Parameter-Centric Analysis Grossly Exaggerates Certainty

William M. Briggs
E-mail: matt@wmbriggs.com*

Abstract The reason probability models are used is to characterize uncertainty in observables. Typically, certainty in the parameters of fitted models based on their parametric posterior distributions is much greater than the predictive uncertainty of new (unknown) observables. Consequently, when model results are reported, uncertainty in the observable should be reported and not uncertainty in the parameters of these models. If someone mistook the uncertainty in parameters for uncertainty in the observable itself, a large mistake would be made. This mistake is exceedingly common, and almost exclusive in some fields. Reported here are measures of the over-certainty mistake made when parametric uncertainty is swapped with observable uncertainty.

Keywords: Model reporting, Posterior distributions, Predictive probability, Uncertainty

1 Over-Certainty

Suppose an observable $y \sim \text{Normal}(0, 1)$; i.e., we characterize the uncertainty in an observable y with a normal distribution with known parameters (never mind how we know them). Obviously, we do not know with exactness what any future value of y will be, but we can state probabilities (of intervals) for future observables using this model.

It might seem an odd way of stating it, but in a very real sense we are infinitely more certain about the value of the model parameters than we are about values of the observable. We are *certain* of the parameters values, but we have uncertainty in the observable. In other words, we know what the parameters are, but we don't know what values the observable will take. If the amount of uncertainty has any kind of measure, it would be 0 for the value of the parameters in this model, and something

*Corresponding author

positive for the value of the observable. The ratio of these uncertainties, observable to parameters, would be infinite.

That trivial deduction is the proof that, at least for this model, certainty in model parameters is not equivalent to certainty in values of the observable. It would be an obvious gaff, not even worth mentioning, were somebody to report uncertainty in the parameters *as if* it were the same as the uncertainty in the observable.

Alas, this is what is routinely done in probability models, see Chapter 10 of [1]. Open the journal of almost any sociology or economics journal, and you will find the mistake being made everywhere. If predictive analysis were used instead of parameteric or testing-based analysis, this mistake would disappear; see e.g. [2, 3, 4, 5]. And then some measure of sanity would return to those fields which are used to broadcasting "novel" results based on statistical model parameters.

The techniques to be described do not work for all probability models; only those models where the parameters are "like" the observables in a sense to be described.

2 Theory

There are several candidates for a measure of total uncertainty in a proposition. Since all probability is conditional, this measure will be, too. A common measure is variance; another is the length of the highest (credible) density interval. And there are more, such as entropy, which although attractive has a limitation described in the final section. I prefer here the length of credible intervals because they are stated in predictive terms in units of the observable, made using plain-language probability statements. Example: "There is a 90% chance y is in (a,b)."

In the $y \sim \text{Normal}(0,1)$ example, the variance of the uncertainty of either parameter is 0, as is the length of any kind of probability interval around them. The variance of the observable is 1, and the length of the $1-\alpha$ density interval around the observable y is well known to be $2z_{\alpha/2}$, where $z_{\alpha/2} \approx 2$. The ratio of variances, parameter to observable, is $0/1 = 0$. The ratio of the length of confidence intervals, here observable to parameter, is $4/0 = \infty$.

We pick the ratio of the length of the $1-\alpha$ credible intervals as observable to parameter to indicate the amount of *over-certainty* . If not otherwise indicated, I let α equal the magic number.

In the simple Normal example, as said in the begining, if somebody were to make the mistake of claiming the uncertainty in the observable was identical to the uncertainty of the parameters, he would be making the worst possible mistake. Naturally, in situations like this, few or none would this blunder.

Things change, though, and for no good reason, when there exists or enters uncertainty in the parameter. In these cases, the mistake of confusing kinds of uncertainty happens frequently, almost to the point of exclusively.

