Notes on Runlengths

John Fricks

February 1, 2006

This note is to explore possible data analysis techniques for run length data from KIF3A/B-GFP, KIF3A/A-GFP and KIF3B/B-GFP in single molecule fluorescence assays. (For simplicity, I will refer to these data sets as *ab*, *aa*, and *bb* respectively.) The method used in Yongrong's thesis consists of minimizing the least square distance between the empirical (observed) cumulative distribution function and the theoretical cumulative distribution function. This method is certainly a dramatic improvement over fitting the probability density function to histograms of the data. And we will see that in simulations, this estimator performs very well. However, we can suggest an estimator which has slightly better performance and also allows us to calculate things such as a confidence interval for the true mean of the exponential distribution (which we will call a).

Each data set consists of independent observations $X_1, ..., X_n$ from a shifted exponential distribution with density

$$f(x) = \frac{1}{a}\exp(-(x - x_0)/a) \tag{1}$$

The x_0 will be treated as fixed, and we will only attempt to estimate a. To simplify matters we will transform the data by subtracting x_0 from the sample to obtain a new sample $Y_1, Y_2, ..., Y_n$ where $Y_i = X_i - x_0$. Each observation, Y_i, of this new sample is drawn from a regular exponential distribution with density

$$g(y) = \frac{1}{a}\exp(-y/a) \tag{2}$$

First, let us find the maximum likelihood estimator for a. This method frequently gives estimators that have small variances, and there is a wealth of asymptotic results that allow for things like confidence intervals. The likelihood function is the joint probability density for the sample (viewed as a function of the parameter–in this case a,

$$L(a) = g(y_1, y_2, ..., y_n) = \prod_{i=1}^{n} g(y_i) = a^{-n} \exp(-\sum_{i=1}^{n} \frac{y_i}{a}) \tag{3}$$

Now, we would like to find a value for a that maximizes this function. Maximizing the logarithm of this function will give us the same value. So, we need to maximize

$$\log L(a) = -n \log a - \frac{\sum_{i=1}^{n} y_i}{a}. \tag{4}$$

Taking the derivative with respect to a and setting the derivative equal to zero, we obtain

$$0 = \frac{-n}{a} + \frac{\sum_{i=1}^{n} y_i}{a^2}. \tag{5}$$

Solving for a gives us

$$a = \frac{\sum_{i=1}^{n} y_i}{n}. \tag{6}$$

Therefore, the maximum likelihood estimator for a is $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. There is another nice feature of this estimator which is that we know the distribution of $\bar{Y}$ exactly. In general, the sum of independent exponential random variables has a gamma distribution. So, $\frac{1}{n} \sum_{i=1}^{n} Y_i$ has a gamma distribution with shape parameter equal to the sample size n and with scale parameter equal to $\frac{a}{n}$. (Thus, the mean of this gamma distribution is a, and the variance is $\frac{a^2}{n}$). Therefore, $\frac{\bar{Y}}{a}$ has a gamma distribution with mean one and variance equal to $1/n$. So, we may create a $100(1-\alpha)$ percent confidence interval using the following expression:

$$P\left(q_{\frac{\alpha}{2}} \leq \frac{\bar{Y}}{a} \leq q_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

where q_r is the rth percentile of a gamma distribution with shape parameter equal to $1/n$ and scale parameter equal to 1. Isolating the parameter a, we obtain

$$P\left(\frac{\bar{Y}}{q_{1-\frac{\alpha}{2}}} \leq a \leq \frac{\bar{Y}}{q_{\frac{\alpha}{2}}}\right) = 1 - \alpha.$$

This estimator also has at least two other attractive features. The first is that it is unbiased; that is, the expected value of the estimator is equal to the parameter we are trying to estimate, a. Another is that this estimator has the minimum variance among all possible unbiased estimators of a. The Cramer-Rao inequality allows us to calculate the smallest possible variance among unbiased estimators of a. If that minimum equals the variance of the sample mean ($Var(\bar{Y}) = \frac{a^2}{n}$), then we know that our estimator has a variance as small or smaller than every other unbiased estimator and is in some way optimal. The formula for the minimum variance when the sample is independent and identically distributed (as in this situation) is

$$\frac{1}{nE(\partial_a \log g(Y_i))^2}$$

where $g(\cdot)$ is the density of Y_i which gives

$$\frac{1}{nE(\partial_a \log \frac{1}{a}\exp(-Y_i/a))^2} = \frac{1}{nE(\frac{Y_i}{a^2} - \frac{1}{a})^2} = \frac{a^2}{n}$$

So, we see that for unbiased estimators of the parameter a, the sample mean has the least variance.

