Assessing Power Output Specifications of PV Modules

This user manual describes Version 1.6, build May 19th, 2012.

APOS photovoltaic StatLab originated in joint research projects of the Institute of Statistics at RWTH Aachen University (Professor Ansgar Steland, Chair of Stochastics) and TÜV Rheinland Immissionsschutz und Energiesysteme GmbH (Dr. Werner Herrmann). We gratefully acknowledge financial support from Solarförderverein Bayern e.V., particularly Professor Dr.-Ing. Gerd Becker and Fabian Flade.

(C) 2009-2012 RWTH Aachen University and Ansgar Steland.

Heads and principal investigators: Professor Dr. Ansgar Steland, Dr. W. Herrmann.
Team: K. Dahmen, W. Herff, A. Johansen, H. Zähle.
Logo design: A. Richau (Berlin). The logo is property of Professor A. Steland.

Meanwhile, APOS was improved and extended. We thank our industrial partners and licensees for their valuable remarks and proposals for improvements. Dr. Andrey Pepelyshev has spend time on careful testing of some extensions.

APOS uses internally Rturbo/P, copyright and all IP by A. Steland, and a modified JAVA library licensed under the LGPL license. The source code can be found in the installation directory. One can easily modify that source, recompile and locate the new jar file in the library folder of the installation directory, in order to use the modified library.

Contents

Introduction **5**

1 Installation **7**

2 Graphical User Interface **8**
2.1 Menu Bar 8
2.2 Data Input 14
2.3 Statistics 18
2.3.1 One-Sample Statistics 18
2.3.2 Two-Sample Statistics 19
2.4 Specifying Parameters 20
2.5 Calculation of the Sampling Plan 28
2.6 Analysis of Lab Samples 31
2.7 Analysis of Arbitrary Lab Samples 33

3 Scenarios **35**
3.1 Setting with flash data tables 36
3.2 Setting without flash data tables 40

4 Histogram Window **44**
4.1 Bandwidth Choice 46
4.2 Choice of the kernel 48
4.3 Additional features 50
4.4 Saving the histogram 55

5 Technical Appendix **56**
5.1 Cases 56
5.2 Theoretical Background of the Decision Rules 57
5.3 Computation of the Sampling Plan 59
5.4 Histogram and kernel density estimator 60

References **63**

Changes

Although at first glance APOS looks similar as the previous version, internally major improvements and extensions have been implemented.
The following major changes are implemented in APOS 2012 Version 1.6:

- Implementation of two new improved methods: KernelICV and DoubleKernel.
- KernelICV is the new default method.
- A PDF report generator has been added to the system.

The following major changes are implemented in APOS 2012 Version 1.5:

- The Modified Procedure is implemented and is the new default method.
- Flasher and lab sample can be compared with the PS bootstrap test.
- Almost all elements of the graphical user interface have now tool tips.
- The user manual can be assessed directly from APOS.
- Time-consuming computation make use of multi-core PCs (up to quad core CPUs).
- Analyses based on lab samples with sample size unequal to the optimal plan implemented.
- Parallel algorithms are implemented to reduce computing time.

Introduction

End users, for instance operators of photovoltaic power plants, and manufactures of photovoltaic (PV) modules have to deal with the fact that often a certain fraction of the modules does not comply with the required quality in terms of the power output. Since inspecting and testing all modules in a photovoltaic lab is infeasible, one has to rely on a random sample of such lab measurements to decide whether a shipment of modules should be accepted or rejected.

As a consequence, producer and end user have to take the risk of false decisions. The end user is exposed to the risk that the shipment is falsely accepted, although the modules are of low quality, whereas the producer is exposed to a false rejection of the shipment, although it is of high quality. Statistical acceptance sampling provides decision rules which allow us to control the corresponding error probabilities associated with the aforementioned risk events.

APOS photovoltaic StatLab 1.6 provides state of the art methods for the statistical acceptance sampling of PV modules. Standard procedures as well as innovative methods are implemented, which take into account additional information contained in flash data tables. The multi-stage decision procedure of APOS photovoltaic StatLab automatically calculates the required sampling plan depending on the parameters defined by the user. The decision procedure is graphically illustrated. Furthermore, the photovoltaic measurements can be analyzed easily.

This user manual describes APOS photovoltaic StatLab Version 1.6 and illustrates the correct application of the software's functionalities, particularly the calculation of the sampling plans, including the analysis of laboratory measurements collected according to the computed sampling plans.

The installation of APOS photovoltaic StatLab is described in Chapter 1. Chapter 2 introduces the graphical user interface and its components. Chapter 3 explains in detail and step by step how to conduct an analysis with APOS photovoltaic StatLab. The graphical visualization of the data sets and their explorative analysis is treated in Chapter 4. The technical appendix summarizes the four basic scenarios and provides mathematical formulas as well as theoretical background on the decision function and the functionalities implemented in the histogram feature.

1 Installation

APOS photovoltaic StatLab 2012 is delivered as an executable installer which installs the software automatically on your computer.

Windows XP/Windows Vista/Windows 7:
After installation the program is available in your program directory located in 'C:\'. It can be started using the menu entry

START → Programs → APOS 2012

Remark: When installing the software under **Windows Vista**, a window pops up. Simply confirm the question to install APOS photovoltaic StatLab on your computer.

APOS photovoltaic StatLab 2012 makes use of the Java Runtime Environment (JRE), which is usually already installed. Otherwise, the installer automatically directs you to an internet of Sun Microsystems which allows you to download and install JRE.

Java is a registered trade mark of Oracle Inc. Windows XP, Windows Vista and Windows 7 are registered trade marks of Microsoft Corporation.

2 Graphical User Interface

After starting APOS photovoltaic StatLab, a graphical user interface appears with a menu bar, two data columns, a panel to specify various parameters and an area where the calculated sampling plan and details of the analysis are shown.

2.1 Menu Bar

Similar as other Windows applications, the user interface has menu bars where actions can be initiated by selecting a menu entry. In the sequel the menu actions of the **File** menu (see Figure 1) are described.

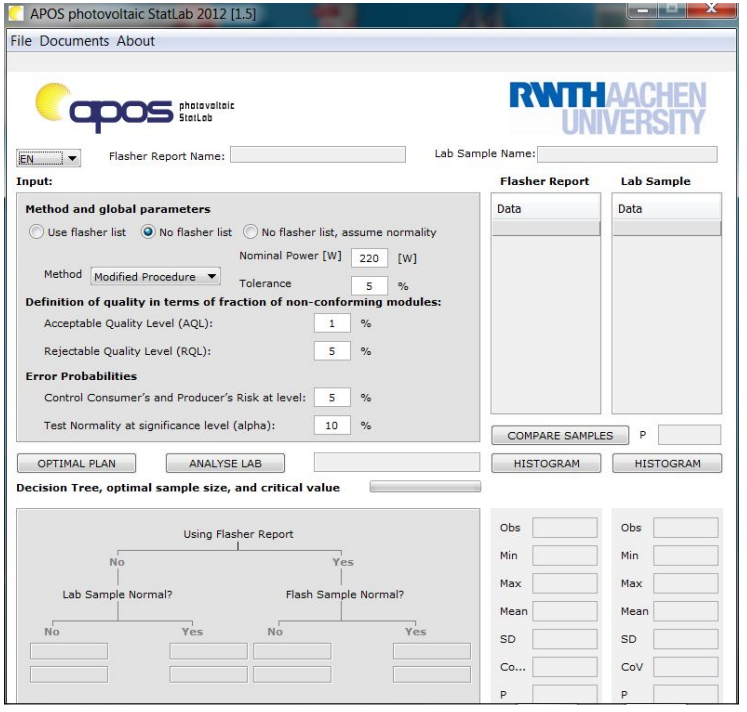

Figure 1: Screenshot of APOS.

(The meaning of the flash data tables and laboratory data tables will be explained below.)

