The New York State Department of Health’s Bureau of Water Supply Protection developed a tool to assess the statistical confidence of water test samples used for estimating positivity rates in large facilities.

This tool, created as a spreadsheet in Microsoft Excel, allows a user to enter the number of individual water samples from a facility, the number of those samples that returned positive results, the overall number of water outlets in that facility—noted as “Population”, and the desired confidence level for analysis. These inputs are then used to calculate confidence intervals, using the aforementioned values, for 10%, 20%, 30%, 40%, and 50% positivity rates (as well as a custom percentage option) for the facility’s overall number of water outlets, which is then compared with confidence intervals for the collected sample. The tool returns the lower confidence interval limits for each of those population value positivity rates (calculated both with a binomial distribution and an F-distribution for smaller samples), the upper and lower confidence interval limits for the sample positivity rate, and a “Yes” or “No” to notify the user if the sample confidence interval and the population confidence intervals are significantly different (i.e. the sample confidence interval and population confidence interval do not overlap). These returned values can help an investigator quickly determine if the results of a sample indicate that the underlying population’s positivity rate is below a threshold proportion; the example that this tool was originally created for was identifying if a healthcare facility had a legionella outbreak using a positivity rate for unusually high presence.

The tool also includes a section to determine the sample size for a desired positivity rate confidence interval, and a section for displaying the p-value used for determining if the sample results are significantly different from a theoretical population result. The positivity rate confidence interval section uses the confidence value from the first part of the calculator and a new input for the desired positivity rate range to display the required sample size for that level of precision. The section for displaying p-values prompts users for a sample size, the number of positive samples, a population size, and the desired number of positives in that population. The tool returns a p-value for determining if the two rates are different, and then interprets that p-value as the “probability that they are different”. This p-value is calculated using a two-proportion z-statistic, defined as:

$$z=\frac{\left(\hat{p}\_{s}-\hat{p}\_{p}\right)}{\hat{p}\_{all}\*\left(1-\hat{p}\_{all}\right)\*\left(\frac{1}{n\_{s}}+\frac{1}{n\_{p}}\right)}$$

where $\hat{p}\_{s}$ is the percent of samples with a positive result, $\hat{p}\_{p}$ is the estimated positivity rate for the population, $\hat{p}\_{all}$ is the pooled positivity rate for the sample and the population, $n\_{s}$ is the number of samples, and $n\_{p}$ is the estimated population size.