IBM SPSS Bootstrapping 23

Note
Before using this information and the product it supports, read the information in "Notices" on page 7.

Product Information

This edition applies to version 23, release 0, modification 0 of IBM® SPSS® Statistics and to all subsequent releases and modifications until otherwise indicated in new editions.

Contents

Chapter 1. Introduction to Bootstrapping

When collecting data, you are often interested in the properties of the population from which you took the sample. You make inferences about these population parameters with estimates computed from the sample. For example, if the *Employee data.sav* dataset that is included with the product is a random sample from a larger population of employees, then the sample mean of $34,419.57 for *Current salary* is an estimate of the mean current salary for the population of employees. Moreover, this estimate has a standard error of $784.311 for a sample of size 474, and so a 95% confidence interval for the mean current salary in the population of employees is $32,878.40 to $35,960.73. But how reliable are these estimators? For certain "known" populations and well-behaved parameters, we know quite a bit about the properties of the sample estimates, and can be confident in these results. Bootstrapping seeks to uncover more information about the properties of estimators for "unknown" populations and ill-behaved parameters.

How Bootstrapping Works

At its simplest, for a dataset with a sample size of *N*, you take *B* "bootstrap" samples of size *N* with replacement from the original dataset and compute the estimator for each of these *B* bootstrap samples. These *B* bootstrap estimates are a sample of size *B* from which you can make inferences about the estimator. For example, if you take 1,000 bootstrap samples from the *Employee data.sav* dataset, then the bootstrap estimated standard error of $776.91 for the sample mean for *Current salary* is an alternative to the estimate of $784.311.

Additionally, bootstrapping provides a standard error and confidence interval for the median, for which parametric estimates are unavailable.

Support for Bootstrapping in the Product

Bootstrapping is incorporated as a subdialog in procedures that support bootstrapping. See “Procedures That Support Bootstrapping” on page 4 for information on which procedures support bootstrapping.

When bootstrapping is requested in the dialogs, a new and separate `BOOTSTRAP` command is pasted in addition to the usual syntax generated by the dialog. The `BOOTSTRAP` command creates the bootstrap samples according to your specifications. Internally, the product treats these bootstrap samples like splits, even though they are not explicitly shown in the Data Editor. This means that, internally, there are effectively *B***N* cases, so the case counter in the status bar will count from 1 to *B***N* when processing the data during bootstrapping. The Output Management System (OMS) is used to collect the results of running the analysis on each "bootstrap split". These results are pooled, and the pooled bootstrap results displayed in the Viewer with the rest of the usual output generated by the procedure. In certain cases, you may see a reference to "bootstrap split 0"; this is the original dataset.

Chapter 2. Bootstrapping

Bootstrapping is a method for deriving robust estimates of standard errors and confidence intervals for estimates such as the mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. It may also be used for constructing hypothesis tests. Bootstrapping is most useful as an alternative to parametric estimates when the assumptions of those methods are in doubt (as in the case of regression models with heteroscedastic residuals fit to small samples), or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors (as in the case of computing confidence intervals for the median, quartiles, and other percentiles).

Examples. A telecommunications firm loses about 27% of its customers to churn each month. In order to properly focus churn reduction efforts, management wants to know if this percentage varies across predefined customer groups. Using bootstrapping, you can determine whether a single rate of churn adequately describes the four major customer types.

In a review of employee records, management is interested in the previous work experience of employees. Work experience is right skewed, which makes the mean a less desirable estimate of the "typical" previous work experience among employees than the median. However, parametric confidence intervals are not available for the median in the product.

Management is also interested in determining what factors are associated with employee salary increases by fitting a linear model to the difference between current and starting salaries. When bootstrapping a linear model, you can use special resampling methods (residual and wild bootstrap) to obtain more accurate results.

