Pedagogy of bootstrapping

P. M. SHANKAR
Department of Electrical and Computer Engineering
Drexel University, Philadelphia, PA, 19104
U S A
shankapm@drexel.edu

Abstract: Demos created to illustrate the bootstrapping procedure and its application to machine vision are presented with special emphasis on the area under the receiver operating characteristics (ROC) curves associated with a sensor. These formed an integral part of the undergraduate course in engineering probability updated to include topics in data analytics. Starting from the basics, the demos offer a step-by-step way to implement bootstrapping and statistical analysis of the areas under the ROC curves. A smaller data set was initially used to articulate all the details before analyzing the larger data set. Students were able to follow the procedures alongside. Results were also verified contemporaneously using commercial software used in ROC analysis. Furthermore, the efficacy of the demos and its effect on the learning process was evaluated through a set of individual assignments to the students followed by the statistical analysis of student surveys given at the beginning and conclusion of the course. They appear to show enhanced understanding of bootstrapping and ROC analysis by the students. The demos and the methodology proposed here could easily be extended to cover other topics of interest to expose the students to didactic aspects of complex concepts in engineering probability.

Key-Words: Bootstrapping. Hypothesis testing. Receiver operating characteristics

1 Introduction

Data analytics is integrated into the undergraduate engineering probability course to expand its scope to include applications of topics in probability and random variables to machine learning, medical diagnostics, signal processing, etc. At Drexel University, author started the initiative last year introducing topics of confusion and transition matrices, receiver operating characteristics (ROC) curves, and parametric hypothesis testing through the analysis of machine vision data [1]. The natural progression of these topics led to the introduction of bootstrapping as a tool to understand and interpret the statistics of metrics of interest such as the population mean, the area under the ROC curve (AUC), or the areas under the ROC curves collected from the same machine by two sensors [2,3]. The pedagogy of bootstrapping and its implementation are often sidelined in publications because of their reliance on bootstrapping software packages [4]-[9]. While software offers practical solutions, the didactic aspects of bootstrapping are lost without access to explicit instructive steps involved. Additionally, concepts of non-parametric hypothesis testing involving z-test, p-value, z_{score}, 95% confidence intervals [10] are also reinforced when bootstrapping is implemented in its basic form. The manuscript reports on demos created for students in an undergraduate course in engineering probability during the Fall quarter of 2018-2019, offering a step-by-step procedure to implement bootstrapping relying only on a uniform random number generator and verifying the results through other means. The demos then formed the basis for weekly exercises for the students. The efficacy of the demos in expanding the knowledge base in data analytics was investigated through student surveys conducted during the first and last week of classes.

2 Background

Inferences and predictions are often made on the basis of a single experiment done a few times. For example, we measure the temperature a few times or at a few locations in a room and offer inferential statistics, namely the mean and variance of temperature for any applications requiring such information. In another experiment involved in the testing of the efficacy of a new medical screening device, we recruit a number of subjects (having no illness and having illness) and undertake the statistical analysis of the efficacy of the device [5, 11,12]. The situation is identical to the case of a sensor used in a machine vision or machine learning system to determine the presence or absence of a target in its field of view. In these cases, the receiver operating characteristics curves (ROC) quantified through the metric, area under the ROC curve (AUC) is used to establish the performance of the new screening device or the sensor [13]. While simple formulas are available to study the statistics of the population mean (measurement of temperature mentioned above), there are very few formulaic tools available to examine the statistics of the AUC to draw inference on its variability

(variance) and its 95% confidence level [14, 15, 16]. In another case, two competing screening (medical) devices may need to be tested to see which one offers a better performance and whether any improvement in capability (quantified through the metric AUC) offered by one of the devices is statistically valid. The interest in machine vision is to test whether one sensor is better than the other in identifying the presence of a target [14, 15, 17, 18, 19]. If there is a way to undertake the experiment multiple times, we may be able to study the inferential statistics. Because of difficulties in recruiting large number of subjects available for screening or the high costs of setting up machine vision testing multiple times, we need to explore other ways to study the efficacy of the devices and sensors.

Computer simulation may be one of the ways to mitigate the issue of limited resources. But we have very little information on the underlying statistics and often we do not know whether the data collected from the sensors or screening of the subjects follow a Gaussian, Rayleigh, Rician, or any other distribution [10]. Therefore, we need to ensure that any simulation undertaken is non-parametric (densities are characterized in terms of its parameters and hence the term parametric is associated with simulations that rely on densities). A simple way to replicate the experiment without making any assumptions on the underlying statistics involves resampling. Bootstrapping allows empirical regeneration of samples of the outcomes of the experiment through continual resampling regardless of the underlying statistics [20].

A number of publications offer insight into bootstrapping. Most of these demonstrate the principle of bootstrapping through examples that examine the statistics of the population mean while others offer explanations on the procedure relying on commercial software for bootstrapping and follow up statistical analysis [4, 5, 6, 8, 9]. Very few, if any, provide a step-by-step description of the mechanics of bootstrapping when formulaic means are absent for verification such as in the case of AUC in a machine vision experiment or comparison of the AUCs in a machine vision experiment involving two sensors. For students in the engineering probability course, the implementation of bootstrapping needs to be illustrated at the very basic level so that they understand and appreciate its usefulness in data analytics to draw conclusions based on the data. This means that the procedure must contain elements of the traditional course material and the possibility of expanding their repertoire of knowledge to cover other topics that would have been very conceptual. Such topics include non-parametric hypothesis testing using z-tests, p-values and confidence intervals [10,20].

