JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Enhancing Excel's Statistical Capabilities: A VBA Approach to Bootstrap Resampling

[1]Debasree Goswami, [2]Jasvinder Goswami

[1]Associate Professor, Department of Statistics, Hindu College, University of Delhi, Delhi, India-7
[2]Associate Professor, Department of Statistics, PGDAV(M) College, University of Delhi, New Delhi, India-65

Abstract : This paper presents two Visual Basic for Applications (VBA) functions designed to facilitate bootstrap resampling in Microsoft Excel, a widely used spreadsheet software. Despite Excel's extensive capabilities, it lacks built-in functions for performing bootstrap resampling, a powerful statistical technique used for estimating the sampling distribution of a statistic. The developed VBA functions address this gap, enabling users to generate bootstrap samples and calculate a range of statistics from these samples directly within Excel. The functions are flexible and user-friendly, accommodating a variety of univariate statistics available in Excel. An application study using R.A. Fisher's "ToothGrowth" dataset demonstrates the practical use of the functions. The development of these functions contributes to the toolbox of statistical methods available in Excel, enhancing its utility for statistical analysis.

Keywords: *Bootstrap Resampling, Microsoft Excel, VBA, Univariate Statistics, Sampling Distribution, Data Analysis, Excel Programming*

I. INTRODUCTION

Bootstrap resampling is a powerful statistical technique introduced by Bradley Efron in 1979 [1] that involves generating a large number of "bootstrap" samples by resampling with replacement from an observed dataset. Bootstrap samples are then used to estimate the sampling distribution of a statistic, from which confidence intervals and hypothesis tests can be obtained. The beauty of this method is when the theoretical distribution of a statistic is complex or unknown, it allows for empirical estimation of the sampling distribution.

However, performing bootstrap resampling is not straight forward in spreadsheet software (in particular MS Excel) due to the lack of built-in functions for this purpose [2]. To address this issue, we present two Visual Basic for Applications (VBA) functions:

- $BootstrapSample(\dots)$ and
- $Bootstrap\_Statistic(\dots)$.

In the former, a given data range is used to generate a predetermined number of bootstrap samples, whereas in the later, a predetermined statistic is computed for a single bootstrap sample.

II. METHODOLOGY

Let the original data set be denoted as $\mathcal{D} = \{x_1, x_2, \dots, x_m\}$, where m is the number of observations. A bootstrap sample $\mathcal{D}^*$ is a set of m observations drawn with replacement from $\mathcal{D}$. The $BootstrapSample(\dots)$ function generates B bootstrap samples $\mathcal{D}_1^*, \mathcal{D}_2^*, \dots, \mathcal{D}_B^*$, where B is the number of bootstrap samples specified (from columns selected) by the user. For each bootstrap sample $\mathcal{D}_i^*$, the $Bootstrap\_Statistic(\dots)$ function calculates a statistic $T(\mathcal{D}_i^*)$, where T is a function [3] specified by the user.

In order to facilitate bootstrap resampling within the MS Excel environment, the details of two VBA functions (listed in the Appendix) are as follows:

- $BootstrapSample(\dots)$: This function generates B bootstrap samples from a given data range, returning a two-dimensional array of bootstrap samples. Each column represents a bootstrap sample $\mathcal{D}_i^*$ for $i = 1,2, \dots B$.
- $Bootstrap\_Statistic(\dots)$: This function generates a single bootstrap sample $\mathcal{D}_i^*$ from a given data range and calculates the specified statistic $T(\mathcal{D}_i^*)$ for that sample. The statistic can be one of several options, including the Average, Median, Standard Deviation, Variance, Skewness, Kurtosis, or Coefficient of Variation. This VBA function also provides a “placeholder”, where the user can introduce another suitable univariate statistic from among the list available in Excel.

