Evaluation of threshold limit methods for sensory data

D. Gallagher and J. Cuppett
Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061-0246, USA (E-mail: dang@vt.edu)

Abstract Current approaches to sensory thresholds, such as geometric means and logistic regression, ignore any formal consideration of uncertainty and variability. Various alternative methods based on approximate confidence and prediction intervals about the logistic regression were examined. All methods tended to provide the same ranking among different analyte/media combinations evaluated. Formal statistical conclusions could be made for thresholds based on interval analyses, but not for geometric mean or logistic regression. Methods based on prediction intervals consistently estimated the highest thresholds. Interval-based methods varied with the level of confidence required, as well as the number of panelists and concentrations tested. The geometric mean method yielded the most consistent estimates across a range of panel sizes.

Keywords Bootstrap analysis; detection limits; geometric mean; logistic regression; uncertainty analysis

Introduction
Taste-and-odour threshold concentrations are important for the water and food industries because they represent the lowest concentration that can be detected reliably by consumers. As sensory thresholds begin to be used in the regulatory and risk assessment fields, it becomes important to understand the statistical basis for determining thresholds. Some attempts have been made to use thresholds in setting secondary maximum contaminant level for methyl tertiary butyl ether (MTBE) (Stocking et al., 2001). Consumers may be considered as sentinels for drinking water contamination with customer complaints serving as an indicator. In both cases, a fundamental understanding of how thresholds are calculated and their statistical interpretation are important in evaluating the sensory data.

Sensory parameters are often determined by alternative forced choice (AFC) tests such as triangle test or one-of-five test. The result is binary – the panelist either correctly selects the different choices or does not. If a range of concentrations is evaluated, the resulting data can be used to estimate the group threshold concentration (van Aardt et al., 2001). The objective of this work was to compare different methods for group sensory threshold determinations, including several that meet a specified confidence level, and to better understand the assumptions and differences among them.

Current methods for sensory threshold determination such as described by ASTM International generally ignore uncertainty, unlike environmental detection limits. For example, compare the ASTM definition of detection threshold to the related concept of the U.S. Environmental Protection Agency (USEPA) detection limit:

ASTM E-1432-91 (1991): A detection threshold is the intensity of the stimulus that has a probability of 0.5 of being detected under the conditions of the test.

USEPA Clean Water Act 136.2(f): A detection limit is the minimum concentration of an analyte that can be measured and reported with a 99% confidence that the analyte concentration is greater than zero.

The detection limit in the latter definition has an explicit consideration of the degree of confidence required for reporting. The threshold limit does not. Meeting a specified confidence levels sets controls for a Type I error (α) or false positive rate (Currie, 1997).
Methods

There are a variety of methods that can be used to determine a threshold. Two commonly used methods are geometric means and logistic regression. Both are defined in ASTM methods.

Geometric mean

The geometric mean threshold ASTM E-679-97 (ASTM, 1997; Meilgaard et al., 1999) first calculates each individual panelist’s threshold concentration as the geometric mean of the highest concentration incorrectly assessed and the lowest concentration correctly assessed (where all subsequent higher concentrations were correct). The group threshold is the geometric mean of the individual panelists’ thresholds. ASTM (1997) indicates that the geometric mean method is a rapid analysis meant to approximate the more formal approach described in ASTM E-1432-91 (1991).

The geometric mean of \( n \) values is calculated as either the \( n \)th-root of the product of the values or equivalently as the antilog of the mean log value:

\[
\text{geometric mean} = \sqrt[n]{\prod_{i=1}^{n} x_i} = e^{\frac{1}{n} \sum_{i=1}^{n} \ln(x_i)} = 10^{\frac{1}{n} \sum_{i=1}^{n} \log(x_i)}
\]

The log of the geometric mean is thus the mean on the log scale. Because a geometric mean implies a log transformation during calculation, it is sometimes considered appropriate for right skewed or log-normal data. While relatively straightforward to calculate, no formal uncertainty considerations are included in the geometric mean threshold. Parkhurst (1998) showed that geometric means perform worse than arithmetic means for mass balance calculations. Crump (1998) found that arithmetic means performed better than geometric means in evaluating exposure assessment during risk analysis.

