**Beyond the Numbers**

How confidence intervals can help focus attention and simplify analysis



Image by the author using [Stable Diffusion](https://stablediffusionweb.com/)

We trust numbers. If a report tells us that a business received 1,000 orders last week, we trust that the business received 1,000 orders. The transactions are a matter of historical record, and their count is a simple arithmetic operation. If anyone asks, “how many?”, we *know* the answer.

It’s much more difficult to understand what 1,000 orders might mean to a business. Someone might ask whether 1,000 orders is “good”, “great”, or merely “okay.” We could compare this week’s result to prior results, giving us a basis for comparison, but we’d still need a way to translate the results (or their differences) into qualitative terms.

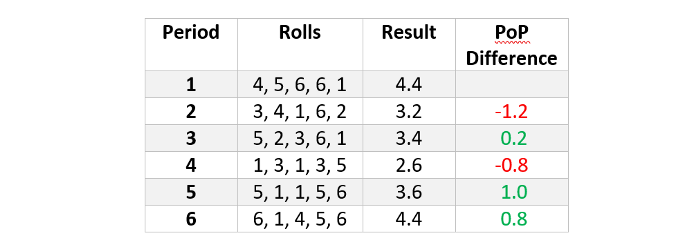
Likewise, someone might ask whether 1,000 orders is enough for the business to achieve its long-term objectives. To answer that question, we’d need to look across multiple weeks of data and separate the predominant trend from the weekly ups and downs. A simple rolling average might do the trick, but we’d still need to decide on the number of weeks to include in our average, which could have a significant effect on our answer.

Despite our trust in numbers, they can tell us only so much: The further we get from “how many?”, the harder it becomes to answer questions and the more context we need to interpret the data.

As the above examples suggest, looking at multiple data points can help. Period-over-period differences, rolling averages, and other calculations help us to see how results are trending. These are often paired with other enhancements, like color coding and up/down arrows, that indicate the direction of the movement but don’t tell us anything about its importance, which can actually lead to the *misinterpretation* of data.

A more effective approach is to use confidence intervals. While it might seem counterintuitive to see confidence intervals with numbers that we’re certain about, there’s a solid statistical basis for the approach and a number of surprising benefits.

Consider a simulated process. For each reporting period, we’ll use the average of five dice as our result. The table below contains the results from the first six periods along with their period-over-period (PoP) differences.



A trend chart makes it easier to see the results (*y*-axis) over time (*x*-axis) and the PoP differences:

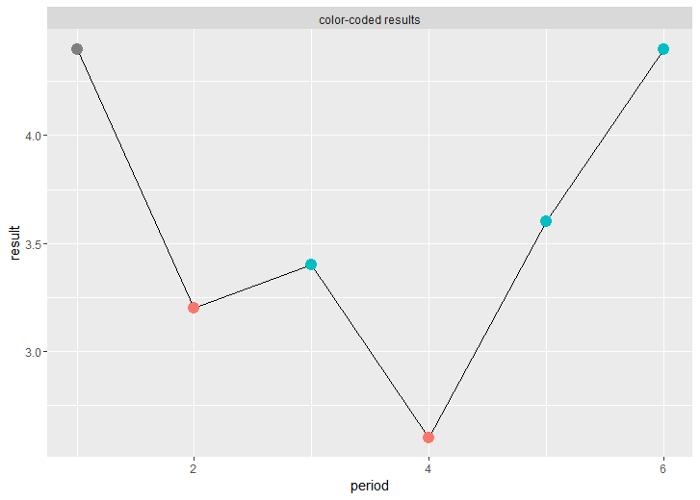


FIgure 1

In both report formats, increases are shown in green and decreases in red. As suggested, though, this common use of color can actually be counter-productive, making *any* change — up or down — seem important.

Imagine interpreting the first few results as they arrive. Period 1’s result is 4.4. At this point, we don’t know if 4.4 is high, low, or typical, but it sets our expectation.