Now, we compare this estimator with a least square fitting of the cumulative distribution function to the empirical cumulative distribution function(as was done previously). If we order our sample from the smallest observation to the largest (which we will write as $Y_{(1)}, ..., Y_{(n)}$), then we can write this estimator as

$$\hat{a} = argmin_a \sum_{i=1}^{n} \left(\frac{i}{n} - (1 - e^{-Y_{(i)}/a}\right)^2$$

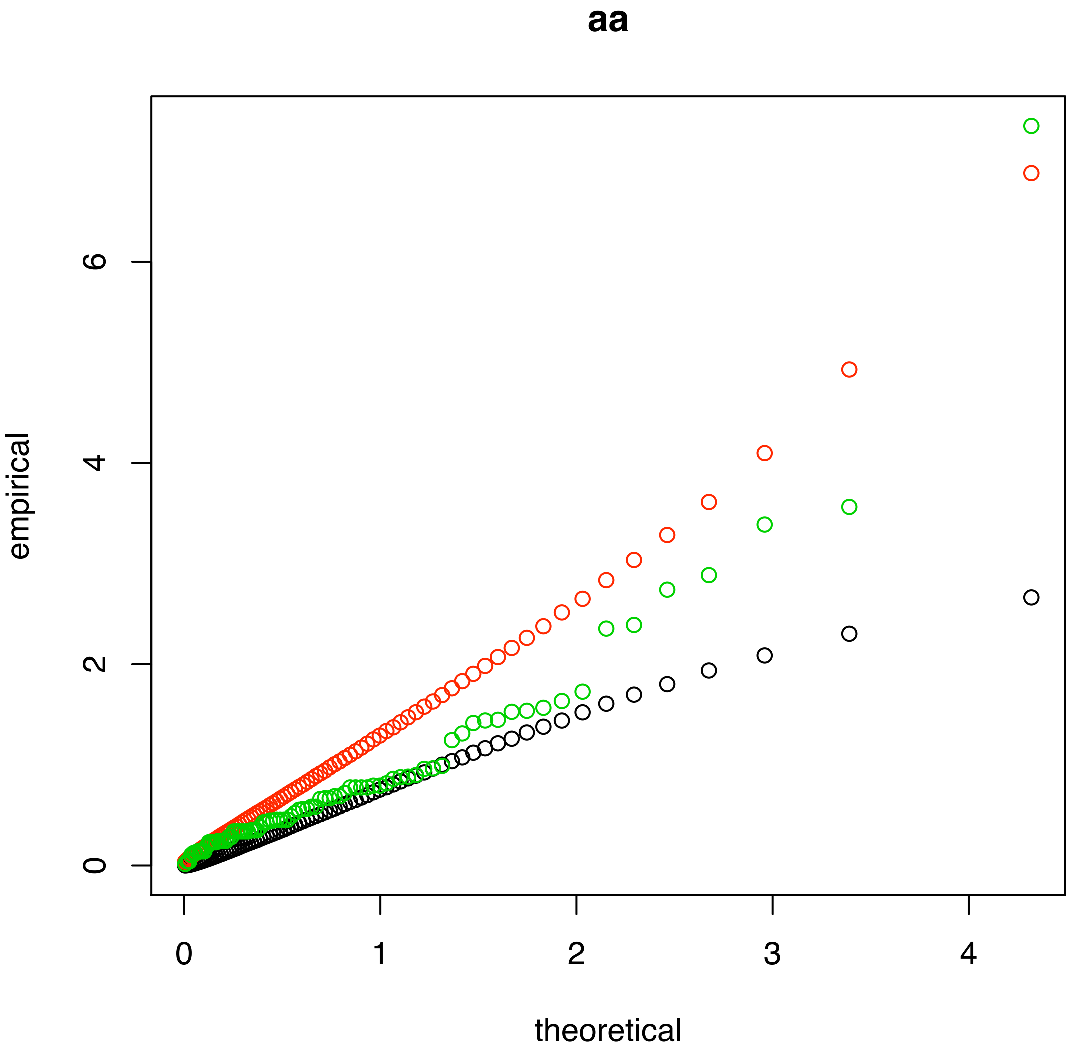
This minimization is done numerically which makes theoretical calculations on this estimator (such as finding $Var(\hat{a})$ or showing that $\hat{a}$ is unbiased) very difficult. However, we can simulate samples and then compare results using the two different methods. So, 10,000 independent random samples were taken each consisting of 45 observations from an exponential distribution with a mean of 0.6. Then, both estimation methods were performed on each sample to obtain 10,000 observations for each of the methods. We will call these $\hat{a}_i$ and $\bar{a}_i$ for $i = 1, ..., 10000$. The mean and variance for the $\bar{a}_i$'s are 0.601 and 0.00792 respectively, and the mean and variance for the $\hat{a}_i$'s are 0.63 and 0.01124 respectively. There are two important things to notice; there is some evidence that $\hat{a}$ is biased since the mean is considerably farther from the population mean of 0.6 than is the sample mean, $\hat{a}$. In addition, the variance of the $\hat{a}_i$'s is larger. So, we see that $\bar{a}$ is centered on the true parameter, a, and is in some way closer to it.

