

The data below are pH and DO data generated by the 314 class (environmental engineering) taught during the fall quarter of 1999. Each student, in sequence, was given a beaker and specific instructions how to fill it with tap water from a lab spigot left running continuously during the experiment. Each student carried out the measurements on the same meters in the same way.

Assuming the pH and DO of the tap water is constant (?), each student should have been attempting to measure the same value.

|                   |      |      |             |
|-------------------|------|------|-------------|
|                   | 7.25 | .91  |             |
|                   | 7.3  | .84  |             |
|                   | 7.2  | 1.15 |             |
|                   | 7.37 | 1.01 |             |
|                   | 7.42 | .73  |             |
|                   | 7.42 | 1.21 | i := 0.. 28 |
|                   | 7.38 | .93  |             |
|                   | 7.4  | 1.35 |             |
|                   | 7.41 | .8   |             |
|                   | 7.42 | 1.34 |             |
|                   | 7.49 | 1.14 |             |
|                   | 7.47 | 1.52 |             |
|                   | 7.40 | 1.22 |             |
|                   | 7.37 | 1.06 |             |
| pH_and_DO_data := | 7.18 | 1.03 |             |
|                   | 7.3  | .93  |             |
|                   | 7.42 | .87  |             |
|                   | 7.45 | .84  |             |
|                   | 7.42 | 2.34 |             |
|                   | 7.35 | 2.38 |             |
|                   | 7.44 | 2.06 |             |
|                   | 7.49 | 1.05 |             |
|                   | 7.42 | 1.2  |             |
|                   | 7.2  | 1.01 |             |
|                   | 7.44 | 1.83 |             |
|                   | 7.38 | .80  |             |
|                   | 7.37 | .58  |             |
|                   | 7.42 | .725 |             |
|                   | 7.42 | .82  |             |

**PRELIMINARY CALCULATIONS**

rename the first column :  $pH_i := pH\_and\_DO\_data_{i, 0}$   $length(pH) = 29$

rename the second column :  $DO_i := pH\_and\_DO\_data_{i, 1}$

=====

compute the log of each pH measurement and store :  $log\_pH_i := ln(pH\_and\_DO\_data_{i, 0})$

compute the log of each DO measurement and store:  $log\_DO_i := ln(pH\_and\_DO\_data_{i, 1})$

=====

compute the mean of the DO data set :  $m_{DO} := mean(DO)$

compute the mean of the logs of the DO data :  $log\_m_{DO} := mean(log\_DO)$

=====

compute the mean of the pH data :  $m_{pH} := mean(pH)$

compute the mean of the logs of the pH data :  $log\_m_{pH} := mean(log\_pH)$

=====

compute the median of the pH data :  $median_{pH} := median(pH)$

compute the median of the DO data :  $median_{DO} := median(DO)$

compute the mode of the pH data :  $mode_{pH} := mode(pH)$

=====

compute the standard deviation of the pH data :  $s_{pH} := stdev(pH)$

compute the standard deviation of the DO data :  $s_{DO} := stdev(DO)$

=====

compute the standard deviation of the logs of the pH data :  $log\_s_{pH} := stdev(log\_pH)$

