Documentation for Program for Rate Estimation and Statistical Summaries (PRESS)

Michael E. Schuckers, C. J. Knickerbocker
Department of Mathematics, Computer Science and Statistics
St Lawrence University, Canton, NY 13617
and
Center for Identification Technology Research,
West Virginia University

schuckers@stlawu.edu
PRESS *v1.0*

September 8, 2004

1 Getting started with PRESS

PRESS is a tool to help researchers to analyze data collected on biometric authentication devices. It has been created through generous funding from the Center for Identification Techology Research (CITeR) and from St. Lawrence University. PRESS is designed to simplify the analysis of bio-authentication data by making it easy for the user to create many of the basic statistical summaries in common usage. These include: confidence intervals, genuine vs. imposter histograms, and ROC curves. In addition PRESS has a tool for determining the number of individuals that need to be tested under certain specified conditions. PRESS handles data either in text or Excel format. PRESS handles data either in text or Excel format. It has been designed by Dr. Michael

Schuckers (St. Lawrence University) and coded by Nona Mramba (University of Maryland) and C. J. Knickerbocker (St. Lawrence University). The version documented here is a free research version and no support other than the documentation provided below is available to non-CITeR members. CITeR members receive free upgrades and advance copies of PRESS.

The documentation presented here is organized in the following way. This section provides some overview topics and gets the user started with the basics of using PRESS. Section 2 described the possible formats for data that is used by PRESS. Common features among the five tabs are found in Section 3. The next Sections 4, 5, 6 and 7 go through the various tabs and describe the functions that each performs. Section 8 describes a tool for converting files to the file type that PRESS uses most efficiently. Finally there is an appendix which gives additional details about the confidence interval methodology used by PRESS.

To begin press, double-click on *PRESS.exe*. The software consists of 5 tabs each performs a different function. The basics steps for each function are given below and illustrated by screenshots of the confidence interval tabs:

Before proceeding it is necessary to define a few terms.

Define an **individual** is an image comparison pair. This is a broad use of this term. This can be either a cross-comparison which is using images or templates from two different persons *or* it can be a comparison of an image against a template for a single person.Define an **attempt** as a single comparison for an individual. Thus there are, generally, a different number of attempts per individual. (It is possible to have the same number of attempts per individual.)

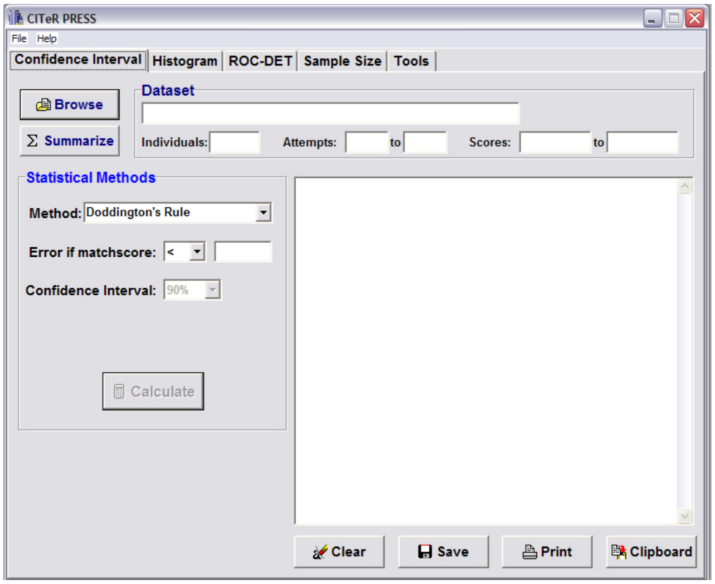

Figure 1: Getting Started

Define a **match score** as the value that the matching algorithm returns as a result of comparing two different biometric images. It is this value that is compared to a threshold for making a determination of correct or incorrect decision. Define **error rate** as the proportion of errors for some group. The observed error rate is the number of errors divided by the total number of attempts. The population error rate is a parameter (in the statistical sense) for some group to which we would like to extrapolate.

Step 1: Click the browse button to select a data set to be analyzed. These files can be either an Excel workbook or a PRESS formatted file (see help Section 2). Figures 1 and 2 illustrate this process.

Step 2: If an Excel spreadsheet is selected a pull down box will appear with

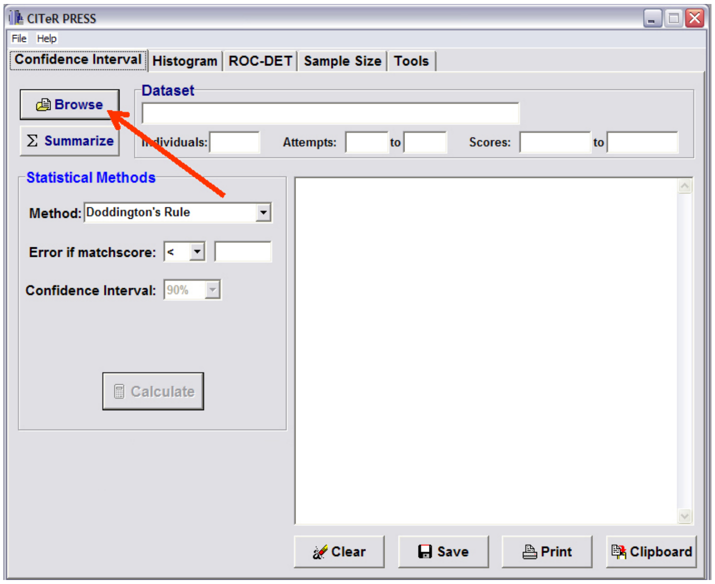

Figure 2: Click here to browse for a file

a list of the available spreadsheets. Select the worksheet containing the data of interest. See Figure 3.