The next period’s result is 3.2, a decrease of 1.2. We still don’t know if 4.4 and 3.2 are good results, but the red reinforces the notion that any decrease is headed in the wrong direction. If this were a real-world business process, we’d be tempted to take corrective action. It’s too early to identify what — if anything — is wrong, but we still might decide to work longer hours, reallocate resources, or whatever makes sense for this particular process.

Period 3’s result is 3.4, a small increase, but it’s reassuring. If we did take corrective action, we might assume that it’s working.

Period 4’s result, though, wipes away the prior period’s gain (and more). Perhaps we need to double down on our efforts or try something new….

Then the opposite occurs: Period 5 wipes away the prior period’s losses (and more). And by the time we get to period 6, we’re back to where we started: 4.4. So far, all the gains and losses have added up to nothing.

Because we’re talking about rolls of dice, that’s actually what we should expect. The individual results are random, so the differences between them are just as random. We can see this more clearly if we continue rolling dice.

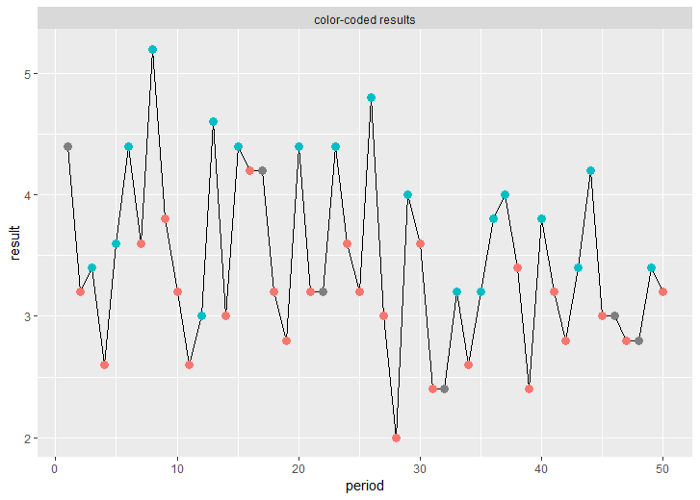


Figure 2

After 50 rolls, the average PoP difference is ±0.8, but the sum of all differences is just -1.2, which averages out to -0.02, which rounds to 0.0. While the PoP differences look big, they eventually cancel each other out.

Let’s consider how we can take advantage of this fact in our reporting.

Because we know the statistical properties of a 5-die roll, we can calculate a 95% confidence interval for our results: 3.5 ± 1.5. That means the average of all results will be 3.5, and we can expect 95% of results to land between 2.0 and 5.0. (More accurately, 95% of results ± 1.5 will contain 3.5.) In this case, 48 of the 50 results (96%) land between those limits.

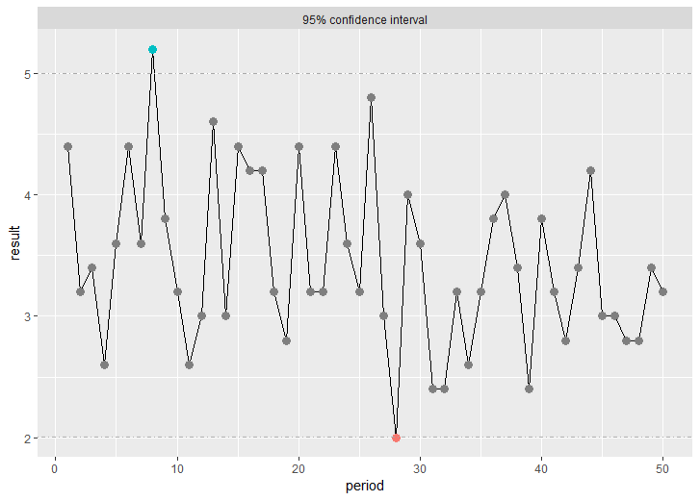


Figure 3

Periods 8 and 28 are the only results that fall outside those limits. That makes them good candidates for further investigation.