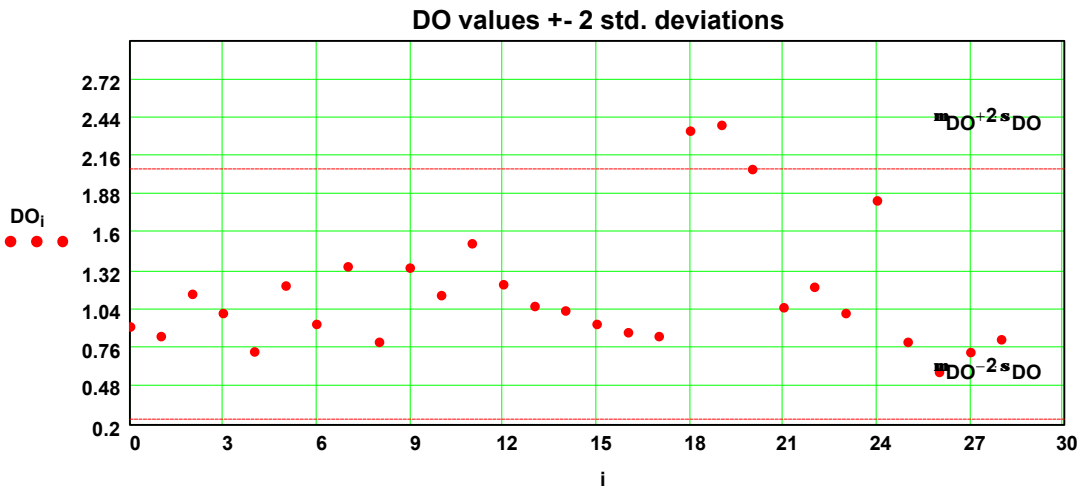
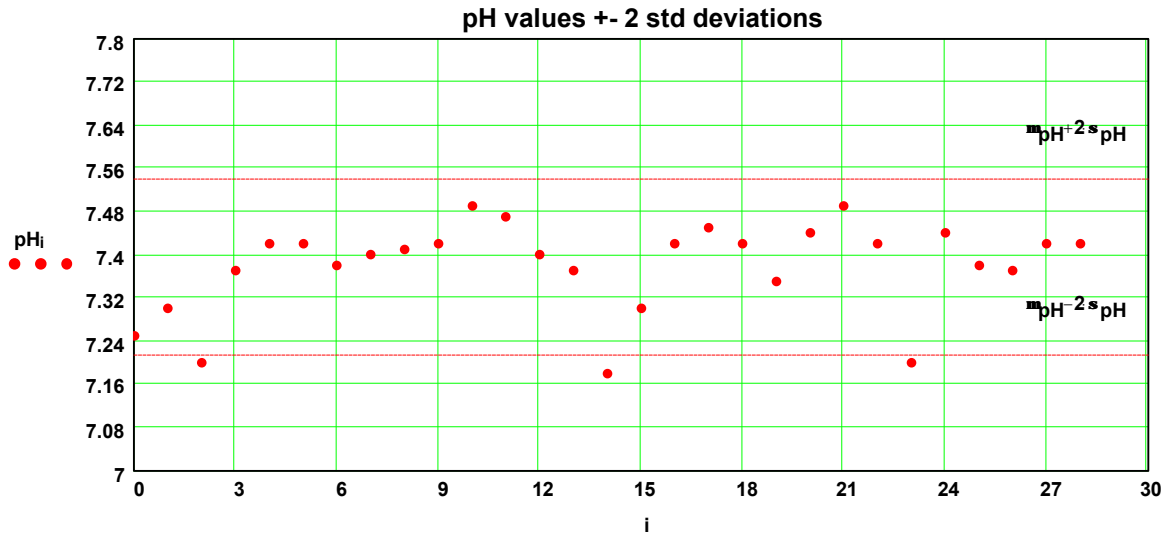
compute the standard deviation of the logs of the DO data :  $log\_s_{DO} := stdev(log\_DO)$

=====

**NOW PLOT THE DATA**

We can plot the data values vs the counter (i) to see if there are trends i.e. did the pH or DO increase or decrease over the time the samples were collected and tested ? By showing the  $\pm 2$  S.D. lines we can check for outliers - assuming normality.

$$\bar{pH} + 2 \cdot s_{pH} = 7.543$$



**INTERVALS, BINS AND HISTOGRAMS**

$n := 100$       number of dividers

$j := 0, 1.. n$       number assigned to each divider

$k := 0, 1.. n - 1$       number of bins

$$\text{interval}_j := \frac{j}{13} + 7$$

each value of interval may be thought of as a "divider" or one side of a bin. A bin is the distance between 2 sequential dividers. In this case the first "bin" would hold values between  $7.0 + 0/15 = 7.0$  and  $7 + 1/15 = 7.0667$ , the second bin would hold values between  $7.0667$  and  $7.2$ .

The "hist" function sorts the data observations into the various bins.

