# PH: Piloting Histogram
# Version 2.0

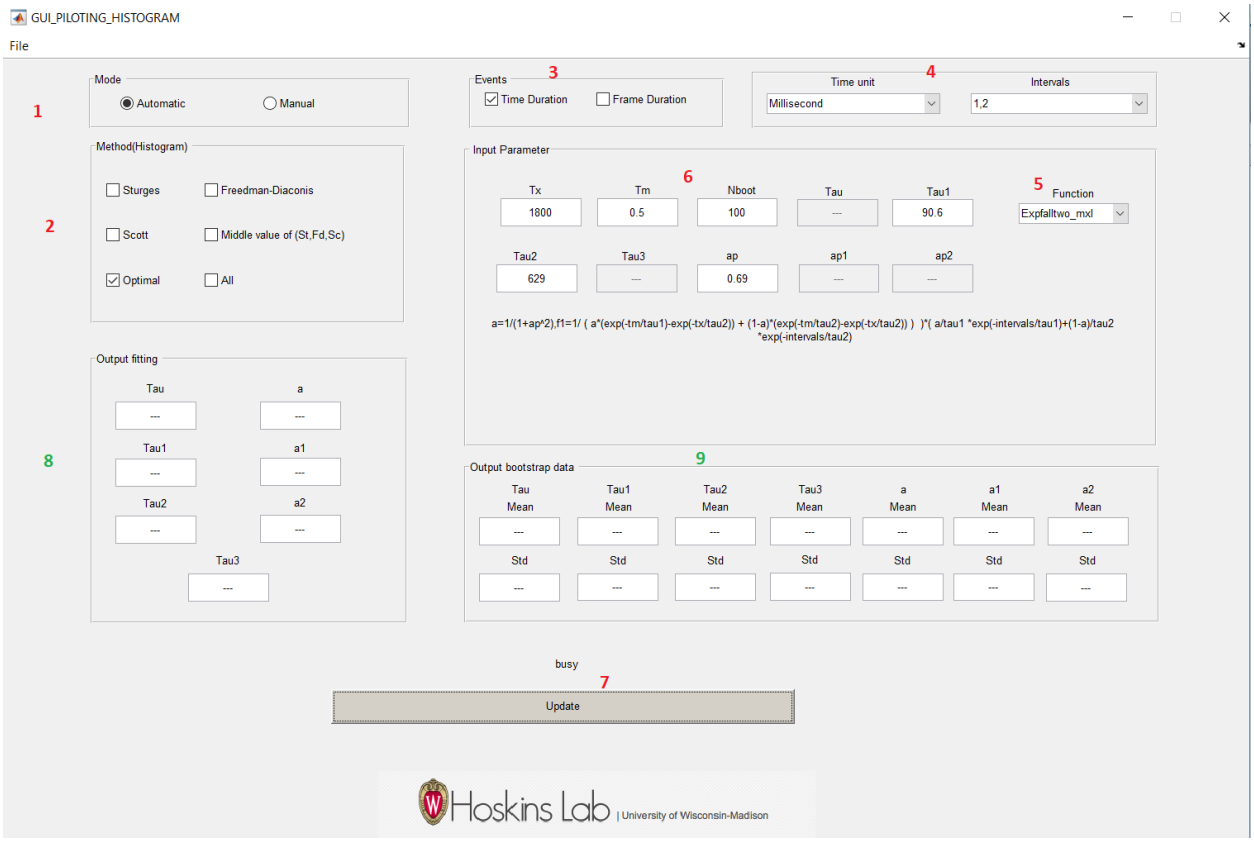



Figure 1.1: GUI_Piloting_Histogram interface (example data (20130808_Rp51a_WTWCE_fit))

**1**-PH software allows the user to work in either automatic or manual mode. If you select automatic mode, the software automatically calculates the ideal number of bins to use in a histogram. If you select the manual mode, the user should enter the edges of the histogram bins (which must contain monotonically increasing values e.g. 0 20 60 120 240 420 600) (Figure 1.1).