- **File actions for flash data tables**

  – Opening of an existing flash data table

  When selecting the menu entry **Open Flasher Report** a window appears which allows you to browse through your filesystem to select a data file. Simply click on the directory and file names, respectively.

  APOS photovoltaic StatLab can handle ASCII data files as well as the common CSV file format. In this way easy data exchange with other programs as *Microsoft Excel* is possible. Data files stored in ASCII format have usually the extension 'txt' or 'dat' and CSV files have the extension 'csv'. Thus, the file name filter defaults to these file name extensions.

  When reading external data files, it is important to note that each data value has to be in new line.

  After reading the data file, the data values are printed in a column headed by **Flasher Report** (see. Section 2.2). Old values will be overwritten.

  – Saving a flash data table

  In order to save flash measurements entered in the **Flasher Report** column, cf. Section 2.2), please use the menu entry **Save Flasher Report As**. A file requester appears which allows you to browse through the filesystem and to specify a file name of a file to which the **Flasher Report** data column will be stored. Enter a new file name into the appropriate dialog field, if you want to store the data into a new data file. You should add one the file extensions

discussed above. If you have selected an existing file, that file will be overwritten with the new data.

To update a data file you can use the menu entry **Save Flasher Report** instead of **Save Flasher Report As**.

- **File actions for laboratory data**
  - Opening an existing lab data file

    In the same way as explained above for flash data, the menu entry **Open Lab Sample** allows you to read laboratory measurements from a data file into the data column headed by **Lab Sample**.

  - Saving laboratory data

    Saving lab data to a file is done in the same way as it is done for flash data. Depending on whether a new file should be created or not, use **Save Lab Sample As** or **Save Lab Sample**.

- **Clear Contents**

  If you have analyzed a case and want to evaluate a new one, you can initialize the program and clear all data fields by selecting the menu entry **Clear Contents**. The program clears all data columns and output fields and resets the decision tree to its initial state. However, the parameters of the Input-area will not be initialized with their default values. In this way, you can analyze various data sets using the same parameter values.

- **Show Manual**

  This menu entry allows you to read the user manual on the screen.

- To quit the program select the menu entry **Quit**.

The menu entry **About** opens a window with information on your license, the software, references to the scientific literature and the authors of the software.

Remarks:

- The default directory for the file requesters is the root directory 'C:'. To ease navigation, subsequent selections start with the most recently selected directory.
- When reading data files, the format of the data values is automatically recognized. That ist, you may use the international format, e.g. 12.34, or the German format, 12,34.
- If the data should be stored using the international format, please make sure that the corresponding selector (upper left corner) is set to **EN**, see Figure 4). To save the data using the German format, select **DE**.

2.2 Data Input

The program distinguishes between flash data tables provided by a producer of PV modules and laboratory data. In case that flash measurements are available, these data values have to be entered in the data column headed by **Flasher Report**.

In a real application, the laboratory measurements are taken *after* having determined an appropriate sampling plan (see Section 2.5). These control measurements have to be entered in the **Lab Sample** data column.

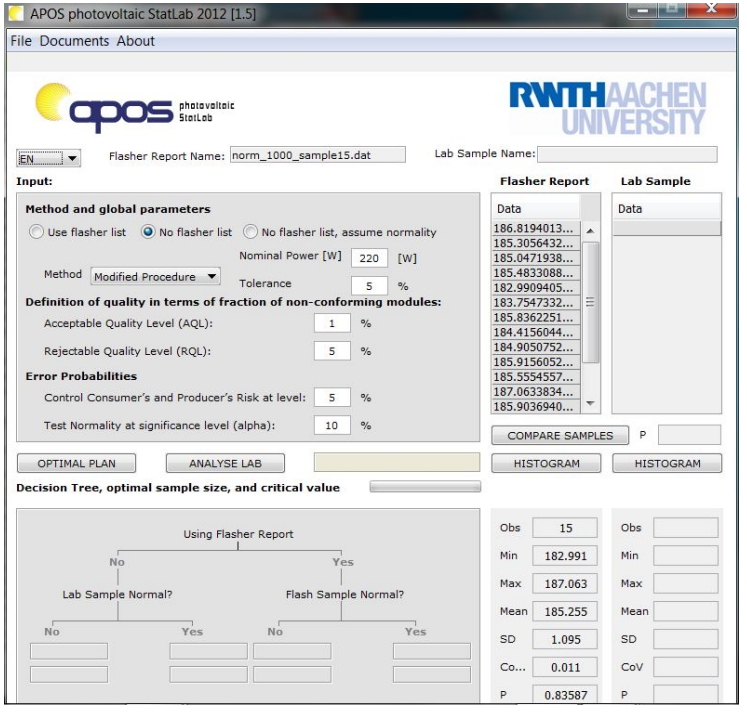

Figure 2: Screenshot of APOS photovoltaic StatLab with flash data

To input data in one of the data columns (see Figure 2), there are two options:

- Input single data

  Double-click the first empty row under Data. The corresponding field changes its color and a cursor appears. Now enter the data value and press the return key. The cursors moves to the next empty field, and you can enter the next value.

- Input data from a data file

  As described in Section 2.1, measurements can be restored from a file using the menu entries **Open Flasher Report** and **Open Lab Sample**, respectively, of the menu **File**. Data values already entered manually will be deleted.

To change values, simply double-click on the corresponding data field.

Remarks:

- Data values are printed on the screen using the international format, e.g. 12.34. To input values manually, you have to use that format, too. However, APOS photovoltaic StatLab can read external data files using the German format, e.g. $12,34$, as well.

- For technical reasons, the row below the last data value has to be empty. Otherwise, the last value will not be used in the calculations. Please check how many observations have been recognized by the system by looking at the field **Obs** which provides the sample size.

- Data files can have the extensions 'csv', 'dat' or 'txt'. The data values should be separated be line delimiters, i.e., each data value should appear in a single row.

- When the data have been read from a file or have been stored to a file, the file name of that file is printed on the screen in the line below the menu bar as indicated by Flasher Report Name: and Lab Sample Name:, respectively. The file names are updated when using the menu actions **Save Flasher Report** or **Save Lab Sample**.

2.3 Statistics

2.3.1 One-Sample Statistics

When data values have been added to the **Flasher Report** column or the **Lab Sample** column, the following statistics are calculated and printed on the screen:

- **Obs** (sample size, number of observations)
- **Min** (minimum)
- **Max** (maximum)
- **Mean** (arithmetic mean)
- **SD** (sample standard deviation)
- **CoV** (coefficient of variation)
- **P** (p-value of the Shapiro-Wilk tests)

The p-value of the flasher report table is used to check the assumption of normality, cf. the parameter **Test Normality at significance level (alpha)** explained in Section 2.4).

The statistic **CoV** is a measure of the variation of the data. It is defined as the ratio of the difference and sum of the maximum and minimum. If the maximum and the minimum are closed to each other relative to the scale of the data, the coefficient of variation is small.

These statistics are automatically updated when data values change.

2.3.2 Two-Sample Statistics

When pressing the button **COMPARE**, a two-sample test is calculated to test the null hypothesis that there is no difference in location between the flasher report table and the lab sample in the sense of the generalized median δ (of differences). The null hypothesis is that the generalized median is equal to 1/2. The PS bootstrap rank test is computed that is related to the Wilcoxon signed rank test that is applied to all between–sample differences. The p-value of the test is calculated. If $p < \alpha$, where α is the significance level

selected by the user, the null hypothesis is rejected and one may conclude that the samples differ in location.

Interpretation

- When the power output distributions of the flash and lab sample have equal shape - but possibly different scales (variances) - the generalized median δ ...
- When the power output distributions are symmetric and of equal shape, δ coincides with the difference of the symmetry centers.
- For identical distributional shapes, the generalized median δ is the location shift.

2.4 Specifying Parameters

You can specify the parameters for your analysis in the gray area (left upper area) headed by **Input** (see Figure 3). These parameters are required to determine a sampling plan. The default values can be changed by clicking into the corresponding fields.