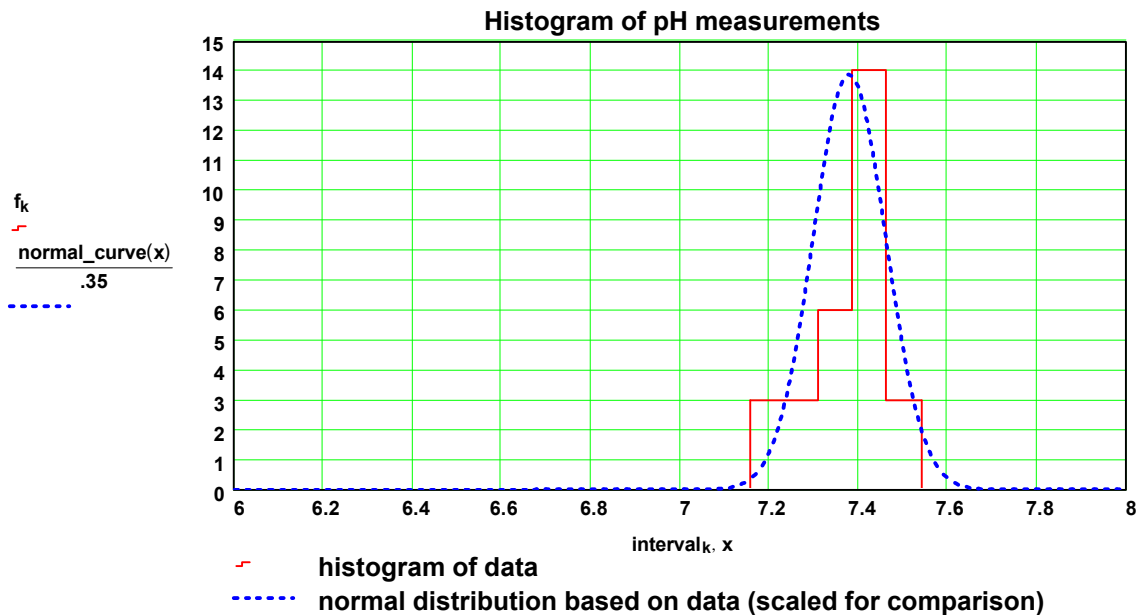
$f := \text{hist}(\text{interval}, \text{pH})$        $f$  is a vector of the number of observations between sequential values of interval

On the same set of axes plot a normal distribution using the "dnorm" function in Mathcad

$x := 6, 6.01 .. 8$

$x$  is a plotting variable spanning the range of the data.  $m_{\text{pH}}$  and  $s_{\text{pH}}$  are the mean and standard deviation, computed from the data

$$\text{normal\_curve}(x) := \text{dnorm}(x, m_{\text{pH}}, s_{\text{pH}})$$



The object of plotting the histogram and the distribution on the same set of axes is to see if they look similar in shape. Note that the y axis for the histogram is the number of observations while that for the normal distribution is simply the value generated from the normal distribution equation. In this case the histogram value was much larger than the equation value so the equation value was scaled to make both plots approximately the same size for easier visual comparison.