Step 3: Having chosen a data set it is necessary to summarize that file. There are two purposes for this. First, it ensures the user is working on the correct data and, second, it gives a cursory statistical summary of the data. Figure 4 illustrates this.

Step 4: The user can now perform the computation for that page. It may be necessary to specify different values for other methods. This is illustrated by Figure 5

For the sample size function, it is not necessary to specify a file, rather the user enters values for a set of parameters. The calculation that is used there is based on the equation published in [Schuckers, 2003].

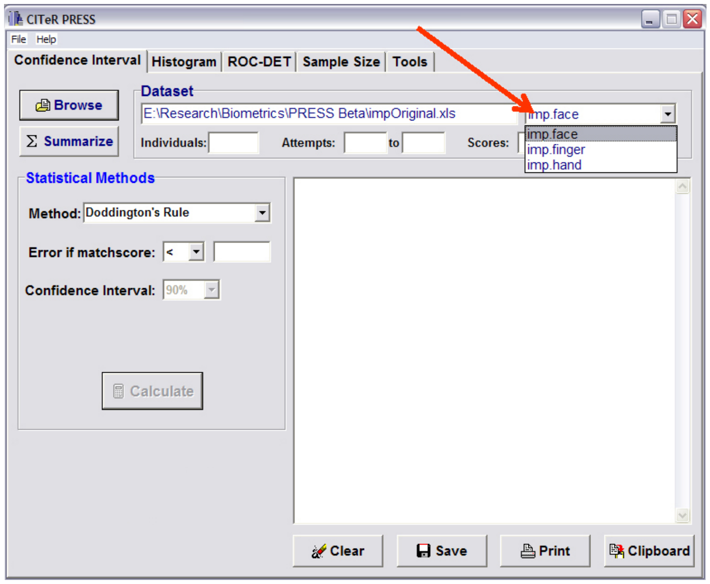

Figure 3: Choosing a worksheet for an Excel file

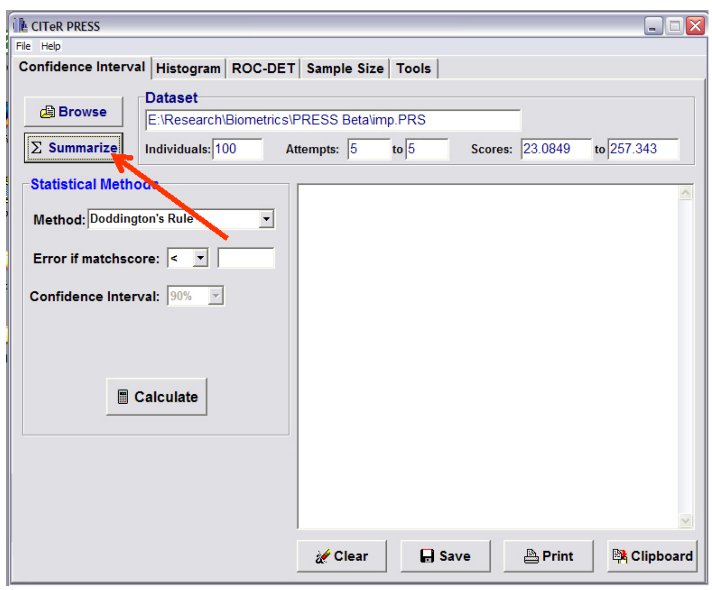

Figure 4: Summarizing the selected dataset

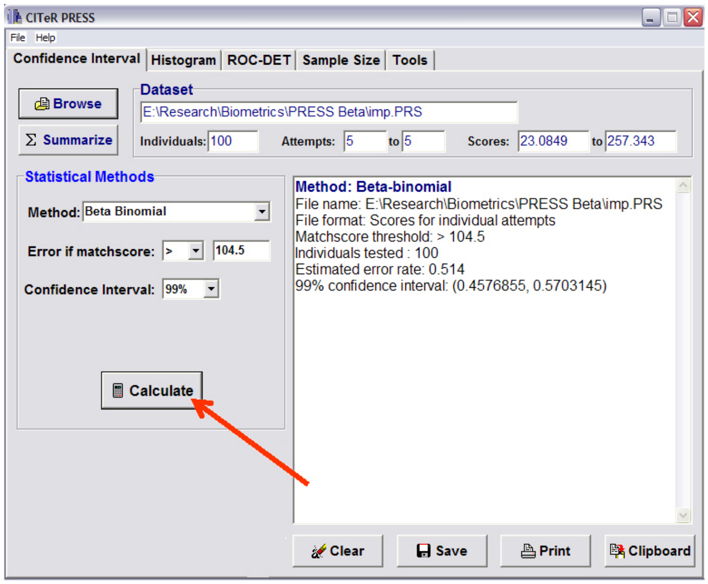

Figure 5: Click here to calculate a confidence interval for the error rate

2 Data Format

PRESS is capable of reading both text files and Excel spreadsheet files. There are various text file formats used by the PRESS software. All use *.prs* as the extension and have similar formats. Below we describe the different forms for the data that are readable by PRESS.