Logistic regression

In logistic regression, the dependent variable is the proportion of correct responses at the given concentration. This proportion may be from multiple replicate testing for an individual, as defined in ASTM E-1432-91. The resulting threshold is the individual and the group threshold is calculated by geometric mean or other techniques. Alternatively, the proportion may represent the results of several individuals, each tested without replication over a range of concentrations. The resulting threshold is a group threshold (van Aardt et al., 2001).

To calculate the threshold, the proportion is transformed by the logit function and regressed against concentration (Collett, 2003). If \( p \) is the probability of detection at a given concentration, the logistic linear model is:

\[
\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x
\]

Maximum likelihood methods are used to estimate the regression parameters \( \beta_0 \) and \( \beta_1 \).

In sensory analysis, thresholds are usually defined for a specified probability of detection – the dependent variable in the logistic regression. In environmental analysis, detection limits are defined at a specified concentration, almost always zero. The concentration is the independent variable in the corresponding calibration curve. Regression analysis assumes that the uncertainty and variability are part of the measurement for the dependent variable, and that the independent variable is known with certainty. Different approaches are needed depending on whether a threshold is defined for a fixed probability of detection or a fixed concentration.
**Probability-based thresholds.** For probability-based threshold definitions such as ASTM (1991), the group threshold concentration is taken as some arbitrary value for the probability of detection, usually 50%. Since the definition is based on a probability of detection, there is no uncertainty in the probability value.

The group logistic analysis is directly analogous to a dose response curve in toxicology, with dose in lieu of analyte concentration, and mortality in lieu of sensory detection. Threshold determination is then equivalent to calculating an LD$_{50}$ (the dose at which 50% of the exposed population is expected to die). This notation is useful because it emphasises that the criterion used for setting the threshold (50% detection) is arbitrary, and other values can be selected. While 50% probability of detection can be a conservative approach for an individual, when applied to a group it may seem unacceptably high. Sensory analysts could calculate thresholds at other probabilities of detection, e.g. Th$_{90}$ for 90% or Th$_{10}$ for 10% probability of detection.

This 50% probability criterion is usually corrected for the probability of guessing by chance as calculated by Abbott’s formula (ASTM 1991; Lawless and Heymann, 1998). This corrects for the type of AFC test used during the evaluation, e.g. there is a 33% chance that a nondiscriminating panelist can guess the correct response in a triangle test, and a 20% chance in a one-of-five test. Abbott’s formula is:

$$p_{\text{criterion}} = \frac{p_{\text{adjusted}} - p_{\text{guess}}}{1 - p_{\text{guess}}}$$

For example, using 50% as the desired criterion and a one-of-five test, where the probability of guessing by chance is 20%, the probability used to define the threshold is 60%.

While Abbott’s formula corrects for the AFC design, the test design itself does not account for the entire variability about the regression. A logistic regression does not calculate an $R^2$ value, but there are several analogues (Collett, 2003) and other goodness of fit indicators. One such indicator usually included in logistic regression outputs is Akaike Information Criterion (AIC). For models with the same scale, number of parameters and data points, AIC increases as the goodness of fit deteriorates.

An alternative to Abbott’s correction is to evaluate the confidence and prediction intervals about the fitted logistic regression. These intervals incorporate the complete variability of the experiment. They will be a function of the number of data points and panelists, the AFC test type and the experimental setting. A 1-$\alpha$ confidence interval represents the region where the true regression curve is believed to fall with 1-$\alpha$ confidence. A 1-$\alpha$ prediction interval represents the region where a new data point is expected to fall with 1-$\alpha$ confidence.

Approximate confidence and prediction intervals can be constructed about the logistic regressions (Piegorsch and Bailer, 1997). Venables and Ripley (2002) show how confidence intervals can be constructed about dose response curves.