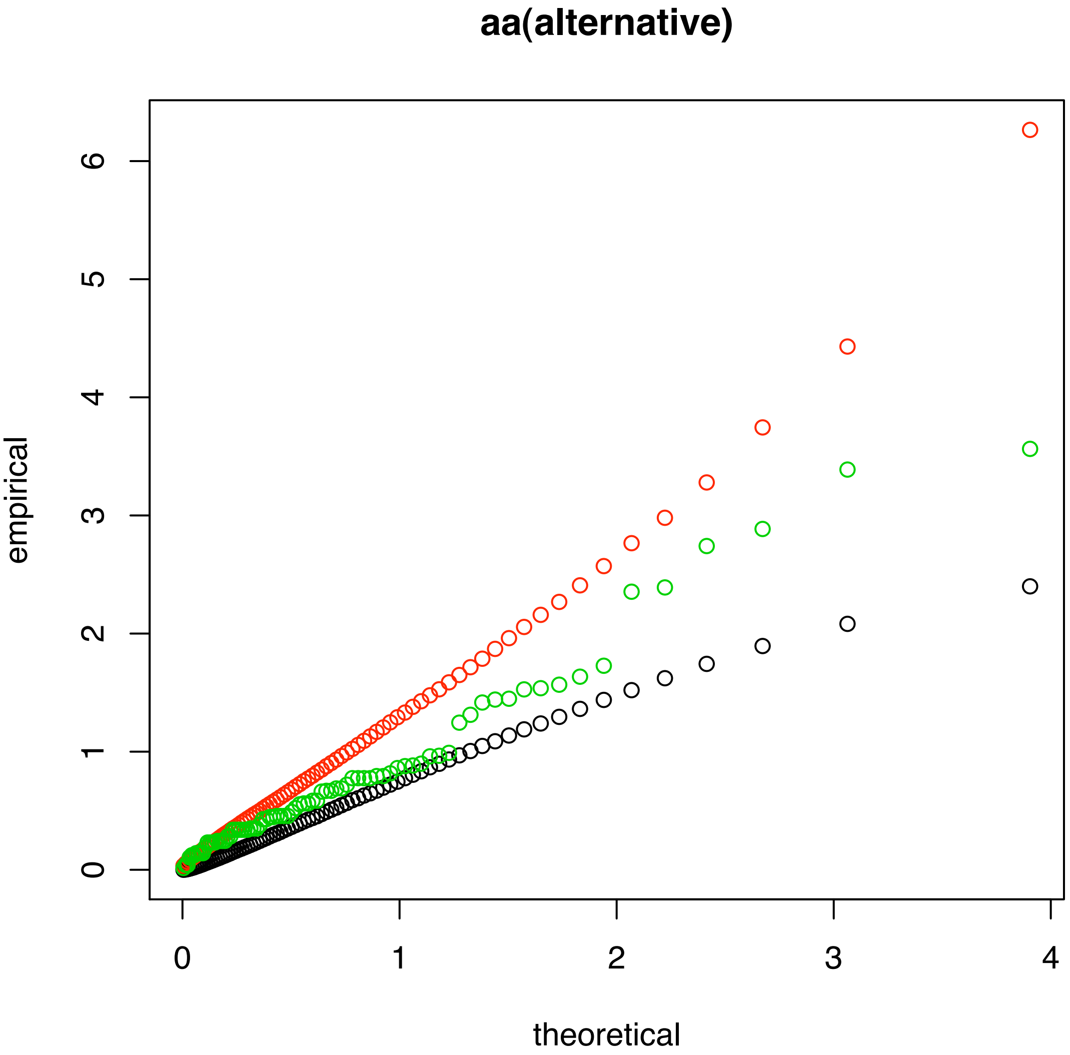
Now, we look at some of the data. First, notice that the estimated value for a correspond fairly closely to the estimate found in the thesis of 0.67, 0.77, and 0.58 for ab, aa, and bb respectively. In addition, we can report the