2.1 Form 1

This format is used for "raw" match scores. Each record or line in the file contains the "attempts" for one individual. For Excel files this corresponds to each row of the worksheet. The attempts are numeric values separated by a blank space or a tab. This data format can also handle processed data and in that case each "attempt" should be registered as either a zero (correct decision) or as a one (error).

Example:
Two individuals with 2 and 3 trials respectively.
12.343 47.24
23.2 24.3323 87.33

Example:
Three individuals with 4,3 and 5 trials respectively.
0 1 0 1
0 1 1
0 0 0 1 0

2.2 Form 2

This format is used for "processed" scores where each record represents the number of "successes" for an individual. Here successes are often the number of errors since we are estimating the overall error rate. Each error is a hit for that rate. The first record of the file represents the number of attempts

Example:
Successes counts for 5 individuals each of which has made 47 attempts.
47
12
24
6
17
22

Example:
Successes counts for 10 individuals each of which has made 8 attempts.
8
2

0
1
3
2
0
1
1
2
2

2.3 Form 3:

This is similar to the previous format except the number of attempts varies with each individual. Thus the data is stored in two columns.

Example:
Success counts for 3 individuals. The first person made 47 attempts with 4 successes, the second person made 41 attempts with 2 successes and the third person made 51 attempts with 5 successes.
4 47
2 41
5 51

As a consequence of this formatting it is may be necessary to adjust the data form. This is particularly the case if all of the individuals in the data have either one or two attempts. If there is only a single column, PRESS assumes the data takes form 2; if there are two columns PRESS assumes the data is in form 3. Because of the way that PRESS reads the data, this type of data , if entered as Form 1 data, will be read as either form 2 or form 3 data. Consider the following:

Example:
Suppose that we test seven people two times each. The raw data is given below using form 1.

56 89
4 24

54 92
52 17
26 48
84 90
12 45

The problem with this is that the data will be read not as raw data with two attempts each but rather as form 3 data with the first person having *not* two attempts but 89 attempts.

Example:
Thus if all the individuals in a particular processed data set have the same number of attempts then the data can be represented in several ways. Consider the following data using form 1 from eight different individuals each tested four times :

0 0 1 0
0 1 1 0
0 0 0 0
0 0 0 0
0 1 0 0
1 0 0 0
0 0 0 0
0 0 0 1

We could represent this data following form 2 as:
4
1
2
0
0
1
1
0
1

Or following form 3 as:
1 4
2 4
0 4

0 4
1 4
1 4
0 4
1 4

Only if all individuals have the same number of attempts is it possible to put data in all three forms. If the number of attempts differs among individuals, then only forms 1 and 3 are appropriate

2.4 Additional formatting notes

The software reads ***all*** the data from a selected worksheet so any data or computations in the worksheet must be removed before PRESS can process the data. Reading data from an Excel spreadsheet is slow. The program can read approximately 1500 data cells per second. Converting the spreadsheet to a PRESS format can speed this up by a factor of 4. The consequences of putting data in an incorrect form are that the confidence intervals created here will be incorrect. These intervals depend on accurately understanding the similarity of errors within an individual and consequently it is important to get the format correct or the confidence intervals will be inappropriate.

3 Common functions

There are several actions that are basic to Windows-based programs and that are used here by several PRESS functions. Thus, we introduce them here rather than repeating them in each of the following sections. In PRESS, these actions are represented by buttons at the bottom of each function.

Print

This button takes the output for a particular PRESS function and sends it to a printer attached to the users computer.

Save

The purpose of this button is to allow the user to save output to a file. For confidence interval output the default is a *.txt* file, while graphical output has bitmap *.bmp* as the default format.

Clipboard

This button makes the output available for pasting into another document. Thus, ROC curves or histogram plots can be pasted into Word documents for example.

Plot

This button is for getting graphical output in the histogram and ROC-DET functions. Changes particular specifications on those functions pages will not result in changes to the graph until the Plot button is clicked.

Clear

Only found in the confidence interval function, this button clears the text window to allow the user to have a less cluttered output window or to start a different analysis.

4 Confidence Interval

Above we used the confidence interval tab to illustrate reading data into PRESS. Here we will add additional details regarding this function. Details on the confidence interval methodology themselves can be found in the Appendix below.

Having found the particular data to be used and having summarized it, it is necessary to take a few more steps. First it is necessary to choose the particular methodology for making a confidence interval. Presently there are six methods that are options: Balanced Repeated Replicates, Best Practices, Beta-binomial, Doddington's Rule, Logit Beta-binomial and Subsets Bootstrap. We strongly suggest that users make themselves aware of the literature regarding these methods and their relative performance. Using one of the two resampling methods will result in an additional dropdown menu appearing for the number of replicates. The options here are 1000, 5000 and 10000. We recommend using 10000, unless the number of individuals being compared is very large say more than 5000. After choosing a method, PRESS needs to know what constitutes an error. Specify this by giving the inequality which determines the decision rule. Next it necessary to choose a confidence level. The current options for this are $90\%, 95\%, 99\%$ and 99.9%. Having followed the above steps, press the calculate button. Doing so, results in PRESS returning the desired confidence interval along with some additional information.