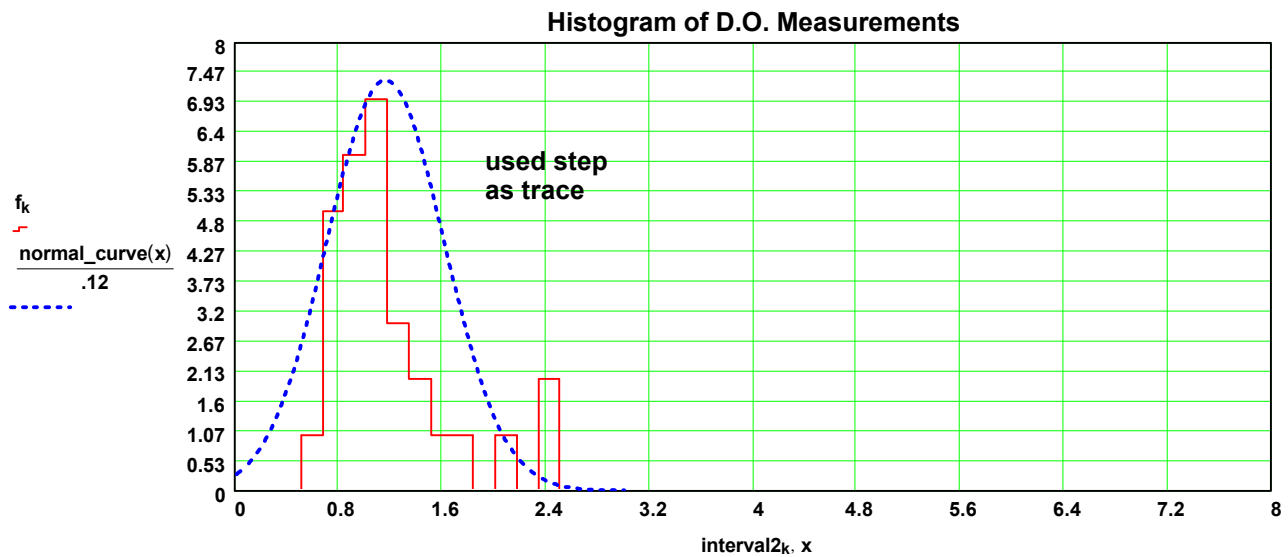
### HISTOGRAM OF DO DATA

$$\text{interval2}_j := \frac{j}{6}$$

f := hist(interval2, DO) see discussion of intervals and bins above

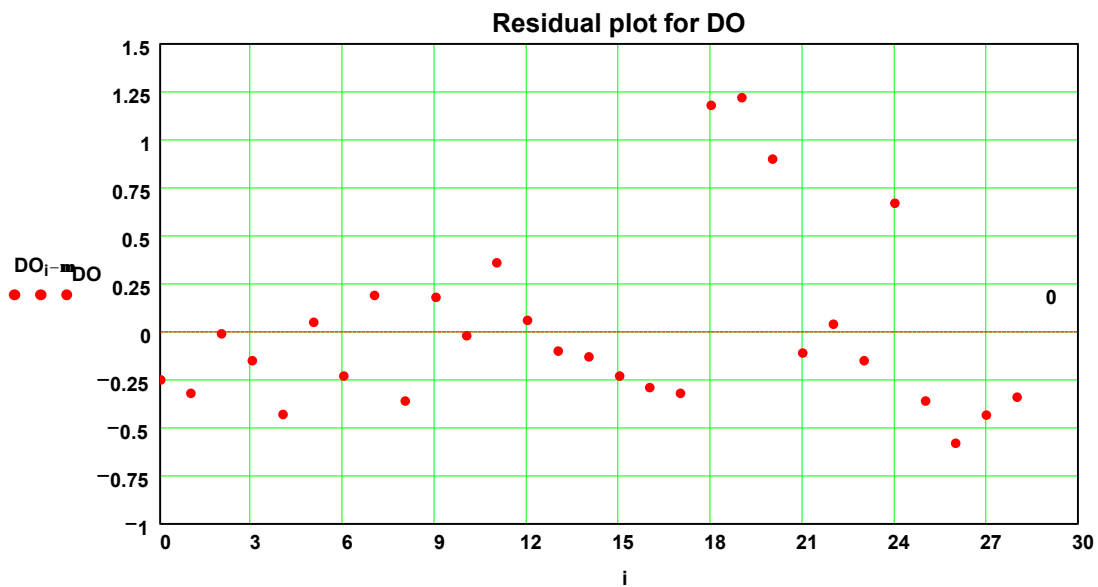
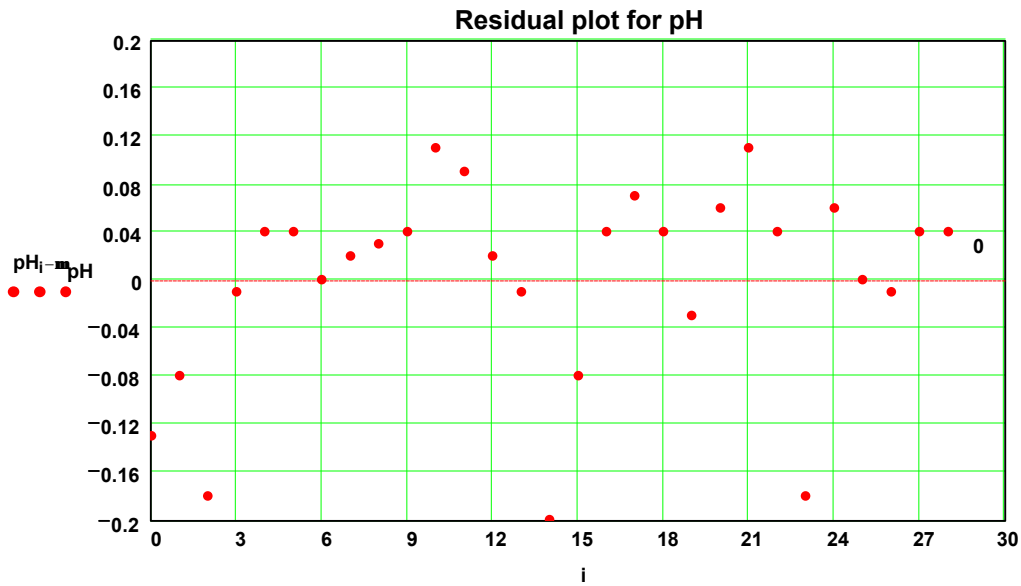
x := .01, .02.. 3

