General Program Description

This program is designed to interpret the results of a sampling inspection, for the purpose of judging compliance with chosen limits. It may also be used to identify outlying values or departure from the assumed (Gaussian or Student's-t) distribution. Uncertainties for the individual values may be entered, and a mean and standard deviation for the set are calculated. The statistical test is based on a selection of the "Consumer's Risk" (CR) and the "Lot Tolerance Percent Defective" (LTPD), so that the stringency of the test may be adjusted. Typical values are CR=0.1 and LTPD=10%. More stringent tests may be made by choice of smaller values of the CR and LTPD. Confidence limits for the fitted distribution may be plotted for identification of outlying points, values that do not fit in the distribution.

Program Interaction and File Handling

The user interacts with the program through menu item selections and dialog boxes containing standard graphical user interface (GUI) components. Files may be saved in the program's CMP format by selecting the "File", "Save File As" menu item. The plot may be saved as a bitmap graphic (BMP) by selecting the "File", "Save Graph As" menu item. Previously saved files may be opened by selecting the "File", "Open" menu item. The graph may be copied to the Windows clipboard, as a BMP graphic, by selecting the "Edit", "Copy Graph" menu item (shortcut "ctrl C" key combination). This may then be pasted as an image into a word processing or presentation program.

Page Setup Options

Although the main purpose of the program is for mathematical interpretation of datasets, the end purpose will probably be the output of the graph for reporting purposes. To give the user some flexibility in what the printed or captured output looks like, some limited choices have been made available through the "File", "Page Setup" menu item. Here choices can be made for font name, font size for various text on the plot, and plot area size. Bold font may be selected separately for the graph title and the rest of the graph's text, although at this time, the Y-Axis font is unaffected by user choices. Font name choices are imported from all fonts on the user's system.

The user has been given the power to make choices that are unusable. Although there is not an "Undo" button, bad choices can be undone through new selections on the setup form. For instance, if a plot has been oversized to the point that the whole graph is not visible, and therefore not printable, the user may select reducing percentages options for the "Plot Area Size" to get back to a usable output. The results may be viewed by clicking on the "Plot Refresh" button, without exiting the "Page Setup" window. The best thing to do here is make some choices, click on the refresh button, and look at the results. When the screen portion looks good, go for the print menu item or the copy graph menu item.

These options were created because the screen and printed output views vary depending on the user's system screen resolution choices and on the video display adaptor and drivers for the display adaptor. The "Page Setup" options are helpful in overcoming these differences.

Program Limitations

The program has been tested to over 50,000 data points and it is believed to be able to handle a full Excel 2000 sheet row limitation of over 65,000 points. If a very large dataset is used, patience is required when opening the "Enable/Disable Data" and "Highlight Data" menu items. It takes a while for the program to load the selection grids, even with today's (2003) super-fast desktop computers. The size of the graph is not saved because it is dependant on the users screen resolution and display adaptor drivers. All graphs open up in the default size. It is a simple matter to resize the graph in the Page Setup window though.

Data Formats and Import Methods

To start a new graph, two methods of importing the data may be used. Data is first put into a spreadsheet type program, such as Excel. The file may then be saved as a comma separated value (CSV) file. This file may be opened in Cumulative Probability Plot using the "File", "Open" menu item. The open dialog box defaults to the program's file format of CMP type. Select CSV as the file type to open and all CSV files in the folder the user are viewing in the dialog will appear. Select the CSV file of interest and press the open button. A default plot of the data will then appear on the screen. The second method involves using the Windows clipboard. In the spreadsheet program, select the data you wish to plot and "copy" it to the clipboard (in most programs the shortcut is the "ctrl C" key combination). In the plotting program, select the menu item "Edit", "Paste Data." A dialog box will appear. If the first row of the data is text, the "label in first row" check box will be checked. This will automatically place the label in the spreadsheet onto the Y-axis label on the plot. If more than one spreadsheet column has been copied, a choice is given of which column to analyze (import into the program). If one of the multiple columns copied is uncertainty data, a selection of which column number contains the uncertainty values can be made here. The data preview section of the dialog box can assist the user in making this selection. Once the choices are correct, press the OK button and a plot of the selected clipboard data appears.