The topic of bootstrapping was introduced after students were exposed to topics in mathematical statistics (marginal, conditional, joint probabilities, and, Bayes' rule), one and two random variables, parametric hypothesis testing (chi square tests), receiver operating characteristics, etc. Students were also familiar with data analytics because they were required to solve one homework problem every week with a unique data set for every student (besides a set of common problems for the class). Students had already done problems in estimating positive predictive values, confusion matrix, area under the receiver operating characteristics curve, chi square testing involving multiple densities to determine the best fit, maximum likelihood estimation of parameters of densities, and statistical analysis of improvement in performance achieved through signal processing algorithms (arithmetic mean, geometric mean, maximum).

At Drexel university, every quarter consists of 10 weeks of classes followed by examinations during the eleventh week. The engineering probability course is offered every quarter (required course for students pursuing baccalaureate degrees in electrical engineering as well as computer engineering) as a 4-credit course with three hours of lecture followed by one hour of recitation every week. For the recitation sessions, the class is split into smaller sections (less than 30 students). During this past quarter, recitations were held in three separate sessions for a class of about 75 students. Lectures and recitations (three different non-overlapping sessions) were covered by the author while teaching assistants were responsible for grading the homework submissions.
The demos created are described next. They were implemented in Matlab (www.mathworks.com). Even though Matlab provides built-in function bootstrap(.) to perform bootstrapping, Matlab was only used for random number generation, general computations, and, plotting.

3 Creation of the demos and results

Three different types of demos were created. The first was a simple one involving the study of the population mean and the other two involved the

statistics of AUC. All the demos were done in class and students were provided with the data sets ahead of the lecture so that they could follow the procedure as the discussion proceeded.

The data set #1 consisted of M = 20 integer samples (numbers 1, 2, 3, 4, and 5) identified as vector $\overline{x}$. The resampling implies picking a number from this set, noting it down, returning the number back to the set. Another number is then taken, noted down and returned back to the set. This process is repeated until we have noted down M new samples. This constitutes the first bootstrap set. Since we are returning the number back to the original set, it can be understood that the numbers (1,2,3,4 or 5) may be repeated. If the original M samples are set in a single column, bootstrapping implies choosing any row randomly each time. In other words, we are choosing numbers randomly between 1 and M (row indices). The row indices for resampling to create the first bootstrap set is the vector $\overline{v}$

$$\overline{v} = ceil\left(rand\left(1, M\right) * M\right) . \quad (1)$$

In eqn. (1), rand(1,M) provides M uniform random numbers in [0,1]. These are scaled by M and ceil(.) provides integers between 1 and M. Use of the indices in eqn. (1) mimics the process of picking a ball from a box containing balls numbered 1, 2, 3, 4, 5, noting its value, putting the ball back, and repeating the procedure M times. The bootstrap set created, $\overline{y}$, is

$$\overline{y} = \overline{x}\left(\overline{v}\right) . \quad (2)$$

This process of creating a new index vector and a corresponding bootstrap set was repeated N (1000) times to generate N bootstrap sets. Fig. 1 shows the original set (bottom row) and 10 bootstrap sets (for convenience each set is shown as a row vector). The column at the right end provides the population mean of each bootstrap set and the original set. N bootstrap sets produce N population means and the mean and variance of the population mean can be calculated.

Figure 2 shows the histogram of the means, mean and variance of the population mean along with 95% confidence interval of the mean (range between 2.5 and 97.5 percentiles). It can be seen that the mean of the population mean and the population mean match (3.2). The population variance (1.9579) is approximately equal to 20 times the variance of the mean (0.089) as expected from the concept of the sample mean [20]. Thus, in this experiment, bootstrapping results were tested against formulaic method for obtaining the mean and variance of the population mean.

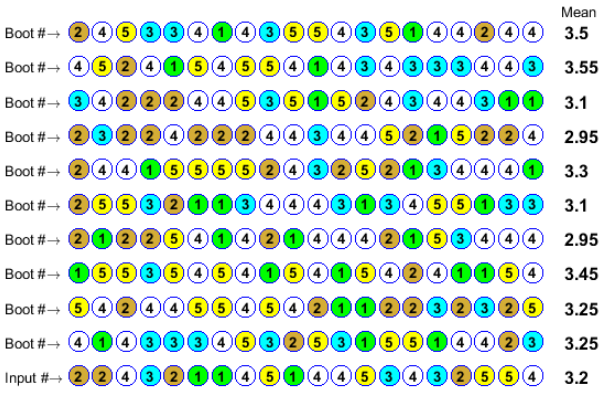

Fig. 1 Original data set (input) and 10 bootstrap sample sets. The mean of each set appears in the last column.

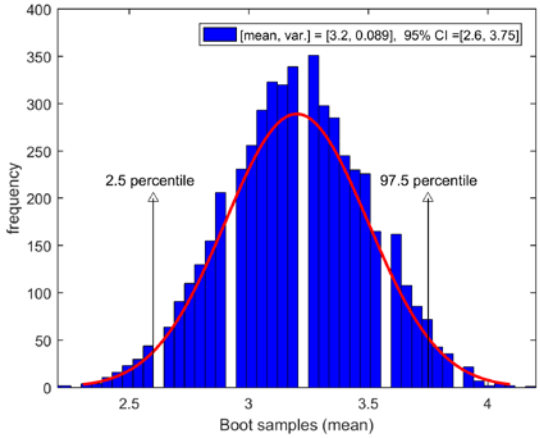

Fig. 2 Histogram of the population mean, its mean, variance and 95% confidence interval of the mean

While the bootstrapping procedure demonstrated with interest in the statistics of the mean is simple and straightforward and results verifiable formulaically, the statistical analysis of AUC does not lend itself to simple formulae. The demos on AUC and comparison of AUCs for two sensors were undertaken in an expanded way, first through the use a smaller size data set (16 samples) to explain all the intricacies and then repeated on a larger size data set (130 samples).

