Why Dichotomization is Needed with Skewed and Non-Linear Data Distributions to Reduce Uncertainty and False Positives and Eliminate False Negatives in Licensing and Regulatory Compliance Decision Making

Richard Fiene PhD and Sonya Stevens, EdD

Edna Bennett Pierce Prevention Research Center at Penn State University

Research Institute for Key Indicators Data Laboratory

July 2025

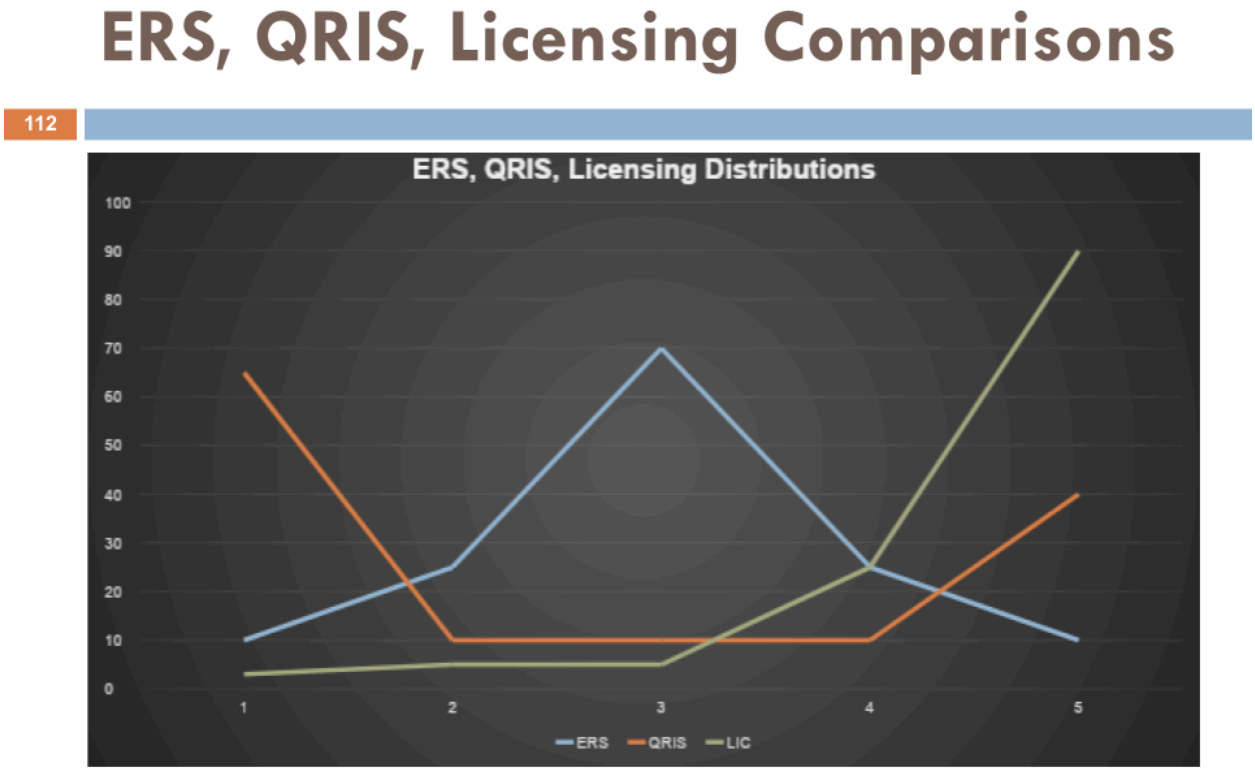
When it comes to making licensing decisions, we want to always be certain that the decision is the correct one. Uncertainty (incorrect citations or issuing a license to a high-risk provider) should be reduced while simultaneously increasing certainty (correct citations or issuing a licensing to a low-risk provider). This brief discusses the dichotomy between certainty and uncertainty as it relates to making licensing decisions about whether a facility is granted a license or not. The goal is to reduce false positives (citing a facility for being out of compliance when they are in compliance) and false negatives (not citing a facility for being out of compliance when they really are out of compliance) by utilizing a statistical data dichotomization technique.

On the surface, making licensing decisions should be very simple and straightforward. Either the facility meets or does not meet the rules/regulations of the oversight agency. However, when one dives deeper into the decision-making process, it often gets complicated.