Nominal Power [W] The nominal power output of the PV modules as provided by the producer.

Tolerance The tolerance is the maximal deviation of the actual power output of a module from the nominal power output, expressed as a percentage. This means, if the tolerance is 10% and the nominal power output is 100 W, a module with power output between 90 W and 100 W is regarded as tolerable. Usually, producer and end user have to agree upon the tolerance. In practice, one often uses the a tolerance of 5%.

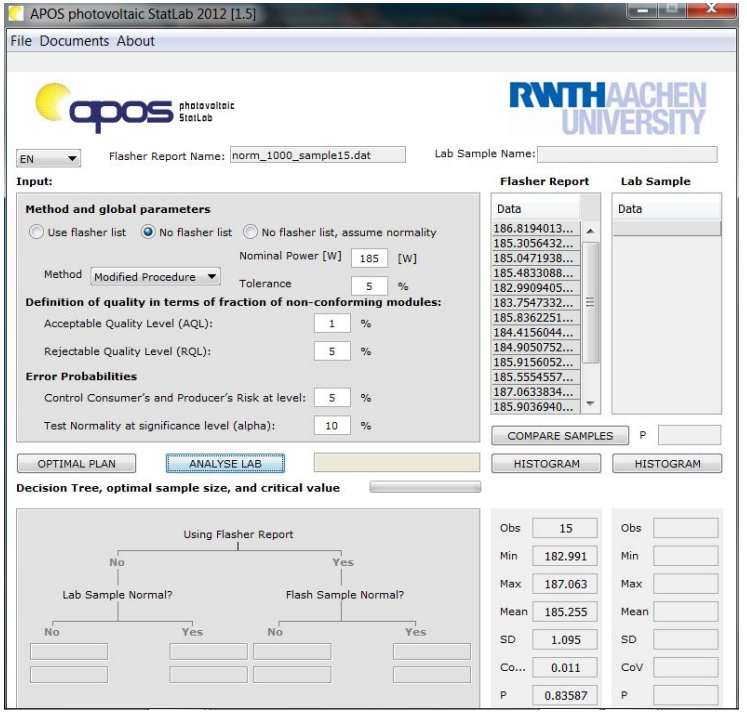

Figure 3: Screenshot with default values except that the Nominal Power is changed to 185.

The radio buttons in the next row of the panel determine the mode of operation of APOS photovoltaic StatLab. There are three modes corresponding to the data situation (what kind of data is available) and how the data should be used. Select one of these modes by clicking on the corresponding button, cf. Section 5.1):

- **Use flasher list**

  The producer has delivered a flash data table. This mode can be selected, if the flash data are already available to the system, either manually or by reading a data file, cf. Section 2.2. Whether or not the software assumes that the flash measurements are normally distributed depends on the flash data and the significance level (see below).

- **No flasher list**

  Use this mode of operation, if the producer has not provided a flash data table or if you want to ignore the flash measurements for an analysis, e.g. for comparisons. In this case no specific assumption for the distribution offthe measurements is assumed.

- **No flasher list, assume normality**

  This mode of operation does not require a flash data table as well. If flash measurements are present, they are ignored and not used for calculations. In contrast to the mode *No Flasher List*, it is assumed that the power output measurements are normally distributed.

Remark: It is crucial for the validity of calculations to select the correct mode.

Selection of the Method

The box to the left is used to select the method used by APOS to calculate a sampling plan in case that a flasher list should be used and the flash measurements are non-normal. The default value is **KernelICV Method**. Alternatively, one may use the **DoubleKernel Method**, which is better for many power output distributions.

Remark:
The **KernelICV Method** as well as the **Double Kernel Method** are always superior to the other methods. The **Modified Procedure** differs from the **KernelICV Method** in that it uses a different (suboptimal) bandwidth selector. However, the **Modified Procedure** is always better than the **Standard Method**, which is based on standard sample quantiles. The Standard Method should only be used if the other procedures lead to unrealistic unplausible result. In such case, compute a sampling plan using the Standard Method, in order to compare the results. For very large flasher report tables the computed sampling plans should be (almost) equal. If in doubt, contact us.

Definition of quality in terms of fraction of non-conforming modules

The two parameters to the right of the Method selector are used to specify the required degree of quality.

Acceptable Quality Level (AQL)	This is the maximal fraction of conforming modules which is in agreement with a 'high quality' (of the lot). If the fraction of non-conforming modules of the lot is less or equal to the AQL, the lot is regarded as being of high quality.
Rejectable Quality Level (RQL)	This is the smallest fraction of non-conforming modules being regarded as 'low quality'. Thus, if the fraction of non-conforming modules of a lot is greater or equal to RQL, the lot is regarded as 'low quality'.

Important Note

These two parameters define your understanding of quality. Their values have a strong influence on the required sample size, but their definition has nothing to do with statistical acceptance sampling. For a manufacturer, AQL and RQL determine the quality, in terms of the fraction of non-conforming modules per shipment, he wants to guarantee to his customers. For an end user, AQL and RQL determine the quality he requests. Typical values for AQL are between 1% and 5%; the RQL typically ranges between 3% and 10%. Obviously, these quality levels satisfy the constraint AQL $<$ RQL make sense.

In general, the effect of AQL and RQL on the required sample size calculated by APOS photovoltaic StatLab is as follows: If the difference $RQL - AQL$ is large, the required sample size of the sampling plan is small, and if $RQL - AQL$ is small, then the required sample size is large.

Error Probabilities

The last two parameters specify the error probabilities used by the program.

Control Consumer's and Producer's Risk at level

If a shipment of modules is of low quality, i.e. if the fraction of non-conforming modules is larger or equal to RQL, then the probability that the shipment is accepted should not exceed that value, which is called consumer risk.
If a shipment is of high quality, i.e. if the fraction of non-conforming modules is smaller or equal to AQL, then the probability that the shipment is rejected should not exceed that value, which is called producer risk.

APOS photovoltaic StatLab controls the consumer and producer risk at the same level to guarantee that interests are taken into account symmetrically. Producer and end user should agree in a contract on that value. Typical values range between 5% and 10%.

Test Normality at significance level (alpha)

This value determines the significance level for the test of the normality assumption. If a flash data table is present, APOS photovoltaic StatLab applies the Shapiro-Wilk test to check the normality assumption for the flash data table.
The null hypothesis of normality is accepted, if the calculated p-value is greater or equal to the significance level. Otherwise, the assumption of normality is rejected. Typical values for the significance level are between 5% and 20%.

2.5 Calculation of the Sampling Plan

When all relevant data have been entered (see Section 2.2) and the parameters have been specified appropriately (see Section 2.4), then one can determine the optimal smapling plan for that setting. This is done by clicking the button **OPTIMAL PLAN** located below the Input-area.
APOS computes the optimal sampling plan. In case that the flasher report table should be taken into account and the Modified Procedure is selected, the calcuation requires substantial computing time and usually takes several seconds or minutes depending on CPU ressources of your PC. During computations, the user interface freezes and you cannot input values or perfrom other actions. A green progress bar shows how computations proceed. When the plan is calculated, the result is printed in the decision tree section.
The decision tree illustrating in the lower left panel illustrating the two-stage decision process will be updated (see Figure 4). The decision tree illustrates the relevant scenario, cf. Section 5.1, and the resulting sampling plan. This means, the decision tree shows whether or not flash measurements have been used for computations and whether or not the assumption of normality has been made.

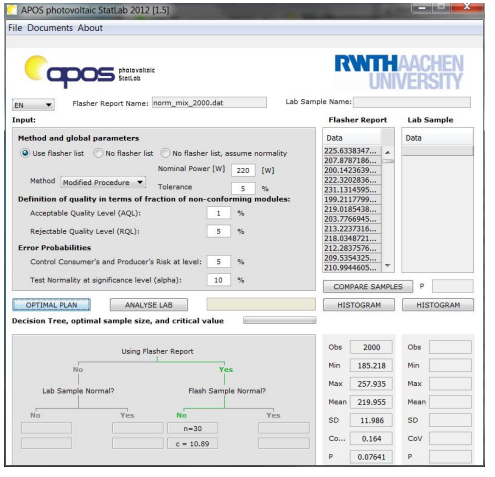

Figure 4: Screenshot after the determination of the optimal sampling plan.