The user may plot two datasets on the same graph, if desired. The second dataset is called "overlay data" in the program. Overlay data is imported to the program through a "File - Open" dialog box by selecting the "Edit", "Input Overlay Data" menu item. Previously saved CMP file data or CSV data may be imported. If a CSV file is imported as overlay, then the dialog box to choose which column to analyze and which column contains uncertainty values appears for selection.

Graph Options

Plot Text and Titles

Graph display options are selected by selecting the "Graph Options" item, under the "Options" menu. The dialog box has four text boxes at the top where text data to be displayed on the graph may be typed in. "Plot Title" displays the text centered above the plot area. "Y Axis Label" displays the text to the left side (Y axis side) of the plot area. "Plot Text" displays any typed text in a movable box within the plot area. "Overlay Text" displays any typed text in a second movable box within the plot area. The latter two may be used to put extra descriptive text about the plot or line labels onto the graph. These boxes may be moved by cursor control.

Y Axis Option Buttons

The Y-axis buttons select whether the data is plotted on a linear or logarithmic scale. If the dataset contains zero or negative values or the uncertainty bars dip below zero with the logarithmic scale option selected, an error message appears and the plot returns to its previous state. Negative and zero containing data values may be deleted through the "Enable/Disable" menu item, allowing the log scale to be turned on.

Distribution Option Buttons

The choice between the Gaussian (normal) distribution and the Student's-t distribution may be useful for small datasets. (This option defaults to the Gaussian distribution for n greater than 29 points.) For small datasets, generally less than 10 points, the values may fit the Student's-t distribution better than the Gaussian distribution. This choice does not change the evaluation of the Test Statistic (Ts), or the value of Cumulative Probability corresponding to the effective LTPD that is achieved by setting the Test Statistic as the Limit. This correspondence is shown by the dotted lines on the graph, with the value of the effective LTPD near the X-axis.

Statistics Box

This box contains 3 check boxes for selecting statistical data to be displayed on the graph and a command button. The first check box determines whether or not the movable text box containing the statistical data (number of points, min, mean, max, sigma, and Ts) is displayed or not. The second check box selects whether or not to display the test statistic lines on the graph (shown as dotted lines). The third checkbox turns uncertainty bars through the plot points on and off, if uncertainty values have been imported.

The Statistics Box Options button brings up another dialog box for more display options for statistical data. Six check boxes give the user the option to display or not display the number of points, min, mean, max, sigma, and Ts respectively. The number of significant digits displayed in the statistics box data may be automatically selected by the program or manually selected by the user utilizing the option button/combo box choices in this dialog box. Finally, labels for the statistics boxes may be keyed in here. This is especially useful if overlay data is displayed to allow the graph viewer to ascertain which data box contains main and which one contains overlay data.

Background Subtraction and Limit Lines

If background has not been subtracted from the dataset, then the program can subtract it, if the user checks the appropriate check box and enters the background value into the accompanying text box. If the dataset has an upper limit, such as a regulatory limit, the user may enter this value and display limit lines, on the graph, at 100%, 90% and 50% of the limit.

Least Squares Fit, Confidence Intervals and Weighting

The Least Squares Fit is chosen by clicking on the “LSQ Fit” box in the Graph Options dialogue box. This then offers a choice for confidence interval bounds to be drawn about the fitted straight line, none, 90% confidence bound, 95%, or 99%. A straight line is fit to the data points. If uncertainties have been provided, a weighted fit may be chosen by clicking on the “Weighting/Weighted” box. (The default is “Un-Weighted”). Points whose uncertainties do not lie within or extend into the confidence bound region should be considered to be outliers. If the confidence bound region does not include the Gaussian line, the distribution of points is poorly represented by a Gaussian (or the Student’s-t distribution, if that has been chosen).