Consider the data collected in a machine vision experiment where a sensor was used to detect the presence of a target in its field of view. The backscattered power or amplitude using a trans-receiver (wireless, infra-red, or acoustic) provided the samples. The data used in this work was created through random number simulation ensuring that the

larger size data satisfied densities such as Rayleigh, Rician, Nakagami, gamma, Weibull etc. observed in wireless and other systems [21]. Sampled values represent data collected when a target was present in the field (hypothesis H_1) or target was absent in the field (hypothesis H_0). Figure 3 (first two columns) displays the data collected with N_1= 10 samples of target absent and N_2= 6 samples of target present. The values are respectively labeled '0' and '1' to indicate the two distinct categories (hypotheses). The step-by-step procedure for the generation of the ROC plot of this machine vision sensor is shown in Fig. 3 with the goal of drawing a plot of the probability of false alarm (deciding that a target is present given hypothesis H_0) and probability of detection (deciding that a target is present given hypothesis H_1) as the threshold for the decision is varied from the maximum of the sample values to '0'. The lowest value of '0' was chosen because the samples represented either power or amplitude and therefore, the range of the values would always be positive.

Target		Unsorted		Sorted		Threshold	Counts			
Absent	Present	Label	Value	Label	Value	T	N_C	N_F	P_D	P_F
1.1428	1.4308	0	1.1428	1	1.5498	1.5498	0	0	0	0
0.3511	1.4577	0	0.3511	1	1.4868	1.4868	1	0	0.167	0
0.9526	0.8309	0	0.9526	1	1.4577	1.4577	2	0	0.333	0
0.8165	1.1524	0	0.8165	1	1.4308	1.4308	3	0	0.5	0
0.8395	1.5498	0	0.8395	0	1.4065	1.4065	4	0	0.667	0
0.9911	1.4868	0	0.9911	0	1.3647	1.3647	4	1	0.667	0.1
1.3645		0	1.3645	0	1.3645	1.3645	4	2	0.667	0.2
1.3647		0	1.3647	1	1.1524	1.1524	4	3	0.667	0.3
0.8563		0	0.8563	0	1.1428	1.1428	5	3	0.833	0.3
1.4065		0	1.4065	0	0.9911	0.9911	5	4	0.833	0.4
		1	1.4308	0	0.9526	0.9526	5	5	0.833	0.5
		1	1.4577	0	0.8563	0.8563	5	6	0.833	0.6
N_1 = 10	N_2 = 6	1	0.8309	0	0.8395	0.8395	5	7	0.833	0.7
		1	1.1524	1	0.8309	0.8309	5	8	0.833	0.8
		1	1.5498	0	0.8165	0.8165	6	8	1	0.8
		1	1.4868	0	0.3511	0.3511	6	9	1	0.9
		[M x 2] matrix		⇓		0.0000	6	10	1	1
		$M = N_1 + N_2 = 16$		Descending order (Value)		M + 1 = 17 Values	1's above T	0's	N_C/N_2	N_F/N_1

Absent / Present

Figure 3 Steps involved in the creation of the receiver operating characteristics curve

The total number of sampled values M is 16. The pooled data in [M x 2] matrix (columns 3 and 4) was sorted in descending order of values (boxed matrix, columns 5 and 6). The threshold set was chosen as the sorted values with the concatenation of '0' at the bottom, resulting in (M+1) threshold values (column 7). The number of 1's above the threshold (N_C) and number of 0's above the threshold (N_F) from the sorted labels column (column 5) were counted and used to obtain probability of detection (P_D) and probability of false alarm (P_F) respectively to generate the ROC curve. Once the ROC curve was obtained, AUC was calculated using trapz(.) command in Matlab. The ROC plot is shown in Figure 4 along with the value of AUC (indicated as Az). None of the steps outlined in Figures 3 and 4 required the use of bootstrapping and students had already seen those steps earlier during the lecture on ROC [1]. Students also had completed individual homework assignments on ROC.

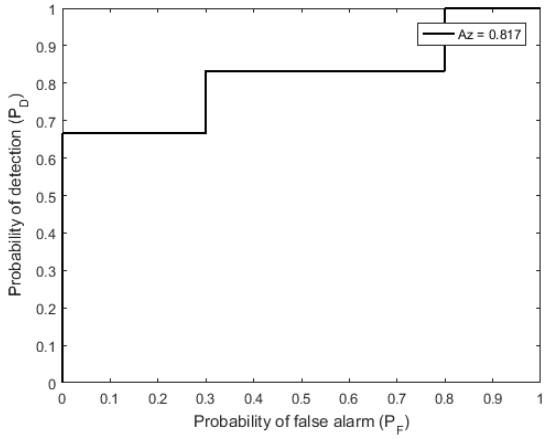

Fig. 4 ROC plot corresponding to the 16-sample set in Fig.3

The next step was to obtain the statistics of the AUC through bootstrapping. This matrix of labels and values (columns 3 and 4) in Fig. 3 was the input to the bootstrapping procedure. The data sets and bootstrapping procedure are illustrated in Table 1. The first two columns in Table 1 constitute the labels and the values. The matrix of these two columns formed the input to the bootstrapping procedure. Initially four bootstrap sets were created and the bootstrap indices are shown in Table 1. The indices were generated as mentioned earlier with M=16. Generation of a single bootstrap set required the choice of rows from the [M x 2] matrix (label and the corresponding value) matching the indices. The four bootstrap sets of the matrix are given in Table 1. Each bootstrap set was now used to obtain the corresponding ROC and AUC. The samples of the ROC plots corresponding to the four bootstrap sets are shown in Fig. 5. The bootstrapping was carried out 5000 times yielding 5000 samples of the AUCs. With the availability of 5000 samples, the mean, standard deviation and 95% confidence interval (between 2.5 and 97.5 percentile as seen earlier) of the mean were then calculated. Fig. 6 displays the result. The very broad range of the 95% confidence interval is a direct manifestation of the very low sample sizes.