Thresholds could thus be defined as the concentration at the lower confidence or prediction interval for a 50% probability of detection. The intervals are calculated for the linear model on the logistic scale, but can be transformed back to the more easily interpreted probability scale for display. Prediction intervals are always wider than confidence intervals, and will always result in higher threshold concentrations.

**Concentration-based thresholds.** Using a probability-based definition for a threshold with a probability of detection of 50% implies that there is no uncertainty in the 50% value. It is, after all, part of the definition. Alternatively, a fixed concentration can be used to define the threshold. This is readily apparent in the USEPA definition of a method detection limit (MDL), which is set at detecting a concentration greater than zero with 99% confidence. If a concentration is used as the basis for a threshold definition, the
uncertainty in the probability of detection must also be considered, as well as the
uncertainty about the regression fit.

Concentration-based thresholds are rarely used in sensory analysis. Most sensory logistic regressions are performed on a log scale for concentration – the regression fits are usually better on the log scale than on a linear scale. However, a linear scale is necessary to evaluate a concentration of zero. There is no mathematical reason linear scales cannot be used and examples of such in the sensory field are available (van Aardt et al., 2001).

Hubaux and Vos (1970) developed a detection limit approach commonly used in the environmental field. The technique has been shown to be both reliable and robust for detection limits (Gibbons, 1994). The method can be extended to use the same logistic regression previously developed with prediction intervals about the regression. The probability of detection for the upper prediction interval at the intercept (concentration = 0) is found first. This probability is extended until it intercepts the lower prediction interval, and the corresponding concentration is the threshold. Each of these methods is illustrated below.

Data sets
Data were collected from the literature for seven compounds/media mixtures. All data were based on AFC tests: copper in distilled and dechlorinated tap water (1-of-5-AFC, 36 panelists) and acetaldehyde in spring water, skim milk, low fat milk, whole milk and chocolate milk (3-AFC, 25 panelists). All statistical calculations were performed in R, version 2.1 (R Development Core Team, 2005).

Uncertainty analysis
Confidence intervals for any of the threshold methods can be calculated by bootstrap resampling (Martinez and Martinez, 2002). Bootstrap resampling uses Monte Carlo simulations that treat the original data set as an estimate of the population. Samples are taken with replacement, and for each sample set the threshold was calculated. Bootstrapping can also be used to evaluate the impact of other parameters, such as the size of the panel. In this work, panelist detection results were sampled with replication 2,000 times. The resulting 2,000 threshold estimates were used to create a cumulative distribution plot, from which appropriate quantiles were determined.

Results and discussion
Table 1 summarizes the calculated thresholds for the different techniques. While the individual thresholds vary widely, the thresholds were generally well correlated. Using the geometric mean as a basis for comparison, the $R^2$ Spearman rank correlation coefficients were 0.93, 0.80, 0.73 and 0.86 for the logistic regression, the confidence interval-based the prediction interval-based, and Hubaux and Vos method respectively. Threshold comparisons among different analyte/media combinations tended to produce the same ranking no matter what method was used.

In general, regressions on the log concentration scale fit better than those on the linear scale. However, the individual geometric means for only one of the seven analyte/media combinations were log normally distributed based on a Shapiro-Wilks test with $\alpha = 0.05$.

Threshold methods based on prediction intervals (prediction intervals and Hubaux and Vos) were much larger than thresholds from the other techniques. A prediction interval estimates where a single new data point would fall, whereas the other techniques estimate some version of the mean for the group. The added uncertainty associated with a single point, particularly during a logistic regression, resulted in a much larger threshold. Figure 1 illustrates each of the four logistic regression-based methods for copper in distilled water.
The conclusions made about each threshold method are different. Using distilled water with 36 panelists and a probability of detection of 50% as an example, the prediction interval-based method indicated there was 95% confidence that a new measurement would be less than 3.13 ppm. The confidence interval-based method indicated there was a 95% confidence that the mean of repeated measures under the same requirements would be less than 0.69 ppm. Based on the Hubaux and Vos results, there was 95% confidence that concentrations greater than 4.4 ppm were above the probability that would be reported if the copper was not present. Because uncertainty is not considered in the geometric mean or logistic regression methods, similar statistical statements cannot be made.

