Suppose you work for a consulting firm which does environmental work. One of your jobs is to measure the pH of a particular waste stream on a regular basis. In the absence of outside, non-random, effects, the pH of this waste stream has a mean and a standard deviation. These have been determined from data already collected. What you want to know is: "What measured pH values would suggest that non-random processes are occurring?" The following procedure, using the pH values collected in the lab, provide a means for detecting when "something may be happening".

Before proceeding here are some rules:

1. Nothing is ever proved (in a mathematical sense) using statistics, we can only talk about probabilities
2. Statistical results are only as good as the data used to obtain them

Mathcad 2001 has a new algorithm for producing histograms that is much easier to use than previous versions. "intervals" is a vector of the interval values for each bar in a desired histogram. pH_data is the data set. "histogram" produces a matrix containing the midpoint of each interval and the number of observations falling within that interval.
\[
\text{soln} := \text{histogram(interval, pH\_data)}
\]

\[
\begin{array}{ccc}
0 & 7.21 & \\
3 & 7.23 & \\
4 & 7.25 & \\
5 & 7.27 & \\
9 & 7.29 & \\
8 & 7.31 & \\
3 & 7.33 & \\
3 & 7.35 & \\
\end{array}
\]

\[
i := 0..7
\]

Name each column of the matrix for ease of identification: \(\text{value}_i := \text{soln}_{i,0}\) and \(\text{number}_i := \text{soln}_{i,1}\)

\[
\text{total} := \sum_i \text{number}_i \quad \text{compute the total number of observations total} = 35
\]

\[
\mu_{\text{pH}} := \text{mean} (\text{pH\_data}) \quad \mu_{\text{pH}} = 7.286
\]

\[
\text{med}_{\text{pH}} := \text{median} (\text{pH\_data}) \quad \text{med}_{\text{pH}} = 7.29
\]

\[
\sigma_{\text{pH}} := \text{stdev} (\text{pH\_data}) \quad \sigma_{\text{pH}} = 0.033
\]

\[
\text{skew} (\text{pH\_data}) = 0.012
\]

Note that the mean and median of this distribution are nearly identical. This is indicative of a normal distribution.

Skew is a measure of the symmetry of the distribution. A value 0 suggests a symmetric distribution. A value other than 1.0 suggests a "lopsided" distribution. Our value is close to zero.

\[
\frac{\sigma_{\text{pH}}}{\mu_{\text{pH}}} = 4.511 \times 10^{-3}
\]

ratio of std. deviation to mean. It appears that the SD of the data set is small compared to the mean value of the data.
Now, if we can demonstrate that our data follow an existing, known, statistical distribution then we can use the equation of that distribution to describe our data and make conjectures about future data values. Of course the most often used distribution is the NORMAL distribution which, when plotted, has the distinctive "bell curve" shape. One qualitative way of doing this is to plot a histogram and see if it is shaped like a bell curve. In the figure below I have used Mathcad 2001’s histogram algorithm to generate the necessary data to do this. On the same set of axes I have used Mathcad’s normal distribution algorithm to produce a normal distribution curve based on the data taken in lab. The actual equation being plotted is shown below

\[ f(pH\_data) := \frac{1}{\sqrt{2 \pi \sigma_{pH}}} \exp \left( \frac{-1}{2 \sigma_{pH}^2} (pH\_data - \mu_{pH})^2 \right) \]

Notice that the shape of the curve is determined entirely by the characteristics of the data (data values, data mean and data SD). Also realize that \( f(pH\_data) \) is simply a number, not a probability.