Table 1 Bootstrapping procedure associated with the 16-sample set from Fig. 3 (four sets of indices and corresponding matrices of [label, value] shown)

input data		indices				bootstrap sets							
Label	value #1	1	2	3	4	#1		#2		#3		#4	
0	1.1428	11	13	16	12	1	1.4308	1	0.8309	1	1.4868	1	1.4577
0	0.3511	7	16	5	7	0	1.3645	1	1.4868	0	0.8395	0	1.3645
0	0.9526	14	6	16	4	1	1.1524	0	0.9911	1	1.4868	0	0.8165
0	0.8165	3	14	9	3	0	0.9526	1	1.1524	0	0.8563	0	0.9526
0	0.8395	2	3	8	5	0	0.3511	0	0.9526	0	1.3647	0	0.8395
0	0.9911	3	14	4	5	0	0.9526	1	1.1524	0	0.8165	0	0.8395
0	1.3645	6	5	14	7	0	0.9911	0	0.8395	1	1.1524	0	1.3645
0	1.3647	14	1	7	1	1	1.1524	0	1.1428	0	1.3645	0	1.1428
0	0.8563	11	4	2	7	1	1.4308	0	0.8165	0	0.3511	0	1.3645
0	1.4065	4	11	6	1	0	0.8165	1	1.4308	0	0.9911	0	1.1428
1	1.4308	12	11	13	16	1	1.4577	1	1.4308	1	0.8309	1	1.4868
1	1.4577	13	10	14	16	1	0.8309	0	1.4065	1	1.1524	1	1.4868
1	0.8309	9	3	1	10	0	0.8563	0	0.9526	0	1.1428	0	1.4065
1	1.1524	3	12	5	13	0	0.9526	1	1.4577	0	0.8395	1	0.8309
1	1.5498	4	1	14	5	0	0.8165	0	1.1428	1	1.1524	0	0.8395
1	1.4868	4	2	9	2	0	0.8165	0	0.3511	0	0.8563	0	0.3511

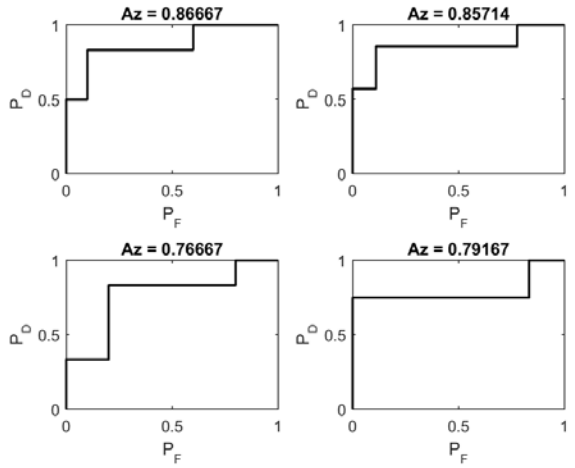

Fig. 5 Four ROC plots associated with the four bootstrap sets in Table 1

The next demo involved the comparison of the performances of two sensors that were part of the same machine vision or machine learning experiment. The goal was to see which sensor was the better of the two in terms of the AUC values. While the first set of measured values was seen in Table 1, Table 2 contains the details of the analysis with two sets. The first three columns in Table 2 represent the binary labels, value#1 (sensor #1) and value #2 (sensor #2). The ROC plots were obtained separately by pairing [label, value#1] and [label, value#2]. Fig. 7 displays the ROC curves corresponding to these two sets. Even though AUC (Az_2) from sensor#2 is higher than AUC (Az_1) from sensor#1, our interest is in assessing whether the difference in AUC values is statistically significant. This requires bootstrapping.

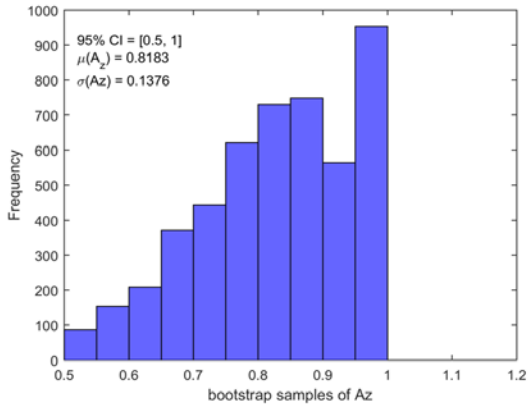

Fig. 6 Histogram of the bootstrap samples of the area under the ROC curve and its statistics for the data in Fig. 3