Example:
Method: Best Practices
File name: $A : test1.xls$
Worksheet name: Genuine
File format: Scores for individual attempts
Matchscore threshold: ≥ 40
Individuals tested : 50
Estimated error rate: 0.0164609
90% confidence interval: (0.003543984, 0.02937783)

Thus for this data on 50 individuals, we can be 90% confident that the population error rate is between 0.00354 and 0.02937 for the threshold specified here. Note that the output from PRESS includes the file (and the worksheet in the case of an Excel file), the format, the decision rule, the number of individuals

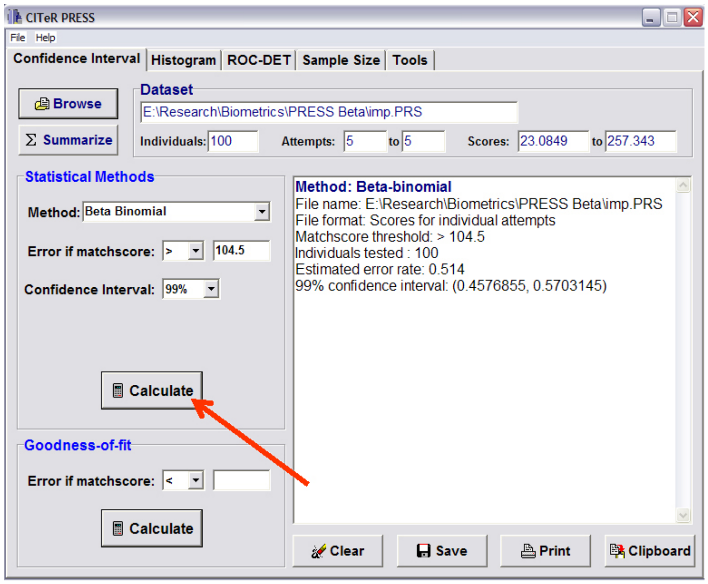

Figure 6: Example Confidence Interval Results

tested and the estimated error rate.

5 Histogram

This tab is meant to show the spread and shape of distributions of match scores. The primary intended usage here is to create a plot to compare genuine and imposter distributions. However this function is flexible enough that it can be used to make histograms of a single distribution. For histogram data, it is necessary to use raw data or the histograms will be relatively meaningless. Upon opening the histogram function, the screen should appear as it does in Figure 7.

As with the confidence interval function it is necessary to specify the location

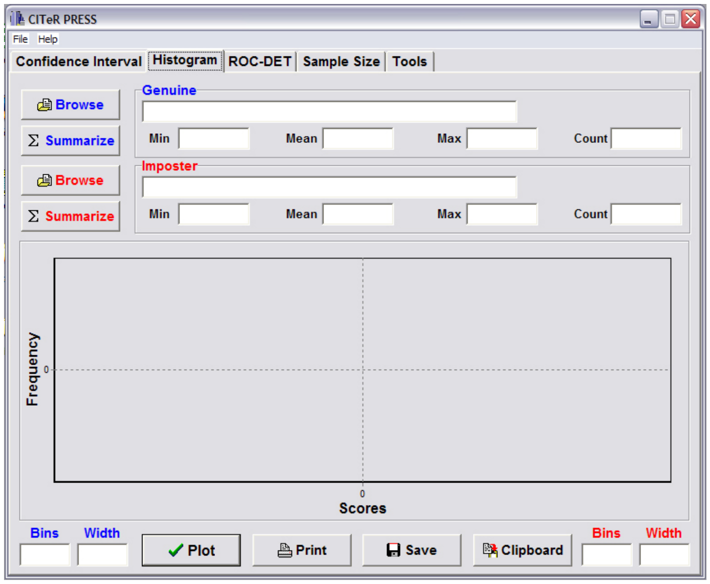

Figure 7: Histogram Function

of the data that will be plotted. Having selected the data and summarized it, simply click on **Plot** to graph the data. Note again that it is possible to only graph a single set of data. An example of a histogram plot can be found in Figure 8

In creating a histogram it is necessary to determine how many bins are to be plotted. PRESS uses a default algorithm for choosing the number of bins. However, we allow the user to alter the number of bins by typing in the number of bins manually. The minimum number of bins that is allowed is 10. Figure 9 illustrates where to enter the number of bins manually.

