data repository associated with this paper.

These percentile confidence bounds do not depend on the characteristics of the grain size distribution, since they are determined by binomial sampling theory. Estimating the corresponding grain size confidence bounds requires the user to map the percentile confidence interval onto the grain size distribution in order to find the grain size confidence interval. The `GSDtools` package will automatically estimate the grain size interval.

Table B1. Upper and lower percentile confidence interval bounds for $\alpha = 0.05$ (95% confidence level)

	n = 100		n = 200		n = 300		n = 400		n = 500	
P	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}
10	4.0	15.8	5.8	14.1	6.6	13.3	7.0	12.9	7.3	12.6
15	7.8	21.8	10.0	19.9	10.9	19.0	11.5	18.5	11.8	18.1
20	12.0	27.6	14.3	25.4	15.4	24.5	16.0	23.9	16.4	23.5
25	16.2	33.2	18.9	30.9	20.0	29.8	20.7	29.2	21.2	28.7
30	20.7	38.7	23.5	36.2	24.7	35.1	25.4	34.4	25.9	34.0
35	25.3	44.0	28.2	41.4	29.5	40.3	30.2	39.6	30.7	39.1
40	30.0	49.2	33.0	46.6	34.3	45.4	35.1	44.7	35.6	44.2
45	34.8	54.3	37.9	51.7	39.2	50.5	40.0	49.8	40.5	49.3
50	39.7	59.3	42.8	56.7	44.2	55.5	45.0	54.8	45.5	54.3
55	44.7	64.2	47.8	61.6	49.2	60.5	50.0	59.7	50.5	59.3
60	49.8	69.0	52.9	66.5	54.3	65.3	55.0	64.7	55.6	64.2
65	55.0	73.7	58.1	71.3	59.4	70.2	60.2	69.5	60.7	69.1
70	60.3	78.3	63.3	76.0	64.6	75.0	65.3	74.3	65.8	73.9
75	65.8	82.8	68.6	80.6	69.8	79.7	70.6	79.1	71.1	78.6
80	71.4	87.0	74.1	85.2	75.2	84.3	75.9	83.7	76.3	83.4
85	77.2	91.2	79.6	89.5	80.7	88.8	81.3	88.3	81.7	88.0
90	83.2	95.0	85.4	93.7	86.3	93.1	86.8	92.7	87.2	92.5

Table B2. Upper and lower percentile confidence interval bounds for $\alpha = 0.10$ (90% confidence level)

	n = 100		n = 200		n = 300		n = 400		n = 500	
P	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}
10	4.8	14.7	6.4	13.4	7.1	12.8	7.5	12.4	7.7	12.2
15	8.8	20.6	10.7	19.0	11.5	18.3	12.0	17.9	12.3	17.6
20	13.1	26.3	15.2	24.5	16.1	23.7	16.6	23.2	17.0	22.9
25	17.5	31.8	19.8	29.9	20.8	29.0	21.3	28.5	21.7	28.1
30	22.1	37.2	24.5	35.1	25.5	34.2	26.1	33.7	26.6	33.3
35	26.7	42.5	29.2	40.3	30.3	39.4	31.0	38.8	31.4	38.4
40	31.5	47.6	34.1	45.5	35.2	44.5	35.9	43.9	36.3	43.5
45	36.3	52.7	39.0	50.6	40.1	49.6	40.8	49.0	41.2	48.6
50	41.3	57.7	43.9	55.6	45.1	54.6	45.8	54.0	46.2	53.6
55	46.3	62.7	48.9	60.5	50.1	59.6	50.8	59.0	51.2	58.6
60	51.4	67.5	54.0	65.4	55.2	64.5	55.8	63.9	56.3	63.5
65	56.5	72.3	59.2	70.3	60.3	69.3	60.9	68.8	61.4	68.4
70	61.8	76.9	64.4	75.0	65.4	74.2	66.1	73.6	66.5	73.2
75	67.2	81.5	69.6	79.7	70.7	78.9	71.3	78.4	71.7	78.1
80	72.7	85.9	75.0	84.3	76.0	83.6	76.5	83.1	76.9	82.8
85	78.4	90.2	80.5	88.8	81.4	88.2	81.9	87.8	82.2	87.5
90	84.3	94.2	86.1	93.1	86.9	92.6	87.3	92.3	87.6	92.1