To understand licensing data, you must first understand what normal distribution and linear relationships look like. When using non-licensing data (such as quality data) to assess decision making distributions we generally see a linear relationship within a normal curve when compared to another group of normally distributed data. The majority of data are somewhere in the middle range with much fewer data points at either end of the data distribution when it comes to a normal curve; and with the linear relationship as one data variable increases the other data variable increases in a corresponding fashion. The following graphic, Figure 1: (ERS: Environmental Rating Scales, QRIS: Quality Rating and Improvement Systems, and Licensing Comparisons) displays this relationship and compares a normally distributed bell shaped curve (ERS-blue) with a bi-modal data distribution (QRIS-orange) and a skewed data

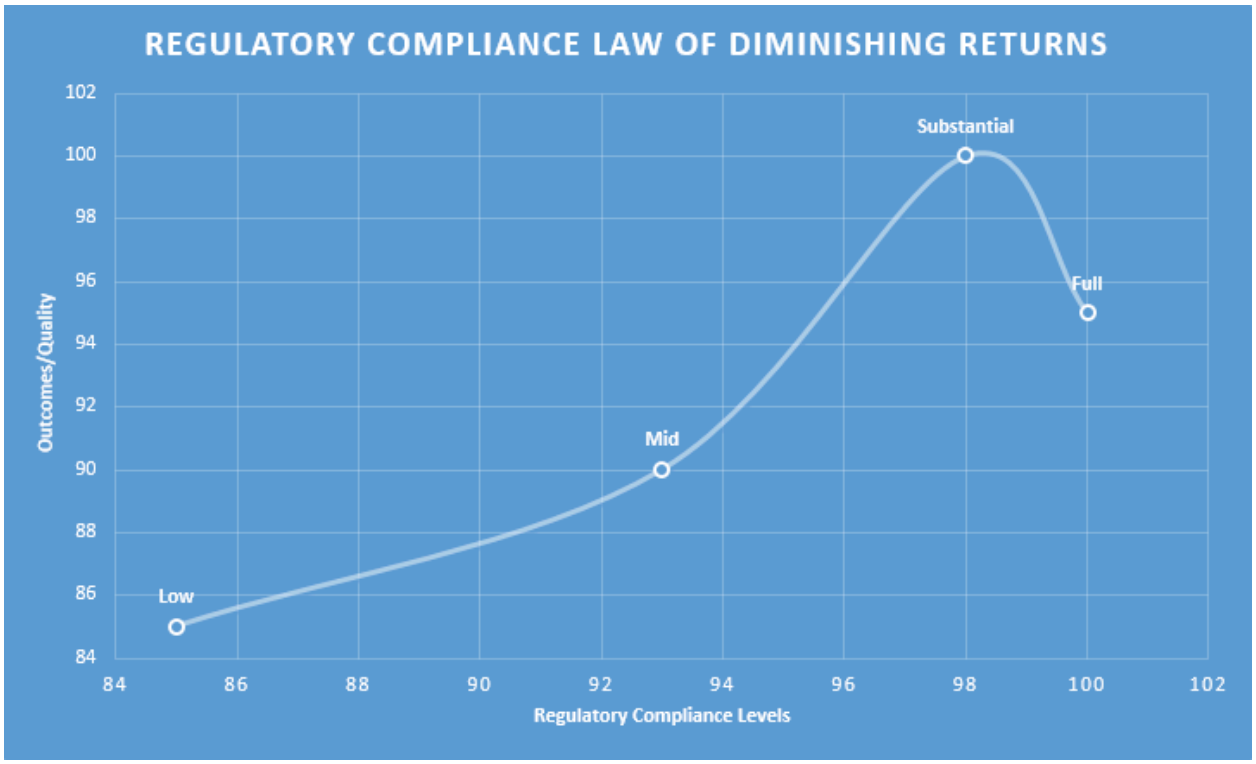
distribution (Licensing-yellow). The ERS data are examples of process quality while Licensing and QRIS data are examples of structural quality.

Figure 1: Comparing Skewed (Licensing), Bi-Modal (QRIS), and Normal Distributions (ERS)