Keep in mind that it doesn’t mean that there will be anything meaningful to find. Again, our example is based on rolling dice, so *all* of these results are random. (With 50 results and a 95% confidence interval, we should expect two or three results to fall outside our confidence interval.) In a real-world business process, though, we wouldn’t know what caused periods 8 and 28 to be so different. We’d need to analyze them to find out (knowing, of course, that there might be nothing).

The real benefit, though, is that we can ignore the results from the other periods because those results are within expectation. Despite appearances, the ups and downs don’t mean anything: They’re a byproduct of the natural variation in the process.

Not having to analyze those periods’ results will save us a significant amount of time. The charts make the difference easy to see. In Figure 2, we color coded every gain and loss, creating the perception that every colored point is meaningful. (It’s lit up like a Christmas tree!) In Figure 3, though, we limited our use of color to just two points, which makes the chart easier to read and focuses our attention on the results that are more likely to be of interest.

Moreover, we’ll reduce the risk of reaching mistaken conclusions about the business. People are very good at finding patterns — even when they don’t exist — so if we analyze every up and down, presuming there’s something to find, there’s a good chance that we *will* find something. And if that leads to mistaken conclusions about what’s working and what’s not, we risk doing more harm to the business than good.

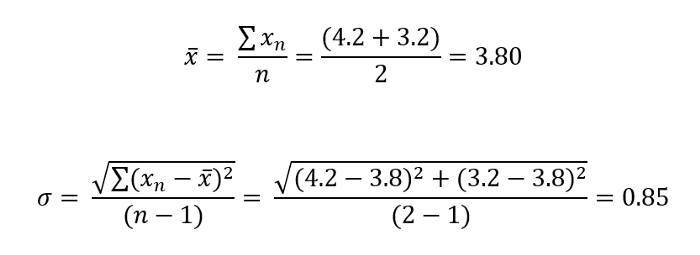
**Real-world complications**

Of course, if this were a real-world business process, we wouldn’t know its statistical properties, so we wouldn’t be able to pre-calculate the limits of our confidence interval. Fortunately, there’s a workaround: We can use the results themselves to infer the statistical properties of our process. Let’s re-use our dice example to see how this works.

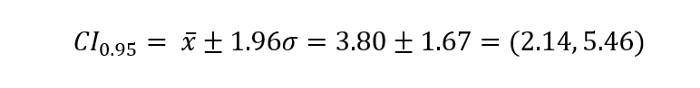
Every period’s result tells us more about the process. After the first period, we have just one result, but it sets our expectation for what results will look like. After the second period, we have two results, so we can update our expectation, and we begin to see how much the results will vary over time.

Statistically, these are the mean (*x\_bar*) or average, which is our expected value, and standard deviation (*σ*), which tells us how much the individual results will vary from the expected value.

After just two periods, we can calculate both:



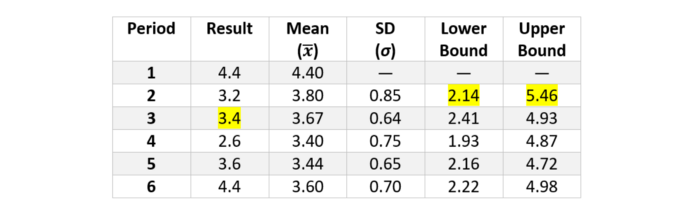
From these, we can calculate a confidence interval that will tell us whether the next period’s result is significantly different from what we’ve seen so far. This particular process is a classic example of the central limit theorem, so it’s safe to assume our results will follow a normal distribution, which means that a 95% confidence level is roughly 2*σ* (i.e., 1.96*σ*).



