2016 ■ Vol. 12 ■ no. 2

GSD: An SPSS extension command for sub-sampling and bootstrapping datasets

Bradley Harding[a,] ✉ and Denis Cousineau[a]

[a]Université d'Ottawa

Abstract ■ Statistical analyses have grown immensely since the inception of computational methods. However, many quantitative methods classes teach sampling and sub-sampling at a very abstract level despite the fact that, with the faster computers of today, these notions could be demonstrated live to the students. For this reason, we have created a simple extension module for SPSS that can sub-sample and Bootstrap data, GSD (Generator of Sub-sampled Data). In this paper, we describe and show how to use the GSD module as well as provide short descriptions of both the sub-sampling and Bootstrap methods. In addition, as this article aims to inspire instructors to introduce these concepts in their statistics classes of all levels, we provide three short exercises that are ready for curriculum implementation.

Keywords ■ statistics teaching; sub-sampling; Bootstrap. **Tools** ■ SPSS.

Acting Editor ■ Sébastien Hélie (Purdue University)

Reviewers ■ one anonymous reviewer

✉ bhard024@uottawa.ca

BH: 0000-0003-1222-8769; *DC*: 0000-0001-5908-0402

10.20982/tqmp.12.2.p138

Introduction

Inference is at the heart of statistics, and while various methods gather insights of the characteristics of the sample, many are often overlooked in statistics classes of all levels. Two of these often neglected concepts are the sub-sampling and Bootstrap methods.

Sub-sampling is a method where one takes a series of observations from a sample in order to create one or many new sub-samples. These sub-samples are then used, for example, to test the reliability of the overall sample. This mimics the standard operation where many samples are taken from a population to infer its characteristics. Sub-sampling is often used in situations where one would want to see whether many small samples are as reliable as one another when inferring the large, original sample. More technically, let n_{ss} be the number of observations taken from the sample of size n_s ; typically, sub-sampling is used in situations where there is a large sample available so that n_{ss} is much smaller than n_s.

Alternatively, the Bootstrap method developed by Efron (Efron, 1979, 1981; Efron & Tibshirani, 1993) is a relative to sub-sampling that has expanded on earlier Jackknife methods (Quenouille, 1949; Tukey, 1958); its purpose is to estimate statistics, very often standard errors, when the normality assumption is not tenable or plausible. The Bootstrap is similar to sub-sampling but it assumes that the sample is the population (or at least the best description of the population's distribution), a reasonable assumption if the researcher has gathered a "good" sample (a sample that was properly gathered from a single population with a sufficient variability of selected scores). To ensure that each sub-sample is different and completely random, each sub-sampled cases are picked from the original sample with replacement. Thus, single cases can be selected multiple times in a sub-sample and others may be absent. The Bootstrap method generates a large number of sub-samples, thereby mimicking the ideal situation where one is able to replicate an experiment, gathering many same-sized samples from the population to assess its characteristics and draw inferences. Additionally, rather than returning a single descriptive statistics to describe the sample (such as the mean, median, standard deviation, skewness, etc.), the Bootstrap returns standard error or confidence interval of a level decided by the user (often 95%) from all sub-samples. As the Bootstrap is non-parametric, it can be used with any population distribution (Chernick, 2012; Yu, 2003). This is of particular interest, as many formulas for standard errors of descriptive statistics require the population to follow a normal distribution for them to be reliable (Harding, Tremblay, & Cousineau, 2014, 2015).

The omission of both these methods in statistics education is evidently unfortunate (for more complete reviews of these concepts, see Chernick (2012) and Yu (2003)). While sub-sampling and the Bootstrap offer many advantages, they are not seen at the same level as other "core" concepts with a longer history in statistics (Yu, 2003). Students could gather intuitions on randomness and sampling by trying these methods for themselves. Instead, students are often quickly taught (if at all) both methods as (1) passive observers or (2) with mathematical arguments. Neither of these two teaching practices foster a learning environment. Additionally, researchers wishing to explore these methods are often required to code the tools themselves (or purchase the BOOTSTRAP add-on for SPSS). We believe that these accessibility constraints may explain their absence in the quantitative methods curriculum. However, as computers have become a necessity in many statistical analyses, it is primordial that the students' education also follows suit. Curriculums and teaching tools must adjust in order to ensure a complete education of the current state in statistics.