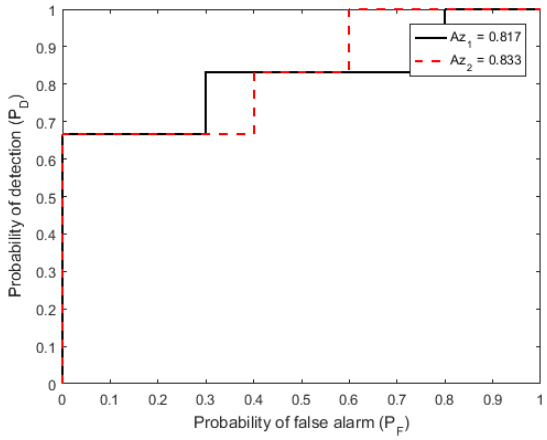

Fig. 7 ROC plots associated with the two data sets from Table 2

Table 2 illustrates the steps involved in performing the bootstrapping of the pared sets. The first three columns of Table 2, namely the [M x 3] matrix consisting of [label, value#1, value#2] with M = 16 was the input. Bootstrapping now involves resampling of the rows of this matrix. Bootstrap indices were generated as in the previous case and the first four bootstrap sets (paired) are shown in Table 2. Each bootstrap set is now used to obtain AUC values just as in Fig. 7. With N-bootstrap sets (N=5000), we now have N samples of the areas, Az_1 and Az_2, corresponding to sensors #1 and #2. With the resampling of rows containing paired values from sensor #1 and sensor #2, we retained the association (if it existed) between the two sets of values and this association would be reflected in the

correlation of the AUCs (if correlation between the two sensors existed).

The assessment of statistical significance entails the use of a z-test requiring the estimation of z_{score}. Following standard notation, the z_{score} of the two correlated data sets (in this case, Az_1 and Az_2) is [10, 20]

$$z_{score} = \frac{\langle Az_1 - Az_2 \rangle}{\sqrt{\mathrm{var}(Az_1 - Az_2)}} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}} = \frac{\Delta\mu}{\sigma_{12}} \quad (3)$$

In eqn. (3), <.> represents the statistical average and

$$\mu_1 = \langle Az_1 \rangle$$
$$\sigma_1^2 = \mathrm{var}(Az_1) = \langle Az_1^2 \rangle - \langle Az_1 \rangle^2 \quad (4)$$
$$\mu_2 = \langle Az_2 \rangle,$$
$$\sigma_2^2 = \mathrm{var}(Az_2) = \langle Az_2^2 \rangle - \langle Az_2 \rangle^2 \quad (5)$$
$$\Delta\mu = \mu_1 - \mu_2 \quad (6)$$
$$\sigma_{12}^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 \quad (7)$$
$$\rho = \frac{\langle Az_1 Az_2 \rangle - \langle Az_1 \rangle \langle Az_2 \rangle}{\sigma_1\sigma_2} \quad (8)$$

Equation (8) represents the correlation coefficient of the areas corresponding to the two sensors. We may perform the z-test in two ways, either as a two-tailed (two-sided) test or a single-tailed (one-sided) test. The former tests whether the difference in AUCs is statistically significant and the latter tests whether one of the sensors performs better than the other (in the present case, a lower or left-tailed test because $\Delta\mu$ is negative). For a two-tailed test, the critical value is 1.96 implying that z_{score} smaller than 1.96 suggest that the null hypothesis is not rejected, i. e., there is no statistically significant difference between the performances of two sensors (measured in terms of the respective AUCs) at a significance level of 0.05. The alternate hypothesis (the difference in areas is statistically significant at 0.05 level) cannot be rejected if z_{score} is larger than 1.96. For a single-tailed test (left-tailed), the critical value is -1.65. If z_{score} is smaller than -1.65, the alternate hypothesis (sensor # 2 performs better than sensor #1) cannot be rejected at a significance level of 0.05. If $\Delta\mu$ is positive, we will undertake a right-tailed test and in this case, z_{score} larger than 1.65 implies that the alternate hypothesis (sensor #1 performs better than sensor # 2) cannot be rejected at a significance level of 0.05. The p-value identified as the probability associated with null hypothesis, p(Null Hypothesis) is expressed as [10]

$$p(Null\ Hypothesis) = \begin{cases} 2\mathrm{Prob}(X > |z_{score}|), & \text{two-tailed} \\ \mathrm{Prob}(X < z_{score}), & \text{left-tailed}\ (z_{score} < 0) \\ \mathrm{Prob}(X > z_{score}), & \text{right-tailed}\ (z_{score} > 0) \end{cases} \quad (9)$$

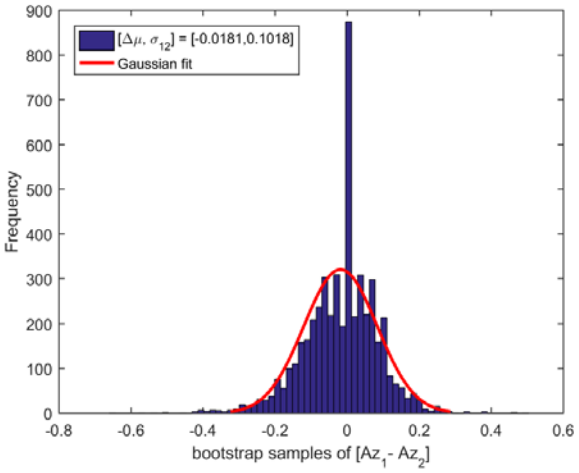

Figure 8 Histogram of the difference in area samples

In eqn. (9), X is a standard normal random variable of zero mean and variance of unity. Fig. 8 displays the histogram of the samples of the AUCs and fig. 9 provides the summary statistics of the comparison of AUCs corresponding to the two sensors. Regardless of whether one-sided or two-sided test is used, it is clear that the null hypothesis, (no difference between the performances of the two sensors) cannot be rejected. Even though the p-values are given, the z_{score} provides answers regarding whether classifier #2 (sensor #2) performs better than classifier #1 (sensor #1) as described in the paragraph above and z-tests are not necessary. The p-value of the two tailed test is approximately twice the p-value of the single tailed test as seen in Fig. 9.

