Background

The 2020 SOTI, like others before it, uses several measures of system reliability to conclude which parts of the Western Interconnection's bulk electric system are becoming more or less reliable over time. Unlike previous years, the 2020 SOTI relies heavily on objective statistical tests for those conclusions. While statistical tools can be powerful, they come with limitations. This appendix explains the objectives and assumptions of the 2020 SOTI analysis so that users may interpret the results appropriately.

Statistical analysis is a tool for identifying patterns in data—differentiating actual trends from variations that are not trends. Frequentist statistics does this by asking how likely is it that the observed pattern would have occurred if there was no trend. If the observed pattern is very unlikely to have occurred without an actual trend, we conclude that there must have been a trend. In other words, the analysis is predicated on assumptions about what *would have been observed* if there was no trend. These assumptions are both speculative and critical. These challenges become more pronounced when the data consists of very few points, (e.g., one point a year for five years). The assumptions behind a statistical analysis are described by a regression model.

Regression Models

This appendix describes two regression models that are used in the 2020 SOTI, though many are possible.

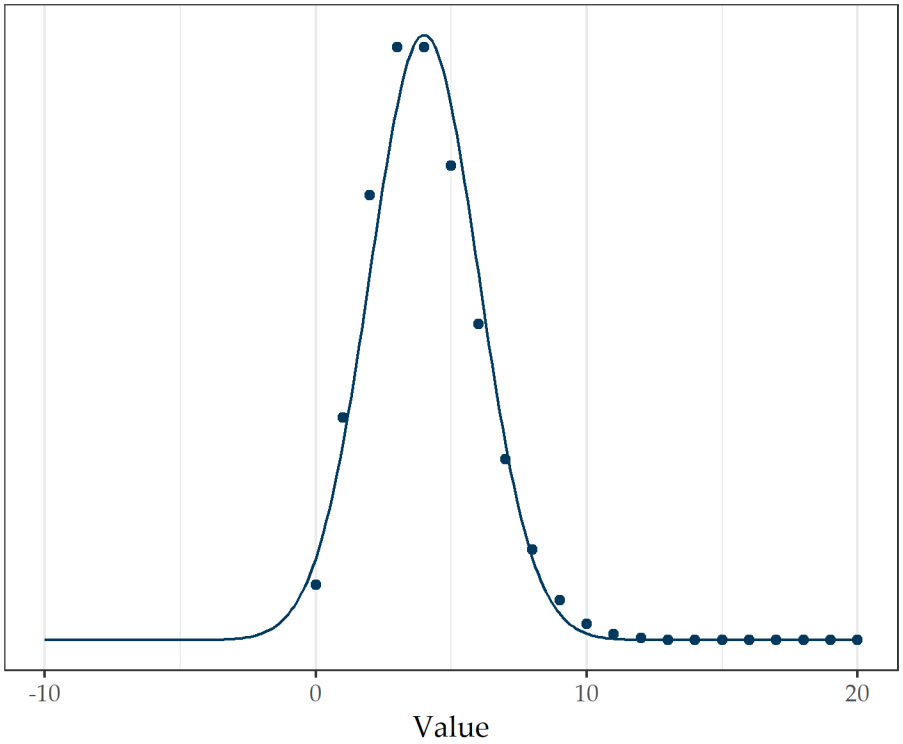
The most common regression approach is Ordinary Least Squares (OLS). This technique is easy to use and consistent with the idea that the "noise" of a measurement, (i.e., the parts of measurements that are not the trend), follow a Gaussian distribution. (Minimizing the sum of the squared errors gives the maximum likelihood estimator for the mean of a Gaussian distribution.) Because many physical phenomena can be shown to eventually converge to Gaussian distributions, this approach is so common it often becomes the default approach.

Despite the general popularity of OLS regression, many phenomena, (in general and in the SOTI), either do not follow Gaussian distributions, or cannot be shown to follow such distributions based on the small number of data points. So, processes that result in a small, discrete number of events may best be described by another model.

A Poisson model is one alternative to a Gaussian model. A Poisson model assumes that events have some unknown likelihood, per unit and unit of time, and asks whether that likelihood has changed. Finding the maximum likelihood estimator for the mean of a Poisson process is more complicated than just minimizing the sum of squared errors, but it is easily accomplished with statistical software.

Figure 1 shows that even Gaussian and Poisson distributions that share the same mean differ in some fundamental ways. While a Gaussian distribution is continuous and extends over all numbers of events, the Poisson distribution is discrete and non-negative. For large numbers of events, the two distributions become more similar, but for small numbers of events the differences are relevant.

Figure 1: Examples of Gaussian and Poisson distributions. Note that the Gaussian distribution allows for fractional and negative values, and consequently may not be a good representation of some phenomena.



Statistical Significance

Using either regression approach, it is possible to calculate the probability that an underlying process without a trend, also known as a null hypothesis, would have produced data as extreme as that observed. This value is known as a "p-value." If the measurements we have seen are very unlikely to have been produced by a system without a trend, we conclude that there must be a trend. This threshold, the boundary between "significant" and "not significant," is specified before the analysis, before producing p-values based on actual data.