normal\_curve(x) := dnorm(x, mDO, sDO)



**PLOT RESIDUALS OF THE DATA (mean - data value)**

The reason for plotting the residuals stems from the realization that normally distributed data are distributed about the data MEAN. Thus the residuals should be normally distributed about 0. When plotted we should see the residuals lying in a band along both sides of the mean.



**DEVELOPING A PROBABILITY PLOT FROM THE DATA ITSELF**

sort the pH data :  $\text{sort\_pH} := \text{sort}(\text{pH})$

sort the DO data :  $\text{sort\_DO} := \text{sort}(\text{DO})$

Before proceeding we need to account for the fact that both sets of data have multiple pH values of the same value. These are called "ties". We deal with them by assigning each tie value an average rank. Example: suppose we have three measurements of 7.3 which show up at rank 10. If we used all three we might rank them 10, 11 and 12. We don't do this, instead we use the value only once and assign it an average rank of 11.

NOTE : There is no easy way to make Mathcad do this for you - it actually requires manual labor !

$\text{jj} := 1..13$

$\text{pH\_data\_wo\_ties} :=$

|      |      |
|------|------|
| 1    | 7.18 |
| 2.5  | 7.2  |
| 4    | 7.25 |
| 5.5  | 7.3  |
| 6    | 7.35 |
| 8    | 7.37 |
| 9.5  | 7.38 |
| 10.5 | 7.4  |
| 11   | 7.41 |
| 15.5 | 7.42 |
| 20.5 | 7.44 |
| 22   | 7.45 |
| 23   | 7.47 |
| 24.5 | 7.49 |

$$\text{ranks}_{\text{jj}} := \text{pH\_data\_wo\_ties}_{\text{jj}, 0}$$

$$\text{ranked\_pH\_wo\_ties}_{\text{jj}} := \text{pH\_data\_wo\_ties}_{\text{jj}, 1}$$

NOTE - When computing probabilities the number of rows in the ORIGINAL DATA SET are used, NOT the number of rows after ties removed