Many procedures support bootstrap sampling and pooling of results from analysis of bootstrap samples. Controls for specifying bootstrap analyses are integrated directly as a common subdialog in procedures that support bootstrapping. Settings on the bootstrap dialog persist across procedures so that if you run a Frequencies analysis with bootstrapping through the dialogs, bootstrapping will be turned on by default for other procedures that support it.

To Obtain a Bootstrap Analysis

1. From the menus choose a procedure that supports bootstrapping and click **Bootstrap**.
2. Select **Perform bootstrapping**.

Optionally, you can control the following options:

Number of samples. For the percentile and BCa intervals produced, it is recommended to use at least 1000 bootstrap samples. Specify a positive integer.

Set seed for Mersenne Twister. Setting a seed allows you to replicate analyses. Using this control is similar to setting the Mersenne Twister as the active generator and specifying a fixed starting point on the Random Number Generators dialog, with the important difference that setting the seed in this dialog will preserve the current state of the random number generator and restore that state after the analysis is complete.

Confidence Intervals. Specify a confidence level greater than 50 and less than 100. Percentile intervals simply use the ordered bootstrap values corresponding to the confidence interval percentiles. For example, a 95% percentile confidence interval uses the 2.5th and 97.5th percentiles of the bootstrap values as the lower and upper bounds of the interval (interpolating the bootstrap values if necessary). Bias corrected and accelerated (BCa) intervals are adjusted intervals that are more accurate at the cost of requiring more time to compute.

Sampling. The **Simple** method is case resampling with replacement from the original dataset. The **Stratified** method is case resampling with replacement from the original dataset, *within* the strata defined by the cross-classification of strata variables. Stratified bootstrap sampling can be useful when units within strata are relatively homogeneous while units across strata are very different.

Procedures That Support Bootstrapping

The following procedures support bootstrapping.

Note:

- Bootstrapping does not work with multiply imputed datasets. If there is an *Imputation_* variable in the dataset, the Bootstrap dialog is disabled.
- Bootstrapping does not work if there are non-integer weight values.
- Bootstrapping uses listwise deletion to determine the case basis; that is, cases with missing values on any of the analysis variables are deleted from the analysis, so when bootstrapping is in effect, listwise deletion is in effect even if the analysis procedure specifies another form of missing value handling.

Statistics Base Option

Frequencies. The following features are supported:

- The Statistics table supports bootstrap estimates for the mean, standard deviation, variance, median, skewness, kurtosis, and percentiles.
- The Frequencies table supports bootstrap estimates for percent.

Descriptives. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the mean, standard deviation, variance, skewness, and kurtosis.

Explore. The following features are supported:

- The Descriptives table supports bootstrap estimates for the mean, 5% Trimmed Mean, standard deviation, variance, median, skewness, kurtosis, and interquartile range.
- The M-Estimators table supports bootstrap estimates for Huber's M-Estimator, Tukey's Biweight, Hampel's M-Estimator, and Andrew's Wave.
- The Percentiles table supports bootstrap estimates for percentiles.

Crosstabs. The following features are supported:

- The Directional Measures table supports bootstrap estimates for Lambda, Goodman and Kruskal Tau, Uncertainty Coefficient, and Somers' d.
- The Symmetric Measures table supports bootstrap estimates for Phi, Cramer's V, Contingency Coefficient, Kendall's tau-b, Kendall's tau-c, Gamma, Spearman Correlation, and Pearson's R.
- The Risk Estimate table supports bootstrap estimates for the odds ratio.
- The Mantel-Haenszel Common Odds Ratio table supports bootstrap estimates and significance tests for ln(Estimate).

Means. The following features are supported:

- The Report table supports bootstrap estimates for the mean, median, grouped median, standard deviation, variance, kurtosis, skewness, harmonic mean, and geometric mean.

One-Sample T Test. The following features are supported:

- The Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Test table supports bootstrap estimates and significance tests for the mean difference.

Independent-Samples T Test. The following features are supported:

- The Group Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Test table supports bootstrap estimates and significance tests for the mean difference.