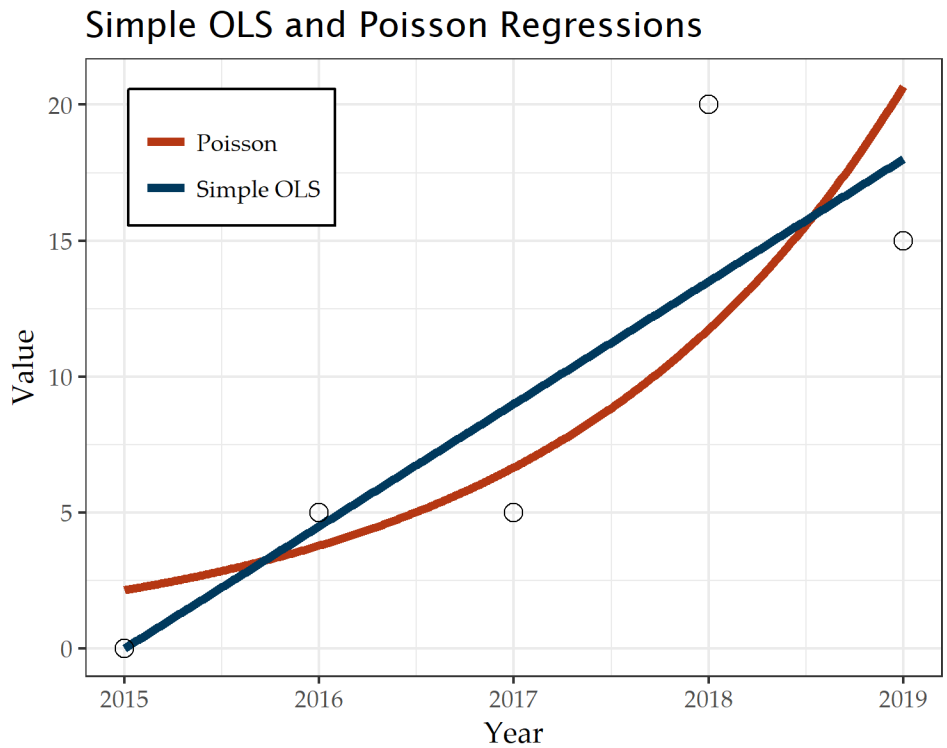
The results of significance tests are binary: either the regression is significant, or it is not. For a specific significance test, there is no middle ground.

One implication of this statistical approach is that even extreme data—data that is less likely than our significance threshold to have happened without a trend—does not *prove* a trend absolutely. Extreme data makes the null hypothesis very unlikely, but not impossible. So, statistical analysis acknowledges the possibility of false positives, results that look like trends, but were not actually trends. The significance threshold represents a tradeoff between the desire to capture all actual trends, and the desire not to inadvertently capture false positives.

Regression Model Comparison

Figure 2 shows that five data points might be fit in two very different ways based on whether we use a Gaussian or Poisson model. The OLS regression finds an annual increase of 4.5 events per year that is **not** significant, whereas the regression based on the Poisson model finds an annual increase of 76% that **is** significant. This example highlights the differences between the two regressions, and arguments can be made for either approach. The deeper lesson is that any conclusion depends on sometimes subtle modeling assumptions.

Figure 2: Comparison of how five hypothetical data points might be fit with two different regression models.



Regression Models used in 2020 SOTI

The 2020 SOTI applies a Poisson model to metrics that are counts and rates. This includes the number of reportable events, generation outages per unit, transmission outages per element, and

misoperations. The SOTI uses an OLS (Gaussian) regression for everything else (generation and transmission mean outage durations).

The 2020 SOTI trend analyses are based on the most recent five years of data if available, or four years otherwise. The decision to base conclusions on the most recent five years of data rather than, for example, the most recent year relative to the year before it, is a compromise between the desire to capture patterns of sufficient duration that they represent material trends in performance and operations, while also being timely. We want the SOTI to be an "early warning device," but not so early that we capture transient behaviors.

A p-value of 0.05 is used as a significance threshold for all SOTI regressions. This means that *if* the actual measurements were as described by the regression model, and *if* there was no actual trend, there would be a 5% chance of observing data as extreme as that observed. The significance threshold is a compromise between the desire to correctly identify actual trends and the desire not to mistakenly conclude that there is a trend from false positives. A 5% false-positive rate is a common significance threshold choice.

One implication of any false-positive rate is that testing several hypotheses, i.e., many combinations of types or causes of events, will produce some false positives. Testing more hypotheses will lead to more false positives. This consideration should prompt caution when interpreting results. The SOTI trend analyses are useful for screening, but should be followed by detailed and extensive analysis before being considered definitive.

Conclusion

The 2020 SOTI uses consistent and objective statistical tests to detect trends in performance data, rather than relying on subjective assessment. This attribute of objectivity can be maintained even if the details of statistical approaches change (e.g., different models, different numbers of years, or different significance thresholds). Nevertheless, minimizing the subjectivity of analysis also minimizes opportunities to be informed by expert opinion. While the results of SOTI analyses are valid and useful, broad conclusions about trends should ultimately be informed by, and complemented with, expert opinion and non-statistical analysis.