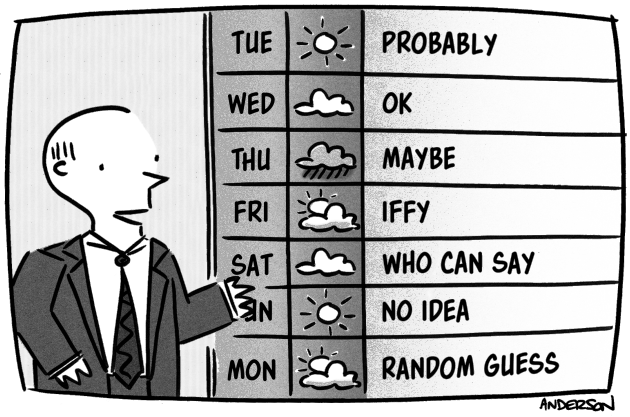
# How to measure the accuracy of forecasts

by Jason Cohen on June 28, 2016

**How do you assess forecasts, when the forecast is only a probability? It's not just about accuracy. Let's dive into the math.**




"And now the 7-day forecast..."

**"There's a 30% chance of rain today."**

And then it didn't rain. So, was the forecast accurate?

Or it did rain. Is the forecast inaccurate?

How do you hold forecasters accountable, when the forecast is itself a probability? The answer appears tricky, but ends up being simple enough to answer with Google Spreadsheets.

It's a journey worth taking because of the value of building better forecasts:

- **Lead scoring**: Putting a value on a new sales lead, predicting the chance of converting to sale, and its dollar value.
- **Predicting churn**: If you could predict the chance that a given customer will churn in the next thirty days, you could be proactive and perhaps avert the loss; do this enough and you're on the road to Product/Market Fit.
- **Predicting upgrades**: If you could predict the chance that a given customer is amenable to an upgrade, you could focus your internal messaging efforts accordingly.
- **Risk assessments**: Establishing better probabilities on risks results in more intelligent investments.

So how do you measure the accuracy of a prediction that is expressed as a probability? Let's return to the meteorologist.

## Accuracy Error

Clearly, a single data point tells you nothing.

Rather, the correct interpretation of "30% chance of rain" is the following: Gather all the days in which the meteorologist predicted 30%. If the meteorologist is accurate, it will have rained 30% of those times. Similarly, the forecaster will sometimes predict 0% or 10% or 50%. So we should "bucket" the data for each of these predictions, and see what actually happened in each bucket.

What is the right math to determine "how correct" the forecaster is? As is often the answer in statistics, we can take the squared difference[1] between the forecast and the actual result.

[1] Why do we square the errors instead of using something simpler like the absolute value of the difference? There are two answers. One is that squaring the differences intentionally exaggerates items which are *very* different from each other. The other is that the mathematics of squared differences is much more tractable than using absolute value. Specifically, you can expand and factor squared differences, and you can use differential calculus. Computing a linear regression line with least-squares, for example, is derived by using calculus to minimize the squared differences, but that same method cannot be applied to linear differences. Nassim Taleb is famously against this practice, but let's not argue the point now.

Suppose we have two forecasters, and we ask: Who is most accurate? "Accuracy Error" is measured by the sum of the squared differences between the forecast and reali-

ty. Whoever has the least total error is the better forecaster.

For example, suppose on some given set of days, forecaster A always predicted a 32% chance of rain, and B always predicted 25%, and suppose in reality it rained on 30% of those days. Then the errors are:

| | Predict | Actual | Squared Diff = Error |
|---|---|---|---|
| A | 32% | 30% | $(0.32 - 0.30)^2 = 0.0004$ |
| B | 25% | 30% | $(0.25 - 0.30)^2 = 0.0025$ |

It feels like we're finished, but alas no. If all we compute is Accuracy Error, we fall for a trap in which we easily forecast with perfect accuracy, while also being utterly useless.

## Discernment

Suppose these meteorologists are in a region that rains 110 days out of every 365. That is, the overall climactic average probability of rain is 30%. A meteorologist would know that. So, a meteorologist could simply predict "30% chance of rain" every single day, no matter what. Easy job!

Our Accuracy Error metric will report that this forecaster is perfect—exactly zero over a whole year of predictions. Because, the prediction is always 30%, and indeed on those days it rained 30% of the time: $(0.30 - 0.30)^2 = 0$ . Except the forecaster isn't perfect; she's not forecasting at all! She's just regurgitating the climactic average.


"But, to be fair, there's a fifty-percent chance of just about anything."

And so we see that, although "accuracy error" does measure something important, there's another concept we need to measure: The idea that the forecaster is being *discerning*. That is, that the forecaster is proactively *segmenting* the days, taking a strong stance about which days will rain and which will not. Staking a claim that isn't just copying the overall average.