For operators of photovoltaic power plants the optimal sampling plan printed on the screen at the bottom of the relevant branch of the decision tree is of primary interest. It is given as a pair of numbers which summarize the decision rule to accept or reject a shipment of PV modules:

$$\text{Sampling Plan:} \qquad (n, c)$$

n : sample size, c: critical value

- The first number, denoted by n, provides the sample size of the test (laboratory) sample. That is, power output measurements of n randomly drawn PV modules are required and have to be entered as described in Section 2.2 in the data column headed by **Lab Sample**.

- The shipment is accepted if and only if the value of the relevant test statistic, cf. Section 5.3), does not exceed the second value, c. The choice of the test statistic and the concrete decision rule depends on the scenario, i.e. on the mode of operation and eventually on the decision of the normality test. For experienced users, the critical value c is reported by APOS photovoltaic StatLab. In combination with the relevant test statistic, you can check the validity of the decision.

  When no flash data table is used and no normality assumption is made, then the following simple rule applies: The number of non-conforming modules of the test sample must be smaller or equal to the number c. Otherwise the shipment is rejected.

2.6 Analysis of Lab Samples

If n laboratory measurements have been entered in the **Lab Sample** column as required by the optimal sampling plan, then these data can be analyzed in order to check whether the lot should be accepted or rejected. This is done by clicking the **ANALYSE LAB** button.

The calculated value of the test statistic and the decision is printed on the screen to the right of that button. Note that the test statistic used by APOS depends on the scenario, see Section 5.3 for details.

Accept: The shipment of PV modules should be accepted.

Reject: The shipment of PV modules should be rejected.

Some important remarks:

- The program checks whether the sample size of the laboratory sample is in agreement with the sample size of the sampling plan. If this is not the case, then a warning message appears after clicking the **ANALYSE LAB**. If these numbers differ, APOS analyses the lab data in a different way as described in the next section. The text field where the result is printed is then colored in red to indicate that the analysis does not guarantee the consumer risk.

- If the mode of operation **No flasher list, assume normality** has been selected and the lab sample is not normally distributed, a remark is given, since then the calculations should be repeated using the mode **No flasher list**. That option does not assume a specific distributional assumption.

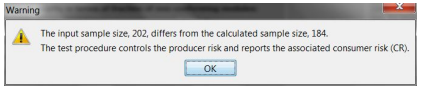

Figure 5: Screenshot with warning

- In order to evaluate another shipment of PV modules, select the menu entry **Clear Contents**, cf. Section 2.1, to reset all fields except the parameter settings.

2.7 Analysis of Arbitrary Lab Samples

As briefly discussed in the previous section, when the sample size of the lab sample differs from the optimal sample size calculated by APOS, APOS behaves differently.
The case of practical interest is that the lab sample size is smaller than the optimal sample size. Then it is not possible to construct a decision rule that guarantees both the consumer's and the producer's risk at the given level. Assuming that the parameters are correctly chosen, APOS implements the following approach: The critical value c of the procedure is adjusted such that the producer's risk is controlled at the level specified by the user. Since the lab sample size is smaller than the sample size required to ensure control over the consumer's risk at the same level, the consumer risk will now be larger than specified. Therefore, APOS calculates an approximation to the consumer risk (CR) and prints the result into an additional field located below the fields for the sampling plan. The lab sample size is then analyzed using the adjusted procedure.

To summarize:

- The critical value is recalculated to ensure that the producer's risk is controlled at the specified level.
- The lab sample is analyzed using the corresponding decision rule.
- The consumer risk CR is calculated.

Remark: If the computed value CR is larger than the specified risk level, low quality is detected with smaller probability, that is less frequently.
Remark: The result of the analysis should be accompanied by the CR value to document the increased consumer risk associated with the analysis due to the small sample size.
As a consequence, the interpretation of the result differs.

3 Scenarios

Depending on the information available to the user, there are various scenarios which require different user input. In the sequel, it is explained in detail how to apply APOS photovoltaic StatLab for the scenarios, cf. Section 5.1.
In order to guarantee correct results, one should proceed according to this user guide.
It is advisable to clear all input and output fields by selecting the menu entry **Clear Content** before each analysis. This ensures that the program is internally initialized. However, the parameters chosen before remain valid and are not reset to their default values.

3.1 Setting with flash data tables

In this scenario a flash data table is available. In order to make a decision whether or not the lot of PV modules should be accepted, the following steps are necessary:

- Input the flash data into the **Flasher Report** data column (cf. Section 2.2).
- Specify the parameters as described in Section 2.4 and select the mode of operation **Use flasher list**.
- Determine the optimal sampling plan by clicking the button **OPTIMAL PLAN**, cf. Section 2.5)
- The required sample size n appears in the decision tree.
- Input n power output measurements of n randomly drawn modules into the data column **Lab Sample**, vf. Section 2.2.
- Analyze the lab sample by clicking **ANALYSE LAB**, cf. Abschnitt 2.7.

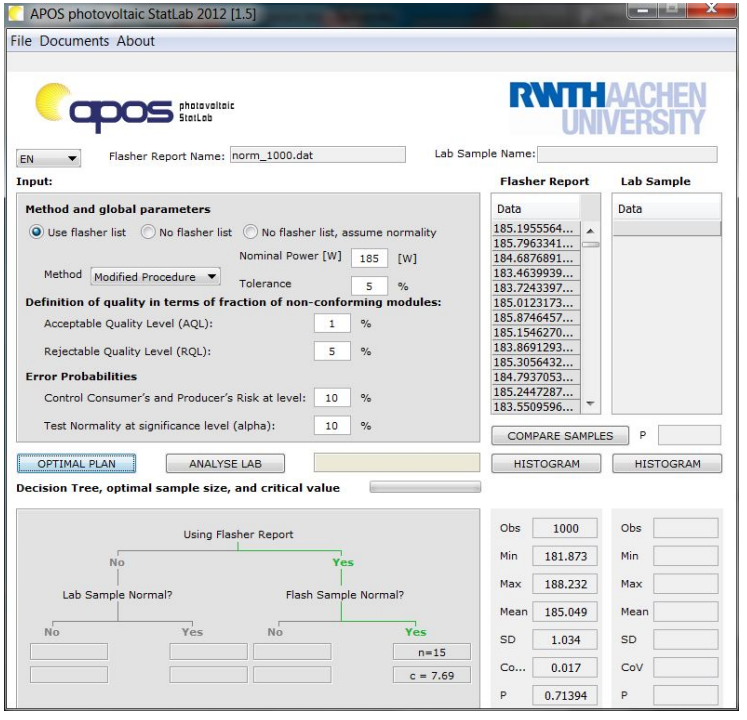

Figure 6: Screenshot after calculating the optimal sampling plan

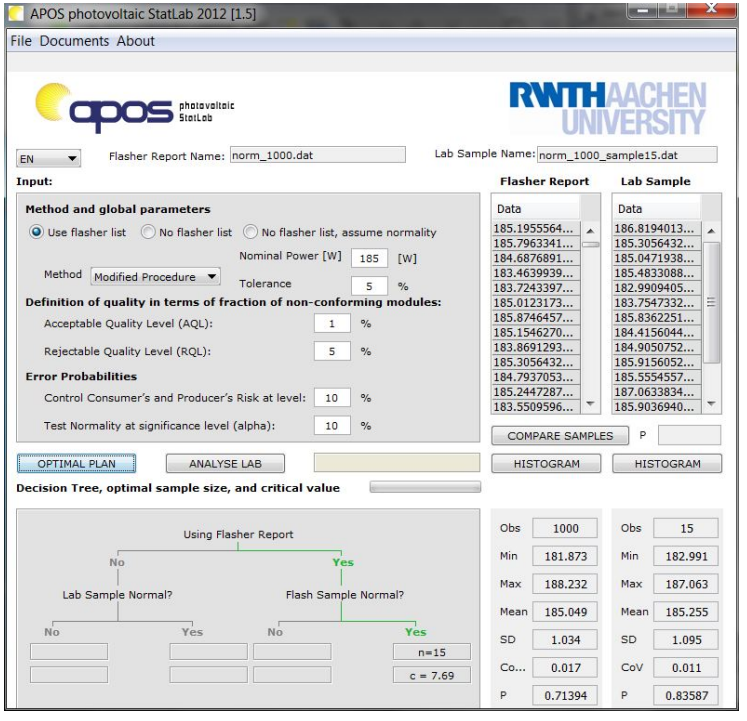

Figure 7: Screenshot after analyzing lab data

- Read the test decision, cf. Section 2.7):
  - **Accept** means that the lot of modules should be accepted.
  - **Reject** means that the lot of modules should be rejected.