Given what we know so far — which isn’t much — we shouldn’t be surprised if period 3’s result is somewhere between 2.14 and 5.46. As it turns out, period 3’s result is 3.4, so it’s within expectation, which means the result is *not* significantly different from prior values.

Now that we have a new result, we can update our understanding of the process by recalculating our cumulative mean, standard deviation, and confidence interval, which we can use to evaluate the next new result.

The table below shows the confidence intervals for the first six periods. Note that we always compare the most recent result with the prior period’s confidence interval, as highlighted below. We can’t include the new result in the calculation of the confidence interval, otherwise we’d be using the result to validate itself.



With every new period, we further refine our understanding of the process and the closer our inferred statistics get to the true parameters. The chart below shows how our inferred limits converge on the theoretical limits and the differences between them diminish.

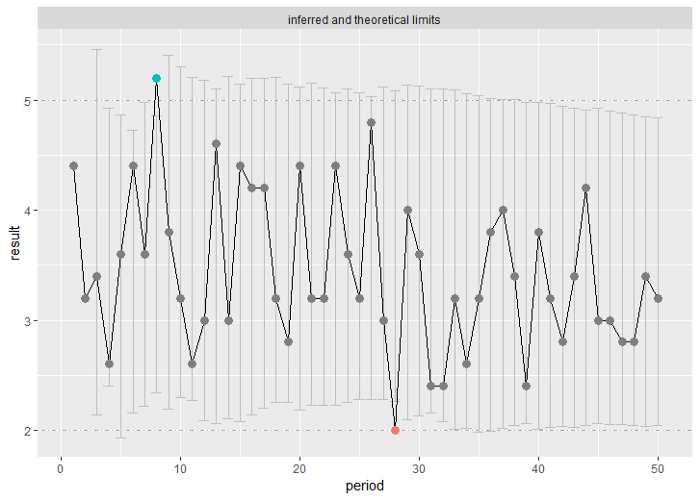


Figure 4

Despite the differences between limits, there are no points where the inferred and theoretical intervals lead to different conclusions. (That can happen, however, as we’ll see later.)

**More real-world complications**

Most real-world scenarios are more complex than rolling dice. A typical business process will include long-term trends (like growth) and seasonal changes (like holiday sales) that our confidence interval needs to account for.

Forecasting provides a convenient solution. The same as before, we use historic results to develop an understanding of the business process. In this case, though, the forecasting algorithm does most of the work. We simply pass our data to the algorithm and ask it to predict one period into the future, a technique known as one-step-ahead forecasting. The forecast will predict the next period’s result (i.e., point estimate) and a confidence interval (i.e., prediction interval) that we can use to evaluate the actual result.

Although the leap from statistics to machine learning might seem huge, forecasting is a logical extension of the concepts described above. In fact, if we use the *mean* method from R’s [*fable*](https://fable.tidyverts.org/) package, we get a chart that’s very similar to the one that we calculated by hand.

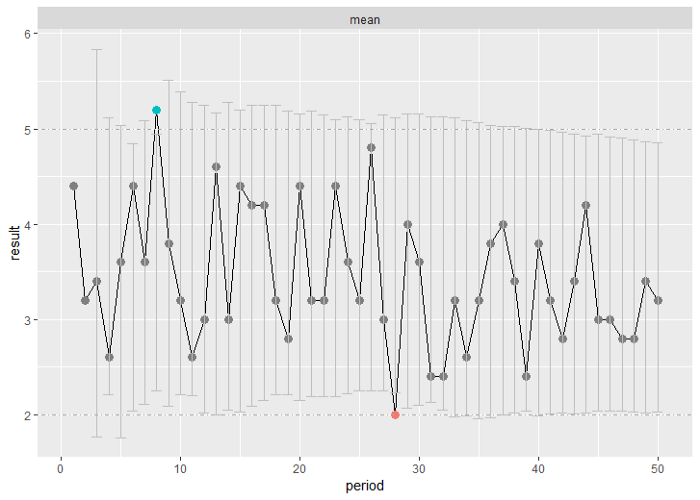


Figure 5

Below is the one-step-ahead forecast built using *fable’s tslm*method (time series linear model), which better illustrates how other algorithms capture trend and seasonality (something the mean method cannot do).