There is a natural tension between accuracy and discernment which becomes apparent when you consider the following scenario:

Suppose forecaster A always predicts the climactic average; thus A has 0 accuracy error but also 0 discernment, and is therefore useless. Now consider forecaster B, who often predicts the climactic average, but now and then will predict 0% or 100% of rain, when he's very sure. And suppose that when he predicts 0% the actual average is 10%, and when he predicts 100% the actual average is 90%. i.e. when "B is very sure," B is usually correct.

B will have a worse accuracy error score, but should have a better discernment score. Furthermore, you would prefer to listen to forecaster B, *even though he has more error than A.* So the idea of "discernment" isn't just a curiosity, it's a fundamental component of how "good" a forecaster is.

How do you compute this "discernment?" We once again use squared differences, but this time we are comparing the difference between *observed results* and the *climactic average*, i.e. how much our prediction buckets differ from the climactic average, and thus how much we're "saying something specific."

For both accuracy error and discernment we use the weighted average for each prediction bucket. Let's start by computing accuracy error for forecaster B, assuming that out of 100 guesses, 8 times he predicted 0%, 12 times he predicted 100%, and the remaining 80 times he predicted the climactic average of 30%:

| Bucket | Actual in this bucket | Accuracy Error |
|---|---|---|
| 0% | 10% | $(0.00 - 0.10)^2 = 0.01$ |
| 30% | 30% | $(0.30 - 0.30)^2 = 0.00$ |
| 100% | 90% | $(1.00 - 0.90)^2 = 0.01$ |

Now we weight the errors by the number of predictions in each bucket, yielding the final Error as the weighted average:

| Bucket | Accuracy Error | # in Bucket | Weighted Error |
|---|---|---|---|
| 0% | 0.01 | 8 | $0.01 \times 8 = 0.08$ |
| 30% | 0.00 | 80 | $0.00 \times 80 = 0.00$ |
| 100% | 0.01 | 12 | $0.01 \times 12 = 0.12$ |
| *Total* | | 100 | 0.20 |
| *Average* | | | $0.20/100$ $= 0.002$ |

Forecaster B's accuracy error of 0.002 is still low. Now we'll compute this new discernment score, which is exactly like accuracy error, except instead of the squared difference between "predicted" and "actual," it's the squared difference between "climactic average" and "actual":

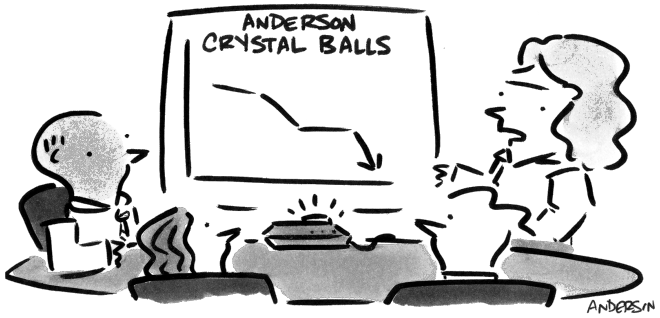
| Bucket | Actual in this bucket | Discernment |
|---|---|---|
| 0% | 10% | $(0.30 - 0.10)^2 = 0.04$ |
| 30% | 30% | $(0.30 - 0.30)^2 = 0.00$ |
| 100% | 90% | $(0.30 - 0.90)^2 = 0.36$ |

And then creating the weighted average, just as with accuracy error:

| Bucket | Discern. | # in Bucket | Weighted Discernment |
|---|---|---|---|
| 0% | 0.04 | 8 | $0.04 \times 8 = 0.32$ |
| 30% | 0.00 | 80 | $0.00 \times 80 = 0.00$ |
| 100% | 0.36 | 12 | $0.36 \times 12 = 4.32$ |
| *Total* | | 100 | 4.64 |
| *Average* | | | $4.64/100$ $0.0464$ |

We might be tempted to conclude "there's more discernment (0.0464) than accuracy error (0.002), therefore this forecaster is better." Is that valid? What is the right way to combine these two numbers for a total "goodness" score?

To answer that, it turns out there's one more concept we need.




"Seriously?! No one saw this coming?!"

## Discernment Error

Consider the life of a forecaster in Antofagasta, Chile, where on average it rains only five days a year (for a grand total of 1.7 *millimeters* of annual rainfall!). At first

glance it seems easy to be a forecaster—just predict "no rain" every day.

Although that forecaster would have low error, she would also be undiscerning. But wait… how could a forecaster *ever* be discerning in Antofagasta? To be discerning you need to make varied predictions. But *reality isn't varied*, so any predictions that *were* varied, would necessarily be *wrong*! In a sense, there's no "space" for discernment, because there's no variation to discern between. It's not fair to ask the forecaster to be more discerning than the variation that is actually available to discern.