The simplest models with parameter uncertainty follow this schema:

$$p(y|\text{DB}) = \int_\theta p(y|\theta,\text{DB})p(\theta|\text{DB})d\theta, \tag{1}$$

where D=$y_1, \ldots, y_n$ represents old measured or assumed values of the observable, and B represents the background information that insisted on the model formulation used. D need not be present. B must *always* be; it will contain the reasoning for the model form $p(y|\theta \mathrm{DB})$, the form of the model of the uncertainty in the parameters $p(\theta|\mathrm{DB})$, and the values of hyperparameters, if any. Obviously, if there are two (or more) contenders i and j for priors on the parameters, then in general $p(y|\mathrm{DB}_k) \neq p(y|\mathrm{DB}_l)$. And if there are two (or more) sets of D, k and l, then in general $p(y|\mathrm{D}_i\mathrm{B}) \neq p(y|\mathrm{D}_j\mathrm{B})$. Both D and B may differ simultaneously, too.

It is worth repeating that unless one can *deduce* from B the form of the model (from the first principles of B), observables do not "have" probabilities. All probability is conditional: change the conditions, change the probability. All probability models are conditional on some D (even if null) and B. Change either, change the probability. Thus all measures of over-certainty are also conditional on D and B.

If D is not null, i.e. past observations exist, then of course

$$p(\theta|\mathrm{DB}) = \frac{p(y|\theta \mathrm{DB})p(\theta|\mathrm{DB})}{\int_\theta p(y|\theta \mathrm{DB})p(\theta|\mathrm{DB})d\theta} \tag{2}$$

The variances of $p(y|\mathrm{DB})$ or $p(\theta|\mathrm{DB})$ can be looked up in the model forms are common, or estimated if not.

Computing the highest density regions or intervals (HDI) of a probability distribution is only slightly more difficult, because multi-modal distributions may not have contiguous regions. We adopt the definition of [6]. The $1-\alpha$ highest-density region R is the subset $R(p_\alpha)$ of y such that $R(p_\alpha) = \{y\colon\ p(y) \geq p_\alpha\}$ where p_α is the largest constant such that $\Pr\big(y \in R(p_\alpha)|\mathrm{DB}\big) \geq 1-\alpha$. For unimodal distributions, this boils down to taking the shortest continuous interval containing $1-\alpha$ probability. These, too, are computed for many packaged distributions. For the sake of brevity, all HDI will be called here "credible intervals."

It will turn out that comparing parameters to observables cannot always be done. This is when the parameters is not "like" the observable; when they are not measured in the same units, for example. This limitation will be detailed in the final section.

3 Analytic Examples

The analytic results here are all derived from well known results in Bayesian analysis. See especially [7] for the form of many predictive posterior distributions.

3.1 Poisson

Let $y \sim \mathrm{Poisson}(\lambda)$, with conjugate prior $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$. The posterior on λ is distributed $\mathrm{Gamma}(\sum y + \alpha, n + \beta)$ (shape and scale parameters). The predictive

posterior distribution is Negative Binomial, with parameters $(\sum y + \alpha, \frac{1}{n+\beta+1})$. The mean of both the parameter posterior and predictive posterior are $\frac{\sum y+\alpha}{n+\beta}$. The variance of the parameter posterior is $\frac{\sum y+\alpha}{(n+\beta)^2}$, while the variance of the predictive posterior is $\frac{\sum y+\alpha}{(n+\beta)^2}(n+\beta+1)$. The ratio of the means, independent of both α and β, is 1. The ratio of the parameter to predictive variance, independent of α, is $1/(n+\beta+1)$.

It is obvious, for finite β, that this ratio tends to 0 at the limit. This recapitulates the point that eventually the value of the parameter becomes certain, i.e. with a variance tending toward 0, while the uncertainty in the observable y remains at some finite level. One quantification of the exaggeration of certainty is thus equal to $(n+\beta+1)$.

