ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

How many replicate tests do I need? – Variability of cookstove performance and emissions has implications for obtaining useful results

Yungang Wang[1], Michael D. Sohn[1], Ashok J. Gadgil[1], Yilun Wang[2], Kathleen M. Lask[3], Thomas W. Kirchstetter[1]

[1]Environmental Energy Technologies Division
Lawrence Berkeley National Laboratory
Berkeley, CA 94720

[2]ISO Innovative Analytics
San Francisco, CA 94111

[3]UC Berkeley, College of Engineering
Applied Science and Technology Program
Berkeley CA 94720

February 2013

The data used in this work were collected during research supported with grant number 500-99-013 from the California Energy Commission (CEC) and the U.S. Department of Energy under contract DE-AC02-05CH11231.

How many replicate tests do I need? – Variability of cookstove performance and emissions has implications for obtaining useful results

Yungang Wang[1], Michael D. Sohn[1], Ashok J. Gadgil[1], Yilun Wang[2], Kathleen M. Lask[3], Thomas W. Kirchstetter[1]

[1]Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94720

[2]ISO Innovative Analytics, San Francisco, CA 94111

[3]UC Berkeley, College of Engineering, Applied Science and Technology Program, Berkeley CA 94720.

Abstract

Almost half of the world's population still cooks on biomass cookstoves of poor efficiency and primitive design, such as three stone fires (TSF). Emissions from biomass cookstoves contribute to adverse health effects and climate change. A number of "improved cookstoves" with higher energy efficiency and lower emissions have been designed and promoted across the world. During the design development, and for selection of a stove for dissemination, the stove performance and emissions are commonly evaluated, communicated and compared using the arithmetic average of replicate tests made using a standardized laboratory-based test, commonly the water boiling test (WBT). However, published literature shows different WBT results reported from different laboratories for the same stove technology. Also, there is no agreement in the literature on how many replicate tests should be performed to ensure "significance" in the reported average performance. This matter has not received attention in the rapidly growing literature on stoves, and yet is crucial for estimating and communicating the performance of a stove, and for comparing the performance between stoves. We present results of statistical

analyses using data from a number of replicate tests of performance and emission of the Berkeley-Darfur Stove (BDS) and the TSF under well-controlled laboratory conditions. We observed moderate variability in the test results for the TSF and BDS when measuring several characteristics. Here we focus on two as illustrative: time-to-boil and $PM_{2.5}$ (particulate matter less than or equal to 2.5 micrometers in diameter) emissions. We demonstrate that interpretation of the results comparing these stoves could be misleading if only a small number of replicates had been conducted. We then describe a practical approach, useful to both stove testers and designers, to assess the number of replicates needed to obtain useful data. Caution should be exercised in attaching high credibility to results based on only a few replicates of cookstove performance and emissions. Stove designers, testers, program implementers and decision makers should all benefit from improved awareness of the importance of adequate number of replicates required to produce practically useful test data.

Keywords: Cookstove; Berkeley-Darfur Stove; Variability; Confidence Interval; Kolmogorov–Smirnov Test; Bootstrap

1. Introduction

About half of the world's population uses biomass as fuel for cooking (IEA, 2004). Exposure to indoor smoke from burning solid fuels leads to an estimated 2 million premature deaths annually and ranks within the top five overall risk factors in poor developing countries (WHO, 2009). This exposure has also been linked to adverse respiratory, cardiovascular, neonatal, and cancer outcomes (Smith et al., 2004; Weinhold, 2011). A 2011 World Bank report notes significant

contributions of biomass cooking to global climate change (World Bank, 2011). The contribution to climate change from black carbon (BC) emission from biomass cooking is a topic of growing interest, especially in terms of climate forcing and melting of glaciers (Hadley et al., 2010; Ramanathan and Carmichael, 2008). Current biomass stoves lead to a large burden of disease, and contribute to adverse impacts on local and the global environment. Hence there is substantial interest in developing and disseminating fuel-efficient biomass stoves with reduced emissions (e.g. DOE 2011). Launched in September 2010, the Global Alliance for Clean Cookstoves (GACC) "100 by 20" goal calls for 100 million homes to adopt clean and efficient stoves and fuels by 2020.

The "three-stone fire" (TSF), is a commonly prevailing cooking method among 3 billion people worldwide. In quantifying the performance of an improved stove, the TSF is commonly used as the baseline. This least expensive class of stove is simply an arrangement of three large stones supporting a pot over an open and unvented biomass fire. TSF is one of the two stoves we tested in this study. We also tested the performance and emissions of the Berkeley-Darfur Stove (BDS) as an exemplar of an improved fuel-efficient biomass cookstove. The BDS was developed at Lawrence Berkeley National Laboratory (LBNL) for internally displaced persons in Darfur, Sudan (http://cookstoves.lbl.gov/). It is an all-metal precision designed natural-convection stove, with design features co-developed by iterative feedback from Darfuri women cooks. The BDS by design accommodates Darfuri traditional round-bottom cooking pots and cooking techniques (Figure 1).

A literature survey of recent laboratory cookstove testing studies shows widely different numbers of replicate tests (Bailis et al., 2007; Jetter and Kariher, 2009; Jetter et al., 2012; MacCarty et al., 2008, 2010; Roden et al., 2009; Smith et al., 2007). The number of replicates

reported in these seven studies range from 1 to 23. However, six out of seven studies have reported results with only 3 replicates. One then can rightly ask: how many replicate tests do I need for useful test results? When asked the purpose of a project, this question becomes more specific. For example, the question might become: how many replicates are needed to estimate with 95% confidence the average "time to boil" within 2 minutes of the true value? This exemplifies perhaps the most frequently asked question in planning stove experiments.

There is no single or simple answer to this question. The answer depends on the experimental design, how many parameters need to be estimated, and the acceptable error of the estimated performance value. The value of the standard error depends on the variability inherent in the experiment, the precision of the measurements, and the number of replicate tests. In most experiments, only limited improvement is possible by modifying the experimental materials, protocol and increasing the precision of the measuring instrumentation. That leaves us with replication as the method to make the experiments more informative. In this study, we investigate the variability of stove performance and emission measurements using BDS and TSF data from the laboratory water boiling test and show how the number of replicates is linked to uncertainty and variability in the experiments and stove performance. We also show how many replicates are likely needed as a function of error tolerance and for various practical performance comparisons, such as "Does Stove A perform better than Stove B?" and "What is the uncertainty in the expected performance of Stove A or Stove B?"

2. Problem statement and causes of variability

Numerous past studies have reported results with only three replicate tests and have shown many different results. MacCarty et al. (2008) conducted three laboratory WBT measurements for each of five major types of biomass cookstoves. Their results showed that TSF used the most energy to boil and simmer the water. The rocket stove used the least energy to complete the task. Time to boil was lowest in the fan-powered stove, followed by the rocket stove. In contrast, Jetter and Kariher, (2009) reported that some rocket stoves used the same amount of energy as the TSF, and surprisingly took longer time to boil water compared to the TSF. And they found no statistical difference in time to boil between the fan-powered stove and the rocket stove with three replicates. These studies indicate that stove comparison results could be misleading if a small number of replicates were conducted. Therefore, the awareness of the existence and implications of the variability is critical when interpreting observations with only a few replicate tests.