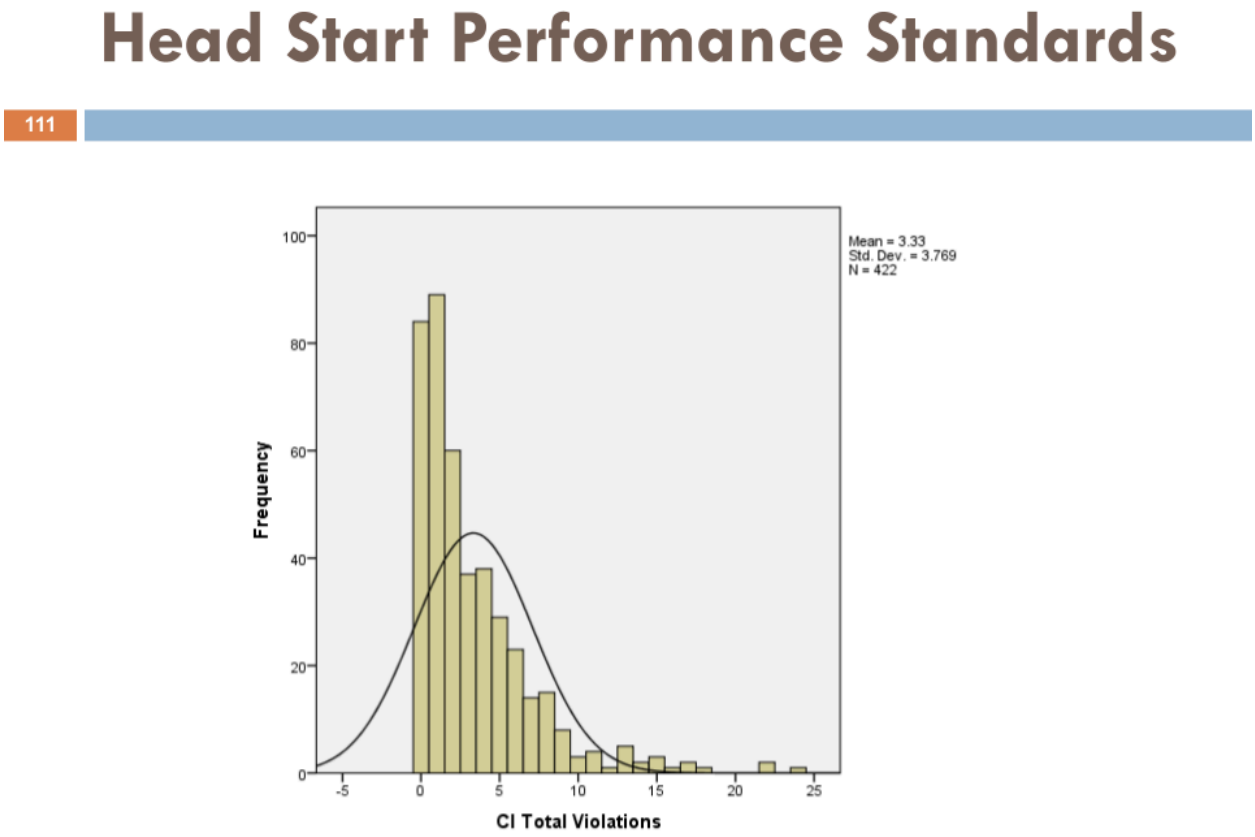
Interestingly, licensing or regulatory compliance data do not follow the same distribution characteristics as with normally distributed bell curves. The data distributions are not normally distributed, and they are generally nonlinear when comparing regulatory compliance data with other data distributions which are normally distributed (such as quality data). In fact, it has been observed that licensing/regulatory compliance data demonstrate a “ceiling effect” or “diminishing returns effect” when compared to normally distributed quality data. Figure 2 below depicts this relationship between process quality (ERS Observations) and structural quality (regulatory compliance).

Figure 2: The Non-Linear Relationship Between Process Quality and Structural Quality



The next graphic (Figure 3: Head Start Performance Standards displays this skewness and ceiling effect (typical of structural quality data distributions) which is not the case in normally distributed data distributions (typical of process quality data distributions). It is clearly evident from the histogram that the majority of programs are at the one end of the data distribution which demonstrates high regulatory compliance with the Head Start Performance Standards (HSPS). The long tail shows the smaller number of programs at a lower level of compliance with HSPS standards. Head Start data being structural quality data show a similar data distribution as one will find with licensing data but the level of skewness with HSPS is somewhat less evident.