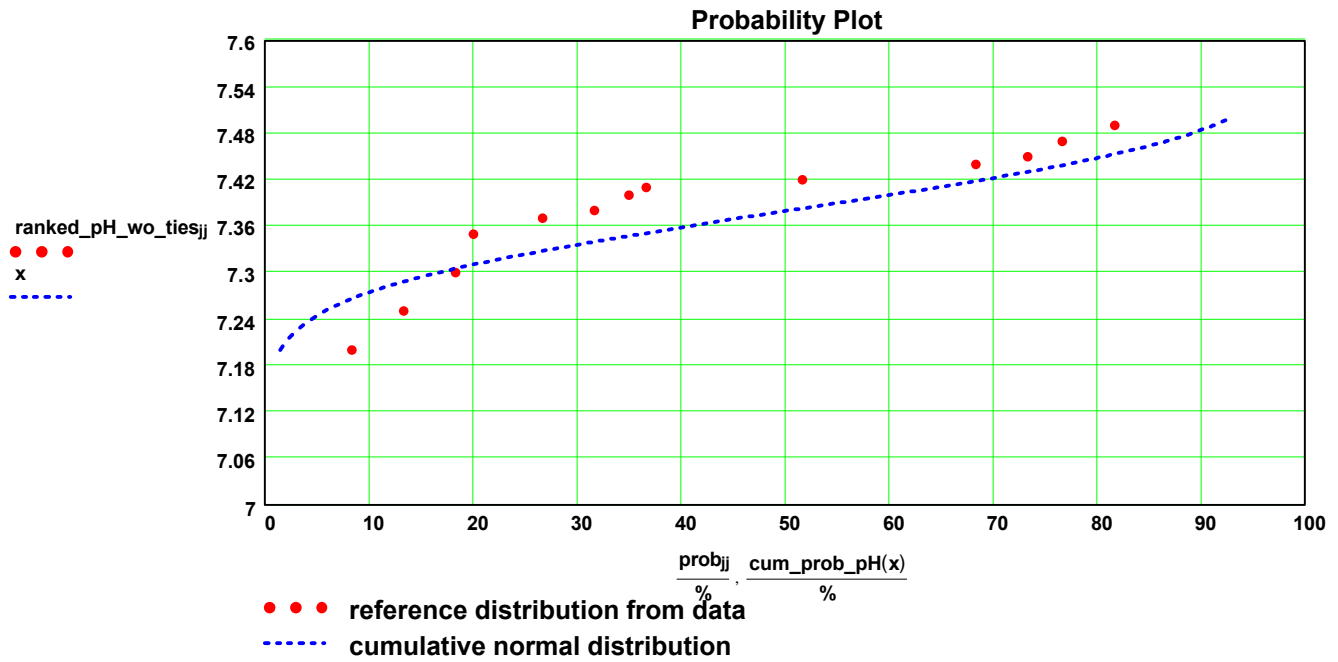
$$\text{prob}_{\text{jj}} := \frac{\text{ranks}_{\text{jj}}}{(\text{rows}(\text{pH}) + 1)}$$

By plotting the ranked data values against their corresponding probabilities we produce what is often called a "reference distribution". It is a perfectly valid cumulative distribution based on the actual data. It makes no assumptions regarding the distribution of the data.

In order to see if the data appear normally distributed we can plot a cumulative normal distribution on the same set of axes using the mean and standard deviation computed from the data.

$$x := 7.2, 7.21 \dots 7.5$$

$$\text{cum\_prob\_pH}(x) := \text{pnorm}(x, \bar{\mu}_{\text{pH}}, s_{\text{pH}})$$



**CONCLUSION - The pH data looks as though it might well be described by a normal distribution**

## CUMULATIVE PROBABILITY PLOT FROM DO DATA

Use the same procedure described above to develop a reference distribution for the DO data. All the caveats mentioned above apply here too.