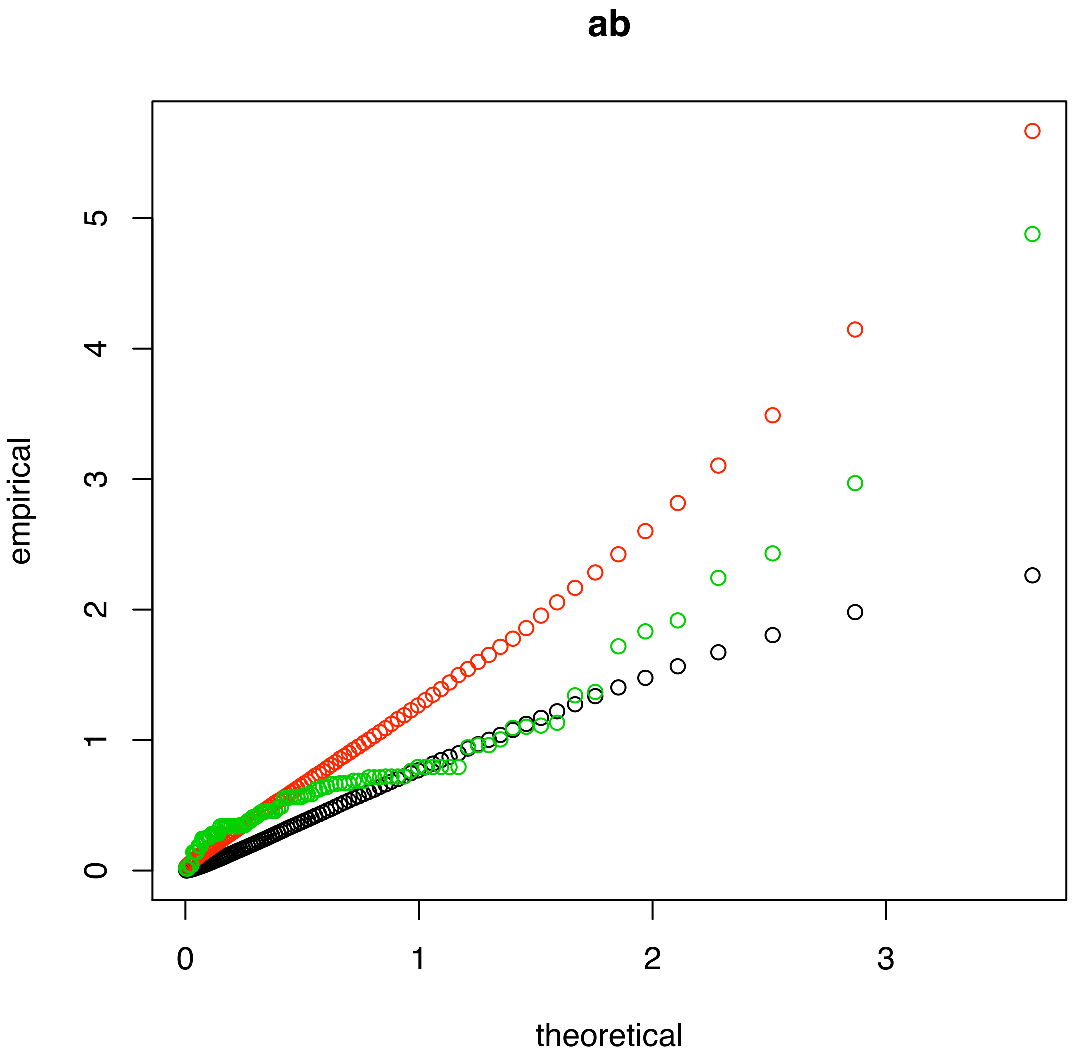
type	$\bar{a}$	lower bound	upper bound
ab	0.691202	0.5708332	0.8543276
aa	0.8449693	0.6890348	1.060860
bb	0.6440058	0.4892249	0.886326