Lot Tolerance Percent Defective (LTPD) and Consumer Risk (CR) Parameters

The program was originally written for and is used at a USDOE site mostly for demonstrating to regulators that radiological facilities are suitable for unrestricted use after remediation and cleanup activities are complete. Originally this was before the MARSSIM guide was written but is still used today as another demonstration of the MARSSIM conclusions reached. The parameters will be explained in those terms.

CR and LTPD, used for testing whether or not a lot of data are acceptable or not, come from recommendations by the State of California, Department of Health Services, Radiological Health Branch for facility release in the 1980s. CR is also sometimes referred to as beta. Their recommended values for the test are beta = 0.1 and LTPD = 10%. This means that, if a lot just passes the acceptance test, there is one chance out of ten (0.1) that 10% of the total number of measured locations would have residual contamination exceeding the limit. This logic may be applied to other statistical data evaluations in regards to applying CR and LTPD choices.

Axes Options

X-Axis Range and Y-Axis Scaling

The programs tries to pick an appropriate X-axis range in the auto mode (default) but the user may expand or contract the range manually with option buttons in this dialog box (99%, 99.9%, 99.99% or 99.999%). The program also defaults to auto-picking the scaling of the Y-axis, but the user may choose their own Y-Min, Y-Max and tick intervals by selecting the manual option and entering values in text boxes.

Y-Label Format, Y-Label Precision, and Data Symbols

Here, the user may select decimal (normal) or scientific notation (e-format) display options for the Y-axis label. The number of decimal places displayed may be changed here also. Three option buttons also allow the user to select circles, squares or triangles for plotting the main data points on the graph. If overlay data is displayed, the program automatically uses the next symbol on the list for that data’s plot.

Eliminating and Highlighting Data

Manipulating Individual Data Points

Zero and negative values and outliers (or any data the user wishes) may be deleted and added in again later by selecting the “Enable/Disable Data” item under the “Edit” menu. The dialog box that appears contains a spreadsheet/database type grid containing the individual data. If the user clicks anywhere on a data row, an x appears denoting that the point will no longer be displayed on the graph. When OK is clicked the graph and statistic boxes are automatically updated for the new number of points. This can be handy for see what the outliers are doing to the line approximation or for viewing datasets, with negative or zero values, on a log Y-scale.

An identical dialog box appears when the user selects the “Highlight/Unhighlight Data” item under the edit menu. Using the same selection process as above, individual data points may be colored in for discussion and description in accompanying report text.

Plot Interpretation

This section will be described in terms of releasing previously contaminated radioactive facilities, the main use for the program at USDOE projects

Statistical analysis is used to convert a large amount of data into a manageable amount of understandable information. This process can involve a variety of techniques, the simplest being to determine the average (or mean) value for a given set of data. This simple determination is improved upon by also calculating the standard deviation of the data about the mean, which gives an estimate of the variability of the data. In many cases, this variability represents variations both in the characteristics or values being measured and in measurement technique fluctuations.

The significance of these quantities (mean and standard deviation) depends upon the distribution assumed for the data. Sometimes there is a theoretically known distribution for a particular measurement process, such as the binominal, or Poisson distribution for counting radioactivity. These distributions are relatively well approximated by the Gaussian, or normal, distribution. In fact, the Gaussian distribution approximates the distribution of many different kinds of measurements and for simplicity is generally assumed to be the proper distribution. The Gaussian distribution is generally seen in the form of a bell-shaped curve, with most values occurring near the mean value and fewer and fewer values existing at increasing distances from the mean, both greater and less than the mean.