**2**- If you select the automatic mode, the software calculates the number of bins in the histogram, using a chosen method. The type of return value depends upon the method chosen. Possible choices for methods are (Figure 1.1):

- Freedman-Diaconis method
  - The Freedman-Diaconis rule[1] is less sensitive to outliers in the data, and might be more suitable for data with heavy-tailed distributions. It uses a bin width, h as

$$h = \frac{IQR(X)}{n^{1/3}}$$

    where n is the number of the data points and X is the data set (in our case X is the event length) and IQR is the interquartile range (midspread or middle 50%) of X.
- Scott
  - Scott's rule [2,3] is optimal if the data is close to being normally distributed. This rule is appropriate for most other distributions, as well. It uses a bin width, h as $h = \frac{3.5*\sigma}{n^{1/3}}$

    where n is the number of the data points and σ is standard deviation of the data set X (in our case X is the event length).

- Sturges
  - Sturges' rule [4] is popular due to its simplicity. It implicitly bases the bin sizes on the range of the data and can perform poorly if $n < 30$, because the number of bins will be small—less than seven—and unlikely to show trends in the data well. It may also perform poorly if the data are not normally distributed (Note: event length is exponential distribution so this method may not work well.) [5]. It calculates the number of bins,

$$\mathrm{k} = (1 + log_2(n))$$

- Middle
  - Uses all three above methods, then picks the middle (median) value
- Optimal

- Optimization principle is to minimize expected Least Square (L2) loss function between the histogram and an unknown underlying density function [6]. This method assumes that samples are drawn from the density independently from each other. The optimal bin width, h* is obtained as a minimizer of the formula, (2M-V) / h^2, where M and V are mean and variance of sample counts across bins with a width h. Optimal number of bins, k is given as

$$k = \frac{(max(X) - min(X))}{h^*}$$

- All (A structure is returned with fields Freedman-Diaconis, Scott, Sturges, Middle, Optimal)

**3**- The software allows the user to select time duration, frame duration, or both (Figure 1.1).

**4**- PH is set up to allow the user to select a time unit and events (1/3/-3) to analyze or any combinations of these events (Figure 1.1).

**5**- PH fits a histogram to a user selected exponential distribution. The distribution functions available are

- 'Expfallone_mxl'(Single exponential distribution)

$$f(t) = \frac{1}{\left(e^{\frac{-t_m}{\tau}} - e^{\frac{-t_X}{\tau}}\right)} \cdot \frac{1}{\tau} \cdot e^{\frac{-t}{\tau}} \quad (1)$$

where $t_m$ is the minimum time.interval that can be resolved in the experiment, $t_X$ is the maximum duration of the experiment and $\tau$ is the rate constant.

- 'Expfalltwo_mxl' (Biexponential distribution)

$$f(t) = \frac{1}{A_1\left(e^{\frac{-t_m}{\tau_1}} - e^{\frac{-t_X}{\tau_1}}\right) + A_2\left(e^{\frac{-t_m}{\tau_2}} - e^{\frac{-t_X}{\tau_2}}\right)} \left(\frac{A_1}{\tau_1} e^{\frac{-t}{\tau_1}} + \frac{A_2}{\tau_2} e^{\frac{-t}{\tau_2}}\right) \quad (2)$$

where $A_2 = 1 - A_1$, $A_1 = \frac{1}{(1+ap^2)}$, $t_m$, is the minimum time interval that can be resolved in the experiment, $t_X$ is the maximum duration of the experiment, $\tau_1$ and $\tau_2$ are the rate constants, and $A_1$ and $A_2$ are the amplitudes.

- 'Expfallthree_mxl' (Triexponential distribution)

$$f(t) = \frac{1}{A_1\left(e^{\frac{-t_m}{\tau_1}} - e^{\frac{-t_X}{\tau_1}}\right) + A_2\left(e^{\frac{-t_m}{\tau_2}} - e^{\frac{-t_X}{\tau_2}}\right) + A_3\left(e^{\frac{-t_m}{\tau_3}} - e^{\frac{-t_X}{\tau_3}}\right)} \left(\frac{A_1}{\tau_1} e^{\frac{-t}{\tau_1}} + \frac{A_2}{\tau_2} e^{\frac{-t}{\tau_2}} + \frac{A_3}{\tau_3} e^{\frac{-t}{\tau_3}}\right) \quad (3)$$

where $A_3 = 1 - A_1 - A_2$, $A_1 = \frac{1}{(1+ap_1{}^2)}$, and $A_2 = \frac{1-A_1}{(1+ap_2{}^2)}$, $t_m$, is the minimum time interval that can be resolved in the experiment, $t_X$ is the maximum duration of the experiment, $\tau_1$, $\tau_2$, and $\tau_3$ are the rate constants, and $A_1, A_2, \text{and } A_3$ are the amplitudes.