For these reasons, we have created an extension command for SPSS, called GSD (Generator of Sub-Sampled Data) which allows new and seasoned SPSS users to easily sub-sample and Bootstrap data. The module can re-sample any SPSS dataset utilizing the Syntax language integrated within the software. Syntax works by "writing" the commands rather than accessing them through the drop-down menus (as is typically taught). We believe this method to be better suited for statistics education as the students must actively decide what parameters to keep and omit, a manipulation that forces students to think critically about what their analyses entail. Additionally, Syntax allows users to save their work; at any time, users can return to their code and directly adjust the script (adjustments using the drop-down menus require the user to restart the drop-down menu process for each modified analysis). For those who are not at ease with learning Syntax, there exists a variety of accessible resources (e.g. *Statistical Computing Seminars Beyond Point and Click: SPSS Syntax* (n.d.); Field (2007); Harding et al. (2014, 2015); in French, Cousineau (2009); online help forums, etc. or Einspruch (2004) for more advanced programming tips). In addition, the "Paste" button at the bottom of most SPSS analyses drop-menus copies the command and writes it directly in the nearest Syntax window along with its sub-commands and parameters.

While writing code may seem like a daunting task to many, the GSD extension command only requires three intuitive sub-commands to sub-sample data of which only one is mandatory. We have used a similar, slightly more complex extension command in our undergraduate and graduate statistic classes to generate random data (GRD; Harding et al. (2014, 2015). With a short learning curve, students were able to use the extension command without any assistance and generated progressively more complex datasets as the semester continued.

Finally, this extension command is appropriate for classroom use as it follows the GAISE College Report's (American Statistical Association et al., 2005) six recommendations. Students are encouraged to manipulate all aspects of the sub-sampling and Bootstrap methods thereby encouraging their statistical literacy and thinking (recommendation 1); GSD is compatible with real data as it can sub-sample or Bootstrap any dataset implemented in SPSS (recommendation 2); GSD encourages students to learn about sub-sampling and Bootstrapping by doing, not reading (recommendation 3); GSD fosters an active learning classroom by allowing the students to explore the characteristics of each method for themselves (recommendation 4); GSD provides the students with the technology to do a process which is otherwise limited by time and prior knowledge of coding interfaces (recommendation 5); the use of this extension module in the classroom allows the teacher to assess each individual's understanding of the methods by observing each student's strategy at accomplishing each task (recommendation 6).

In this article, we will show how users can utilize GSD for both re-sampling methods. To ensure a smooth learning of this new command we will be sub-sampling from two specific datasets: *SmallSample.sav* and *LargeSample.sav*, which are available on the journal's web site. The large sample follows a normal distribution with a mean of 100 and a standard deviation of 15 to mimic IQ scores sampled from a normal population; in this sample, there are 5000 observations. The small sample has 50 observations and follows a non-normal distribution. For the following examples, we assume that these files are saved on the desktop of a Windows OS computer. The frequency distribution plots of these datasets are illustrated in Figure 1.