Although credible intervals for both parameter and predictive posteriors can be computed easily in this case, it is sometimes an advantage to use normal approximations. Both the Gamma and Negative Binomial admit normal approximations for large n. The normal approximation for a Gamma$(\sum y + \alpha, n+\beta)$ is Normal$((\sum y+\alpha)/(n+\beta), (\sum y+\alpha)/(n+\beta)^2)$. The normal approximation for a Negative Binomial $(\sum y + \alpha, \frac{1}{n+\beta+1})$ is Normal$((\sum y+\alpha)/(n+\beta), (n+\beta+1) * (\sum y+\alpha)/(n+\beta)^2)$

The length of the $1-\tau$ credible interval, equivalently the $z_{\tau/2}$ interval, for any normal distribution is $2z_{\tau/2}\sigma$. Thus the ratio of predictive to parameter posterior interval lengths is independent of τ and to first approximation equal to $\sqrt{n+\beta+1}$. Stated another way, the predictive posterior interval will be about $\sqrt{n+\beta+1}$ times higher than the parameter posterior interval. Most pick a β of around or equal to 1, thus for large n the over-certainty grows as $\sqrt{n}$. That is large over-certainty by any definition.

Also to a first approximation, the ratio of length of credible intervals also tends to 0 with n. Stated another way, the length of the credible interval for the parameter tends to 0, while the length of the credible interval for the observable tends to a fixed finite number.

3.2 Normal, σ^2 Known

Let $y \sim \text{Normal}(\mu, \sigma^2)$, with σ^2 known, and with conjugate prior $\mu \sim \text{Normal}(\theta, \tau^2)$. The parameter posterior on μ is distributed Normal with parameters $\sigma_n^2\left(\frac{\theta}{\tau^2} + \frac{n\bar{y}}{\sigma^2}\right)$ and $\sigma_n^2 = \sigma^2\tau^2/(n\tau^2+\sigma^2)$. The posterior predictive is distributed as Normal, too, with the same central parameter and with the spread parameter $\sigma_n^2 + \sigma^2$.

The ratio of parameter to predictive posterior variances is $\sigma_n^2/(\sigma_n^2+\sigma^2)$, which equals $\tau^2/((n+1)\tau^2+\sigma^2)$. This again goes to 1 in the limit, as expected. The ratio of credible interval lengths, predictive to the posterior, is the square root of the inverse of that, or $\sqrt{n+1+\tau^2/\sigma^2}$. As with Poisson distributions, this gives over-certainty which also increases proportionally to $\sqrt{n}$.

3.3 Normal, σ^2 Unknown

Let $y \sim \text{Normal}(\mu, \sigma^2)$, with a Jeffrey's prior over both parameters, which is proportional to $1/\sigma^2$. Then the marginal parameter posterior for μ (considering σ^2 to be of no direct interest) is a scaled T distribution with parameters $(\bar{y}, s^2/n)$ and with $n-1$ degrees of freedom. The predictive posterior is also a scaled T distribution also with $n-1$ degrees of freedom, and with parameters $(\bar{y}, s^2(n-1)/n)$.

For modest n, a normal approximation to the scaled T is sufficient. Thus the ratio of parameter to predictive posterior variances is equal to $1/(n-1)$. As before, this tends to 0 with increasing n. The ratio of the length of credible intervals is obvious, which again shows over-certainty rises proportionally to about $\sqrt{n}$.

Consider conjugate priors instead. Conditional on σ^2, the distribution of μ is a Normal with parameters $(\theta, \sigma^2/\tau)$. And the distribution of σ^2 is and Inverse Gamma with parameters $(\alpha/2, \beta/2)$. Then the conditional parameter posterior of μ is distributed as a scaled T with $\alpha + n$ degrees of freedom and with parameters $((\tau\theta + n\bar{y})/(\tau + n), (n-1)^2 s^2/\theta)$. The predictive posterior is also a scaled T with $\alpha + n$ degrees of freedom and with the same central parameter, but with a spread parameter equal to parametric posterior but multiplied by $\tau + n$.