The literature generally shows that even under carefully controlled conditions, stove test results (fuel efficiency, time-to-boil, emissions of particulates and gases) show moderate to high test-to-test variability. Others have reported (e.g. Chen et al., 2012), and we have observed, substantial variability in our own high frequency (1 Hz) measurements of stove emissions (Kirchstetter et al., 2013). There are many possible causes of this variability, and we list a few here. Stove efficiency and emissions are generally a function of thermal power, and owing to the discrete nature of fuel-feeding events, a stove's thermal power invariably varies, also contributing to temporal variability within a test, which can translate into test-to-test variability. Despite due care, the ratio of bark to sapwood to hardwood for various pieces of fuelwood can be different, and thus with different burn characteristics. Furthermore different pieces of fuelwood may have different surface to volume ratios, contributing to different rates of burning. Lastly,

even reasonably experienced and careful stove testers can demonstrate some variability in the way they tend the fire in the stove from test to test, and within a test (Granderson et al., 2009). All these (and other uncontrolled factors) together give rise to what we lump together as variability in the test-to-test replicate results for a stove under controlled laboratory conditions.

3. Approach

The question of "how many replicate tests do I need" is not novel. It is a well researched question in classical statistical theory, but has not received much attention from the current stove research community. We briefly summarize here the statistical background relevant to answer the question.

3.1 Probability density function and cumulative distribution function

Technically, for a continuous random variable, the probability density function (PDF) describes the probability that a value will be within a certain range of the sample. However, as this range is evaluated by integrating, it can be chosen to be quite small, so for most practical purposes, the PDF may be considered the probability of obtaining a particular value. (Ellison, 2009). Graphically, if the PDF is a curve, the cumulative distribution function (CDF) is the area under that curve. It is used to evaluate the probability as an area; the larger the included range, the greater the probability. Because of this, the CDF over the entire range is equal to 1. For a normal (or Gaussian) distribution, the CDF curve is a normal ogive curve, which is a smooth, even S-shaped curve (Ellison, 2009). Any skewing in the distribution away from the Gaussian will lead to one half of the S to be elongated or distorted.

3.2 Standard error and confidence interval for an average

The standard error is the measure of the magnitude of the experimental error of an estimated statistic (e.g. average). For the sample average $\bar{x}$ from n replicate tests, the standard error $\sigma_{\bar{x}}$ is $\sigma/\sqrt{n}$, where σ is the standard deviation of the n replicates. The standard deviation (or variance) refers to the variation of observations within individual experimental units, whereas the standard error refers to the random variation of an estimate (made with only n replicates) from the "true" value that should be obtained as the number of replicates is increased to a very large number, tending to be infinity. The standard deviation σ is calculated by:

$$\sigma = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} \tag{1}$$

where $x_i = 1,2, \ldots n$ are the individual measurements used to calculate the average. A convenient way to calculate the sample standard deviation is using the "STDEV" function in Excel. Replication will not reduce the standard deviation but it will reduce the standard error. The standard error on the mean (often called the expected value) can be reduced by increasing the number of replicates. In practical term, this means that our goal is to achieve a standard error small enough to make convincing and useful conclusions, but not too small. If the standard error is large, the experiment is inclusive, whereas if it is smaller than necessary, resources have been wasted.

The confidence interval is used to indicate the reliability of an estimate made from a given number of replicates. The $(1-\alpha)100\%$ confidence interval for the average $\bar{x}$ has the form $\bar{x} \pm E$, where E is called the half-length, since a segment of the length of $2E$ centered on $\bar{x}$, provides the full confidence interval. E is related to α, σ, and n (the number of replicates) by the

following equation.

$$E = Z_{\alpha/2}\sigma/\sqrt{n} \tag{2}$$

Where $Z_{\alpha/2}$ is a dimensionless number that can be looked up in standard handbooks (e.g. Berthouex and Brown, 2002). The number of replicates that will produce this interval half-length is

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2 \tag{3}$$

This assumes random sampling. It also assumes that n is large enough that the normal distribution can be used to define the confidence interval. To apply equation (3), we must specify $E, \alpha\ (or\ 1-\alpha), and\ \sigma$. Values of $(1-\alpha)$ that might be used are shown in the top row with corresponding values of Z in the bottom row of Table 1.

When the measurements are assumed to be normally distributed but the number of replicates is small (by small, textbooks suggest less than 30) and the population standard deviation is unknown, a Student's t-distribution is used. To calculate the number of replicates n, the coefficient t_p is used in place of $z_{\alpha/2}$shown in equation (3). With this replacement, equation (3) can be used to obtain n. A selection of t-values is listed in Table 2. The t value decreases as n increases, but there is little change once n exceeds 5. Beyond n of 5 (examining Eq. 2), the greatest gain in narrowing the confidence interval comes from the decrease in $1/\sqrt{n}$ and not from the decrease in t. An exact solution of the number of replicates for small n (less than 30) requires an iterative solution, but a good approximate solution is obtained by using a rounded value of t = 2.1 or 2.2, which covers a good working range of n = 10 to n = 25. When analyzing

data we carry three decimal places in the value of t, but that kind of accuracy is misplaced. The greatest uncertainty lies in the value of the specified σ (refer to Eq. 2), so we can conveniently round off t to one decimal place.

The alert reader would notice by now that the number of desired replicates, n, depends on knowledge of σ, via Eq. 3. However, σ is not known in advance when start a new measurement. That is another reason not to be unreasonably precise about this calculation of n. The number of replicates you calculate should usually be rounded up, not just to the next higher integer, but to some even larger convenient number. For example, if you calculate an n equals to 11, you might well decide to conduct 15 or 20 measurements to allow for possible loss of information (e.g. from failed tests). If you find the number of replicates too small after completing the experiments and during data analysis, it is expensive to go back to collect more experimental material. In other words, the calculated n is guidance and not a limitation. Additional information about confidence interval estimation and experiment sizing can be found in Berthouex and Brown (2002), Spiegel et al. (2008), and Taylor (1997).

3.3 Bootstrapping

All the preceding discussion was predicted on the assumption of Gaussian distribution of underlying population. What if the distribution is not Gaussian? Bootstrap is a powerful statistical method that allows estimation of the variability of many properties of the data without making any assumptions about the shape of the original distribution *F*. Efron (1977) provides an accessible explanation, with examples, of the bootstrap method. The key principle of Bootstrapping is to simulate repeated observations from the unknown distribution ***F*,** using repeated sampling of the obtained single set of data. Bootstrapping can be implemented by

constructing a number of resamples of the observed dataset. Each resample is obtained by random sampling *with replacement* from the original dataset (Varian, 2005). Increasing number of resamples can reduce the impact of random sampling errors, but it cannot increase the amount of information existing in the original dataset (Efron and Tibshirani, 1993).