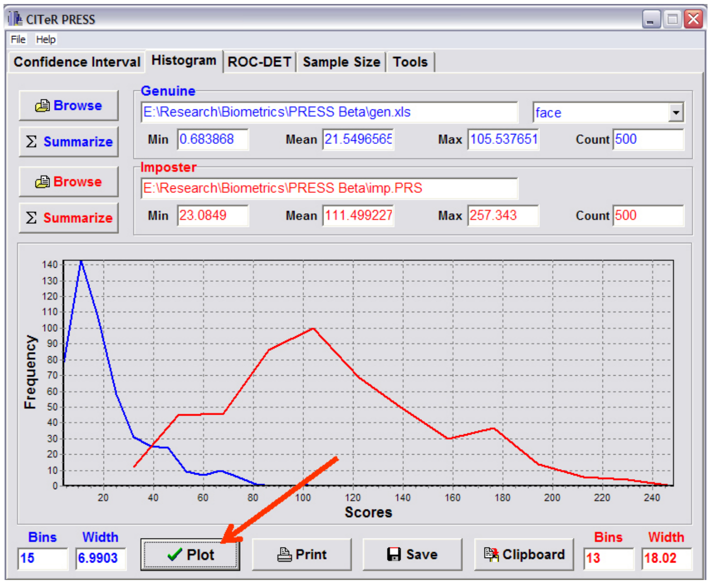

Figure 8: Plotted Histogram

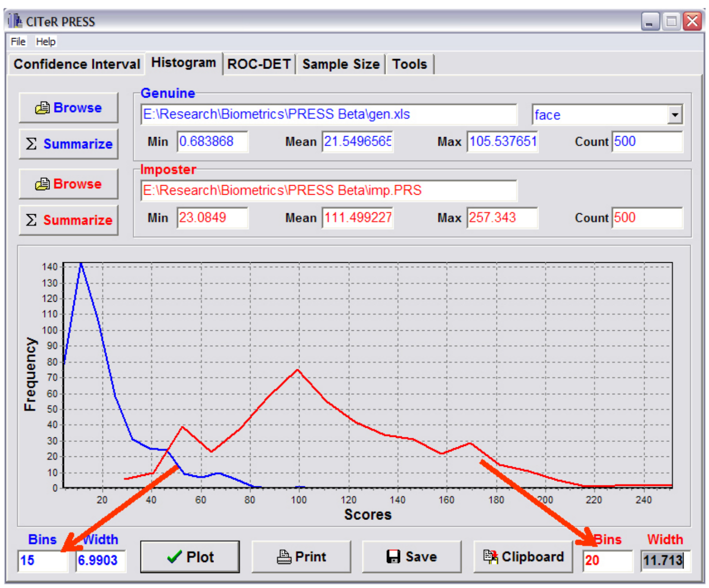

Figure 9: Bin entries for changing bin widths

6 ROC-DET Curves

The purpose of a receiver operating characteristic (ROC) curve or a Detection Error Tradeoff (DET) curve is to show the overall performance of a device across many threshold values. To permit flexibility PRESS does not differentiate between ROC curves, DET curves, or any of the other variations on this plot. PRESS allows the user to specify the graph and the scale for making these plots. It is left to the reader to recognize the axes and interpret the graph based on this.

To plot the basic curve it is necessary to import data for both a set of imposters and a set of genuine users. In addition it is necessary to specify a decision rule for the data and the number of thresholds to be checked. Figure 10 shows where to specify this information. Having told PRESS which data sets to use, the graph is made by pressing the plot button (Figure 11).

There are several dynamic features of this graph. First along with the graph the equal error rate or ERR is output with the data, Figure 12. Second, it is possible to alter the axes. One way to do this is to alter the scale of the axes. This can be done by choosing the radio buttons marked *FAR Axis* and *FRR Axis*. See Figure 13. Third, there are two options for the characters to be plotted. The ROC curve can either be plotted using points for a set of thresholds or line segments connecting those points can be plotted. Figure 14 shows how to toggle between these two choices. In addition to these it is possible to switch the X and Y axes by clicking on the *Switch Axes* checkbox (Figure 15). Finally, the user can "zoom in" on any portion of the ROC/EER graph by holding the

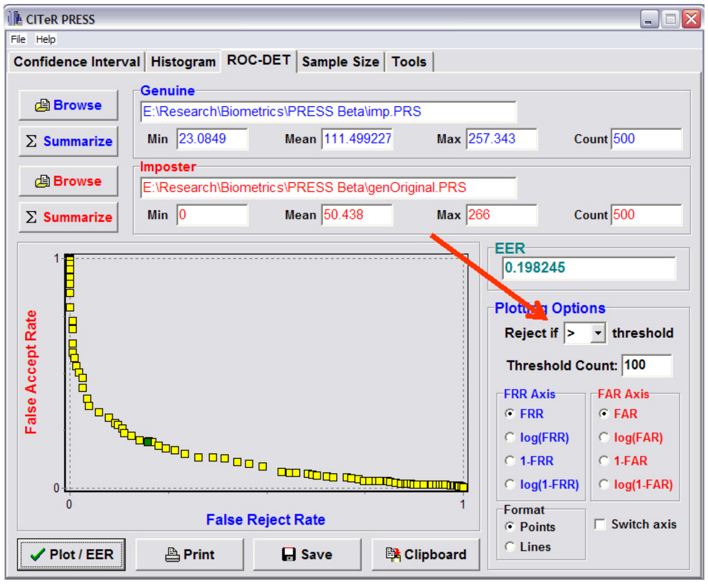

Figure 10: Choosing a decision rule

right mouse button down and drag it from the upper left corner to the lower right corner of the desired rectangle. The bounds of the highlighted region will be the bounds for the new graph. The original size can be restored by holding the left mouse button down and dragging from right to left. By holding the right mouse button down the use can shift the curve around the graph area.