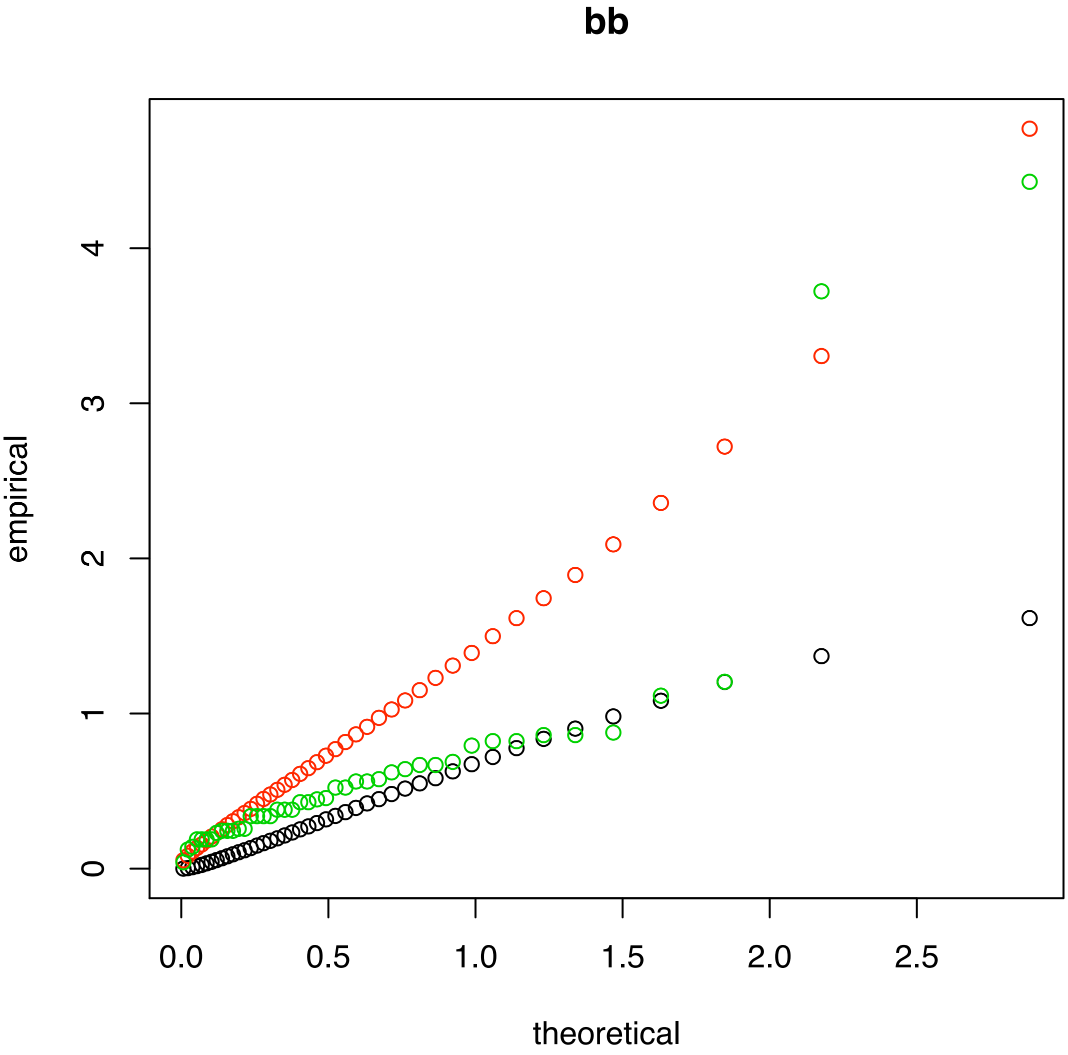
upper and lower bound of a 95% confidence interval for the parameter using the method outlined above.