3.4 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (K-S) test quantifies if two cumulative distribution functions (CDFs) are from the same population or not. It does so by exploring the maximum distance between the two CDFs. Corder et al. (2009) provide a good summary of the K-S test. The null hypothesis of a K-S test poses that the two samples are from the same population, and the research hypothesis poses either that they generally differ, leading to a two-tailed probability estimate, or that they differ in a specific direction, leading to a one-tailed estimate (Wall, 2003). The K-S test can be used to compare a sample distribution and a reference distribution or to compare two sample distributions.

The K-S test is a nonparametric statistical test and is only limited by the condition that it must be applied to continuous distributions. Unlike the t-test and other parametric tests, which require assuming Gaussian distribution, continuity is the primary requirement for application of K-S test making it a very useful tool with unknown distributions. Also for small and medium samples, it is more effective to use the K-S test over other nonparametric "goodness-of-fit" tests, such as the chi-square test or the Wilcoxon test. The different research hypotheses of the K-S test also provide directional flexibility which the chi-square test cannot provide (Wall, 2003).

4. Methods

4.1 Laboratory testing

Laboratory tests of BDS and TSF were performed at the LBNL cookstove testing facility. Concentrations of $PM_{2.5}$ (particulate matter less than or equal to 2.5 micrometers in diameter), CO/CO_2, BC, and several other co-pollutants emitted from BDS and TSF were simultaneously measured. The DustTrak (TSI 8534) used to measure $PM_{2.5}$ mass concentrations was calibrated against gravimetrically determined concentrations. The CO/CO_2 concentrations were measured in a single instrument by nondispersive infrared absorption spectroscopy (NDIR analyzer, CAI 600 series). A cookstove smoke-specific calibration was developed for the BC aethalometer measurements. The results were compared with particle light-absorption coefficients measured with a photoacoustic absorption spectrometer (PAS) and BC concentrations measured using a thermal-optical analysis method. The moisture content of each piece of fuel wood was measured using a moisture meter (Delmhorst, J-2000). Soft (pine and fir) and hard (oak) woods were used in an equal number of tests with both stove types. Soft wood pieces were saw-cut to approximately 15 cm long with a square cross-section of approximately 4 cm^2 and hard wood pieces were hatchet-cut to a similar size but irregular shape.

The BDS and TSF were compared using the international water boiling test (WBT) protocol (version 3.0, http://www.pciaonline.org/node/1048). The WBT is intended to provide a method to compare the performance and emissions of different stoves in completing a defined standardized task (Bailis et al., 2007). In this test, a fire is ignited and maintained by periodic addition of fuelwood to bring water in a pot to boil and subsequently maintain it on simmer for 15 minutes, whereupon the fire is extinguished and the mass of remaining fuelwood is measured. One of the main metrics in this test is the time to boil, which is the amount of time it takes to

bring water to a boil under specific conditions. The detailed testing methodology and results are given by Kirchstetter et al. (2013).

4.2 Data analysis

Stove performance is strongly influenced by the skill of the person tending the stove. Dozens of tests were practiced by trained stove testers on both TSF and BDS, and these data were discarded before performing the tests to produce the data reported in this paper. This ensured that the variability observed in the test results was not being primarily influenced by increasing skill of the tester in tending the stove. There were 20 and 21 tests completed for TSF and BDS for data analysis, respectively. All instrumentation discussed above operated properly during these 41 tests. The statistical analysis was performed using Statistical Analysis System (SAS Institute Inc., version 9) and R (http://www.r-project.org/).

5. Results and discussion

5.1 Data overview

The stove performance and emission results of 21 BDS tests and 20 TSF tests are comprehensively presented in Kirchstetter et al. (2013). The moisture content and dry mass of the soft and hard woods were similar to each other and were the same for TSF and BDS tests. The moisture content of soft wood (9 ± 2%) and hard wood (10 ± 2%) pieces was essentially the same. The dry mass of soft (20 ± 9 g) and hard wood (26 ± 13 g) pieces was similar. The completion of tests with softwood (10 tests) required about 90% of the time duration and 90% of the wood mass compared to those with hard wood (10 tests). The ratio of efficiency of the BDS

and the TSF, measured in time and dry wood consumed for test completion, was essentially the same for both wood types.

The data of time to boil and $PM_{2.5}$ emission factor for TSF and BDS are selected for the statistical analysis in this study. The histogram plots of these data are shown in Figure 2 and Figure 3. The CDF plots for the same data are shown in Figure 4 and Figure 5. On average, cooking tests with the BDS were completed in 74% of the time for TSF (30.3 minutes vs. 41.0 minutes). There was less variation in time to boil with the BDS, as indicated by a narrower spread in the CDF curves for BDS compared to TSF (Figure 4). The average $PM_{2.5}$ emission factor for the BDS tests was 80% of that for the TSF (3.1 g/kg-wood burned vs. 3.9 g/kg-wood burned). $PM_{2.5}$ shows large test-to-test variability. The distributions of BDS and TSF $PM_{2.5}$ data overlap substantially, but the questions to answer are whether the BDS performs differently than the TSF, e.g. cooks faster, and emits less $PM_{2.5}$.

5.2 Number of replicate tests to estimate the mean

We next computed the number of replicate tests needed to estimate the mean within a user-defined level of confidence. For example, suppose the analyst desires to compute the expected boil time of the BDS within a range of plus or minus 2 minutes. Suppose also that the analyst desires the certainty of that estimate to be 95%. In words, the analyst is saying, "I would like to know the number of replicate tests needed to compute the average boil time of the BDS within a range of 4 minutes, and that I want to know that range with a confidence of 95%." Figure 6 shows the number replicates needed for three probability levels (0.1, 0.05, and 0.01), which correspond to confidences of 90%, 95%, and 99%, respectively. We compute the number of replicates using Equation (3). The *x-axis* represents the number of replicates ranging from 1 to

25. The *y-axis* represents the width of the confidence interval about the mean, which is twice of the E value in Equation (2). As can be seen in the figure, the smaller the confidence interval about the mean desired, the larger the number of replicates required.

As the 0.05 probability in Figure 6 shows, if the width of the confidence interval for the mean time to boil is 4 minutes at the probability of 0.05, 7 replicates are required. Note that the standard deviation for the underlying distribution in Equation (2) is calculated based on the original 21 replicate tests. If only two replicates are conducted, the width of the confidence interval about the mean is 38 minutes at the probability of 0.05 (191 minutes for the probability of 0.01, 19 minutes for the probability of 0.10). When the number of replicates increases to 5, the width shrinks to 5.3 minutes at the probability of 0.05 (8.8 minutes for the probability of 0.01, 4.1 minutes for the probability of 0.10). The width of confidence interval about the mean is relatively stable when the number of replicates is greater than 15. The similar trend is observed for the BDS $PM_{2.5}$ emission factor data. The width of the confidence interval about the mean BDS $PM_{2.5}$ emission factor is enormous for $n < 5$, and becomes steady when $n > 10$.