The demos created so far offered students a reasonably clear set of steps involved in bootstrapping regardless of whether we are looking at the statistics of the population mean, AUC or comparison of the performance of two sensors (or diagnostic devices). However, the analysis based on AUCs done with a smaller data size offers only a very limited view. Students were also provided with larger sets of data from two sensors. In this case, the data consisted of 70 values of target absent and 60 values of target present. Fig. 10 displays the two ROC plots corresponding to the two sensors.

Classifier # 1 Area #1, $\langle Az_1 \rangle = \mu_1 = 0.8165$, $\sigma(Az_1) = \sigma_1 = 0.1402$

Classifier # 2 Area #2, $\langle Az_2 \rangle = \mu_2 = 0.8347$, $\sigma(Az_2) = \sigma_2 = 0.1199$

correlation coefficient ($[Az_1, Az_2]$) = $\rho = 0.7044$

$$z_{score} = \frac{mean(Az_1 - Az_2)}{std.\ dev(Az_1 - Az_2)} = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}} = \frac{\Delta\mu}{\sigma_{12}}$$

$\Delta\mu$ = -0.0181 σ_{12} = 0.1018

z_{score} = -0.1782 (z_{score} is negative !)

z-test (two-sided): Alternate Hypothesis: Statistically significant difference between #1 and #2
Null hypothesis: no difference between #1 and #2
p(Null Hypothesis) = 0.8586

z-test (one-sided) [left tail, $z_{score} < 0$]: Alternate Hypothesis: #2 better than #1
Null Hypothesis: no difference between # 2 and #1
p(Null Hypothesis) = 0.42928

Fig. 9 Summary statistics of the analysis of the areas under the ROC curves from the two data sets

The bootstrapping (N=5000) procedure was now undertaken as described earlier and results are shown in Figs. 11, 12 and 13. Fig. 11 displays the histogram of the bootstrap samples of AUCs from the sensors, their mean and the respective standard deviations. Fig. 12 displays the statistics of the difference in AUCs along with the Gaussian fit and the mean and standard deviation of the difference in AUCs. Fig. 13 displays the summary of the statistical analysis. The value of the correlation coefficient is also provided showing that the sensors are correlated. Fig. 13 shows that alternate hypothesis (using either test) could not be rejected at a significance level of 0.05. Appropriate theoretical aspects in Fig. 9 and Fig. 13 were generated automatically in Matlab to offer a formal and complete picture of the analysis without the need for any supplementary materials or notes. This automated generation of the summary results allowed the students to see the variability in the statistics of AUCs as the demos were run repeatedly (even though the basic conclusions did not change from iteration to iteration).

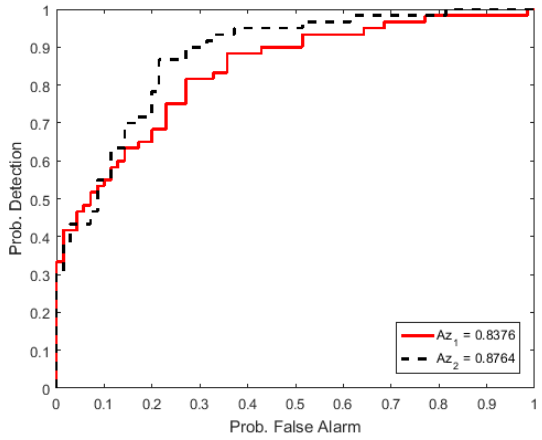

Figure 10 ROC plots associated with the two larger data sets

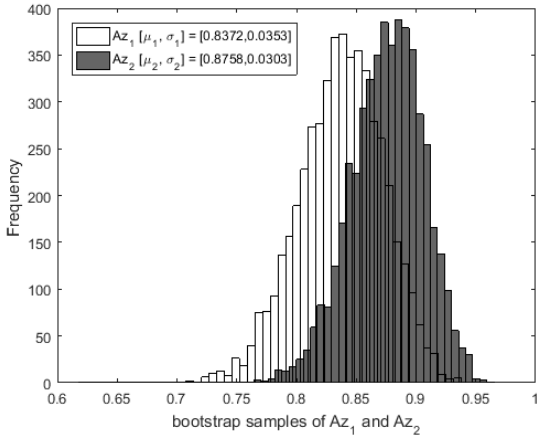

Fig. 11 Histograms of the bootstrap samples of the two areas

The step-by-step procedure of bootstrapping demonstrated through so far still lacked verification from external sources. While the first demo of the study of the population mean does not require any extra proof (the mean and standard deviation of the population have formulaic answers as explained earlier), the analysis of the AUC of a single classifier (sensor) and the comparison of the AUCs from two classifiers (sensors) would not be pedagogically complete without verification of the results from other sources that do not use bootstrapping.

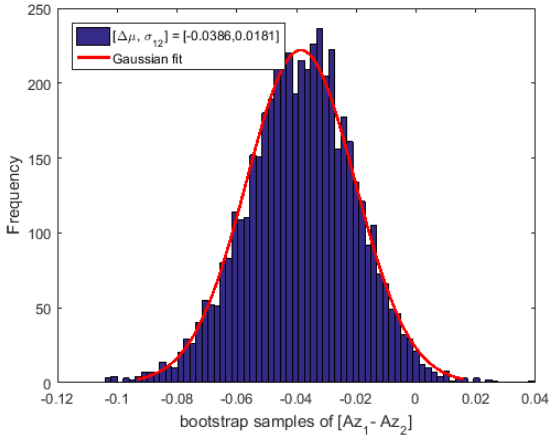

Fig. 12 Histogram of the difference in area samples corresponding to the data sets used in Figure 10

Formulae for estimating the standard deviation of the AUC and the z_{score} for the comparison AUCs obtained from two classifiers are available in

literature [14, 15]. While the statistical analysis of the AUC of a single sensor is relatively straightforward, the comparison of the performance of two correlated sensors is far more complex requiring access to a table of correlation values provided [15] or advanced statistical methods [14]. A simpler solution was to use commercial software for verification purposes that utilized both these methodologies.