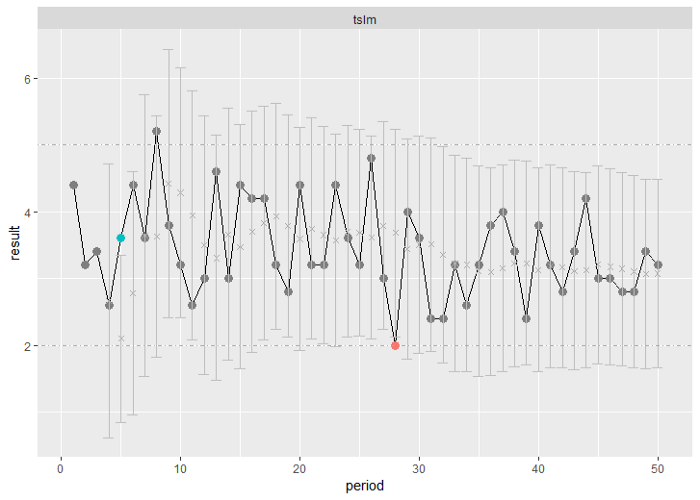


Figure 6

The first few results (periods 1–4) follow a negative trend, which the algorithm expects to continue, so the first few prediction intervals (periods 4–5) are lower than the intervals predicted by the *mean* method.

Period 5 goes against that early trend, however: The forecast predicted a value of 2.1 (shown as a grey x), but the actual result was 3.6. The algorithm quickly adapts with a higher point estimate for period 6 and a broader prediction interval.

Once the algorithm sees enough increases and decreases, it recognizes them as natural variation in the process, and the limits of our prediction intervals again converge on the theoretical limits.

Period 5 is particularly interesting, though, because it’s also an example of a false positive. It lands *outside* the limits of our inferred interval but *inside* the limits of the theoretical interval. False positives aren’t generally a concern, however, because we’ll probably want to investigate any result that’s significantly different from what we’ve seen before.

Period 8 is the opposite, an example of a false negative. Anytime the limits of our prediction interval are wider than those of the theoretical interval, we risk missing results that warrant further examination. A simple fix is to calculate a second interval based on a lower confidence level (e.g., 80%) to identify results that are *somewhat* surprising, but still worthy of our attention.

In fact, having multiple confidence intervals is a simple way to establish what qualitative terms (like “good” “great” and “okay”) might mean for your business. The [68–95–99.7 rule — Wikipedia](https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule), based on integer multiples of *σ* (1*σ*, 2*σ*, and 3*σ*), is a convenient starting point, but you can use any set of limits that makes sense for your process.

**Final considerations**

Before you use any of these techniques, you’ll need to ensure that the variance in your process is consistent, regardless of the size or sequence of results (see [Homoscedasticity and heteroscedasticity — Wikipedia](https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity)). If you go with forecasting, you’ll also need to make decisions about algorithms and accuracy measures (see [Forecasting — Wikipedia](https://en.wikipedia.org/wiki/Forecasting)). And once you start tracking results, you’ll want to monitor them for changes in trend that indicate the process has been fundamentally altered (see [Change detection — Wikipedia](https://en.wikipedia.org/wiki/Change_detection)).

Fortunately, there are well established ways of addressing these issues. And, in most cases, the effort is worth it. Confidence intervals give us a powerful tool for interpreting results. They leverage the full history of available data to put the most recent result into context, making it easier to assess if that result is something we should investigate or ignore. Conversely, looking at business results without confidence intervals is like reading the words on a page but ignoring the sentences: You might be able to make sense of individual numbers, but you’ll miss out on their combined meaning.