3.2 Setting without flash data tables

In contrast to the scenario above, in this scenario no flash data table is available. In this case, the program can not infer from the data whether or not the measurements are normally distributed. Therefore, the user has to specify whether the calculations should be done assuming normality.

Similar as above, the following steps have to be done:

- Specify the parameters as described in Section 2.4. Particularly,
  - select the mode of operation **No flasher list**, if no specific distribution for the measurements should be assumed.
  - select the mode of operation **No flasher list, assume normality**, if the power output measurements are normally distributed.
- Determine the optimal sampling plan by clicking **OPTIMAL PLAN**, cf. Section 2.7.
- The required sample size n is printed in the decision tree.
- Input power output measurements of n randomly drawn modules into the data column **LabSample**, cf. Section 2.2)
- Analyze the lab sample by clicking the button **ANALYSE LAB**, cf. Section 2.7.
- Read the test decision, cf. Section 2.7:

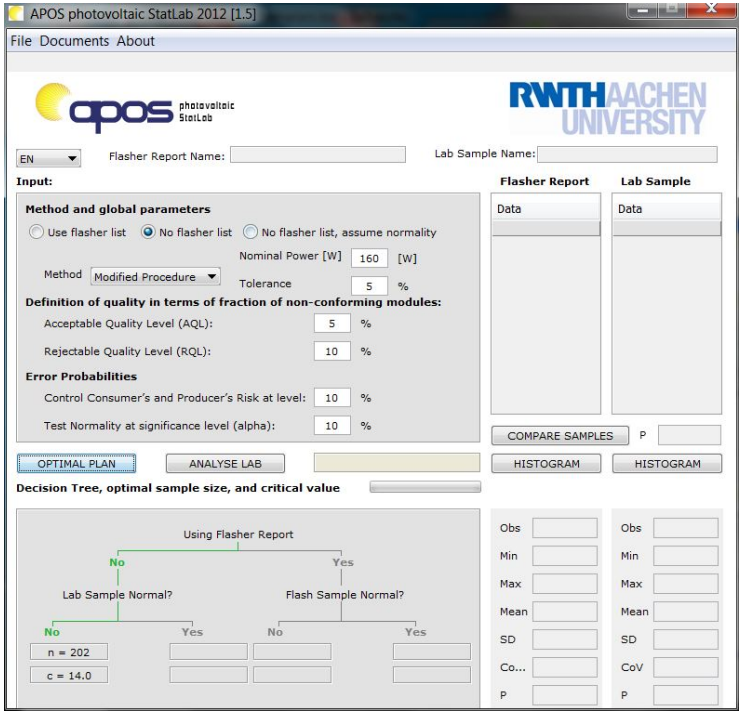

Figure 8: Screenshot with optimal sampling plan

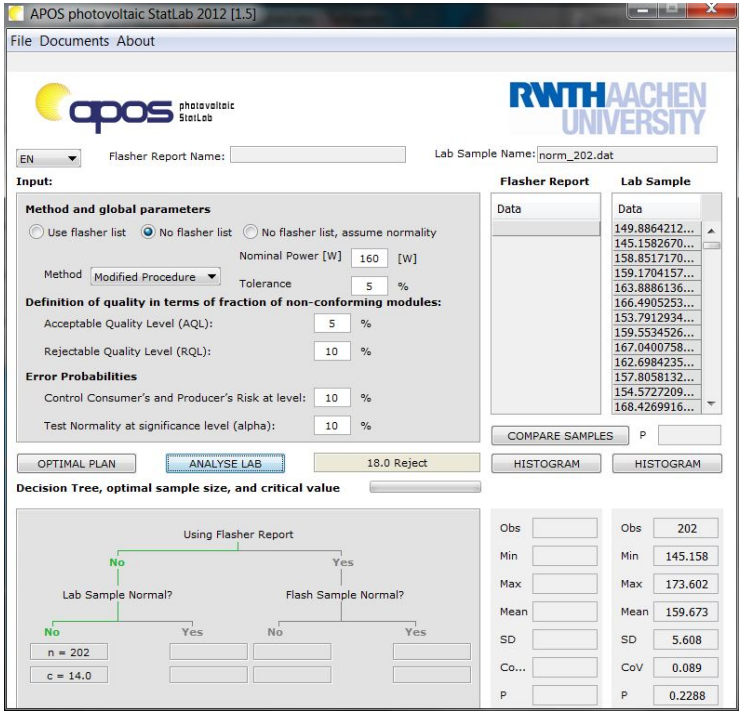

Figure 9: Screenshot after analyzing lab data

- **Accept** means that the user should accept the shipment of PV modules.
- **Reject** means that the user should reject the shipment of PV modules.

Recall that in the present scenario no flash data are available. Therefore, the corresponding data column **Flasher Report** should be empty; indeed, it is ignored when performing calculations. For the same reason, it is not necessary to specify the parameter **Test Normality at significance level (alpha)**, since that parameter is not used by the program in this case.

It is worth mentioning that the computed p-value can be used to check whether or not the normality assumption is justified (for the lab sample).

4 Histogram Window

Given there are samples available in the **Flasher Report** or **Lab Sample** column, cf. Section 2.2, one can press the button **HISTOGRAM** to obtain a histogram plot of the corresponding data set augmented by additional data analysis visualizations. The output is shown in a new window.

The histogram is calculated using bins (intervals) of width 1 W. As a consequence, area as well as height of each bin are proportional to the relative frequencies. To ease interpretation, the relative frequency of the bins can be read off the vertical axis, cf. Section 5.4.

The graphic is augmented by a blue curve representing a kernel density estimate. That estimate uses the Gaussian smoothing kernel and the asymptotically optimal bandwidth choice. The kernel as well as the bandwidth can be chosen manually by the user.