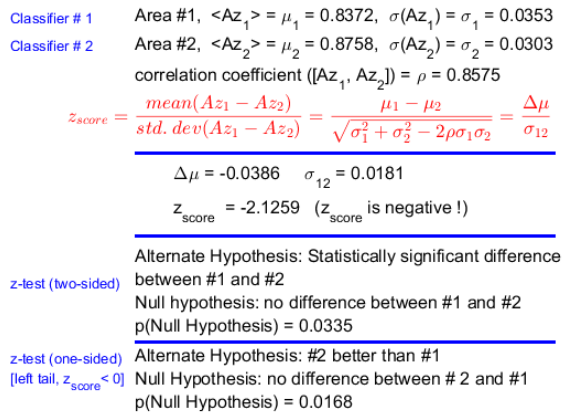

Fig. 13 Summary statistics of the analysis of the areas under the ROC curves from the two data sets in

Medcalc® software (www.medcalc.org) was used in this work. It provided the option of statistical analysis undertaken using formulae suggested by both research groups for the estimation of standard deviation AUC and z_{score} for the comparison of AUCs from two sensors [14, 15]. The results from Medcalc (shown during the lecture) are provided in the supplementary materials (Pages 2-5). They clearly show that the statistical parameters estimated from the bootstrapping procedure demonstrated in this work align very well with the formulaic results (Medcalc only performs a two-sided z-test). The larger data set used in this work is also included in the supplementary materials (Pages 8-11).

4 Discussion and Conclusions

Demos specifically created to elucidate the concepts and applications of bootstrapping were illustrated. The use of the smaller size data made it possible for the students to follow the inner workings of the bootstrapping procedure and the associated statistical explorations while highlighting the perils of having smaller sizes resulting in higher values of the standard deviations for the AUCs. The inclusion of the formulaic analysis of both smaller size and larger size data using Medcalc established two important points, the validation of the results of bootstrapping undertaken here and the simplicity offered by the bootstrapping procedure which relied only on a uniform random number generator. It was also clear that the ensuing statistical analysis could be carried out through the use of simple tools in Matlab.

The demos also offered an opportunity for the students to reinforce some of the concepts learned earlier in the course such as the z_{score} (covered in connection with the Gaussian random variables), correlation of random variables, variance of the sum and difference of correlated random variables, Gaussian fits, and, learn newer topics such as the z-test and the association between p-values and significance levels. Following the demo, students were assigned homework problems (individual data sets) on bootstrapping. They were allowed to use the Matlab command bootstrap(.) to create bootstrap sets for the analysis. Homework problems on paired sensor analysis were not assigned because of the lack of a large number of paired sets for individual students and it is expected that such a homework problem will be ready for the next academic year.
While it was possible to observe the reactions of the students during the lecture from successful completion of the individual homework assignments and subsequent interactions outside the class, an indirect assessment of the students' views was undertaken. Students were surveyed in week#1 in seven areas in the form of seven questions with Likert style scores of 1, 2, 3, and 4 corresponding to their level of knowledge [very little, some, well, very well], with questions on ROC (Q#5) and bootstrapping (Q#6). The same survey was given on the last day of classes. Even though the course had 73 students, 68 students were present during the survey in week#1 and 66 students were present on the last day of classes. Surveys were anonymous and the survey questionnaire is included in the supplementary materials (Page 6).

For Q#5, the mean and standard deviation of the score were 1.134 and 0.4198 respectively in week#1 and 2.8336 and 0.8392 respectively in week#10. For Q#6, the mean and standard deviation of the score were 1.1765 and 0.4869 respectively in week#1 and 2.4848 and 0.8813 respectively in week#10. These values suggest an enhanced understanding of the topics during the instruction. Even though the views of the students are subjective, they still offered some positive perspective on the data analytics portion of

the course. Since the surveys were anonymous, it was not possible to undertake a paired t-test to see the increased average scores were statically significant. Simple t-tests and Wilcoxon rank sum tests [10] showed that the alternate hypothesis that students gained understanding could not be rejected (p-value <0.001). The bar charts displaying the survey results pertaining to Q#5 and Q#6 are included in the supplementary materials (Page 7).

The demos created to articulate the details of bootstrapping to expand the topics in data analytics appear to have been beneficial to the students. Author is planning to incorporate additional topics on non-parametric hypothesis testing (z-tests, t-tests, Wilcoxon rank sum test etc.) during the upcoming years using similar methodology developed in this work.

References:

[1] P. M. Shankar, Pedagogy of Bayes' rule, confusion matrix, transition matrix, and receiver operating characteristics, *Computer Applications in Engineering Education*, 2019, DOI: 10.1002/cae.22093

[2] T. C. Hesterberg, What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum, *The American Statistician*, Vol. 69, 2015, pp. 371-386.

[3] A. G. Munoz-Repiso, and F. J. Tejedor, The incorporation of ICT in higher education. The contribution of ROC curves in the graphic visualization of differences in the analysis of the variables, *British J. Educational Technology*, Vol. 43, 2012, pp. 901-919.

[4] P. Bertail, S. J. Clémençcon and N. Vayatis, On bootstrapping the ROC curve. Proc. of the 21st International Conference on Neural Information Processing Systems (NIPS 2008), pp. 137-144.