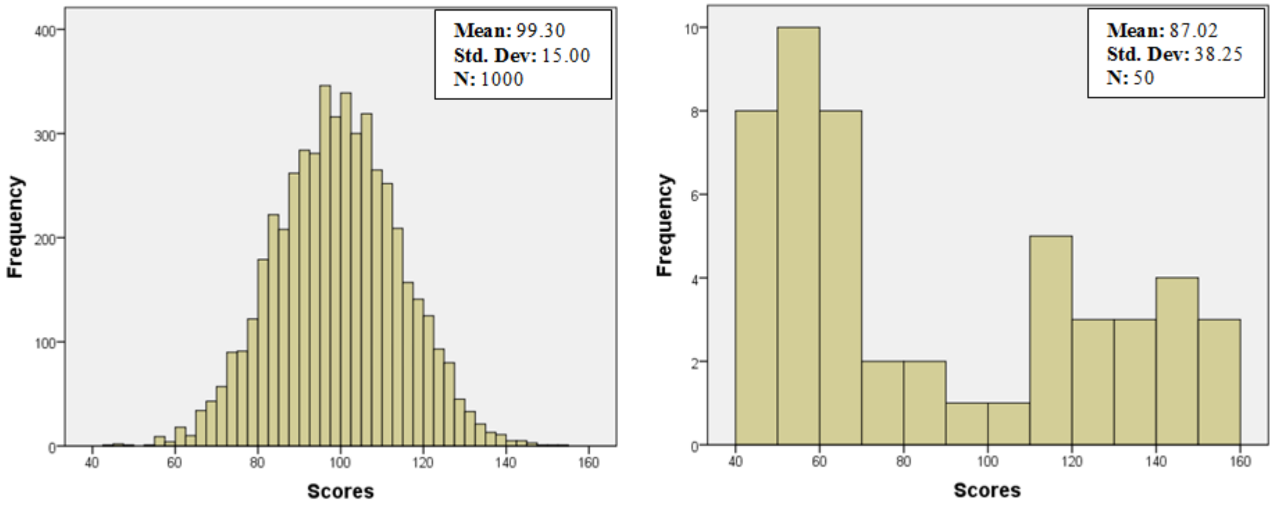
The GSD extension command is available on the journal's web site; Appendix A indicates how to install it in SPSS.

Using GSD for sub-sampling

GSD has three sub-commands to sub-sample data, two of which are presented here:

- `/FROM = file`: This sub-command specifies which SPSS dataset file is to be sub-sampled. For this sub-command to work, one must specify the complete path,

Figure 1 ■ Histograms of the two datasets used as samples for the GSD command. Both datasets were generated with GRD (Harding, Tremblay, & Cousineau, 2014, 2015). The histogram on the left has a sample size of 5000 and is used for sub-sampling whereas the histogram on the right has a sample size of 50 and is used for Bootstrapping. Descriptive statistics are located in the top-right corner of each plot.



including both the file name and extension (.sav). By default, if this sub-command is omitted (or `/FROM = *` is used), GSD will sub-sample from the opened dataset (or the most recently opened dataset if multiple datasets are currently opened).

- `/CASES` = n_{ss}: This sub-command sets the number of cases to sub-sample. `/CASES` is the only mandatory sub-command to run GSD. As GSD samples with replacement, each case can be selected more than once with equal probability.

For example, in Listing 1, one would sub-sample twenty-five observations (`/CASES=25`) from the LargeSample.sav dataset located on the desktop (`"c:\users\USERNAME\desktop\LargeSample.sav"`, where `USERNAME` must be adapted to your workstation):

Note the period at the end of the command: it tells SPSS that the command is complete and that GSD is ready to run (when this happens, the command's name will turn from red to blue). Place your cursor within the command's listing (or highlight it completely) and press "Ctrl-R" to run it.

In Figure 2, we present the frequency plot for four sub-samples taken from LargeSample.sav that have 25 cases each. As seen, while each sub-sample is different from one another, they do somewhat resemble one another as well as the distribution from which they were sub-sampled.

In the top-right corner of each panel, we present the mean, the standard deviation, the standard errors of the mean ($SE_{\overline{X}}$), as well as the sub-sample size for each sub-sample. The SE for each sub-samples are small which suggests that they are reliable estimates of the population, or more appropriately in this case, the sample. These SE of the mean were obtained by using the well-known formula $s/\sqrt{n_{ss}}$, in which s is the sub-sample's standard deviation and n is the sub-sample's size. However, what is less known from this equation is that it is an estimate based on the assumption that the population is normally distributed (see Harding et al., 2014), which is true of the population used to generate the *LargeSample* dataset. However, when the data analyst suspects that the population is not normally distributed, the above formula to estimate the standard error of the mean should not be trusted and an alternative estimation method must be used. This is precisely the purpose of the Bootstrap.

Using GSD to Bootstrap Data

Because the reliability of many descriptive statistics depends on the sample size, the Bootstrap generates sub-samples that are the same size as the original sample (therefore, $n_{ss} = n_s$).

To generate multiple sub-samples, a third sub-command is used:

- `/REPLICATIONS` = n_r: This sub-command dictates how many sub-samples are required from the original dataset. Each replication is completely random and independent.