7 Sample Size Calculations

This function differs from the previous three in that it does not require PRESS to read data in. Instead the point of this tab is to determine the number of individuals needed to create a confidence interval with a certain upper bound.

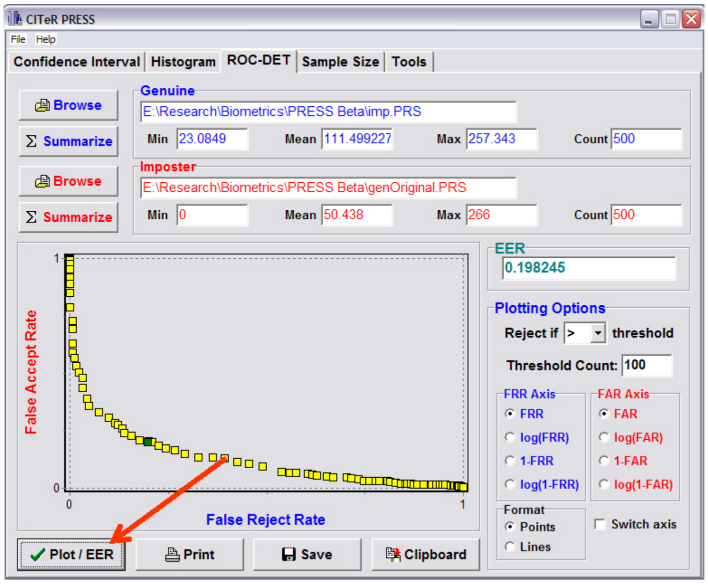

Figure 11: Plot Button

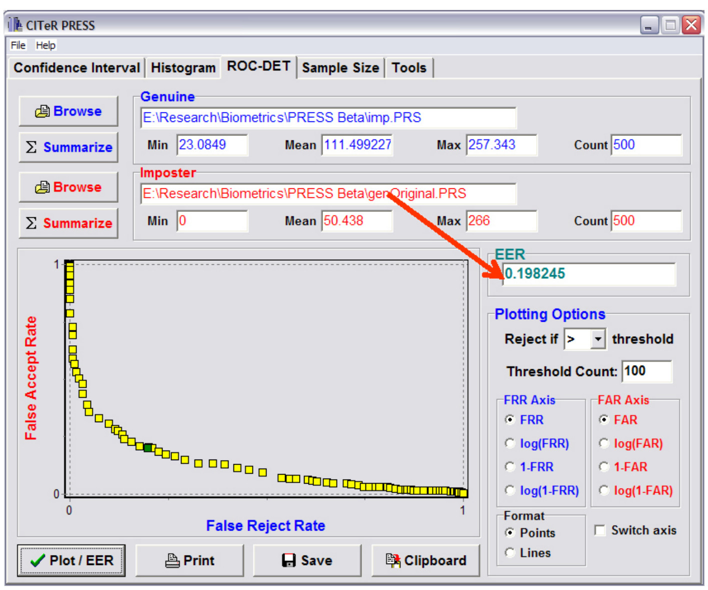

Figure 12: Location of the outputted EER

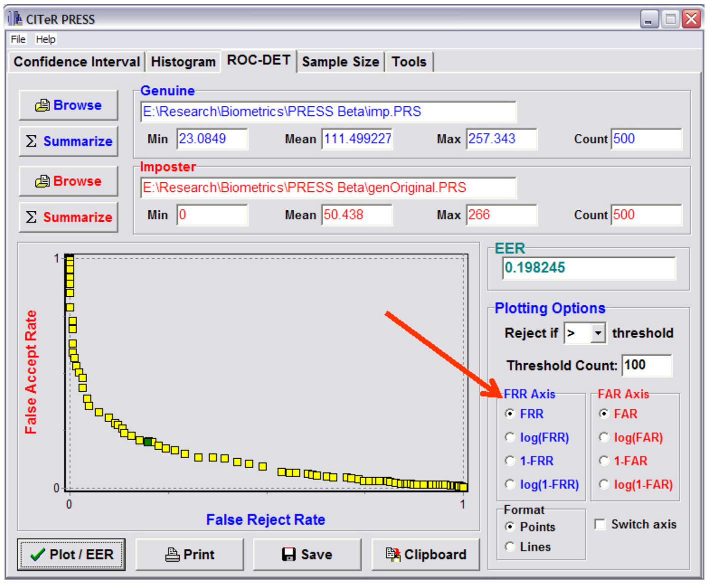

Figure 13: Switching Axes

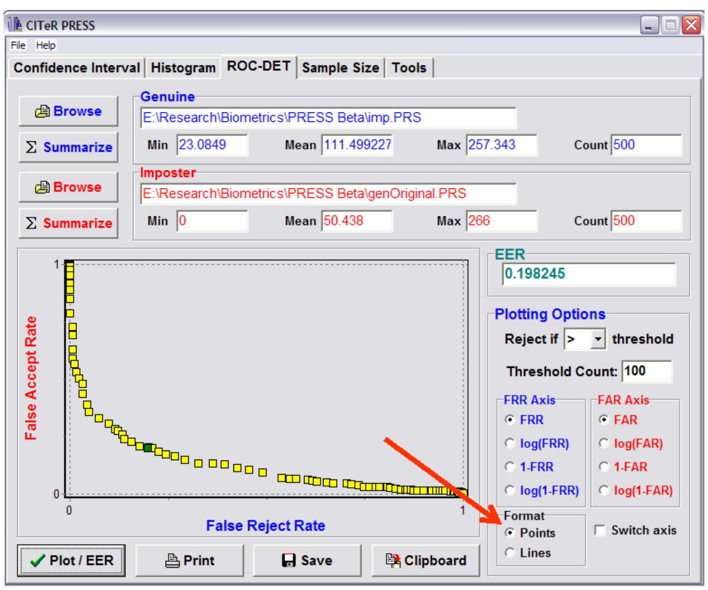

Figure 14: Choosing points or lines

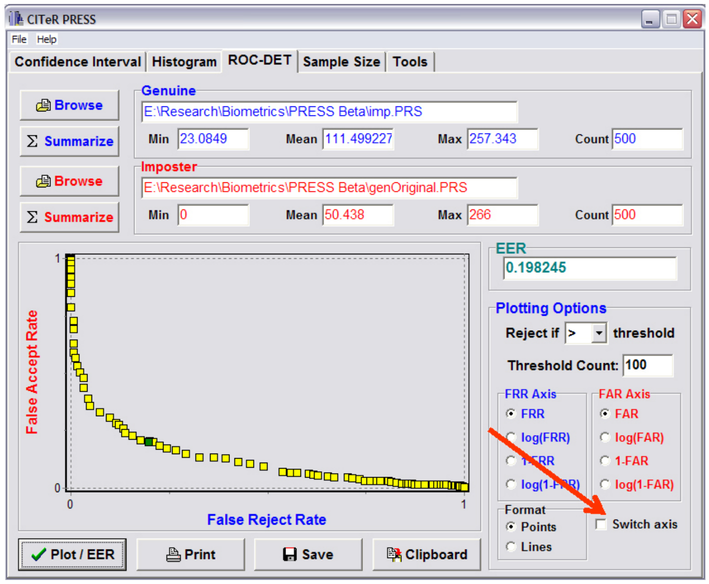

Figure 15: Switch Axes

The confidence interval is specified in this way for two reasons. First, it is based on the logit beta-binomial method which results in an assymmetric confidence interval. Second, because of the way in which biometric authentication devices are used, it is somewhat natural to set an upper bound on what the range of values for the error rate.