III. APPLICATION

In the application phase of our study, we make use of the "ToothGrowth" dataset, which was originally compiled by R A Fisher [4]. This dataset consists of 60 data points distributed across three variables: "len", "supp", and "dose". The "len" variable represents the length of odontoblasts, which are cells involved in tooth growth. The "supp" variable indicates the supplement type used, either

"OG" for orange juice or "VC" for ascorbic acid. The "dose" variable denotes the daily dosage. We focus on the “len” variable for our illustration.

- The steps to generate $B = 10$ bootstrap samples from the "len" variable using the $BootstrapSample(\dots)$ function are:

Step 1: Enter the data for the "len" variable in cells $B3:B62$ in Excel.
Step 2: Select the range of cells $C3:L62$. This range includes 60 rows and 10 columns, which will hold the 10 bootstrap samples.
Step 3: With the range $C3:L62$ still selected, type the formula $= BootstrapSample(B3:B62)$ into the formula bar.
Step 4: Press $Ctrl + Shift + Enter$ as required for entering an array formula.
Upon completion of these steps, Excel will surround the formula with curly braces "{ }" to indicate that it's an array formula. The selected cells $(C3:L62)$ will be filled with the bootstrap samples (Figure 1).

Each bootstrap sample is a column in the resulting two-dimensional array and denoted as $\mathcal{D}_i^*$ for $i = 1,2,\dots 10$ (see Figure 1). Following the generation of these bootstrap samples, users can then proceed with further analysis such as computing the statistic $\hat{T}(D_i^*)$, calculating bias, standard error (S.E.), or any other statistical measures as required.

- The steps to calculate the statistic, say, "Median" of a bootstrap sample from the "len" variable using the $Bootstrap\_Statistic(\dots)$ function are:

Step 1: Enter the data for the "len" variable in cells $B3:B62$ in Excel.
Step 2: Select a cell where we want the Median of a bootstrap sample to be displayed.
Step 3: In the selected cell, type the formula $= Bootstrap\_Statistic(B3:B62, "Median")$ and Press Enter (Figure 2).
Upon completion of these steps, the selected cell will display the Median of the bootstrap sample, which is a single estimate based on resampling the original data. To study the sampling distribution of the Median, this process can be repeated to generate a large number of bootstrap sampled medians (Figure 2). Each repetition will generate a new bootstrap sample and calculate its Median. By repeating this process multiple times, a distribution of bootstrap sampled medians can be generated, which approximates the sampling distribution of the Median (Figure 3). This distribution can provide insights into the estimator’s variability and bias.

IV. CONCLUSION

In conclusion, the implementation of bootstrap resampling in Excel using the $BootstrapSample(\dots)$ and $Bootstrap\_Statistic(\dots)$ functions has been demonstrated to be a practical and effective tool for statistical analysis. These functions provide a straightforward and accessible way to perform bootstrap resampling and calculate various statistics from bootstrap samples.

Sl. No.	"len"	BootStrap Samples									
		1	2	3	4	5	6	7	8	9	10
1	4.2	16.5	10	9.7	9.7	5.2	21.5	7.3	23	25.8	33.9
2	11.5	11.2	20	27.3	11.5	26.4	23.6	9.4	17.3	26.4	26.4
3	7.3	9.7	11.2	26.4	18.5	17.3	23.3	32.5	23	25.2	25.5
4	5.8	21.5	27.3	20	26.7	29.5	11.5	25.5	8.2	18.5	17.3
5	6.4	11.2	15.2	25.8	19.7	26.4	16.5	25.5	27.3	15.5	24.8
6	10	26.4	11.2	23.6	15.2	16.5	24.5	11.2	15.2	4.2	17.3
…	…	…	…	…	…	…	…	…	…	…	…
57	26.4	14.5	27.3	7	25.2	22.5	32.5	23.6	15.2	27.3	4.2
58	27.3	16.5	10	26.7	17.6	23.6	7	29.5	19.7	11.2	27.3
59	29.4	9.7	22.5	7	27.3	16.5	8.2	5.2	25.5	15.2	14.5
60	23	33.9	8.2	23.3	14.5	9.4	25.5	23.6	11.2	27.3	21.2

Figure 1: Application of the $BootstrapSample()$ function to generate $B = 10$ bootstrap samples generated from the "len" data.