Now, we could use more quantitative means (chi square test for example) to determine if the data can be described by a normal distribution but I’m going to say it can as a result of "eyeballing" the plot above and noticing that the two shapes are similar, that is, the histogram looks like a normal distribution. If I am correct this means that the distribution of repeated, independent, measurements of the pH of Folk Lab tap water can be described by a normal distribution. Other characteristics of the data which suggest a normal (or at least symmetric) distribution are:

1. median = mean (approximately)
2. skewness near zero
STOP - what is meant by "independent" measurements? Independence implies that the value obtained in any single measurement is not influenced by (correlated with) other measurements. In the case of the pH data this is why only one person was allowed into the lab at a time to perform the test, each person was given identical instructions, individuals were not allowed to see the results until after the measurement was made, and all were told not to speak to others until after all had run the test. This minimized the chance of one person's measurement being influenced by the actions and observations of those before him/her. Thus, the pH measurements taken were (hopefully) independent.

O.K. - if we believe our data can be described by a normal distribution AND we believe we have collected enough data to give reasonably correct estimates of the sample mean and sample standard deviation of the data set, then how do we obtain probabilities from this curve.

THE PROBABILITY OF A FUTURE MEASUREMENT FALLING BETWEEN ANY 2 VALUES IS GIVEN BY THE AREA BENEATH THE NORMAL CURVE BETWEEN THOSE VALUES

Look again at the normal curve plotted from the data. The area under the curve between pH = 7.23 and where the normal curve coincides with the x axis is, qualitatively speaking, small. This implies that the probability of getting a pH values less than 7.23 solely by chance is small (BUT NOT ZERO). A similar line of reasoning can be employed to conclude that the probability of getting a pH value > 7.37 is small, BUT NOT ZERO.

So... can we quantify what we mean when we say "small"? Yes, we can by using the qnorm algorithm. This algorithm returns the pH value corresponding to a specified probability of occurrence, \( p \), between 0 and 1.

\[
\text{qnorm}(p, \mu_{pH}, \sigma_{pH})
\]
\textbf{The cumulative normal probability distribution}

A cumulative normal probability distribution is obtained by plotting the area between \(-\infty\) and any pH value. This area is then plotted against the pH value. The "pnorm" algorithm in Mathcad does this.

\[
\text{pnorm}(.05, \mu_{\text{pH}}, \sigma_{\text{pH}}) = 7.232 \\
\text{pnorm}(.95, \mu_{\text{pH}}, \sigma_{\text{pH}}) = 7.34 \\
\text{pnorm}(.5, \mu_{\text{pH}}, \sigma_{\text{pH}}) = 7.286
\]

5% of future pH values would be expected to be less than 7.232 by chance alone. That is, 5% of the area under the curve lies to the left of the value 7.232.

5% of future pH values would be expected to be greater than 7.34 by chance alone. That is, 5% of the area under the curve lies to right of 7.34.

50% of future pH values would be expected to be greater than 7.286 by chance alone. This is the mean value and equals \(\mu_{\text{pH}}\).

With a cumulative normal distribution we can read the probability of occurrence of any specified pH value directly from the plot. For example, getting a pH measurement less than 7.23 has a corresponding probability of .057 or about 5.7%. The corresponding probability of getting a pH value less than 7.34 is .95 or 95%.
SO WHAT, IF ANYTHING, DOES ALL THIS MEAN?

Well, this kind of analysis can be used to alert us to non-random changes in a process. These alerts always come with a specified probability and are NEVER a guarantee that changes have or have not occurred.

Go back to our environmental technician reading the pH of a waste stream. Lets assume he has performed the exact analysis described above and obtained the same numbers. On his next trip he measures a pH value of 9.0! Our analysis tells us that the probability of getting such a value solely by chance is VERY SMALL BUT NOT ZERO. So...do we conclude that something dire has occurred? Perhaps, but not before we check out and recalibrate the pH meter (and perhaps the technician) and take another measurement or two....Perhaps the technician had a hangover and simply misread the meter.

On the other hand, suppose the technician measures a value of 9.0 but, knowing what the previous data values looked like, decides that something MUST be wrong and replaces the measured value with the mean value from previous data, or, no value at all. Is this reasonable? NO! - not without actually determining that something IS wrong...otherwise the actual value must be recorded. Never second guess the data....Note that recording the mean, or no value at all, rather than the measured value (without first determining if there is a problem) is an example of a non-independent measurement - the technician’s actions are totally influenced by the other measurements.