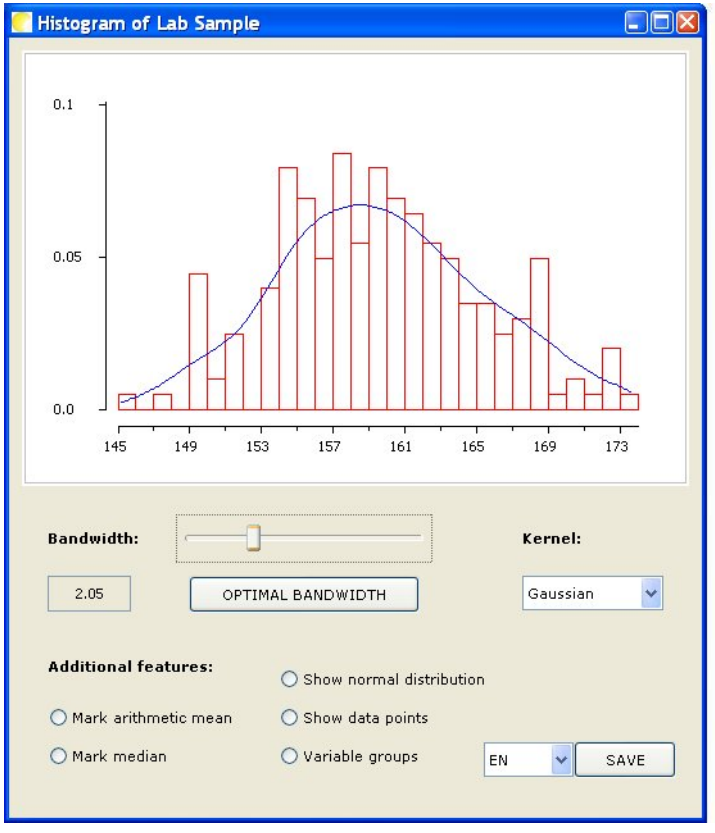

Figure 10: Screenshot of the histogram window

4.1 Bandwidth Choice

The bandwidth currently used is printed in the field **Bandwidth:**. You can modify that value using the slider; the step size is 0.01. The bandwidth controls the data fidelity of the density estimate. when selecting a small bandwidth, for each x-value only a few data points in a local vicinity are used. Larger bandwidths lead to estimates using more neighboring data points. Thus, too small bandwidths yield wiggly curves, whereas too large bandwidths lead to an oversmoothed curve. As a consequence, the quality of a kernel density estimate crucially depends on the bandwidth choice.

Click the button **OPTIMAL BANDWIDTH** in order to compute a statistically optimal bandwidth. APOS photovoltaic StatLab computes the unbiased cross-validated bandwidth. Depending on the sample size, this computer-intensive calculation can take some time. The result is printed in the field **Bandwidth** and the slider is updated accordingly. In general, the cross-valiated bandwidth should be the method of choice to select the bandwidth. Nevertheless, it is advisable to change the bandwidth using the slider in order to investigate the effect of the bandwidth on the result and to compare the density estimate with the histogram. In rare cases the cross-validated bandwidth selection method can yield misleading results.
Cross-validation is implemented for the Gaussian kernel, which is often used in practice. Therefore the kernel is automatically set to the Gaussian kernel when clicking the button **OPTIMAL BANDWIDTH**, and eventually the selected entry in the field **Kernel:** changes to **GAUSSIAN**.

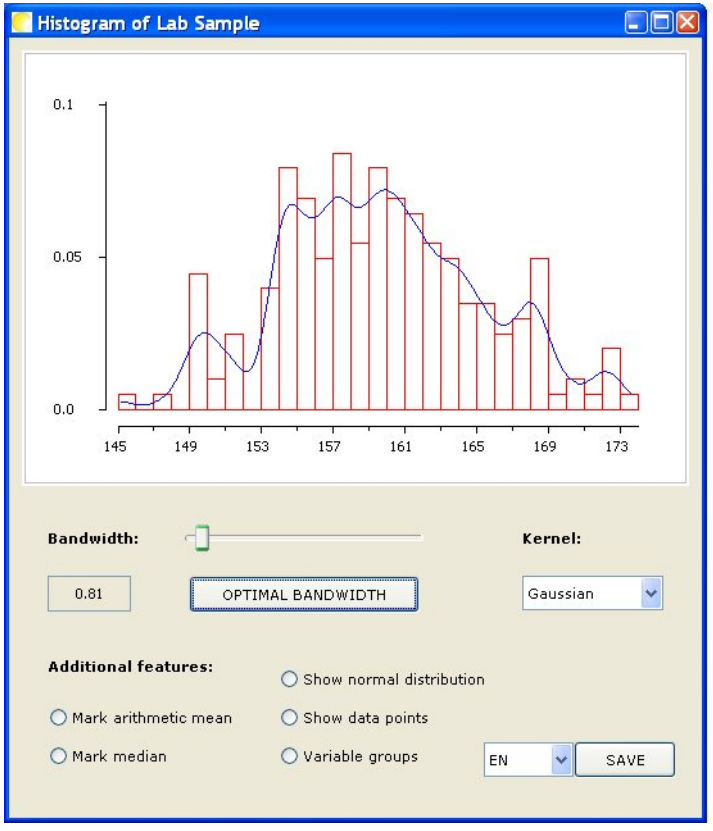

Figure 11: Screenshot after the calculation of the optimal bandwidth

4.2 Choice of the kernel

The selector **Kernel:** at the right side of the panel allows the user to select one out of five kernel functions. The following kernels are implemented:

- **Gaussian**
- **Rectangular**
- **Triangular**
- **Epanechnikov**
- **Biweight**

The mathematical definitions of these kernel functions are provided in the appendix. When selecting another kernel, the density estimate is updated accordingly.

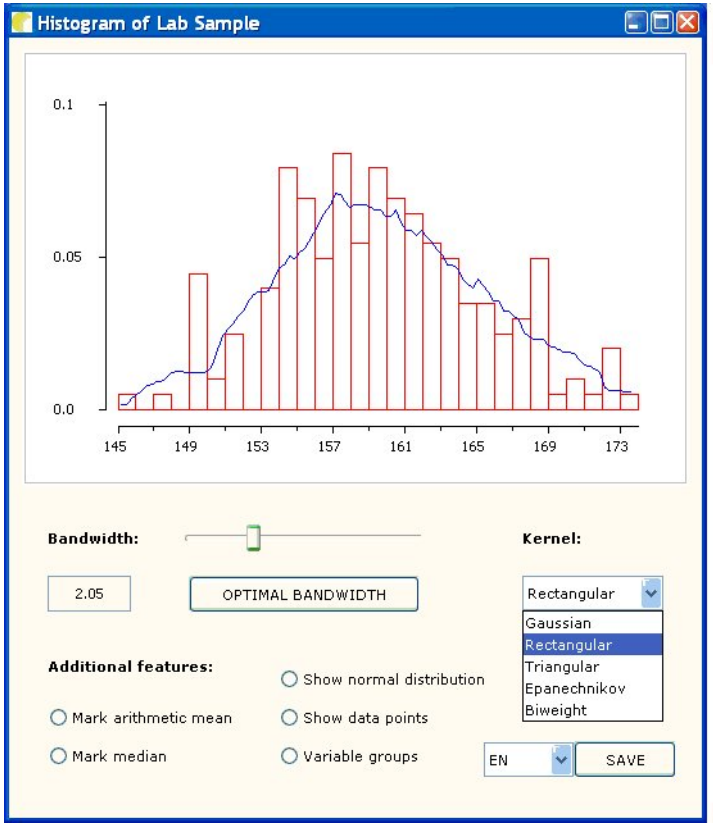

Figure 12: Screenshot using the **Rectangular** kernel.

4.3 Additional features

The buttons of the **Additional features** panel allow to overlay additional graphics. It is also possible to change the definition of the bins (class intervals) for the computation of the histogram. Each feature can be activated and deactivated, respectively, by clicking the button.

Mark arithmetic mean The arithmetic mean of the data set is added to the plot. It is marked by a vertical green line.

Mark median The median of the data set, a robust measure of location, is added to the plot. The median is marked by a vertical blue line.

Show normal distribution A normal density with mean equal to the sample mean and variance equal to the sample variance is added to the plot. The intensity of the curve depends on the p value of the Shapiro-Wilk test for normality. The more the p value indicates that one should rely on the normality assumption, the better the curve is visible. When selecting this feature, a scale is shown which allows to read the p value.
(This feature is disabled, when the sample of the data set is less than 5.)

Show data points All observations of the data set are shown as small grey vertical lines below the horzontal scale.