Sl.No.	50 x 5 = 250: Bootstrap Sampled Medians				
1	22.50	18.05	19.40	17.30	19.85
2	20.75	20.60	22.45	21.50	21.50
3	21.50	23.00	18.50	18.80	15.50
4	19.85	19.25	17.60	20.60	18.05
5	17.45	19.25	17.30	18.50	19.70
6	14.50	17.45	16.50	17.60	21.50
. . .	. . .	. . .	. . .	. . .	. . .
48	21.50	16.90	20.75	17.30	20.00
49	20.75	20.75	21.35	18.80	15.85
50	17.30	20.60	19.70	19.85	17.30

Figure 2: Application of the Bootstrap_Statistic( ) function to generate 250 samples and calculate Median values for the "len" data.

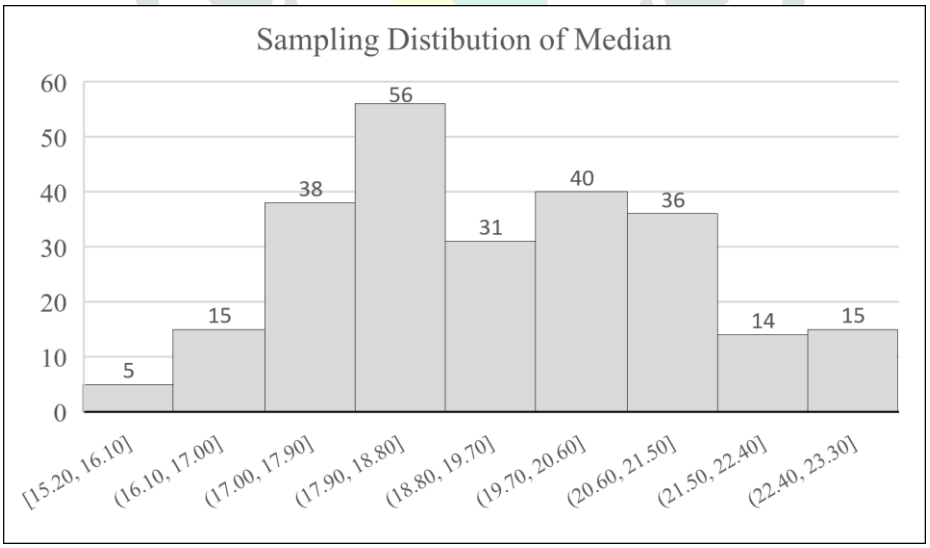

Figure 3: Histogram plot illustrating the sampling distribution of 250 bootstrap sample median values.