[5] C. D. Brown and H. T. Davis, Receiver operating characteristics curves and related decision measures: A tutorial, *Chemometric and Intelligent Laboratory Systems*, Vol. 80, 2006, pp. 24-38.

[6] N. Hu and R. B. Dannenberg, Bootstrap learning for accurate onset detection, *Machine Learning*, Vol. 65, 2006, pp. 457-471.

[7] H. Liu, G. Li, W. G. Cumberland, T. Wu, Testing statistical significance of the area under a receiving operating characteristics curve for repeated measures design with bootstrapping, *Journal of Data Science*, Vol. 3, 2005, pp. 257-278.

[8] A. Moise, B. Clément, P. Ducimetière, P. and M. G. Bourassa, Comparison of receiver operating curves derived from the same population: a bootstrapping approach, *Computers and Biomedical Research*, Vol. 18, 1985, pp. 125-131.

[9] S. Yue and P. Pilon, A comparison of the power of the t-test, Mann-Kendall and bootstrap tests for trend detection, *Hydrological Sciences–Journal–des Sciences Hydrologiques*, Vol. 49, 2004, pp. 21-37.

[10] V. K. Rohatgi and A. K. Saleh, *An Introduction to Probability and Statistics*, Wiley, 2001.

[11] J. Eng, Teaching receiver operating characteristic analysis: An interactive laboratory exercise, *Academic Radiology*, Vol. 19, 2012, pp. 1452-1456.

[12] D. J. Hand and R. J. Till, A simple generalization of the area under the ROC curve for multiple class classification problems, *Machine Learning*, Vol. 45, 2001, pp. 171-186.

[13] T. Fawcett, An introduction to ROC analysis, *Pattern Recognition Letters*, Vol. 27, 2006, pp. 861-874.

[14] E. R. DeLong, D. M. DeLong and D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics*, Vol. 44, 1988, pp. 837-845.

[15] J. A. Hanley and B. J. McNeil, A method for comparing areas under the receiver operating characteristic curves derived from the same cases, *Radiology*, Vol. 148, 1983, pp. 839-843.

[16] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, Vol. 143, 1982, pp. 29-36.

[17] D. D. Dorfman, K. S. Bernbaum, and R. V. Lenth, Multireader, multicase receiver operating characteristic methodology: A bootstrap analysis, *Academic Radiology*, Vol. 2, 1995, pp. 626-633.

[18] A. D. Kester and F. Buntinx, Meta-analysis of ROC curves, *Medical Decision Making*, Vol. 20, 2000, pp. 430-439.

[19] N. A. Obuchowski, New methodological tools for multiple-reader ROC studies, *Radiology*, Vol. 243, 2007, pp. 10-12.

[20] B. Efron and R. Tibshirani, Bootstrap methods for standard errors, confidence Intervals, and other measures of statistical accuracy, *Statistical Science*, Vol. 1, 1986, pp. 54-75.

[21] P. M. Shankar, *Fading and Shadowing in Wireless Systems*, Second Edition, Springer, 2017.

Table 2 Bootstrapping procedure associated with the data collected from two sensors

input data			bootstrap sets											
Label	value#1	value#2	# 1			#2			#3			#4		
0	1.1428	1.1328	0	0.8395	0.8455	1	1.4577	1.4277	0	1.4065	1.4134	0	0.8165	0.4045
0	0.3511	1.0526	0	1.1428	1.1328	1	1.5498	1.5147	0	1.3645	1.2685	0	1.3647	1.351
0	0.9526	0.4511	1	1.4308	1.5308	0	1.3645	1.2685	0	0.9526	0.4511	0	1.3645	1.2685
0	0.8165	0.4045	0	0.8395	0.8455	0	0.3511	1.0526	0	0.9911	0.9345	0	0.8165	0.4045
0	0.8395	0.8455	1	1.4308	1.5308	0	0.3511	1.0526	1	1.5498	1.5147	0	1.3647	1.351
0	0.9911	0.9345	0	1.1428	1.1328	0	1.4065	1.4134	0	0.8395	0.8455	1	1.5498	1.5147
0	1.3645	1.2685	0	0.9911	0.9345	0	0.9911	0.9345	0	0.8563	0.8293	0	1.3647	1.351
0	1.3647	1.351	0	1.3647	1.351	1	1.5498	1.5147	0	0.9911	0.9345	0	1.1428	1.1328
0	0.8563	0.8293	0	0.8395	0.8455	0	0.8395	0.8455	0	1.3647	1.351	0	0.8165	0.4045
0	1.4065	1.4134	0	0.8395	0.8455	0	0.9911	0.9345	1	1.4868	1.5241	0	1.1428	1.1328
1	1.4308	1.5308	0	1.1428	1.1328	0	1.3645	1.2685	0	0.8165	0.4045	1	1.4577	1.4277
1	1.4577	1.4277	1	1.4868	1.5241	0	1.4065	1.4134	0	0.8165	0.4045	1	1.5498	1.5147
1	0.8309	1.1224	0	1.3645	1.2685	0	1.3645	1.2685	1	1.4577	1.4277	1	1.4868	1.5241
1	1.1524	0.9233	0	0.8165	0.4045	0	0.8563	0.8293	0	0.8395	0.8455	1	1.4308	1.5308
1	1.5498	1.5147	1	1.4868	1.5241	0	0.9526	0.4511	0	0.9526	0.4511	1	1.1524	0.9233
1	1.4868	1.5241	0	0.8563	0.8293	0	0.8395	0.8455	0	0.8395	0.8455	0	0.8395	0.8455