Paired-Samples T Test. The following features are supported:

- The Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.
- The Test table supports bootstrap estimates for the mean.

One-Way ANOVA. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Multiple Comparisons table supports bootstrap estimates for the mean difference.
- The Contrast Tests table supports bootstrap estimates and significance tests for value of contrast.

GLM Univariate. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the Mean and standard deviation.
- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.
- The Contrast Results table supports bootstrap estimates and significance tests for the difference.
- The Estimated Marginal Means: Estimates table supports bootstrap estimates for the mean.
- The Estimated Marginal Means: Pairwise Comparisons table supports bootstrap estimates for the mean difference.
- The Post Hoc Tests: Multiple Comparisons table supports bootstrap estimates for the Mean Difference.

Bivariate Correlations. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates and significance tests for correlations.

Notes:

If nonparametric correlations (Kendall's tau-b or Spearman) are requested in addition to Pearson correlations, the dialog pastes `CORRELATIONS` and `NONPAR CORR` commands with a separate `BOOTSTRAP` command for each. The same bootstrap samples will be used to compute all correlations.

Prior to pooling, the Fisher *Z* transform is applied to the correlations. After pooling, the inverse *Z* transform is applied.

Partial Correlations. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.

Linear Regression. The following features are supported:

- The Descriptive Statistics table supports bootstrap estimates for the mean and standard deviation.
- The Correlations table supports bootstrap estimates for correlations.
- The Model Summary table supports bootstrap estimates for Durbin-Watson.
- The Coefficients table supports bootstrap estimates and significance tests for the coefficient, B.
- The Correlation Coefficients table supports bootstrap estimates for correlations.
- The Residuals Statistics table supports bootstrap estimates for the mean and standard deviation.

Ordinal Regression. The following features are supported:

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Discriminant Analysis. The following features are supported:

- The Standardized Canonical Discriminant Function Coefficients table supports bootstrap estimates for standardized coefficients.
- The Canonical Discriminant Function Coefficients table supports bootstrap estimates for unstandardized coefficients.
- The Classification Function Coefficients table supports bootstrap estimates for coefficients.

Advanced Statistics Option

GLM Multivariate. The following features are supported:

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Linear Mixed Models. The following features are supported:

- The Estimates of Fixed Effects table supports bootstrap estimates and significance tests for the estimate.
- The Estimates of Covariance Parameters table supports bootstrap estimates and significance tests for the estimate.

Generalized Linear Models. The following features are supported:

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

Cox Regression. The following features are supported:

- The Variables in the Equation table supports bootstrap estimates and significance tests for the coefficient, B.

Regression Option

Binary Logistic Regression. The following features are supported:

- The Variables in the Equation table supports bootstrap estimates and significance tests for the coefficient, B.

Multinomial Logistic Regression. The following features are supported:

- The Parameter Estimates table supports bootstrap estimates and significance tests for the coefficient, B.

BOOTSTRAP Command Additional Features

The command syntax language also allows you to:

- Perform residual and wild bootstrap sampling (`SAMPLING` subcommand)

See the *Command Syntax Reference* for complete syntax information.

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 10504-1785
U.S.A.

For license inquiries regarding double-byte (DBCS) information, contact the IBM Intellectual Property Department in your country or send inquiries, in writing, to:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
1623-14, Shimotsuruma, Yamato-shi
Kanagawa 242-8502 Japan

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Licensees of this program who want to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact:

IBM Software Group
ATTN: Licensing
200 W. Madison St.
Chicago, IL; 60606
U.S.A.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The licensed program described in this document and all licensed material available for it are provided by IBM under terms of the IBM Customer Agreement, IBM International Program License Agreement or any equivalent agreement between us.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurements may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

All statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Each copy or any portion of these sample programs or any derivative work, must include a copyright notice as follows:

© your company name) (year). Portions of this code are derived from IBM Corp. Sample Programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at www.ibm.com/legal/copytrade.shtml.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Index

B

Printed in USA