One difficulty with calculating a sample size for testing is that there are actually two sample size components: the number of individuals, n, and the number of times each individual is tested, m. Since it is not possible to simultaneously solve for m and n, PRESS uses a conditional solution. Specify appropriate values for the error rate, the intra-individual correlation, the maximum error rate - this is the upper bound -, the confidence level and the number of trials - attempts - per individual. We then find n via the following equation, given the

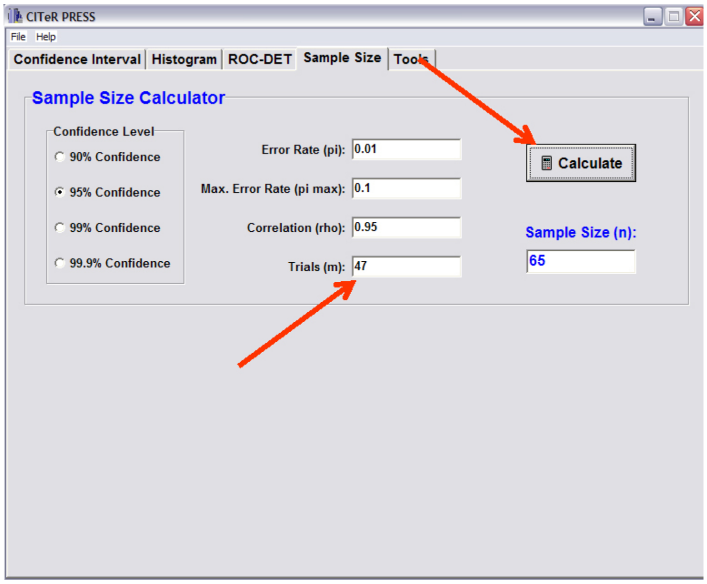

Figure 16: Calculation of the number of individuals for testing

other quantities,

$$n = \left[\left(\frac{z^2_{1-\frac{\alpha}{2}}}{logit(\pi_{max}) - logit(\pi)} \right)^2 \frac{1 + (m-1)\rho}{m\pi(1-\pi)} \right] \tag{1}$$

where n, m, π, π_{max} and α represents the number of individuals, the number of trials, the error rate, the desired upper bound for the confidence interval and one minus the confidence level respectively. See [Schuckers, 2003] for further details.