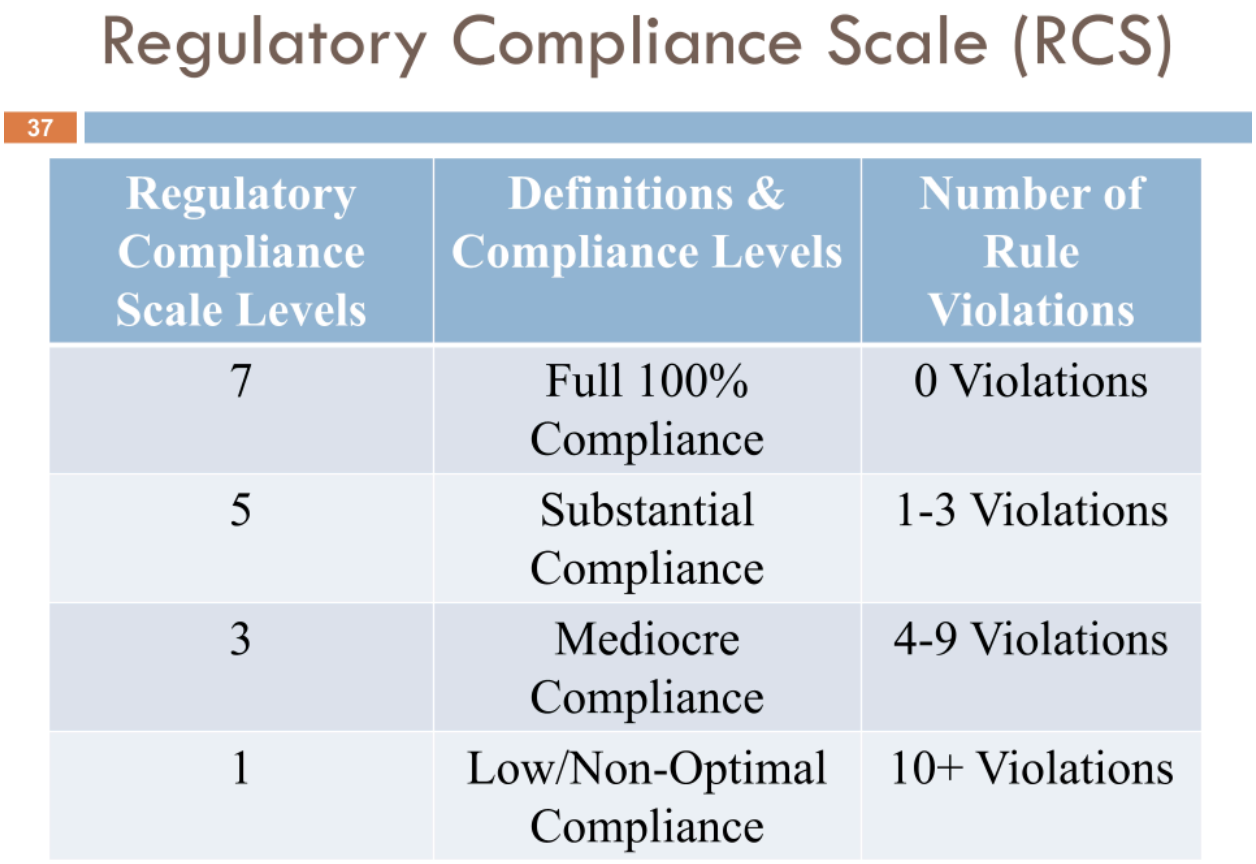
Figure 3: Skewed Data Distribution Example Demonstrated by Compliance with the Head Start Performance Standards



Because licensing data are nonlinear in nature and demonstrate this ceiling effect, it increases the chances for false negatives and false positives in the licensing decision making process. With a ceiling effect it is very difficult to distinguish between the high performing facilities (those that are in high regulatory compliance and high quality) and the mediocre performing facilities (those that are having difficulty with regulatory compliance and their quality levels are marginal).

To reduce false positives and hopefully eliminate false negatives, it is recommended to dichotomize, or group, the data distribution to high and low regulatory compliance groups as depicted in the following Regulatory Compliance Scale (RCS)(Figure 4) which uses a rating scale from 1 through 7. When this is done, the scaling technique fits more closely with the normal program quality scaling of the ERS: Environmental Rating Scales.

Figure 4: The Regulatory Compliance Scale: Moving Frequency Data to Ordinal Groupings