Listing 1 ■ GSD instructions to sub-sample data. USERNAME must be adapted to your computer

```
GSD
  /FROM = "c:\users\USERNAME\desktop\LargeSample.sav"
  /CASES = 25.
```

The number of sub-samples in a Bootstrap analysis should be quite large: 1000 replications is a minimum; many recommend 5000. Furthermore, with the faster computers available now, it is common to see estimates based on many tens of thousands replications. Bootstrap estimates based on fewer than 1000 sub-samples would definitely need to be justified.

A standard error can be obtained from Bootstrapping a descriptive statistic on a large number of sub-samples and computing the standard deviation of these statistics. A 95% confidence interval can be likewise obtained by computing the interval in which 95% of the statistics lie (i.e., the 2.5% percentile for the lower bound, and the 97.5% percentile for the upper bound of the interval). Both standard error and confidence interval can be obtained from the same sub-samples.

To Bootstrap the dataset using GSD, one must ensure that /CASES is the same size as the original sample's size (here we use the *SmallSample.sav* dataset saved on the desktop that has 50 cases) and follow the steps given in Listing 2.

In Listing 2, sub-samples are taken from the sample *SmallSample.sav* illustrated in Figure 1.

- First the GSD command generates many replications as was introduced in Listing 1 (here, 1000, but try with 5000) each with $n_{ss} = n_s$; The replication number is kept in a new variable called replication (note that REPLICATIONS is the name of the sub-command and "replication" is the name of the newly generated variable);
- The OMS (Output Management System) command selects what descriptive statistics of the following instruction (in this case the MEANS command) will be captured in a new dataset and where to save this new dataset (Pfister, Schwarz, Carson, & Jancyzk, 2013);
- MEANS computes the mean of all the replications;
- OMSEND command signifies that the OMS command is done;
- The GET command gets the new, *BootstrapMeans.sav* dataset. In this file, each case (each line) is the descriptive statistic of one sub-sample (here, each line is a mean); there must be 1000 lines as we requested 1000 replications.
- (optional) The GRAPH command produces a frequency distribution plot of means (shown in Figure 3); on this plot, every datum represents one of the 1000 computed Bootstrap means.
- MEANS and FREQUENCIES compute the standard error as well as the 95% confidence interval (shown in the top right corner of Figure 3).