The above method is only the first step in selecting a suitable number of replicate tests. In real laboratory testing conditions, instrument malfunctions, determination of the point and duration that water simmers can be questionable. Other unpredictable events can also occur. These factors should be taken into consideration beyond the statistical inference when determine the number of replicate tests. More replicate tests should be planned than required by the statistical estimation to compensate for these unusual occurrences. This also increases the margin of safety in case the variability in the underlying distribution, represented by the standard deviation (σ) in Equation (2), is larger than anticipated. A conservative margin of 100% is recommended based on our abundant stove laboratory testing experience.

5.3 Number of replicate tests to compare two stoves

We next examine how many replicate tests are needed to confirm whether the performance of two stoves is indistinguishable, within a level of confidence. In essence, we test whether the underlying statistical distribution of the two stoves for the mean boil time or emission factor are the same. Figure 7 shows the probability as a function of the number of replicates calculated using the K-S test.

On the *x-axis* is the number of replicates. For every replicate number, we generated 50,000 bootstrap samples using the original 21 replicate tests for the BDS and 50,000 bootstrap samples using the original 20 TSF replicate tests. For each pair of samples, we compute the probability (p-value) that they come from the same distribution. We then compute the ratio, or probability, of the number of pairs that come from the same distribution divided by 50,000 with a confidence of 95%. The y-axis shows the resulting probability. For example, when the number of replicates is greater than 6, the probability that the BDS and the TSF time to boil data are from two different distributions is greater than 95%. For the $PM_{2.5}$ emission factor data, 30 replicates are required to ensure that at least 95% chance the BDS and the TSF samples are drawn from two different distributions.

6. Conclusions

Our results show moderate inherent variability among the TSF and BDS time to boil and $PM_{2.5}$ emission measurements. We demonstrate using these data as examples that some stove laboratory testing results could be misleading if only a small number of replicate tests were conducted. However, there are costs associated with increasing the number of replicates. Our

analysis suggests that the variability in BDS time to boil and $PM_{2.5}$ emission data requires on the order of 20 replicates to obtain useful data. The average value of any measured parameter should be always reported together with the number of replicates conducted and the uncertainty (e.g. standard deviation or confidence interval). Cautions must be exercised in the interpretation of results based on only a few replicates. A brief survey of recent peer-reviewed journal papers reporting results of stove testing suggests that inadequate replicates of tests are common in the published literature.

The implications of these results include the following. (1) In the stove design and laboratory testing phase, researchers needs to conduct a relatively large number of replicates to ensure with some confidence that the improvements of stove performance and emission levels are truly achieved. (2) In the stove field testing phase, even more tests are required because of the less controlled testing environment, which would lead to larger inherent variability within replicates. (3) In the stove dissemination and adoption phase, decision makers and policy analysts should take into consideration of the variability and confidence intervals of the laboratory and field testing results prior to any decisions.

Acknowledgements

The data used in this work were collected during research supported with grant number 500-99-013 from the California Energy Commission (CEC). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CEC. Kathleen M. Lask was supported with National Defense Science and Engineering Graduate (NDSEG) Fellowship. The authors gratefully acknowledge Douglas Sullivan, Jessica Granderson, Chelsea Preble, and

Odelle Hadley of Lawrence Berkeley National Laboratory for their support of this project, as well as the many students, interns, and researchers who, before us, contributed to the development of the Berkeley-Darfur Stove.

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof, or The Regents of the University of California.

References

Bailis, R., Berrueta, V., Chengappa, C., Dutta, K., Edwards, R., Masera, O., Still, D., Smith, K. R., 2007. Performance testing for monitoring improved biomass stove interventions:

experiences of the household energy and health project. Energy for Sustainable Development 11 (2), 57-70.

Berthouex, P. M., Brown, L. C. 2002. Statistics for Environmental Engineers. Second Edition. Lewis Publishers.

Chen, Y., Roden, C. A., Bond, T. C., 2012. Characterizing biofuel combustion with patterns of real-time emission data (PaRTED). Environmental Science & Technology 46, 6110-6117.

Corder, G., Foreman, D., Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach. (Wiley, 2009)

DOE, Biomass cookstoves technical meeting: Summary report (DOE, Alexandria, VA, 2011). http://www1.eere.energy.gov/biomass/pdfs/cookstove_meeting_summary.pdf (accessed on 2/4/2013).

Efron, B., 1979. Bootstrapping methods: Another look at the jackknife. The Annals of Statistics 7 (1): 1-26.

Efron, B., Tibshirani, R., 1993. An introduction to the Bootstrap. Boca Raton, FL: Chapman & Hall/CRC. ISBN 0-412-04231-2.

Ellison, S., Barwick, V., Duguid Farrant, T., Practical Statistics for the Analytical Scientist: A Bench Guide, 2nd ed., (Royal Society of Chemistry, 2009)\

Granderson, J., Sandhu, J. S., Vasquez, D., Ramirez, E., Smith, K. R., 2009. Fuel use and design analysis of improved woodburning cookstoves in the Guatemalan Highlands. Biomass and Bioenergy 33, 306-315.

Hadley, O. L., Corrigan, C. E., Kirchstetter, T. W., Cliff, S. S., Ramanathan, V., 2010. Measured black carbon deposition on the Sierra Nevada snow pack and implication for snow pack retreat. Atmos. Chem. Phys., 10, 7505-7513.

IEA, 2004. Energy and development. World Energy Outlook 2004. IEA Publications, Paris.

Jetter, J., Kariher, P., 2009. Solid-fuel household cook stoves: Characterization of performance and emissions. Biomass and Bioenergy 33, 294-305.

Jetter, J., Zhao, Y., Smith, K. R., Khan, B., Yelverton, T., DeCarlo, P., Hays, M. D., 2012. Pollutant emissions and energy efficiency under controlled conditions for household biomass cookstoves and implications for metrics useful in setting international test standards. Environmental Science & Technology 46, 10827-10834.

Kirchstetter, T., Preble, C., Hadley, O., Gadgil, A., 2013. Quantification of black carbon and other pollutant emissions from a traditional and an improved cookstove. Lawrence Berkeley National Laboratory (LBNL) Report, number: LBNL-6062E.

MacCarty, N., Ogle, D., Still, D., Bond, T., Roden, C., 2008. A laboratory comparison of the global warming impact of five major types of biomass cooking stoves. Energy for Sustainable Development 12 (2), 56-65.

MacCarty, N., Still, D., Ogle, D., 2010. Fuel use and emissions performance of fifty cooking stoves in the laboratory and related benchmarks of performance. Energy for Sustainable Development 14, 161-171.

Milton, J. S., Arnold, J. C. 1995. Introduction to probability and statistics: Principles and applications for engineering and the computing sciences, 3rd ed., McGraw-Hill.

Ramanathan, V., Carmichael, G., 2008. Global and regional climate changes due to black carbon. Nature Geoscience 1, 221 – 227.

Roden, C. A., Bond, T. C., Conway, S., Benjamin, A., Pinel, O., MacCarty, N., Still, D., 2009. Laboratory and field investigations of particulate and carbon monoxide emissions from traditional and improved cookstoves. Atmospheric Environment 43, 1170-1181.