- 'Expfallone_all_events_mxl', 'Expfalltwo_all_events_mxl', and 'Expfallthree_all_events_mxl' use the same formulas as listed above for their respective exponential distributions but address molecules that last to the end of the recording (+3 in Glimpse annotation). A separate vector list for spots that do not vanish should be available and labeled ‘mxintervals’. If this is not found, using one of these methods will return an error message.

**6**- Based on the selected distribution function, the user should enter different initial parameters (Figure 1.1).

1. Tx= maximum duration of the experiment.
2. Tm= minimum time interval that can be resolved in the experiment.
3. Nboot =number of times the selected function will be fitted on the random subset of input data.
4. Fitting Parameters: For Single exponential distribution enter the initial guess for Tau (see Eq.1). For bi-exponential distribution fitting enter the initial guesses for Tau1, Tau2, and ap (see Eq.2). For tri-exponential distribution fitting enter the initial guesses for Tau1, Tau2, Tau3, ap1 and ap2 (see Eq.3)

**7**- Pressing the “Update” button will run the software. The user should select the file that one wish to analyze as well as the output directory for saving the output data (One subfolder will be created here for the output data) (Figure 1.1).

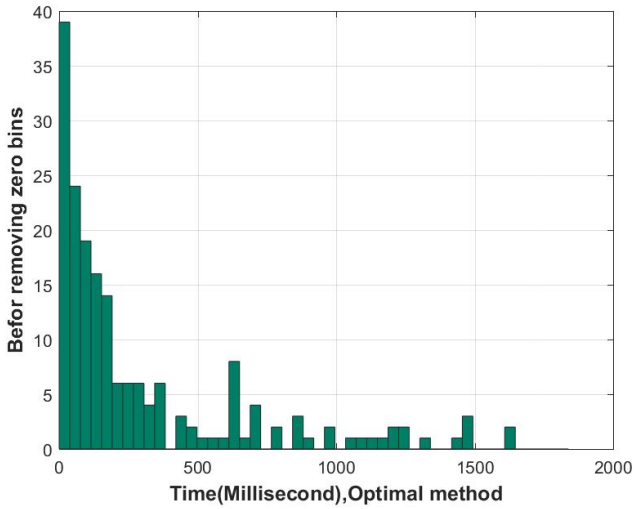
## Output

In parts **8** and **9** the user can see the results of the fitting and bootstrapping analysis (Figure 1.1). Part **8** displays the output of the fitting.

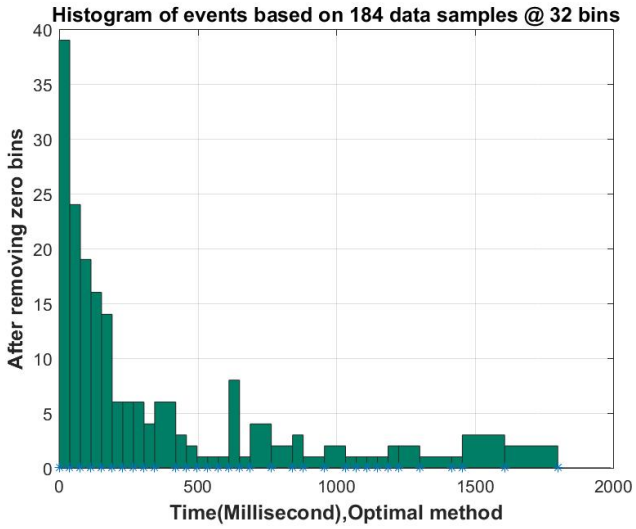
For nboot =1000, Bootstrap analysis will create 1000 values for the fitting parameter like Tau for single exponential distribution fit. These 1000 values are next fit with a normal distribution function to obtain a mean and standard deviation values for Tau. Similarly, for double exponential distribution fit 1000 values of Tau1, Tau2 and ap are calculated. Each of these values are fit with a normal distribution function to obtain the mean and

stand deviation values for Tau1, Tau2 and ap respectively. Part **9** gives these mean and standard deviation (Std) values determined by bootstrapping.