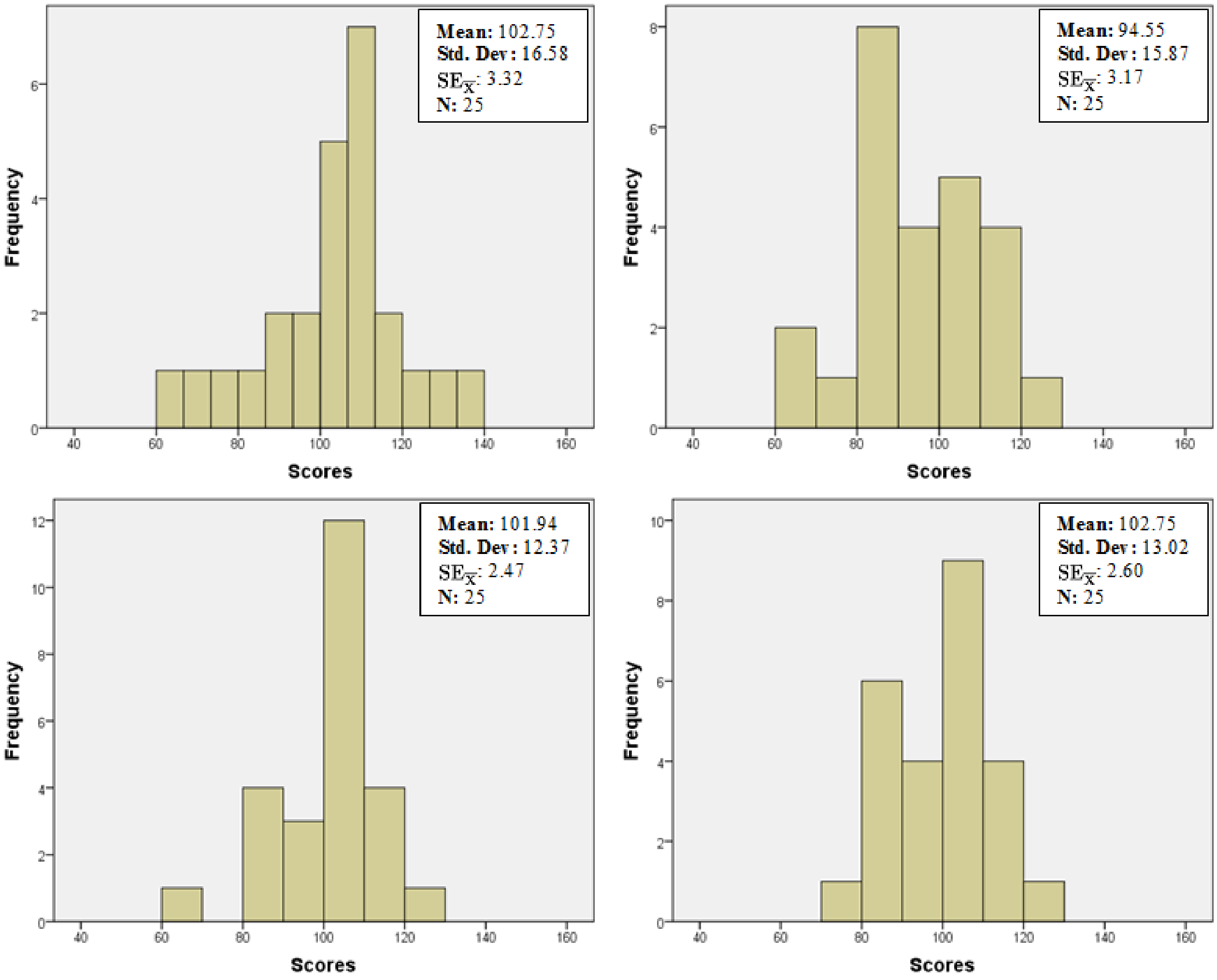
In this example, the sample's distribution is not normal. It is therefore imperative that the Bootstrap is used to interpret the descriptive statistics and standard error. In this new distribution, the mean of 1000 sub-samples is 86.64 with a 95% confidence interval of [76.61, 97,74]. As seen, although the Bootstrap mean estimate is nearly undistinguishable from the original sample's mean (seen in Figure 1) the Bootstrap's standard error (5.24) is smaller than the one obtained from the formula assuming normality ($38.25/\sqrt{50} = 5.41$). Additionally, Figure 3 has longer tails which shows that while the majorty of means are located around the mean of the original sample, a smaller proportion of sub-sample mean values are much further away. As a consequence, the 95% CI obtained from the Bootstrap is wider, but safer to use, as it does not assume normality.

Graphical user interface for GSD

While we advocate the use of Syntax as a teaching tool for various statistical analyses, the use of drop-down menus has its place. For example, if one wants to sub-sample a dataset with immediate feedback or sub-sample datasets without worrying about replicating the analysis, drop down menus are arguably easier to use. For this reason, we have also opted to develop a graphical user interface (GUI), available in the "Utilities" drop-menu. [1] In this menu, illustrated in Figure 4, users have access to the three GSD commands as well as the option to print debug information, the command's version, and the code GSD utilizes. By default, GSD sub-samples the opened dataset although it is possible to choose a specific dataset using the "Browse" option. Again, the only sub-command needed to run GSD is the number of cases.

[1] This dialog menu, *GSD_GUI.spd* is freely available on the journal's website. Installation of the GUI is covered in Appendix A.

Figure 2 ■ Four sub-samples of 25 cases that were sub-sampled from the large sample presented in Figure 1. The standard error of the mean as well as descriptive statistics for each sub-sample are presented in the top-right corner of each plot.



Conclusion

The GSD extension command allows users to sub-sample and Bootstrap data from any previously saved SPSS dataset. As both of these methods are rarely taught in introductory statistics classes, we thought best to create a free user-friendly extension command for SPSS (this software being a standard for many quantitative methods curriculums) in order to familiarize both new, and not-so-new, students of statistics to these methods.