Smith, K. R., Mehta, S., Maeusezahl-Feuz, M. (2004) Indoor smoke from household solid fuels. In Comparative quantification of health risks: global and regional burden of disease due to selected major risk factors, M. Ezzati, A.D. Rodgers, A.D. Lopez, and C.L.J. Murray eds., World Health Organization, Geneva, Switzerland.

Smith, K. R., Dutta, K., Chengappa, C., Gusain, P. P. S., Berrueta, V., Masera, O., Edwards, R., Bailis, R., Shields, K. N., 2007. Monitoring and evaluation of improved biomass cookstove programs for indoor air quality and stove performance: Conclusions from the household energy and health project. Energy for Sustainable Development 11 (2), 5-18.

Spiegel, M. R., Lipschutz, S., Liu, J., Mathematical Handbook of Formulas and Tables, 3rd ed. (McGraw-Hill, 2008).

Taylor, J. R. 1997. An Introduction to Error Analysis, 2nd ed. (University Science Books, 1997).

Varian, H., 2005. Bootstrap tutorial. Mathematics Journal 9, 768-775.

Wall, J. V., Jenkins, C. R., Practical Statistics for Astronomers, (Cambridge University Press, 2003).

Weinhold, B., 2011. Indoor PM Pollution and Elevated Blood Pressure: Cardiovascular Impact of Indoor Biomass Burning. Environmental Health Perspectives, 119 (10), A442.

WHO, Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks (WHO, Geneva, 2009); www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_Front.pdf (accessed on 2/4/2013).

World Bank, Household Cookstoves, Environment, Health and Climate Change: A New Look at an Old Problem (63217, World Bank Washington, DC, 2011); http://climatechange.worldbank.org/content/cookstoves-report (accessed on 2/4/2013).

Table 1. Summary of Z values.

$1 - \alpha = 0.99$	$1 - \alpha = 0.95$	$1 - \alpha = 0.90$
$z = 2.56$	$z = 1.96$	$z = 1.64$

Table 2. Student's t-distribution critical values.

n (Number of replicates)	n – 1 (Degrees of Freedom)	$t_{.995}$ (One sided) or $t_{.99}$ (Two sided)	$t_{.975}$ (One sided) or $t_{.95}$ (Two sided)	$t_{.95}$ (One sided) or $t_{.90}$ (Two sided)
1	-	-	-	-
2	1	63.657	12.706	6.314
3	2	9.925	4.303	2.920
4	3	5.841	3.182	2.353
5	4	4.604	2.776	2.132
6	5	4.032	2.571	2.015
7	6	3.707	2.447	1.943
8	7	3.500	2.365	1.895
9	8	3.355	2.306	1.860
10	9	3.250	2.262	1.833
11	10	3.169	2.228	1.812
12	11	3.106	2.201	1.796
13	12	3.054	2.179	1.782
14	13	3.012	2.160	1.771
15	14	2.977	2.145	1.761
16	15	2.947	2.132	1.753
17	16	2.921	2.120	1.746
18	17	2.898	2.110	1.740
19	18	2.878	2.101	1.734
20	19	2.861	2.093	1.729
21	20	2.845	2.086	1.725
22	21	2.831	2.080	1.721
23	22	2.819	2.074	1.717
24	23	2.807	2.069	1.714
25	24	2.797	2.064	1.711
26	25	2.787	2.060	1.708
27	26	2.779	2.056	1.706
28	27	2.771	2.052	1.703
29	28	2.763	2.048	1.701
30	29	2.756	2.045	1.699

Figure 1. Schematic of the Berkeley-Darfur Stove. (1) A tapered wind collar that increases fuel-efficiency in the windy Darfur environment and allows for multiple pot sizes; (2) Wooden handles for easy handling; (3) Metal tabs for accommodating flat plates for bread baking; (4) Internal ridges for optimal spacing between the stove and a pot for maximum fuel efficiency; (5) Feet for stability with optional stakes for additional stability; (6) Nonaligned air openings between the outer stove and inner fire box to accommodate windy conditions; and (7) Small fire box opening to prevent using more fuel wood than necessary.