Table 1 Summary of threshold evaluations

<table>
<thead>
<tr>
<th>Analyte</th>
<th>Media</th>
<th>Thresholds, ppm</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Geo-metric Mean</td>
<td>Logistic</td>
</tr>
<tr>
<td>Copper</td>
<td>Distilled water</td>
<td>0.48</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Dechlorinated</td>
<td>0.39</td>
<td>0.74</td>
</tr>
<tr>
<td>Acet-aldehyde</td>
<td>Chocolate milk</td>
<td>17.2</td>
<td>12.92</td>
</tr>
<tr>
<td></td>
<td>Whole milk</td>
<td>4.23</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>Lowfat milk</td>
<td>4.09</td>
<td>4.28</td>
</tr>
<tr>
<td></td>
<td>Skim milk</td>
<td>3.01</td>
<td>2.37</td>
</tr>
<tr>
<td></td>
<td>Spring water</td>
<td>0.15</td>
<td>0.31</td>
</tr>
</tbody>
</table>

1Using a probability of 50% detection and Abbott’s correction
2Based on $\alpha = 0.05$
3Hubaux and Vos logistic regression performed on linear concentration scale. All other logistic regressions performed on log concentration scale

The conclusions made about each threshold method are different. Using distilled water with 36 panelists and a probability of detection of 50% as an example, the prediction interval-based method indicated there was 95% confidence that a new measurement would be less than 3.13 ppm. The confidence interval-based method indicated there was a 95% confidence that the mean of repeated measures under the same requirements would be less than 0.69 ppm. Based on the Hubaux and Vos results, there was 95% confidence that concentrations greater than 4.4 ppm were above the probability that would be reported if the copper was not present. Because uncertainty is not considered in the geometric mean or logistic regression methods, similar statistical statements cannot be made.
The AIC goodness of fit parameters can be compared where the same analyte, number of panelists and concentrations tested were used. The copper results indicate that a simple adjustment depending on the AFC test type is not adequate to capture uncertainty. The logistic regression fit for copper in distilled water was much better (lower AIC) than the fit for copper in tap water, even though the same AFC test was used. The threshold methods that used confidence or prediction intervals incorporated this variability; Abbott’s correction did not.

The threshold methods that use confidence or prediction intervals will yield different results depending on the level of confidence \((1-\alpha)\) required. As the required confidence increases, so will the threshold because the intervals will widen. Figure 2 illustrates the confidence interval threshold for copper in distilled water as the required confidence level changes. The vertical line is for a 95% confidence level, the same used in Table 1.

Bootstrapping was used to evaluate the uncertainty in the threshold estimates as well as the sensitivity to the number of panelists. Figure 3a shows the 95% confidence intervals about the thresholds estimated for copper in distilled water using confidence interval-based threshold. Higher thresholds yielded greater uncertainty. Figure 3b indicates that the geometric mean stabilised to a constant value even with very small panels. The logistic regression threshold decrease substantially as the panel increased in size from 20 to 36 panelists. Threshold methods based on levels of confidence are sometimes criticized because they vary as the number of panelists and tested concentrations change (Currie,
but Figure 3b indicates that the logistic regression threshold also varied with sample size, even though no formal level of confidence was involved. Confidence and prediction based methods were even more sensitive to changes in the number of panelists.

Conclusions
Current threshold determination methods ignore uncertainty and are limited in the statistical conclusions that can be drawn. The interpretability of a threshold depends on how it was calculated. Current threshold methods were designed to evaluate consumer accept ance. As thresholds become more important in risk assessment and regulations, a more fundamental statistical understanding of their interpretation is needed. This work discusses some of the issues to be considered.

Acknowledgements
Drs. Andrea Dietrich and Susan Duncan provided sensory data. Portions of this material are based upon work supported by the National Science Foundation under Grant No. 0329474. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References