As this extension command was developed mainly as a teaching tool, here are suggestions of applied classroom exercises:

- Have the students sub-sample a large dataset with increasingly larger sub-sample sizes. They should be able to see that the larger the sub-sample is, the more it resembles the original sample – have them note what they think is an appropriate size to gather inferences from.
- Sub-sample a dataset with sub-samples of small sizes, then increasingly larger sub-samples and have students note how standard deviation evolves. Many students have the false belief that standard deviation "shrinks" with increasingly large sample sizes. Invite them to repeat sub-sampling many times so that they do not forge an opinion based on a single sub-sample.
- Have the students create a sample that is rather small (this can be as simple as asking the class their height, age, how far they live from certain landmarks, etc.,

Listing 2 ■ GSD instructions to bootstrap the mean, obtaining Bootstrap estimates of the standard error and 95% confidence interval

```
GSD
  /FROM = "c:\users\USERNAME\desktop\SmallSample.sav"
  /CASES = 50
  /REPLICATIONS = 1000.

OMS
  /SELECT TABLES
  /IF COMMANDS = [’ MEANS’]
      SUBTYPES = [’ Report’]
  /DESTINATION FORMAT = SAV
   OUTFILE = ’c:\users\USERNAME\desktop\BootstrapMeans.sav’.

MEANS Scores BY replication
  /CELLS = MEAN.

OMSEND.

GET FILE = "c:\users\USERNAME\desktop\BootstrapMeans.sav".
GRAPH
  /HISTOGRAM Scores.
MEANS Scores
  /CELLS = stddev.
FREQUENCIES Scores
  /FORMAT notable
  /PERCENTILES 2.5 97.5.
```

for example using televoting, T. Groulx and Cousineau (submitted)) and have them Bootstrap the statistics of these measures. From here, one could compare the size of the confidence interval with the variability of the original datasets and note results.

Statistics have evolved considerably in the last hundred years. Yet, educational approaches remain seemingly stagnant (Stuart, 1995). Statistics teachers of tomorrow should have the opportunity to integrate sub-sampling methods in their core classes in order to create a new breed of researchers who conceptualize sampling and sub-sampling more deeply. While statistics have evolved, it is important that its education does as well.

Authors’ note

We would like to thank David Marso for his feedback and contributions to the sub-sampling and Bootstrapping code. This project was funded by the first author’s NSERC scholarship.

References

American Statistical Association, M., A., Cobb, G., Cuff, C., Garfield, J., Gould, R., . . . Witmer, J. (2005). *Guidelines for Assessment and Instruction in Statistics Education [GAISE] College Report*. Alexandria, VA: American Statistical Association.

Chernick, M. R. (2012). Resampling methods. *WIREs Data Mining Knowledge and Discovery*, *2*, 255–262. doi:10.1002/widm.1054

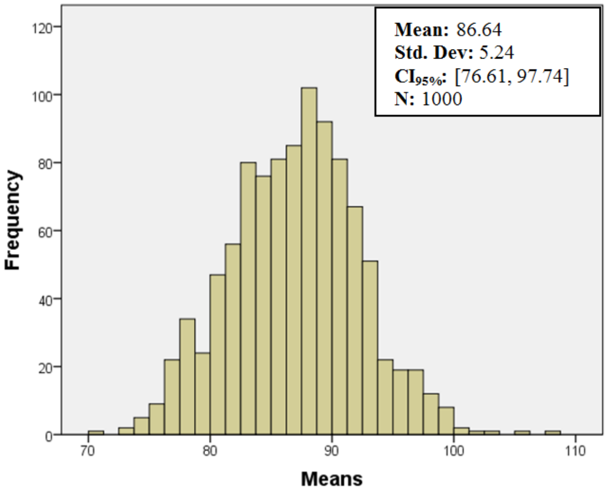
Cousineau, D. (2009). *Panomara des statistiques pour psychologues (1st edition)*. ISBN: 9782804108113. Bruxelles: Groupe de Boeck.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, *7*, 1–26. doi:10.1214/aos/1176344552

Efron, B. (1981). Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika*, *63*, 589–599. doi:10.1093/biomet/68.3.589

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Figure 3 ■ Histogram of means from the 1000 Bootstrapped sub-samples taken from the small sample presented in Figure 1. The 95% confidence interval of the means as well as the Bootstrapped sample's descriptive statistics are presented in the top-right corner of the plot (Note the change in scale on the horizontal axis).