V. APPENDIX

```
Function BootstrapSample(data_range As Range) As Variant
    Dim data() As Variant
    Dim bootstrap_samples() As Variant
    Dim i As Integer, j As Integer
    Dim rng As Range
    Dim no_of_bootstrap_samples As Integer

    ' Copy data from range to array
    data = data_range.Value

    ' Get the number of bootstrap samples from the selected range
    no_of_bootstrap_samples = Application.Caller.Columns.Count

    ' Check if the number of rows selected is more than the data range
    If no_of_bootstrap_samples > UBound(data, 1) Then
        BootstrapSample = CVErr(xlErrNA)  ' Return error if the number of rows selected is more than
the data range
        Exit Function
    End If

    ' Initialize array for bootstrap samples
    ReDim bootstrap_samples(1 To UBound(data, 1), 1 To no_of_bootstrap_samples)

    ' Generate bootstrap samples
    For i = 1 To no_of_bootstrap_samples
        ' Resample data with replacement
        For j = 1 To UBound(data, 1)
            Set rng = data_range.Cells(Int((data_range.Cells.Count) * Rnd + 1))
            bootstrap_samples(j, i) = rng.Value
        Next j
    Next i

    ' Return array of bootstrap samples
    BootstrapSample = bootstrap_samples
End Function
```

```
Function Bootstrap_Statistic(data_range As Range, name_of_statistic As String) As Variant
    Dim data() As Variant
    Dim bootstrap_sample() As Variant
    Dim i As Integer
    Dim rng As Range
    Dim stat As Variant

    ' Copy data from range to array
    data = data_range.Value

    ' Initialize array for bootstrap sample
    ReDim bootstrap_sample(1 To UBound(data, 1))

    ' Generate bootstrap sample
    For i = 1 To UBound(data, 1)
        ' Resample data with replacement
        Set rng = data_range.Cells(Int((data_range.Cells.Count) * Rnd + 1))
        bootstrap_sample(i) = rng.Value
    Next i

    ' Calculate statistic
    Select Case name_of_statistic
        Case "Average", "AVERAGE", "average", "Ave", "AVE", "ave"
            stat = Application.WorksheetFunction.Average(bootstrap_sample)
        Case "Median", "MEDIAN", "median", "Med", "MED", "med"
            stat = Application.WorksheetFunction.Median(bootstrap_sample)
        Case "StDev", "STDEV", "stdev", "StDev.S", "STDEV.S", "stdev.s", "StDev.P", "STDEV.P", "stdev.p"
            If name_of_statistic = "StDev.P" Or name_of_statistic = "STDEV.P" Or name_of_statistic = "stdev.p" Then
                stat = Application.WorksheetFunction.StDev_P(bootstrap_sample)
            Else
                stat = Application.WorksheetFunction.StDev_S(bootstrap_sample)
            End If
        Case "Var", "VAR", "var", "Var.P", "VAR.P", "var.p", "Var.S", "VAR.S", "var.s"
            If name_of_statistic = "Var.P" Or name_of_statistic = "VAR.P" Or name_of_statistic = "var.p" Then
                stat = Application.WorksheetFunction.Var_P(bootstrap_sample)
            Else
                stat = Application.WorksheetFunction.Var_S(bootstrap_sample)
            End If
        Case "Skew", "SKEW", "skew", "Skew.P", "SKEW.P", "skew.p"
            If name_of_statistic = "Skew.P" Or name_of_statistic = "SKEW.P" Or name_of_statistic = "skew.p" Then
                stat = Application.WorksheetFunction.Skew_p(bootstrap_sample)
            Else
                stat = Application.WorksheetFunction.Skew(bootstrap_sample)
            End If
        Case "Kurt", "KURT", "kurt", "Kurtosis", "KURTOSIS", "kurtosis"
            stat = Application.WorksheetFunction.Kurt(bootstrap_sample)
        Case "Coefficient of Variation", "COEFFICIENT OF VARIATION", "coefficient of variation", "CV", "cv"
            stat = (Application.WorksheetFunction.StDev_S(bootstrap_sample) /
Application.WorksheetFunction.Average(bootstrap_sample))
        ' Add more cases here for other statistics
        Case Else
            stat = CVErr(xlErrValue)  ' Return error if statistic is not recognized
    End Select

    ' Return statistic
    Bootstrap_Statistic = stat
End Function
```

References

[1] B. Efron, "Bootstrap methods: Another look at the Jackknife," *Annals of Statistics,* pp. 1-26, 1979.

[2] W. R. Thomas, "Bootstrap on a shoestring: Resampling using spreadsheets," *The American Statistician,* vol. 48, no. 1, pp. 40-42, 1994.

[3] B. Efron and R. J. Tibshirani, An introduction to the bootstrap, CRC press, 1994.

[4] "ToothGrowth: The Effect of Vitamin C on Tooth Growth in Guinea Pigs," RDocumentation, [Online]. Available: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/ToothGrowth. [Accessed 25 July 2023].