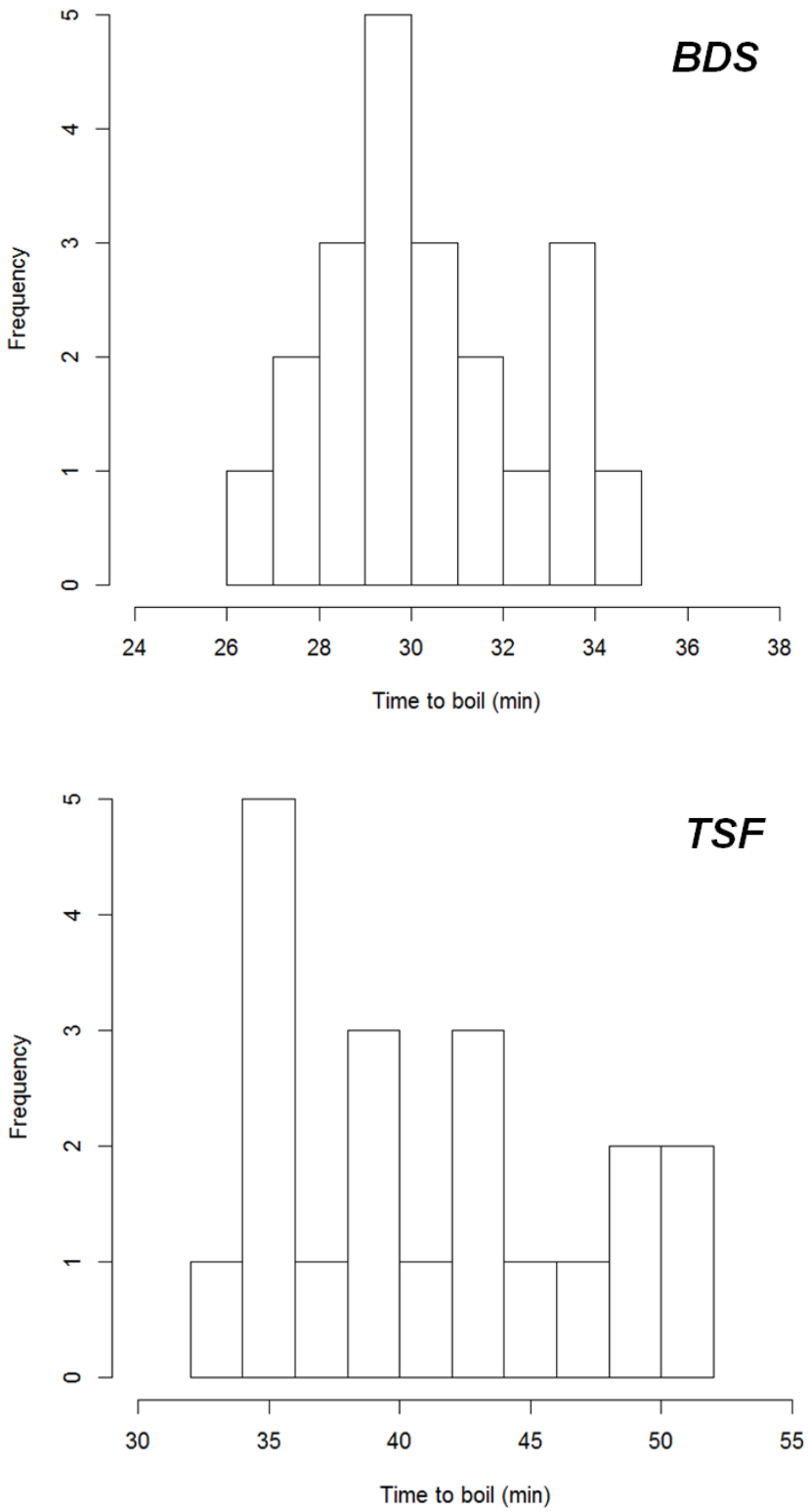

Figure 2. Histogram of time to boil data for the BDS and the TSF.

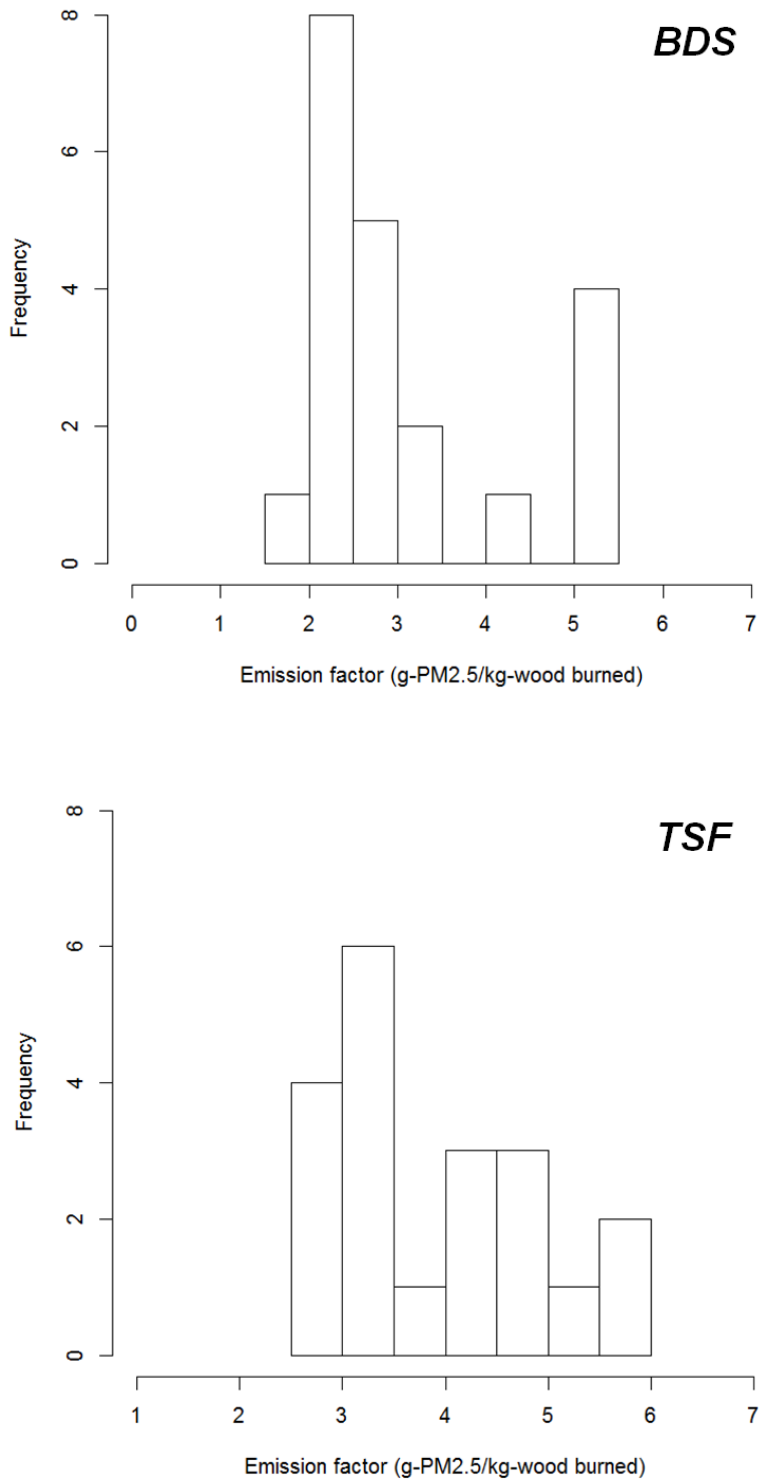

Figure 3. Histogram of $PM_{2.5}$ emission factor data for the BDS and the TSF.