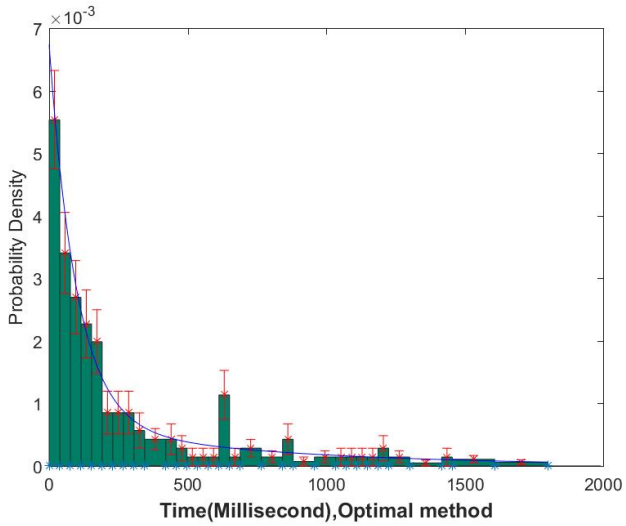
In addition, the output figures and data (result.mat) will be saved in the output directory. Example output figures are shown below.



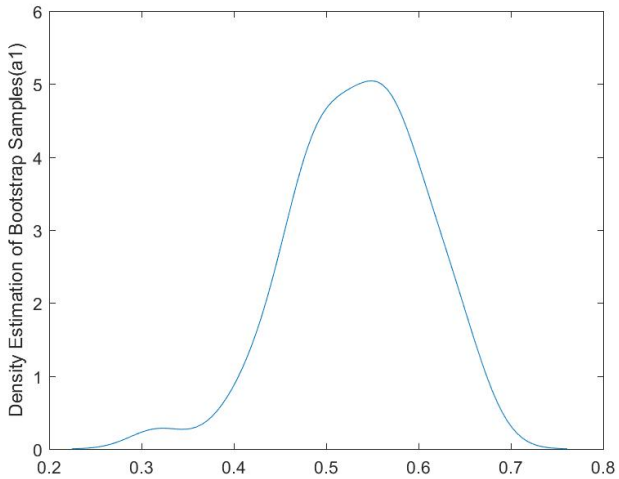

1-Histogram before removing zero bins



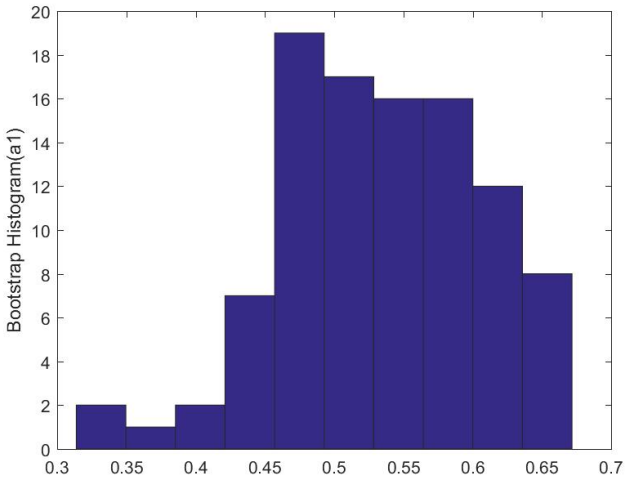

2- Histogram after removing zero bins




3-Probability density with a distribution function fit plotted on top.




4-Density estimation of the bootstrapping samples. We use it to examine the shape of our bootstrap distribution.




5- Bootstrap Histogram. We usually use the histogram to examine the shape of our bootstrap distribution. A histogram divides sample values into many intervals and represents the frequency of data values in each interval with a bar.

**References:**

[1] Freedman, David, and Persi Diaconis. "On the histogram as a density estimator: L 2 theory." Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57.4 (1981): 453-476.

[2] Scott, David W. Multivariate density estimation: theory, practice, and visualization. John Wiley & Sons, 2015.

[3] Scott, David W. "On optimal and data-based histograms." Biometrika 66.3 (1979): 605-610.

[4] Sturges, Herbert A. "The choice of a class interval." Journal of the american statistical association 21.153 (1926): 65-66.

[6] Hyndman, Rob J. “The problem with Sturges’ rule for constructing histograms.” Online publication available at https://robjhyndman.com/papers/sturges.pdf

[5] Shimazaki, Hideaki, and Shigeru Shinomoto. "A method for selecting the bin size of a time histogram." Neural computation 19.6 (2007): 1503-1527.