Compare that with forecaster in Portland, Oregon, USA where it rains 164 days out of the year—about 45%. And there's no "rainy season"—it's just chaotic. Here even just predicting 55% or 35% here and there could still be highly accurate but increase discernment. And a world-class forecaster has the space to create a significant amount of discernment.

So it's not quite fair to ask "How discerning is the forecaster?" Instead we should ask "How discerning is the forecaster, *compared with* how much discernment is even *possible*?"

The maximum amount of discernment possible, given the climactic average $c$, is[2]:

$$c(1-c)$$

In general, the closer the climactic average is to 0% or 100%, the less discernment there can be. The maximum possible discernment of 0.25 is available when the climactic average is 50%, i.e. it's a coin flip.

[2] For the mathematically curious: Maximum discernment happens when the forecast is only 100% or 0%, and when it is completely accurate. If there were $N$ predictions, the positive case will happen $Nc$ times, and the negative case $N(1-c)$ times. Discernment in the positive bucket is $(1-c)^2$ and in the negative bucket $(0-c)^2 = c^2$. The weighted average is $\frac{1}{N}[Nc(1-c)^2 + N(1-c)c^2]$. Factoring out the common $Nc(1-c)$ and canceling with $\frac{1}{N}$, we're left with $c(1-c)[(1-c)+c] = c(1-c)$.

In the 30% example, the maximum possible discernment is $0.30(1-0.30) = 0.21$. Forecaster B's discernment of 0.04 is therefore not too impressive—plenty of room for improvement. Although still of course B is better than A, who had no discernment whatsoever.

In the case of the desert, with a climactic average of $5/365 = 0.013\%$, there's only 0.0128 potential discernment available.

In any case, this allows us to compute a metric that is comparable to error, but for discernment:

[Discernment Error] = [Maximum Discernment] − [Discernment]

That is, if you have no discernment, that's another type of "error"—your forecast is lacking the descriptive power that would come from being maximally discerning. The more discernment you demonstrate, the less of this "discernment error" you exhibit, and the better your forecast is. Just like the less accuracy error you have, the better your forecast is.

## Putting it all together: The Forecasting Error Score

It turns out[3] you can simply add accuracy error to discernment error, and arrive at a rigorous metric:

[3] See "Further Reading" below for the mathematical justification.

[Forecast Error] = [Error] + [Discernment Error]

Or, writing out each component:

[Forecast Error] = [Error] + [Maximum Discernment] − [Discernment]

Here's a way to see why this math works: Every forecaster's baseline is to guess the climactic average. That will get you a total score of $c(1-c)$, because you have no accuracy error, but maximum discernment error.

From there, forecasters will try to deviate from the climactic average. The more they put themselves on the line, the less discernment error they rack up, however they also have to be right! The best forecasters outperform the climactic-average (discernment) by more than the accuracy error they introduce. The overall score tells you who is better. Lower is better, since the score is a "total error."

It is possible to do worse than the climactic average—to make guesses, but be wrong. You know that's happening when the total error is larger than the baseline $c(1-c)$.

Indeed, an alternate and equivalent scoring method divides Forecast Error by $c(1-c)$, to create a metric where 1 means "equal to the baseline," 0 means "perfect forecast," and the amount greater or less than 1 indicates proportionally how much better or worse the forecast is from baseline. This has the virtue of the reader not needing to know the value of $c(1-c)$ in order to understand whether a forecaster is better or worse than the baseline, and by how much.

While the total score is useful, the individual components of accuracy error and discernment are also useful because they help you analyze what's going on:

| Accuracy Error | Discern | Meaning | Explanation |
|---|---|---|---|
| ↑ | ↓ | **Failure** | You're not segmenting the population, and yet you're still worse than guessing the climactic average. |
| ↓ | ↓ | **Useless** | You're only accurate because you're just guessing the average. |
| ↑ | ↑ | **Try Again** | You're making strong predictions, so at least you're trying, but you're not guessing correctly. |
| ↓ | ↑ | **Ideal** | You're making strong predictions, and you're correct. |

Now that you know how to measure forecasts, it's time for you to build some forecasting models. So go try to better understand your customers and prospects, and use this math to know whether iterations of your model are improving.

## Further Reading

- Glenn Brier's original paper proposing this method in 1950, but without this three-component breakdown that was discovered by Allen Murphy in 1973.
- Brier score on Wikipedia: A more formal explanation including the three-component breakdown (which are labelled and explained differently from my exposition, but which are mathematically identical).
- Stein's Paradox: An estimator that's always better than the historical average, but in a way that apparently can't be true.