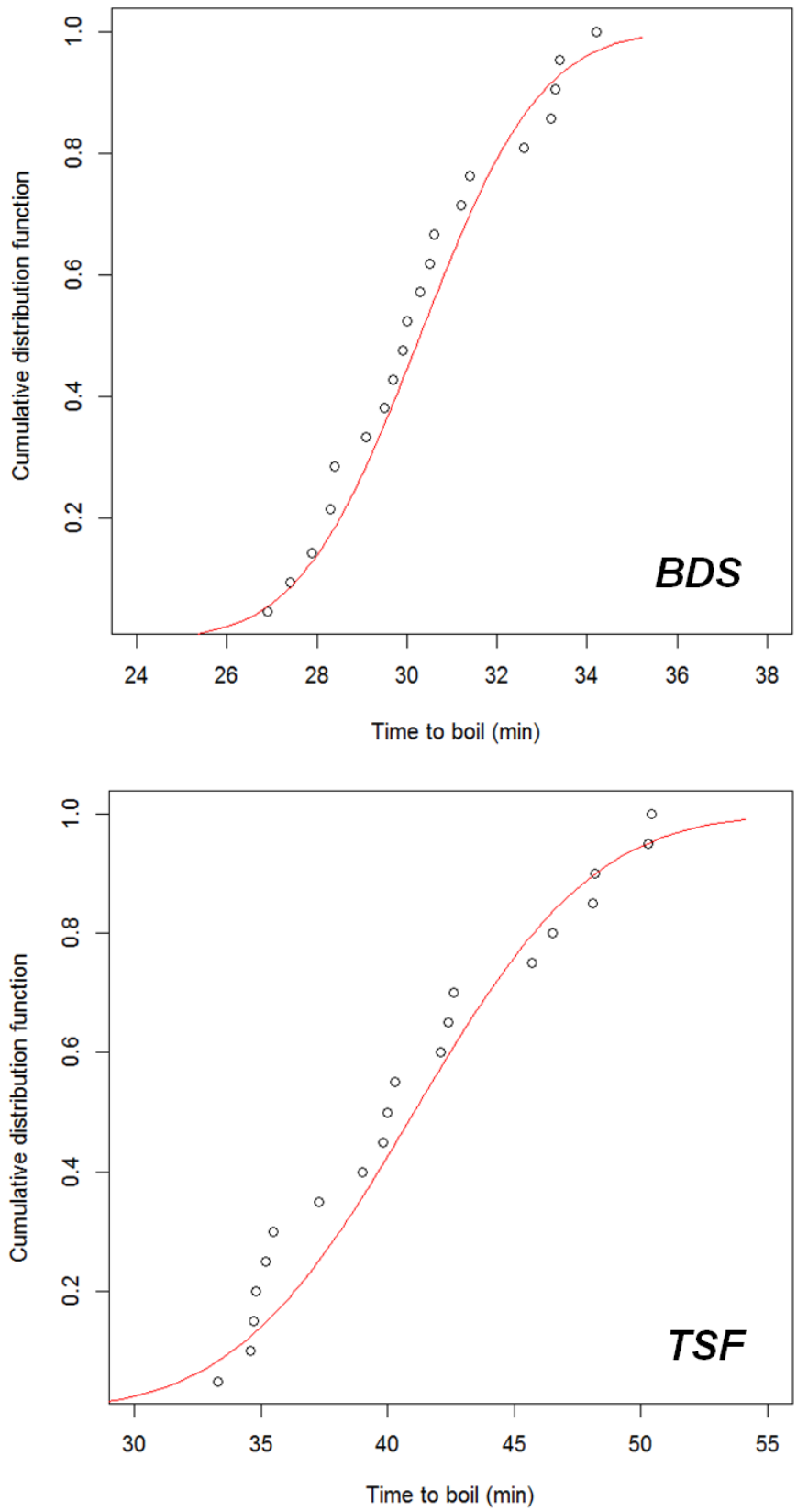

Figure 4. Cumulative distribution function (CDF) of time to boil data for the BDS and the TSF.

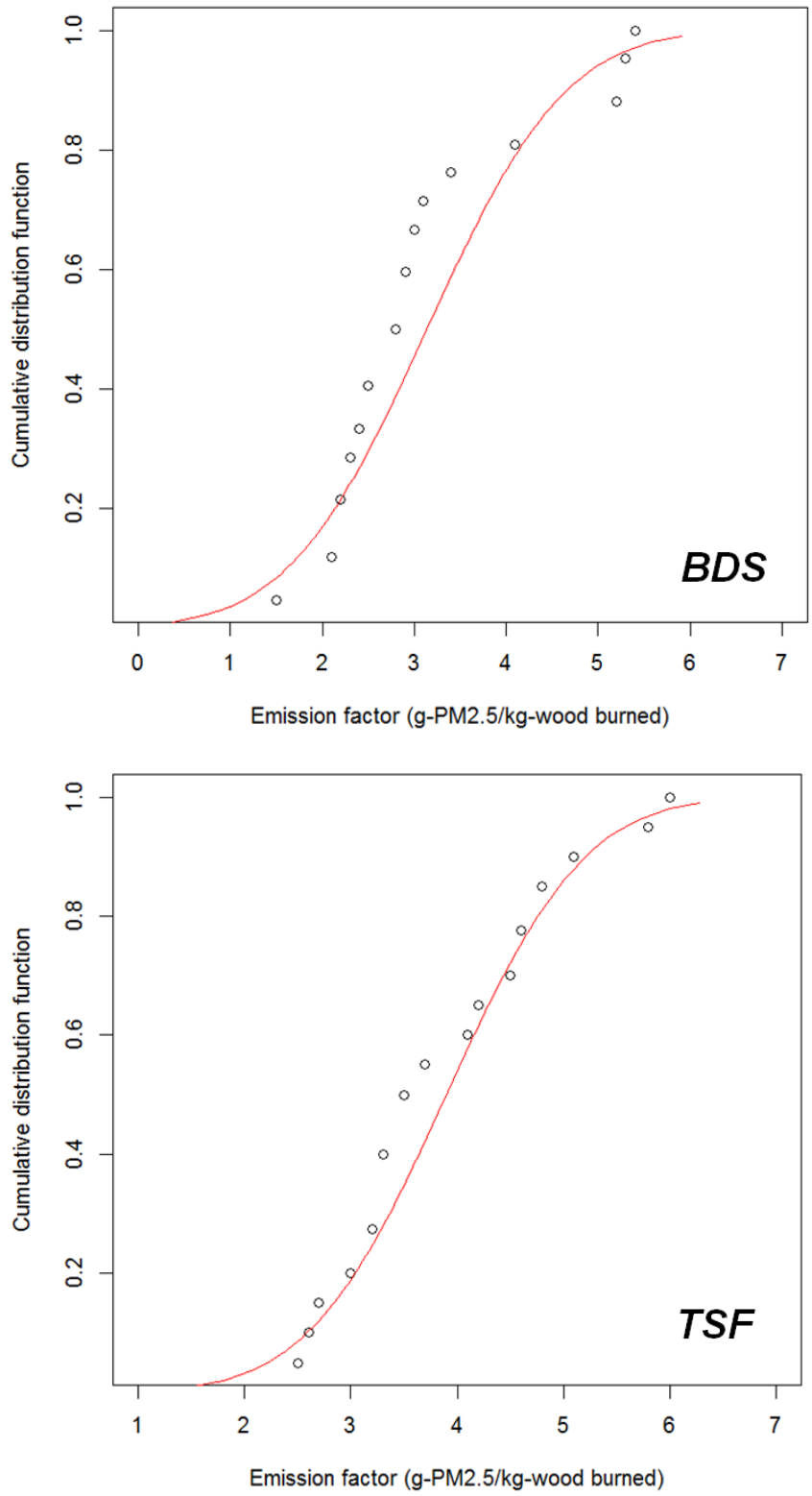

Figure 5. Cumulative distribution function (CDF) of $PM_{2.5}$ emission factor data for the BDS and the TSF.