|                    |      |      |  |
|--------------------|------|------|--|
|                    | .58  | 1    |  |
|                    | .725 | 2    |  |
|                    | .73  | 3    |  |
|                    | .8   | 4.5  |  |
|                    | .82  | 6    |  |
|                    | .84  | 7.5  |  |
|                    | .87  | 9    |  |
|                    | .91  | 10   |  |
|                    | .93  | 11   | ties assigned an average rank as described above |
|                    | 1.01 | 12.5 |  |
|                    | 1.03 | 14   |  |
|                    | 1.05 | 15   |  |
| DO_data_wo_ties := | 1.06 | 16   |  |
|                    | 1.14 | 17   |  |
|                    | 1.15 | 18   |  |
|                    | 1.2  | 19   |  |
|                    | 1.21 | 20   |  |
|                    | 1.22 | 21   |  |
|                    | 1.34 | 22   |  |
|                    | 1.35 | 23   |  |
|                    | 1.52 | 24   |  |
|                    | 1.83 | 25   |  |
|                    | 2.06 | 26   |  |
|                    | 2.34 | 27   |  |
|                    | 2.38 | 28   |  |

$k := 1 \dots \text{rows}(\text{DO\_data\_wo\_ties}) - 1$

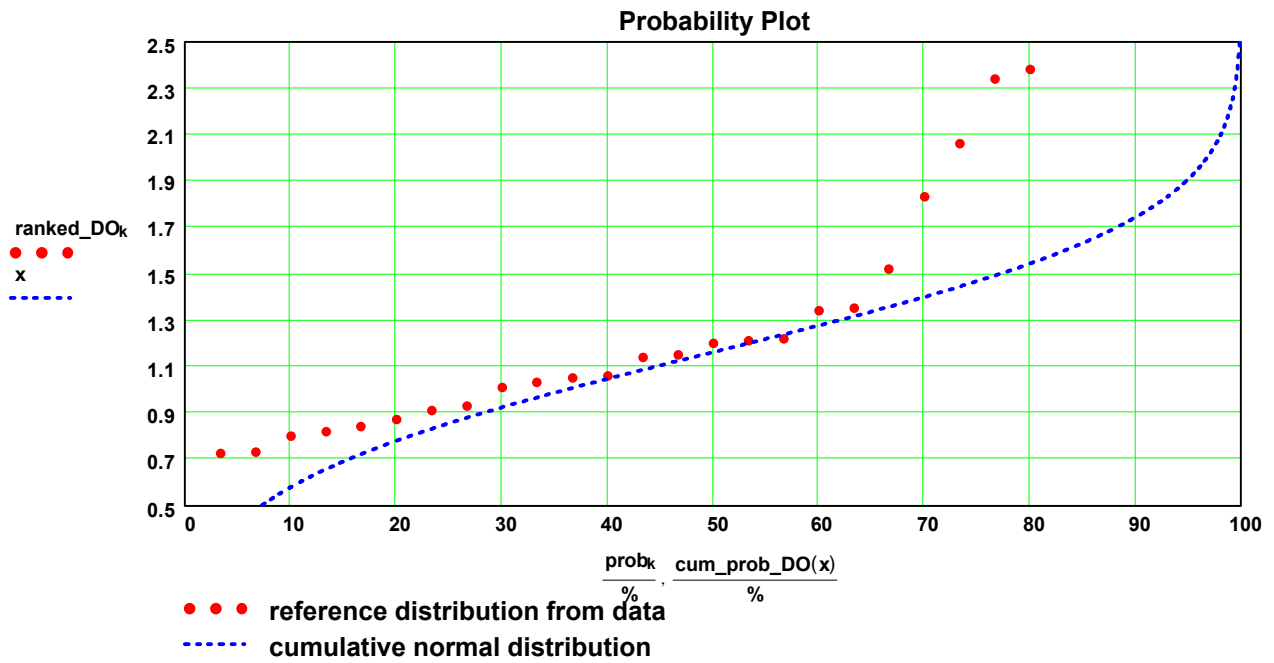
$\text{DO\_wo\_ties}_k := \text{DO\_data\_wo\_ties}_{k,0}$

$\text{ranked\_DO} := \text{sort}(\text{DO\_wo\_ties})$

$$\text{prob}_k := \frac{k}{\text{rows}(\text{pH}) + 1}$$

$x := .5, .501 \dots 2.5$

$\text{cum\_prob\_DO}(x) := \text{pnorm}(x, \mu_{\text{DO}}, \sigma_{\text{DO}})$



**CONCLUSION - The DO data does not seem to fit a normal distribution very well**