8 Tools

This function in PRESS takes Excel files and converts them to PRESS format files. By doing this the files are more quickly summarized in PRESS so that

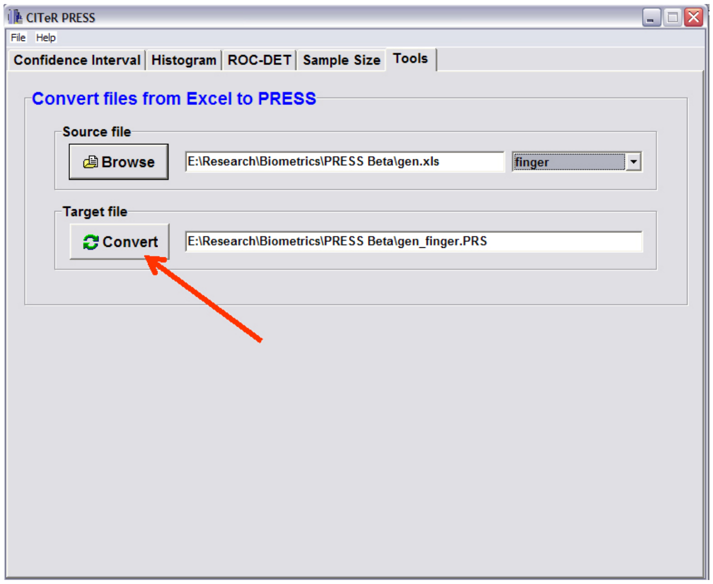

Figure 17: Converting a file to *.prs* format

they can be analyzed faster. Browse for the file that you would like to convert. PRESS will give a default name to the as yet uncoverted file. Change this if you desire a different name for the file, but be sure to leave the file with a ***.prs*** extension. (Figure 17 shows this process.) Then click on the convert button. PRESS will convert the file *converted* and store the *converted .prs* in the same directory as the original. To change a text file to PRESS format simply change the extention to *.prs* from *.txt*.

9 Appendix: Confidence Interval Details

In this section are details regarding the different methods for creating confidence intervals that are available in PRESS. Further details and a comparison of the properies of these methods can be found in [Schuckers et al., 2004]. We strongly recommend that users make themselves aware of the details for the particular methodologies that they are using.

9.1 Balanced Repeated Replicates

The Balanced Repeated Replicate is a non-parametric approach for creating a confidence interval for the mean error rate. This approach creates replicate datasets to get an estimated sampling distribution of the estimated error rate. Appropriate percentiles are then chosen to give the endpoints of the confidence interval. A replicate dataset is created by taking a single sample from each individual, then combining these samples for all individuals. From each replicate dataset, a single estimate of the error rate is calculated and the collection of these is the distribution of estimated error rates. The Balanced Repeated Replicate is considered to be a resampling method as is the Subset Bootstrap.

9.2 Best Practices

Best Practices gives a confidence interval for the mean error rate. The confidence interval is created by taking the estimated error rate and adding and subtracting a quantile of the normal distribution to a standard error based on the variability of the individual error rates. Statistically it should be asymptotically equivalent - as the number of total attempts gets large - to the Beta-binomial. The difference between these two methods being their slightly different ways of estimating the intra- individual correlation. Effectively, Best Practices is using a method of moments approach, while the Beta-binomial approach is using an analysis of variance approach to estimate this quantity.

9.3 Beta-binomial

Beta-binomial gives a symmetric confidence interval for the mean error rate. The confidence interval is created by taking the estimated error rate and adding and subtracting a quantile of the normal distribution to a standard error based on the variability of the individual error rates. Statistically it should be asymptotically equivalent - as the number of total attempts gets large - to the Best Practices. The difference between these two methods being their slightly different ways of estimating the intra-individual correlation. Effectively, Best Practices is using a method of moments approach, while the Beta-binomial approach is using an analysis of variance approach to estimate this quantity. The version we present here is an updated version of that which appears in the IJIG article below.

9.4 Doddington's Rule

Doddington's Rule gives a 90% confidence interval for the mean error rate. The confidence interval is created by taking the estimated error rate and adding and subtracting 30% of that estimated error rate.

9.5 Logit Beta-binomial

The Logit Beta-binomial approach gives an asymmetric confidence interval for the mean error rate. The transformed confidence interval is created by taking the log odds ratio or the logit of the estimated error rate and adding and subtracting a quantile of the normal distribution to a standard error for the logit of the error rate based on the variability of the individual error rates. Those endpoints are then returned to the original error rate scale by the inverse logit function. Asymmetry results from the logit transformation not being linear.

9.6 Subset Bootstrap

The Subset Bootstrap is a non-parametric approach for creating a confidence interval for the mean error rate. This approach creates replicate datasets to get an estimated sampling distribution of the estimated error rate. Appropriate percentiles are then chosen to give the endpoints of the confidence interval. A replicate dataset is created by taking a number of samples equal to the number of attempts for that individual with replacement from each individual, then combining these samples for all individuals. From each replicate dataset, a single estimate of the error rate is calculated and the collection of these is the distribution of estimated error rates. The Subset Bootstrap is considered to be a resampling method as is the Balance Repeated Replicate.

References

[Schuckers et al., 2004] Schuckers, M. E., Hawley, A., Livingstone, K., Mramba, N., and Knickerbocker, C. J. (2004). A comparison of statistical methods for evaluating matching performance of a biometric identification device- a preliminary report. In Jain, A. K. and Ratha, N. K., editors, *Biometric Technol-*

ogy for Human Identification, volume 5404 of *SPIE Proceedings Series*, pages 144–155.

[Schuckers, 2003] Schuckers, M. E. (2003). Estimation and sample size calculations for correlated binary error rates of biometric identification rates. In *Proceedings of the American Statistical Association: Biometrics Section [CD-ROM]*, Alexandria, VA. American Statistical Association.