Einspruch, E. (2004). *Next steps with SPSS*. Thousand Oaks, CA: SAGE Publications, Inc.

Field, A. (2007). *Discovering statistics using SPSS*. Thousand Oaks, CA: SAGE Publications, Inc.

Harding, B. & Cousineau, D. (2014). GRD: an SPSS extension command for generating random data. *The Quantitative Methods for Psychology*, *10*(2), 80–94.

Harding, B. & Cousineau, D. (2015). GRD 2.0: an extended SPSS extension command for generating random data. *The Quantitative Methods for Psychology*, *11*(3), 127–138.

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: a review and evaluation of standard error estimators using monte carlo simulations. *The Quantitative Methods for Psychology*. *10*(2), 107–123.

Harding, B., Tremblay, C., & Cousineau, D. (2015). The standard error of the pearson skew. *The Quantitative Methods for Psychology*. *11*(1), 32–36.

Pfister, R., Schwarz, K., Carson, R., & Jancyzk, M. (2013). Easy methods for extracting individual regression slopes: comparing SPSS, R, and Excel. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 72–78.

Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Soc. Series B*, *11*, 18–84. doi:10.1017/S0305004100025123

Statistical computing seminars beyond point and click: SPSS syntax. (n.d.). UCLA: Institute for Digital Research and Education. Retrieved from http://www.ats.ucla.edu/stat/spss/seminars/spss_syntax/

Stuart, M. (1995). Changing the teaching of statistics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *44*(1), 45–54. doi:10.2307/2348615

T. Groulx, J. & Cousineau, D. (submitted). IVote: A simple system to conduct polls and quizz in class settings. *The Quantitative Methods for Psychology*.

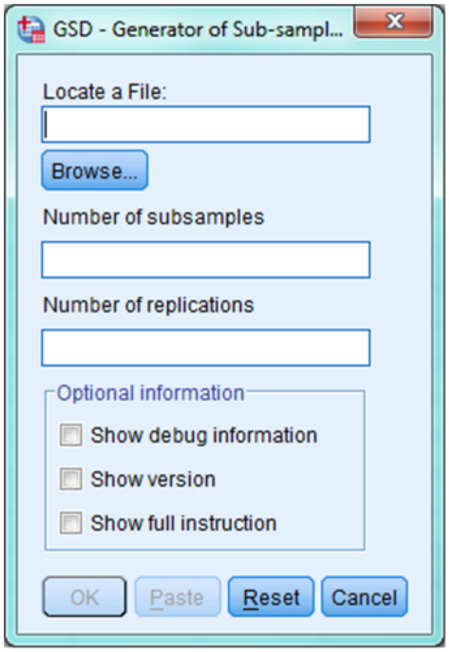
Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, *29*, 614.

Yu, C. H. (2003, January). Resampling methods: concepts, applications, and justification. *Practical Assessment, Research & Evaluation*, *8*, 19. Retrieved from http://PAREonline.net/getvn.asp?v=8%5C&n=19

Appendix A: Installation of the GSD extension command and its GUI

As this is an extension command for SPSS, it is not native to the software and requires an installation prior to use. The installation process begins by first downloading the *GSD.spe* and *GSD _GUI.spd* files available on this journal's web site. Additionally, the Python Essentials plugin (found on the IBM website) must be installed. SPSS 22 and above have the Python Essentials plugin directly integrated within the software, therefore no installation is necessary. For versions of SPSS prior to 22, installing the correct plugin is crucial for the extension command to function. Visual details of the installation are available in Figure 5.

Figure 4 ■ The Graphical User Interface created for GSD



To install the GSD module, one must:

1. Go to the “Utilities” drop-down menu, click the “Install Extension Bundle”;
   (a) The user will then be brought to a file browser dialog that will allow the user to locate the GSD.spe file;
   (b) Click Open once the GSD.spe file is located;
2. Restart SPSS;
3. Once SPSS is restarted, it is possible to verify the installation by clicking on the “View installed Extension Bundles” in the “Utilities” drop-menu (Figure 5);

Once the installation is completed, open a Syntax window to begin using the GSD extension module.