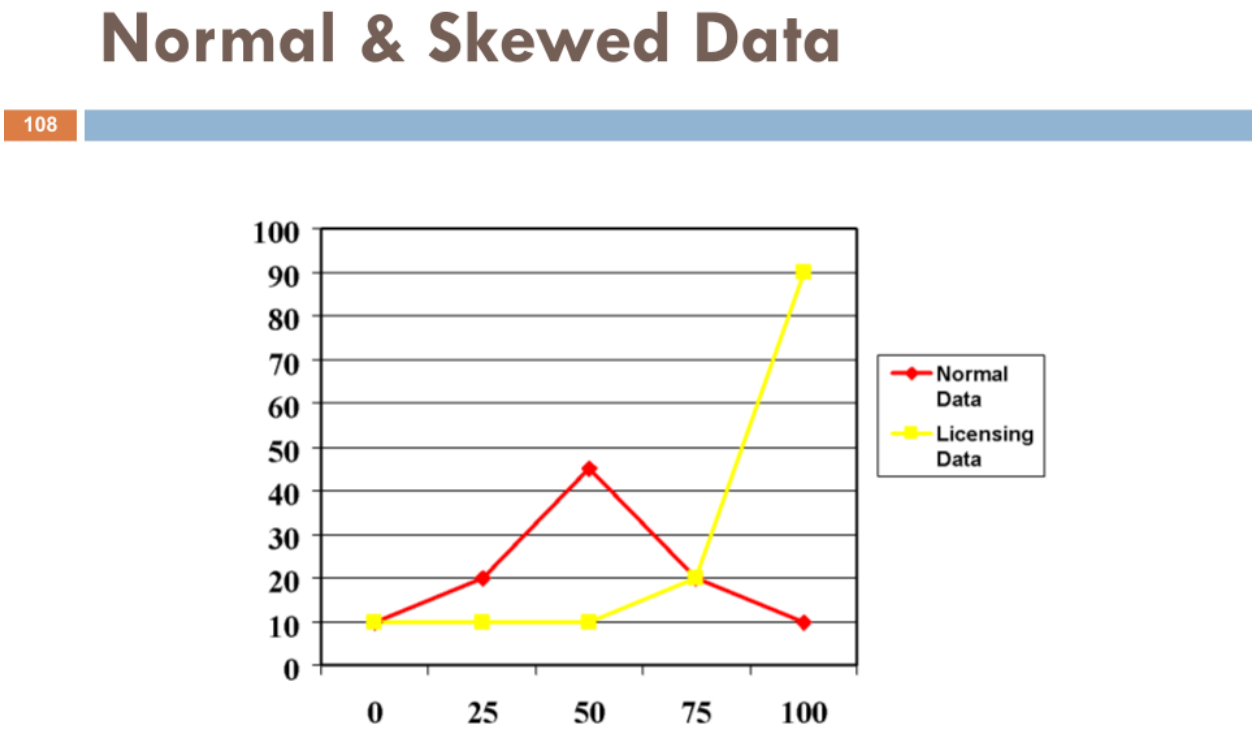


Moving violation count data to regulatory compliance buckets enhances the certainty of the licensing decision making which will reduce the uncertainty associated with false positives. Likewise, if the high group contains only full compliance with all rules it demonstrates limited uncertainty by eliminating, or reducing, false negatives. This is also true when substantial compliance is used for licensing decision making in place of full compliance and the following buckets can be used: high, mediocre, low.

Statistically, dichotomization is generally not recommended nor desirable with normally distributed data, but when data distributions are severely skewed it provides a safeguard from false negatives and false positives from occurring (See the below graphic (Figure 5) which depicts normally and skewed data distributions). In practical terms, this means selecting the top and bottom 10% of the data distribution, the very best facilities (in the above graphic this would be all 7 rated programs) and the very

worst facilities (in the above graphic this would be all 1 rated programs) and then eliminating or not using the middle 80% of facilities in analyses for licensing decision making (in the above graphic this would be all 3 and 5 rated programs).

Figure 5: Examples of Normal and Skewed Data Distributions to be Used for Dichotomization



By using this approach, agencies can improve the certainty of decision making dramatically when it comes to determining which facilities receive a license and which do not. This dichotomization approach should only be used in situations in which the data distribution is nonlinear (ceiling effect) and severely skewed (which is the case with all licensing or regulatory compliance data distributions) but is not recommended to be used with normal data distributions which are linear in nature when compared to other linear data distributions and no ceiling effect is present because of skewed data (which is not the case in process quality data as depicted in the graphic below (Figure 6:

ECERS: Early Childhood Environmental Rating Scale Total Scores which essentially depicts a normally distributed bell curve).

For additional detailed information about licensing and regulatory compliance data distributions, dichotomization, reducing false positives and reducing and/or eliminating false negatives, and regulatory compliance buckets/scaling, please go to the following research website that has several papers and articles describing these associated methodologies, http://rikinstitute.com or contact Dr Fiene directly at rfiene@rikinstitute.com.

Figure 6: ECERS Total Scores Data Distribution as an Example of Normally Distributed Data