Table B3. Upper and lower percentile confidence interval bounds for $\alpha = 0.20$ (80% confidence level)

	n = 100		n = 200		n = 300		n = 400		n = 500	
P	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}
10	5.7	13.5	7.1	12.5	7.6	12.1	8.0	11.8	8.2	11.6
15	10.0	19.2	11.5	18.0	12.2	17.5	12.6	17.2	12.9	17.0
20	14.4	24.7	16.1	23.4	16.9	22.8	17.3	22.5	17.6	22.2
25	19.0	30.1	20.8	28.7	21.6	28.1	22.1	27.7	22.4	27.4
30	23.6	35.4	25.6	33.9	26.5	33.2	26.9	32.8	27.3	32.5
35	28.4	40.7	30.4	39.1	31.3	38.4	31.8	37.9	32.2	37.6
40	33.2	45.8	35.3	44.2	36.2	43.5	36.7	43.0	37.1	42.7
45	38.1	50.9	40.2	49.3	41.2	48.5	41.7	48.1	42.0	47.8
50	43.1	55.9	45.2	54.3	46.1	53.5	46.7	53.1	47.0	52.8
55	48.1	60.9	50.2	59.3	51.1	58.5	51.7	58.1	52.0	57.8
60	53.2	65.8	55.3	64.2	56.2	63.5	56.7	63.0	57.1	62.7
65	58.3	70.6	60.4	69.1	61.3	68.4	61.8	67.9	62.2	67.6
70	63.6	75.4	65.6	73.9	66.4	73.2	66.9	72.8	67.3	72.5
75	68.9	80.0	70.8	78.7	71.6	78.0	72.1	77.6	72.4	77.4
80	74.3	84.6	76.1	83.4	76.8	82.8	77.3	82.4	77.6	82.2
85	79.8	89.0	81.5	88.0	82.2	87.5	82.6	87.1	82.8	86.9
90	85.5	93.3	87.0	92.4	87.6	92.0	87.9	91.8	88.2	91.6

Table B4. Upper and lower percentile confidence interval bounds for $\alpha = 0.33$ (67% confidence level)

	n = 100		n = 200		n = 300		n = 400		n = 500	
P	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}	P_{lower}	P_{upper}
10	6.5	12.4	7.7	11.8	8.1	11.5	8.4	11.3	8.6	11.2
15	11.0	18.0	12.3	17.2	12.8	16.8	13.1	16.6	13.3	16.5
20	15.6	23.4	17.0	22.5	17.6	22.1	17.9	21.8	18.2	21.6
25	20.3	28.7	21.8	27.7	22.4	27.3	22.8	27.0	23.0	26.8
30	25.0	34.0	26.6	32.9	27.3	32.4	27.6	32.1	27.9	31.9
35	29.8	39.2	31.5	38.0	32.1	37.5	32.5	37.2	32.8	37.0
40	34.7	44.3	36.4	43.1	37.1	42.6	37.5	42.3	37.8	42.0
45	39.6	49.4	41.3	48.2	42.0	47.6	42.4	47.3	42.7	47.1
50	44.6	54.4	46.3	53.2	47.0	52.6	47.4	52.3	47.7	52.1
55	49.6	59.4	51.3	58.2	52.0	57.6	52.5	57.3	52.7	57.1
60	54.7	64.3	56.4	63.1	57.1	62.6	57.5	62.3	57.8	62.0
65	59.8	69.2	61.5	68.0	62.1	67.5	62.6	67.2	62.8	67.0
70	65.0	74.0	66.6	72.9	67.3	72.4	67.6	72.1	67.9	71.9
75	70.3	78.7	71.8	77.7	72.4	77.3	72.8	77.0	73.0	76.8
80	75.6	83.4	77.0	82.5	77.6	82.1	77.9	81.8	78.2	81.6
85	81.0	88.0	82.3	87.2	82.8	86.8	83.1	86.6	83.3	86.5
90	86.6	92.5	87.7	91.8	88.1	91.5	88.4	91.3	88.6	91.2

Code and data availability. The source code for the R package, as well as the analysis code and the data used to create all of the figures in this paper are available online (doi:10.5281/zenodo.3234121). The R package used to estimate the confidence intervals for grain size distributions is also available online (doi:10.5281/zenodo.3229387).

Author contributions. B.C. Eaton drafted the manuscript, created the figures and tables, and wrote the code for the associated modelling and analysis in the manuscript; R.D. Moore developed the statistical basis for the approach, wrote the code to execute the error calculations, reviewed and edited the manuscript, and helped conceptualize the paper; and L.G. MacKenzie collected the laboratory data used in the paper, tested the analysis methods presented in this paper, and reviewed and edited the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The field data used in this paper was provided by BGC Engineering, by undergraduate students in a field course run by one of the authors, and by an NSERC-funded research project being conducted by two of the authors. The paper benefited greatly from the input of one anonymous reviewer, and from Dr. K. Bunte. We would also like to acknowledge the contributions of numerous graduate students at UBC who tested our R package and commented on the revised version of the manuscript.

References

Beschta, R. L. and Jackson, W. L.: The intrusion of fine sediments into stable gravel bed, Journal of Fisheries Resource Board of Canada, 36, 204–210, 1979.

Buffington, J. M. and Montgomery, D. R.: A procedure for classifying textural facies in gravel-bed rivers, Water Resources Research, 35, 1903–1914, 1999.

Bunte, K. and Abt, S. R.: Sampling frame for Improving pebble count accuracy in coarse gravel-bed streams, Journal of the American Water Resources Association, 37, 1001–1014, 2001a.