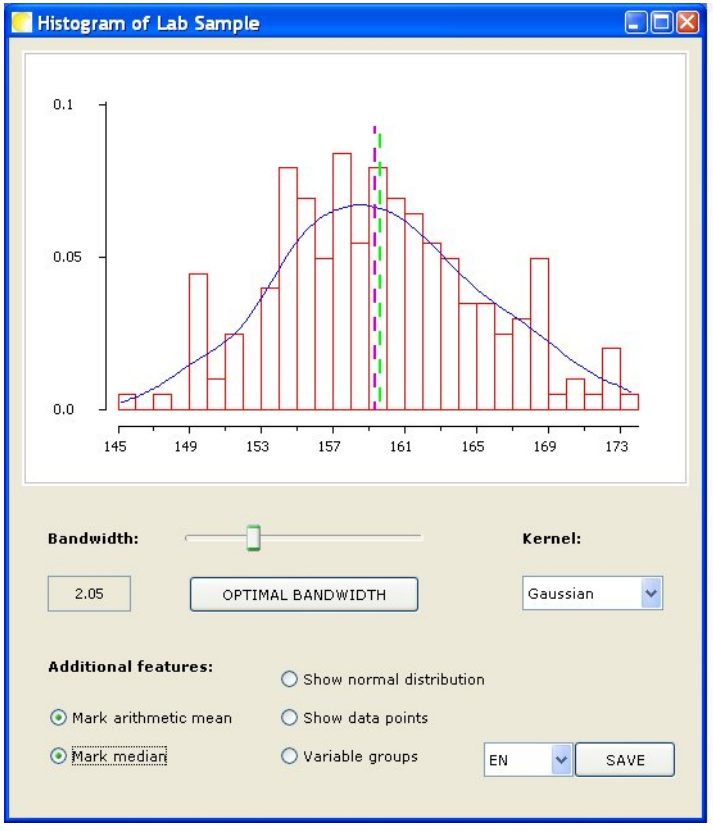

Figure 13: Screenshot with **Mark arithmetic mean** and **Mark median** activated.

Variable groups	The histogram is no longer calculated using equidistant class intervals (bins) of size 1 W. Instead, the quantiles of the normal distribution are used.

Remarks:

- The size of the graphic window can be changed to enlarge the plot. The size of the plot is changed accordingly.
- The title of the graphics window shows whether the flash data table or the lab sample are visualized.
- No histogram is shown if the data set consists of one observation or if all observations attain the same value.

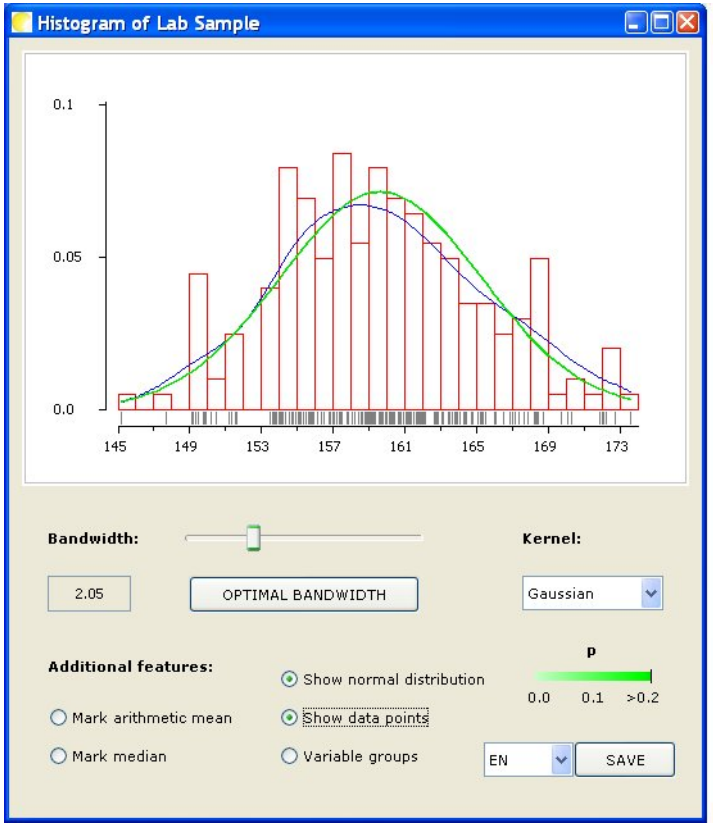

Figure 14: Screenshot with **Show normal distribution** and **Show data points** activated.

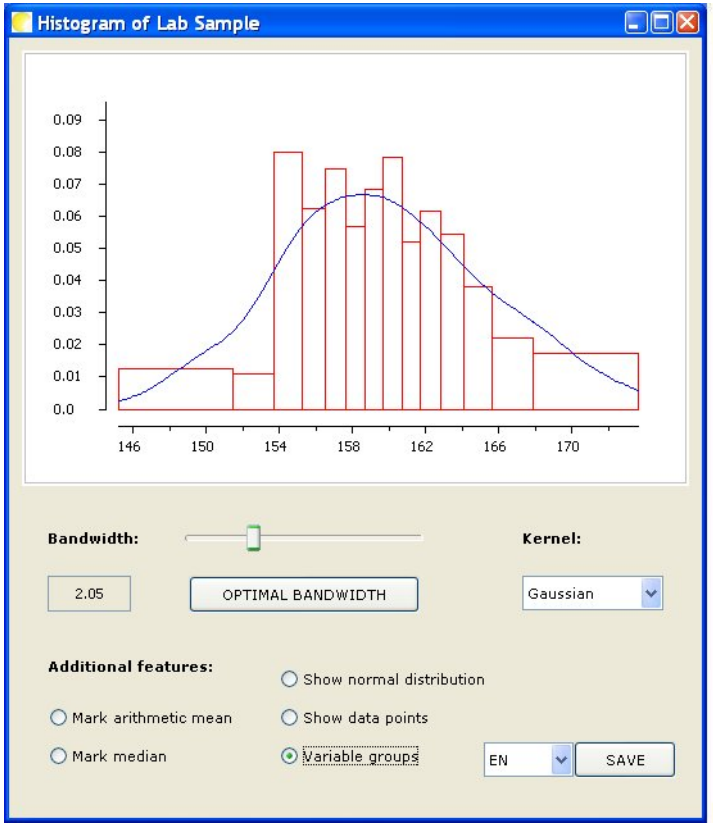

Figure 15: Screenshot with **Variable groups**

4.4 Saving the histogram

The button **SAVE** in the lower right part of the panel allows to save the actual histogram in a file. When clicking the button a file requester appears to select the file. The following quantities are stored in the file:

- **lower bound** of the bins
- **upper bound** of the bins
- **absolute frequency**
- **relative frequency**
- **density** (ratio of relative frequency and interval length.

In order to use a new file, simply enter the file name in the corresponding text field. It is advisable to add a file name extension as 'txt', 'dat' or 'csv'. Otherwise, select an existing file which is then overwritten.

The format of the values is used according to the language selector to the left of the **SAVE** button. Use **EN** for the international format and **DE** for the German format.

5 Technical Appendix

This appendix describes the mathematical formulas used by APOS photovoltaic StatLab, cf.[3, 4, 9]. For background on statistics, statistical quality control and acceptance sampling, we refer to [8], [5] and [2], among others.

5.1 Cases

As indicated by the decision tree of APOS photovoltaic StatLab, the program distinguishes four scenarios. First, we have to the following two cases:

- **No flasher list**

  The end user has no additional flasher list information on the power output of the production lot.

- **With flasher list**

  The end user has access to a flash data table from the modules' manufacturer.

In the first case, one can work with a distribution-free approach or rely on methods assuming normality of the data. In the second case, APOS photovoltaic StatLab provides a nonparametric approach based on an empirical distribution function as well as a method assuming normality. As a consequence, there are four scenarios. It is important to stress that each scenario has its own decision rule.