Obviously, the ratio of parametric to predictive posterior variances is $1/(\tau + n)$, which again tends to 0 with n. Using the same normal approximation shows the credible interval ratio gives an over-certainty multiplier of $\sqrt{\tau + n}$.

The choice of Jeffrey's improper or the conjugate prior makes almost no difference to amount of over-certainty, as expected.

3.4 Regression, σ^2 Known

A regression model for observable y with predictor measures x is $y = x\beta + \varepsilon$, where the uncertainty in ε is characterized by a Normal distribution with parameters $(0, (\lambda I)^{-1})$, where λ is a scalar and I the identity matrix. The parameter posterior for β is a Normal distribution with parameters $(x'x + \lambda\sigma^2 I)^{-1} x'y$ and $\sigma^2(x'x + \lambda\sigma^2 I)^{-1})$. The predictive posterior for a new or assumed x, which we can write as w (a single vector), is also a Normal distribution, with parameters $w'(x'x + \lambda\sigma^2 I)^{-1} x'y$ and $\sigma^2(1 + w'(x'x + \lambda\sigma^2 I)^{-1} w)$.

Now as $\lambda \to 0$ the prior more resembles a Jeffrey's prior. Then the "ratio" of parametric to predictive variances is $\sigma^2(x'x)^{-1}(\sigma^2)^{-1}(1 + w'(x'x)^{-1}w)^{-1} = (x'x)^{-1}(1 + w'(x'x)^{-1}w)^{-1}$. The quantity $(1 + w'(x'x)^{-1}w)$ will be some scalar a, thus the ratio becomes $(x'x)^{-1}/a$. The ratio therefore depends on the measures x and their inherent variability. This will become clearer in the numerical examples.

3.5 Other Models

There are a host of models where the calculation pickings are easy and analytic. However, with complex forms, analytics solutions are not readily available. In these cases, we use standard simulation approaches. This will be demonstrated below for a general regression example, which introduces an extension of over-certainty computation. The same limitation mentioned at the beginning about these techniques only working for situation where the parameters are "like" the observable still applies.

4 Numeric Examples

4.1 Poisson

The length of years (rounded up) served by each Pope of the Catholic Church was collected by the author. These ran from 1 year, shared by many Popes, to 36 years, which was the longest and was for St Peter, the first Pope. The mean was 8.3 years. There were 263 past Popes (and one current one). A natural first model to try is the Poisson, as above. The hyperparameters chosen were the commonly picked $\alpha = \beta = 1$. The model is shown in Fig. 1.

The spikes are the frequency of observations (a modified histogram), the dashed red line the Gamma parameter posterior, and the black solid the Negative Binomial predictive posterior. The peak of the parameter posterior is not shown, though the center of the distribution is clear.

It is clear the certainty in the λ parameter is not anywhere close to the certainty in the future values of length of reign. That is, the certainty in the parameter is vastly stronger than the certainty in the observable.

The length of the 95% credible interval for the parameter posterior is 0.69 years, while it is 11 years for the predictive posterior, a ratio of 15.9. This is very close to the normal approximation value of $\sqrt{n+\beta+1} = 16.3$.

It is clear from the predictive posterior that the model does not fit especially well. It gives too little weight to shorter reigns, and not enough to the longest. Clearly better alternatives are available, but they were not attempted. I kept this poor model because the graph makes an important point.

It would be wrong to say "every time", but very often those who use probability models never check them against observables in this predictive manner. An estimate of the parameter is made, along with a credible (or confidence) interval of that parameter, and it is that which is reported, as stated above. Predictive probabilities of observables, the very point of modeling, is usually forgotten. Because of this, many poor models are released into the wild.