To install the GUI, one must:

1. Go to the “Utilities” drop-down menu, click the “Install Custom Dialog”;
   (a) The user will then be brought to a file browser dialog that will allow the user to locate the GSD_GUI.spd file;
   (b) Click Open once the GSD_GUI.spd file is located;
2. Restart SPSS;
3. Once SPSS is restarted, it is possible to verify the installation by clicking on the “Utilities” drop-menu; GSD should be located at the top of the list.

If there are any difficulties with the installation, ensure that the correct file is installed, that SPSS has been restarted, and that the Python Essentials plugin has been correctly installed (if necessary). Additionally, for the GUI to work, the GSD module must already be installed. For a more complete installation guide, consult Appendix 2 in Harding et al.’s (2014) article.

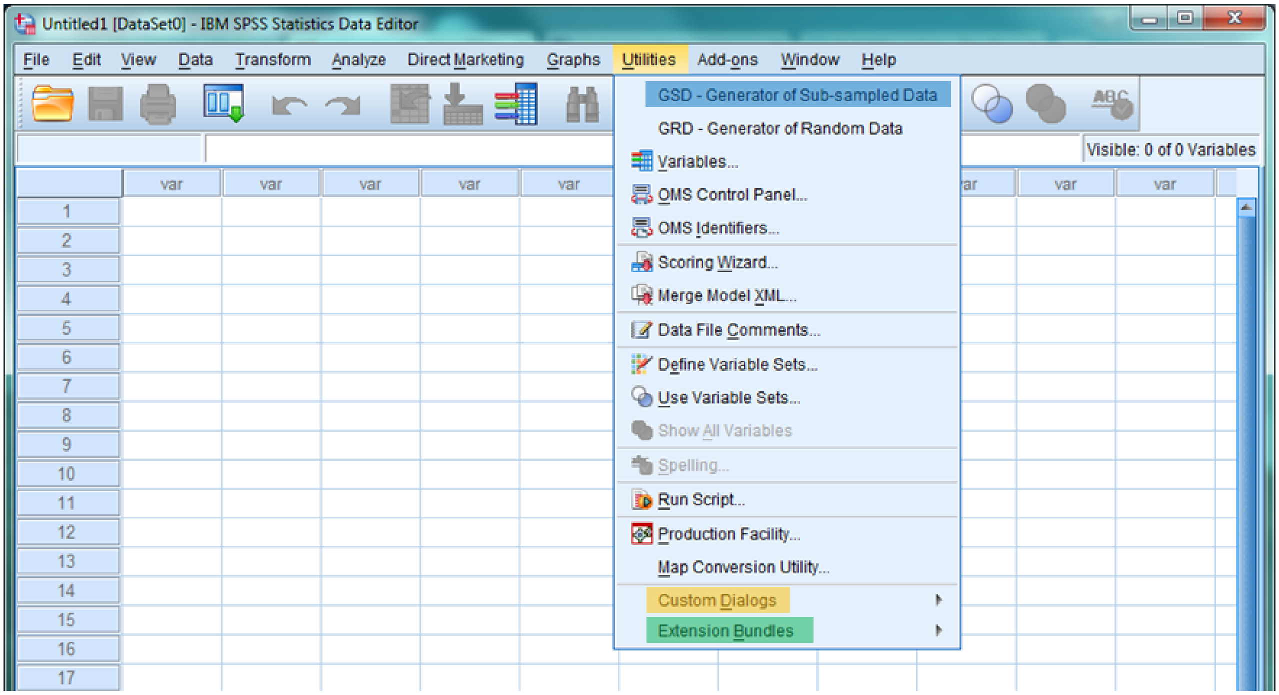
Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal’s web site.

Citation

Harding, B. & Cousineau, D. (2016). GSD: an SPSS extension command for sub-sampling and bootstrapping datasets. *The Quantitative Methods for Psychology*, *12*(2), 138–146. doi:10.20982/tqmp.12.2.p138

Figure 5 ■ Where to find the "Install Extension Bundle" window (highlighted in green) as well as the "Install Custom Dialog" window (highlighted in orange) when browsing the SPSS data editor window. Finally, the location of the GSD GUI is highlighted in blue.



Received: 02/02/2016 ~ Accepted: 30/03/2016