5.2 Theoretical Background of the Decision Rules

The operator of a photovoltaic power plant is interested in a valid decision rule to decide whether he should accept or reject a shipment of PV modules. The (asymptotically) optimal rules for the four scenarios are as follows, cf. Section 5.1):

Let n denote the sample size and $P_1, \ldots, P_n$ the observed power output measurements of the PV modules. The end user (operator) accepts the shipment, if

$$T_n \begin{cases} \leq c & \text{, for the distribution-free scenario employing attribute sampling,} \\ \geq c & \text{, for the remaining three scenarios.} \end{cases}$$

Here c denotes the critical value which APOS photovoltaic StatLab outputs below the required sample size n. Further, T_n denotes the corresponding test statistics. The test statistic is given by

$$T_n = \begin{cases} \sum_{i=1}^{n} \mathbb{1}_{\{P_i < \mu_0 - \varepsilon\}} & \text{, for the scenario without a flash data report and not assuming a specific distribution,} \\ \dfrac{\overline{P}_n - (\mu_0 - \varepsilon)}{S_n} \sqrt{n} & \text{, for the scenario without a flash data table assuming normal data,} \\ \dfrac{\overline{P}_n - (\mu_0 - \varepsilon)}{S'_N} \sqrt{n} & \text{, for the scenario with a flash data table,} \end{cases}$$

where (cf. Section 2.4):

μ_0 : nominal power output (APOS:**Nominal Power**)

ε : tolerance (APOS: **Tolerance**)

$P_1, \ldots, P_n$: observed mean power output of the sampled modules

$\overline{P}_n = \frac{1}{n} \sum_{i=1}^{n} P_i$: mean (APOS: **Mean**) of the lab measurements

$S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(P_i - \overline{P}_n\right)^2}$: sample standard deviation (APOS: **SD**) of the lab measurements

$P'_1, \ldots, P'_N$: flash data table

$\overline{P}'_N = \frac{1}{N} \sum_{i=1}^{N} P'_i$: mean (APOS: **Mean**) of the flash data

$S'_N = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(P'_i - \overline{P}'_N\right)^2}$: sample standard deviation (**SD**) of the flash data

Notice that in general the sample size N of the flash data table is much more larger than the sample size n of the lab sample. Thus, S'_N tends to be closer to the true standard

deviation than S_N.

5.3 Computation of the Sampling Plan

For Modified Procedure the computation of the sampling plan requires to solve nonlinear equations of the form

$$\int_{-\infty}^{x} \hat{f}_h(u)\, du = p$$

for given $p \in (0, 1)$. Here $\hat{f}_h$ is the kernel density estimator discussed in the next section with cross-validated bandwidth choice. The integral can be calculated explicitly, but the nonlinear equation has nevertheless be solved by a numerical algorithm. APOS applies the method of bisection to do so.

For certain flasher lists, it may happen that the bisection algorithm does not converge. APOS is able to detect such cases and then applies a more time consuming brute force method to determine a sufficiently accurate solution by de- or increasing x by increments $\pm\Delta$, where $\Delta = 0.1, 0.01, 0.01$. The method has been tested with difficult real data sets and works stably.

5.4 Histogram and kernel density estimator

In order to represent graphically the distribution of measurements following a continuous distribution (density), it is common to use a histogram. The data are classified using a collection of intervals which forms a partition of the range. Then the (relative) frequencies of each class are calculated. Above each interval a rectangle is drawn whose area is proportional to the corresponding relative frequency. Suppose the classes are given by the intervals $[g_0, g_1], (g_1, g_2], \ldots, [g_{m-1}, g_m]$ with corresponding relative frequencies $p_1, \ldots, p_m$. Then the height of the ith rectangle equals the ratio of the relative frequency p_i and the interval's length $g_i - g_{i-1}$. The resulting function which maps x to the height of the rectangle x belongs to, is regarded as an estimate of the unknown underlying density function.

Although density functions are often continuous, the histogram always yields a discontinuous estimate. Therefore one should calculate additionally a kernel density estimator which produces a continuous function estimate.

If $x_i, \ldots, x_M$ denote the M observed values, then the Rosenblatt Parzen kernel density estimator is defined by

$$\hat{f}_h(x) = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{h}\, k\left(\frac{x - x_i}{h}\right).$$

Here the bandwidth parameter $h > 0$ (APOS: **Bandwidth**) determines the degree of smoothing. The function k is called kernel function (APOS: **Kernel**). APOS photovoltaic StatLab implements the following kernels:

$$\textbf{Gaussian} \qquad k(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\,x^2\right)$$

$$\textbf{Rectangular} \qquad k(x) = \frac{1}{2\sqrt{3}}\,\mathbb{1}_{|x|<\sqrt{3}}$$

$$\textbf{Triangular} \qquad k(x) = \frac{1}{\sqrt{6}}\left(1-\frac{|x|}{\sqrt{6}}\right)\,\mathbb{1}_{|x|<\sqrt{6}}$$

$$\textbf{Epanechnikov} \qquad k(x) = \frac{3}{4\sqrt{5}}\left(1-\frac{x^2}{5}\right)\,\mathbb{1}_{|x|<\sqrt{5}}$$

$$\textbf{Biweight} \qquad k(x) = \frac{15}{16\sqrt{7}}\left(1-\frac{x^2}{7}\right)^2\,\mathbb{1}_{|x|<\sqrt{7}}$$

Having fixed a kernel, one still can vary the bandwidth. Small bandwidths localize the estimation. Asymptotic theory suggests the bandwidth choice $h \sim n^{-1/5}$. However, APOS photovoltaic StatLab allows the calculation of an optimal bandwidth. Here the method of least squares cross-validation is used, see [7] and [1], which minimizes the quantity

$$\int \left(\hat{f}_h(x) - f(x)\right)^2 dx = \int \hat{f}_h^2(x)\,dx - 2\int \hat{f}_h(x) f(x)\,dx + \int f^2(x)\,dx$$

Since the density f is unknown, one minimizes the expression

$$\int \hat{f}_h^2(x)\,dx - \frac{2}{M}\sum_{i=1}^{M} \hat{f}_{h,i}(x_i)$$

where

$$\hat{f}_{h,i}(x) = \frac{1}{M-1} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{1}{h}\, k\left(\frac{x - x_j}{h}\right).$$

For the Gaussian kernel that quantity is given by

$$\frac{1}{M^2} \sum_{i,j=1}^{M} \varphi_{(0,2h^2)}\left(x_i - x_j\right) - \frac{2}{M} \sum_{i=1}^{M} \hat{f}_{h,i}(x_i)$$

where $\varphi_{(0,2h^2)}$ stands for the density function of the normal distribution with mean 0 and variance $2h^2$.

References

[1] Bowman, A.W. (1984). An alternative method of cross-validation for the smoothing of density estimation. *Biometrika*, **71**:353-360.

[2] Casella, G. and Berger, R.L. (2002). *Statistical Inference*, 2nd Edition,Wadsworth Duxbury, Pacific Grove.

[3] Herrmann, W., Althaus, J., Steland, A. and Zähle, H. (2006). Statistical and Experimental Methods for Assessing the Power Output Specification of PV Modules, *Proceedings of the 21st European Photovoltaic Solar Energy Conference*, 2416-2420.

[4] Steland, A. and Herrmann, W. (2009). Evaluation of photovoltaic modules based on sampling inspection using smoothed empirical quantiles. *Progress in Photovoltaics*, **18(1)**:1-9.

[5] Montgomery D. *Introduction to Statistical Quality Control*. 5th ed., Wiley: New York, 2005.

[6] Steland, A., Akram, M. and Padmanabhan, A.R. (1982). Resampling methods for the nonparametric and generalized Behrens-Fisher problem. *Sankhya Series A*, **73-A(2)**: 267-302. *Scandinavian Journal of Statistics*, **9**:65-78.

[7] Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**:65-78.

[8] Steland, A. (2010). *Basiswissen Statistik*, 2nd edition, Springer-Verlag, Berlin/Heidelberg.

[9] Steland A. and Zähle H. (2009). Sampling inspection by variables: Nonparametric setting. *Statistica Neerlandica*, **63(1)**:101-123.