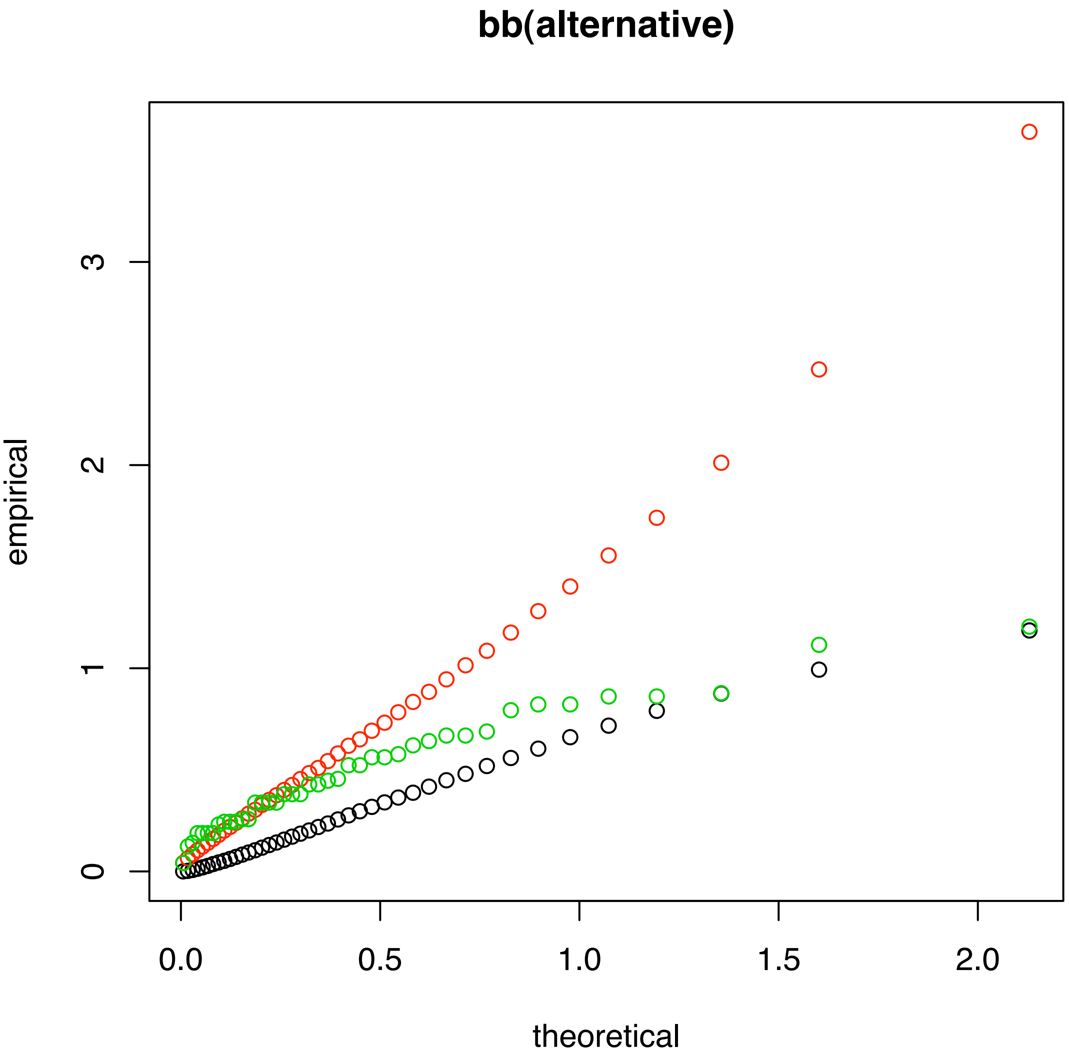
One of the great advantages of the curve fitting method is that one has a visual representation of the quality of the estimation procedure as presented in figure 3.9 of Yongrong's thesis. An alternative graphical approach is to look at a qq plot of the data. This graphs the theoretical quantiles for an exponential distribution (with a mean equal to the sample mean) with the empirical quantiles from the data. If the data truly comes from an exponential distribution and one has a sample which is very large, then the qq plot will be very close to a straight line. Below one can see the qq plots for the data. A 95% confidence envelope is also designated on the plots. One thing to notice is that in the plot for the *aa* data, the largest value matches up poorly with the theoretical quantiles. This may imply that this observation is an outlier. After deleting this data point, we get the plot labeled $aa(alternative)$ where the plot appears much more contiguous. Deleting this point changes the mean from 0.845 to 0.766, a rather dramatic shift; the confidence interval for a becomes $[0.624, 0.963]$. A similar phenomena appears in the *bb* data with two values that appears too extreme from the theoretical quantiles. After removing these two points, the mean shifts from 0.644 to 0.48, and the new confidence interval is $[0.363, 0.667]$. The *ab* data seems to reasonably follow an exponential distribution. It is not completely clear that these possible outliers should be deleted, but it might be good to consider whether there may be an occasional observation which is not consistent with the other observations.


aa
empirical
theoretical
0
2
4
6
0
1
2
3
4


aa(alternative)
empirical
theoretical
0
1
2
3
4
5
6
0
1
2
3
4


ab
empirical
theoretical
0
1
2
3
4
5
0
1
2
3


bb
empirical
0
1
2
3
4
theoretical
0.0
0.5
1.0
1.5
2.0
2.5


bb(alternative)
empirical
theoretical
0
1
2
3
0.0
0.5
1.0
1.5
2.0