However, it is difficult to derive the bell-shaped curve from experimental data unless the data are specifically selected to demonstrate the curve, and deviations from the distribution are difficult to see. A better version is the so-called “cumulative probability function” utilized in this program, which forms an S-shaped curve when plotted in the usual manner. This can be further improved by adjusting the abscissa (the “X” values in an X-Y graph) so that the “S” curve becomes a straight line. This is a standard statistical technique and is the basis for special graph paper used for probability analysis of data. The parameters of the Gaussian distribution (the mean and the standard deviation) are determined by the usual calculation methods:

Mean = $\overline{X}$ = $\frac{\sum X_i}{N}$ Standard Deviation = s = $\left[\frac{\sum\left(X_i-\overline{X}\right)^2}{N-1}\right]^{1/2}$

where X_i represents the individual data values, and N is the number of points.

Where the data are not well-represented by a Gaussian distribution (and this is true in most cases) the departure is readily apparent; the data points do not lie along a straight line representing the Gaussian distribution. In most cases, this departure takes a single typical form. Much of the data lies along the theoretical straight line, with a few points at either extreme lying somewhat above it. This form can usually be interpreted as showing a large number of uncontaminated measurements where the variability is due to random fluctuations in the measurements themselves, with the balance being locations that harbor more or less residual contamination.

If the contaminated area is large, there will be many points departing from the curve. In these cases, the points will not fit the theoretical straight line. If most of the region in question is contaminated, the distribution will be dominated by the contaminated data points, in a line of points generally sloping from the lower left to the upper right, fitting more or less closely, a theoretical straight line.

This program is used to provide a sampling inspection test. It uses a standard quality control technique called inspection by variables, in which the distribution of the measured values is used to predict the probability that other unmeasured values would exceed a specified limit. The standard test method requires calculating the mean ($\overline{X}$) and the standard deviation(s). Then, depending the values chosen for certain parameters that reflect the performance of the test in accepting bad lots, or rejecting good lots, the necessary number of samples is determined and a multiplier, k, is computed so that the inequality $\overline{X}+ks<U$ where U is the acceptance limit, representing an acceptable lot. The parameters used in the program to calculate k are the CR and LTPD discussed elsewhere in the help system. This value of $\overline{X}+ks$, that is compared to the limit is what is called the Test Statistic, or Ts. The value of Ts is a point near the upper end of the observed data distribution. If this value is less than the acceptance limit U, the lot has passed the Sampling Inspection by Variables Test, according to the criteria chosen for CR and LTPD.

The usual manner of applying this inspection method is to use tables giving the value of the sample size (N) and multiplier (k) for the selected values of CR and LTPD. The program uses the number of measured values (N) in the lot to compute k, and this value is used to calculate $\overline{X}+ks$. The formula for calculating k is:

$$k=\frac{K_2+\sqrt{{K_2}^2-ab}}{a} \quad \text{with} \quad a=1-\frac{{K_{beta}}^2}{2(N-1)} \quad \text{and} \quad b={K_2}^2-\frac{{K_{beta}}^2}{N}$$

The value of K_2 is that for the variable of a Gaussian distribution corresponding to the LTPD value, and the value of K_{beta} is that for the Gaussian variable corresponding to CR.

Plot Printing

When the “File”, “Print” menu item is selected, the standard Windows printer selection dialog box appears. Here the user may select basic printer options, as well as which printer to send the graph to. The program has been tested and works fine with network printers. “Portrait” or “Landscape” printing may be selected in a separate dialog box, by clicking the “File”, “Page Setup” menu item. At this time, the graph is a bitmap screen capture of exactly what appears in the main window. A neat little trick is to “restore down” the program window and manually manipulate the window size to hide say the filename and date at the bottom of the graph*. Anything hidden will not print on the graph. Vector graphics and system fonts are planned for future versions of the program for a more professional output.

*Restore down is accomplished by clicking on the middle of the 3 buttons in the top right corner of the window. This button toggles the window between “window maximized” and “window restored down” states. In the “restored down” state, the window may be dragged around the desktop and manually resized by dragging any side of the window in or out with the mouse. The only thing that prints (or gets copied to graph with Ctrl C) is what appears in the window, so anything that you hide by resizing the window does not print or get copied to the clipboard.