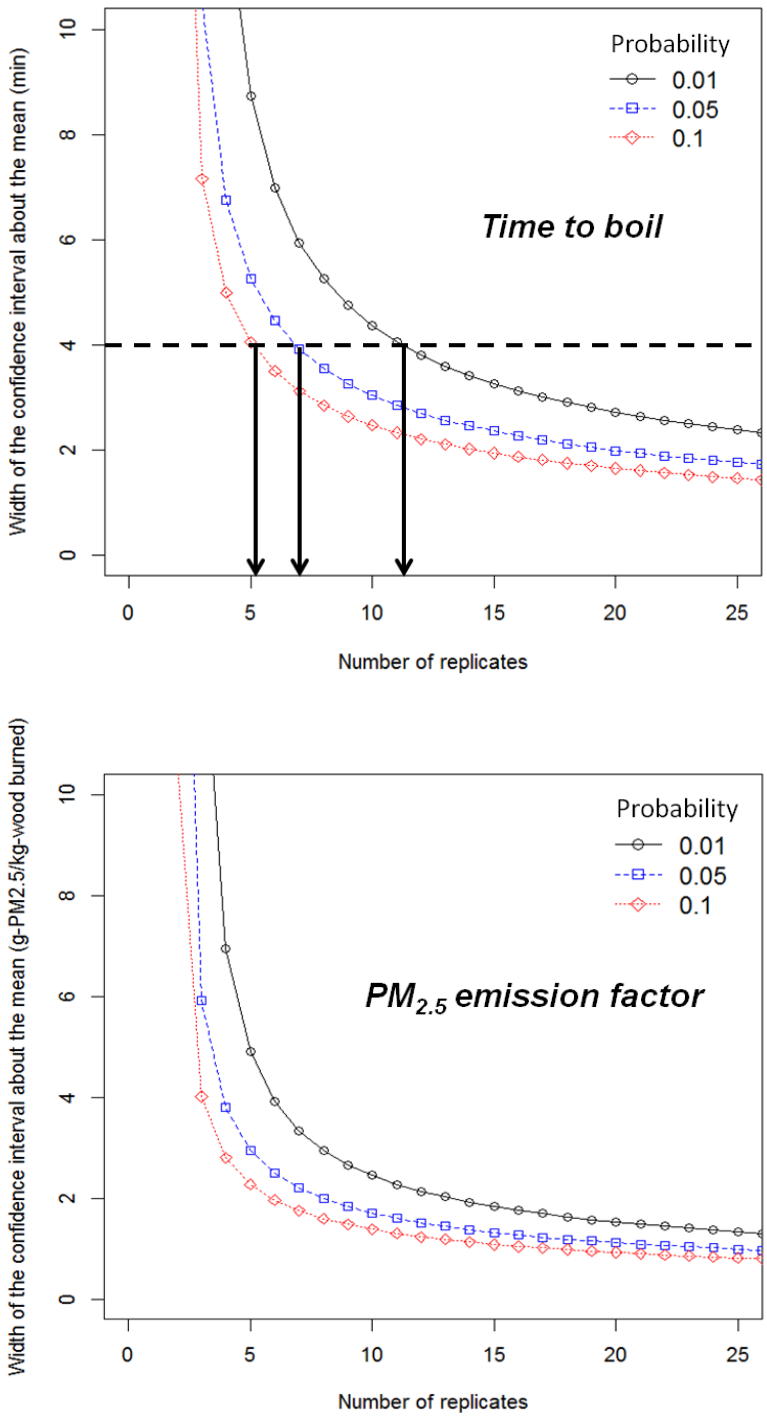

Figure 6. The width of the confidence interval about the mean as a function of the number of replicate tests at three probability levels (0.1, 0.05, and 0.01) for the BDS time to boil data and $PM_{2.5}$ emission factor data. For example, if the width of the confidence interval for the mean time to boil is 4 minutes at probability levels of 0.1, 0.05, and 0.01, 5, 7 and 12 replicates are required, respectively as indicated by the black horizontal dash line and the black vertical arrows.

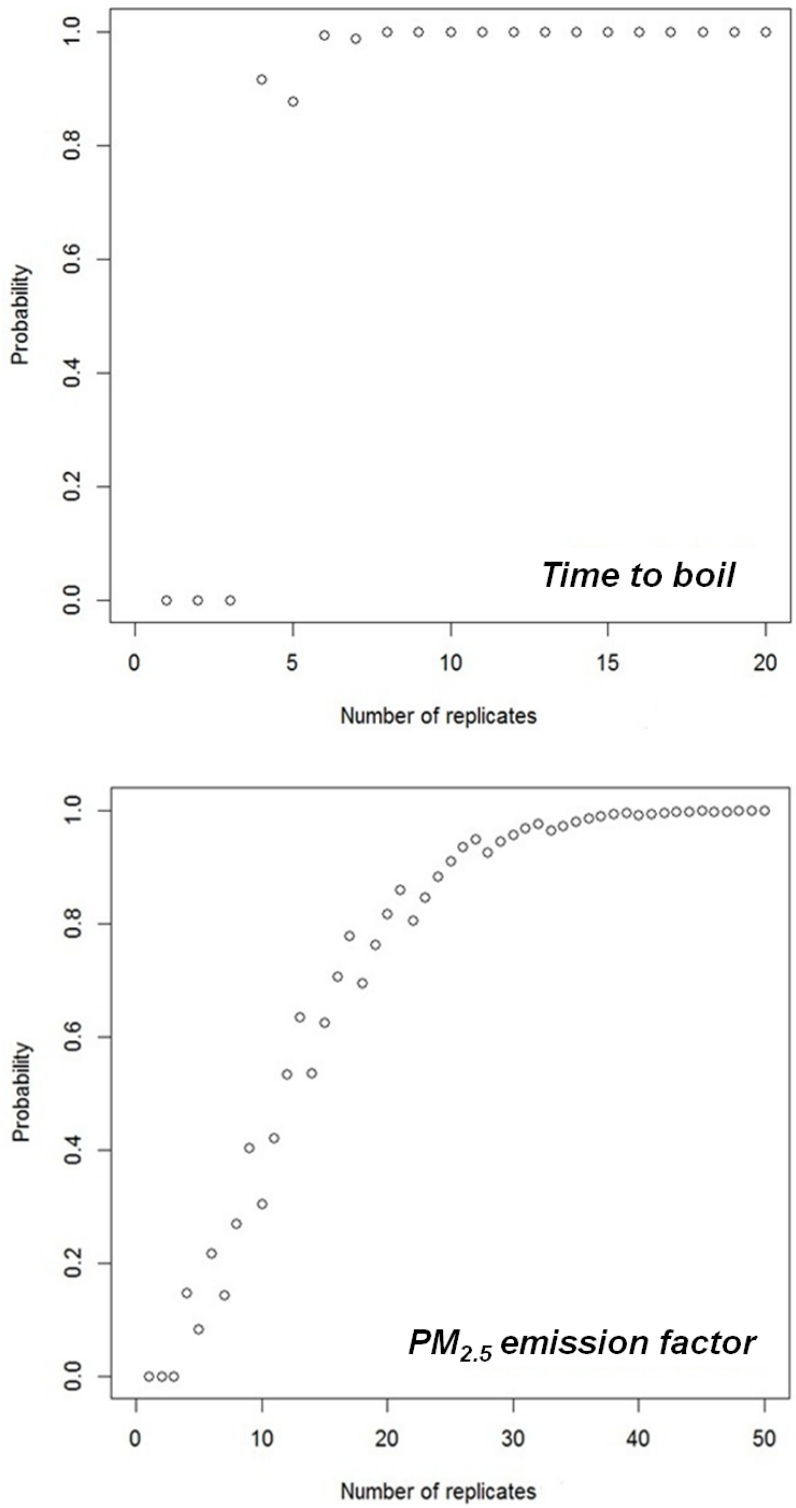

Figure 7. Kolmogorov-Smirnov test result showing the probability of the BDS and the TSF bootstrap samples are drawn from two different distributions as a function of the number of replicate tests for the time to boil and $PM_{2.5}$ emission factor data.