Bunte, K. and Abt, S. R.: Sampling surface and subsurface particle-size distributions in wadable gravel-and cobble-bed streams for analyses in sediment transport, hydraulics, and streambed monitoring, Gen. Tech. Rep. RMRS-GTR-74, US Department of Agriculture, Rocky Mountain Research Station, Fort Collins, CO, 2001b.

Bunte, K., Abt, S. R., Potyondy, J. P., and Swingle, K. W.: Comparison of three pebble count protocols (EMAP, PIBO and SFT) in two mountain gravel-bed streams, Journal of the American Water Resources Association, 45, 1209–1227, 2009.

Church, M., McLean, D. G., and Wolcott, J. F.: River bed gravels: sampling and analysis, in: Sediment Transport in Gravel-bed Rivers, edited by Thorne, C., Bathurst, J., and Hey, R., pp. 43–88, John Wiley & Sons Ltd., 1987.

Daniels, M. D. and McCusker, M.: Operator bias characterizing stream substrates using Wolman pebble counts with a standard measurement template, Geomorphology, 115, 194–198, 2010.

Eaton, B. C., Andrews, C., Giles, T. R., and Phillips, J. C.: Wildfire, morphologic change and bed material transport at Fishtrap Creek, British Columbia, Geomorphology, 118, 409–424, 2010a.

Eaton, B. C., Moore, R. D., and Giles, T. R.: Forest fire, bank strength and channel instability: the ~~'~~‘unusual’ response of Fishtrap Creek, British Columbia, Earth Surface Processes and Landforms, 35, 1167–1183, 2010b.

Efron, B.: Computer Age Statistical Inference, Cambridge University Press, 2016.

Fripp, J. B. and Diplas, P.: Surface sampling in gravel streams, Journal of Hydraulic Engineering, 119, 473–490, 1993.

Green, J. C.: The precision of sampling grain-size percentile using the Wolman method, Earth Surface Processes and Landforms, 28, 979–991, 2003.

Hey, R. D. and Thorne, C. R.: Accuracy of ~~Surface Samples from Gravel Bed Material~~surface samples from gravel bed material, Journal of Hydraulic Engineering, 109, 842–851, 1983.

~~Hyndman, R. J. and Fan, Y.: Sample quantiles in statistical packages, The American Statistician, 50, 361–365, 1996.~~

~~Kondolf, G. M.: Application of the pebble count: notes on method, purpose, and variants, Journal of the American Water Resources Association, 33, 79–87, 1997.~~

Kondolf, G. M. and Li, S.: The pebble count technique for quantifying surface bed material size in instream flow studies, Rivers, 3, 80–87, 1992.

Kondolf, G. M., Lisle, T. E., and Wolman, G. M.: Bed Sediment Measurement, chap. 13, pp. 347–395, John Wiley, 2003.

Latulippe, C., Lapointe, M. F., and Talbot, T.: Visual characterization technique for gravel-cobble river bed surface sediments, Earth Surface Processes and Landforms, 26, 307–318, 2001.

Leopold, L. B.: An improved method for size distribution of stream bed gravel, Water Resources Research, 6, 1357–1366, 1970.

Lisle, T. E. and Hilton, S.: The Volume of fine sediment in pools – an index of sediment supply in gravel-bed streams, Water Resources Bulletin, 28, 371–383, 1992.

Lisle, T. E. and Hilton, S.: Fine bed material in pools of natural gravel bed channels, Water Resources Research, 35, 1291–1304, 1999.

Marcus, W. A., Ladd, S. C., Stoughton, J. A., and Stock, J. D.: Pebble counts and the role of user-dependent bias in documenting sediment size distributions, Water Resources Research, 31, 2625–2631, 1995.

Meeker, W. Q., Hahn, G. J., and Escobar, L. A.: Statistical Intervals, John Wiley & Sons, Hoboken, New Jersey, USA, 2nd edn., 2017.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282 – 290, 1970.

Olsen, D. S., Roper, B. B., Kershner, J. L., Henderson, R., and Archer, E.: Sources of variability in conducting pebble counts: their potential influence on the results of stream monitoring programs, JAWRA Journal of the American Water Resources Association, 41, 1225–1236, 2005.

Petrie, J. and Diplas, P.: Statistical approach to sediment sampling accuracy, Water Resources Research, 36, 597–605, 2000.

Phillips, J. C. and Eaton, B. C.: Detecting the timing of morphologic change using stage-discharge regressions: a case study at Fishtrap Creek, British Columbia, Canada, Canadian Water Resources Journal, 34, 285–300, 2009.

Rice, S. and Church, M.: Sampling surficial fluvial gravels: The precision of size distribution percentile estimates, Journal of Sedimentary Research, 66, 654–665, 1996.

Wicklin, R.: Simulating data with SAS, SAS Institute, 2013.

Wilcock, P. R. and McArdell, B. W.: Surface-based fractional transport rates: Mobilization thresholds and partial transport of a sand-gravel sediment, Water Resources Research, 29, 1297–1312, 1993.

Wolman, M. G.: A method of sampling coarse river-bed material, EOS, Transactions American Geophysical Union, 35, 951–956, 1954.