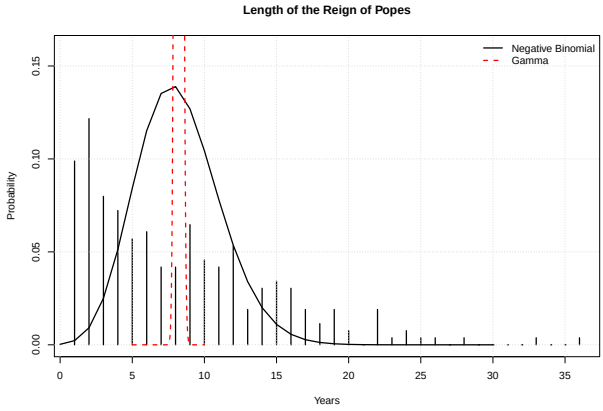

Fig. 1 The spikes are the frequency of observations (a modified histogram), the dashed red line the Gamma parameter posterior, and the black solid the Negative Binomial predictive posterior. The peak of the parameter posterior, scaled to the frequency, is not shown because it is so far off the scale, though the center of the distribution is clear.

4.2 Normal, σ Unknown

We skip over the normal example with σ known and move to where σ is unknown. The rape rate per 100,000 for each of the 50 United States in 1973 was gathered from the built-in R dataset `USArrests`. A normal model was fit to it using Jeffrey's prior. The result is in Fig. 2.

The histogram is superimposed by the parameter posterior (red, dashed), and predictive posterior (black, solid). The models fit is in the ballpark, but not wonderful. It is clear uncertainty in the parameters is much greater than in the observables.

The lenth of the parameter posterior credible interval is 5.3 (rapes per 100,000), while for the predictive posterior it was 37.3. The ratio of predictive to parameter is 7. In other words, reporting only the parameter uncertainty results is seven times too certain.

Why one would do this model is also a question. We have the rates, presumably measured without error, available at each state. There is no reason to model unless one wanted to make predictive statements of future (or past unknown) rates, conditional on assuming this model's adequacy.

There is also the small matter of "probability leakage", [8], shown a the far left of the predictive distribution, which we leave until the regression examples. Probability leakage is probability given to observations known to be impossible, but which is information not told to B. This will be clear in the examples.

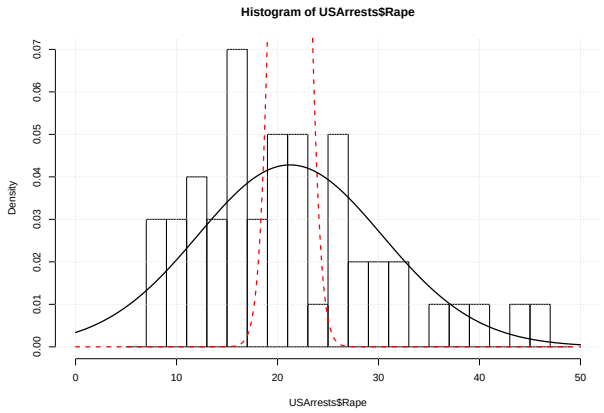

Fig. 2 The histogram is superimposed by the parameter posterior (red, dashed), and predictive posterior (black, solid). The models fit is in the ballpark, but not wonderful. It is clear uncertainty in the parameters is much greater than in the observables.

4.3 Regression 1, σ^2 Known

We use here the beloved `cars` dataset from R, which measured the speed of 50 cars (in mph) and how long it took them to stop (in feet). Speeds ranged from 4 to 25 mph, and distances from 2 to 120 feet. In order for σ to be known, we cheated, and used the estimate from the ordinary linear regression of speed on distance. Fig. 3 shows the results.

Now this data set has been used innumerable times to illustrate regression techniques, but I believe it is the first time it has been demonstrated how truly awful regression is here.

In each panel, the predictive posterior is given in black, and the parameter posterior is given in dashed red. In order to highlight the comparisons, the parameter posterior was shifted to the peak of the predictive posterior distributions. The parameter posterior is of course fixed—and at the "effect" size for speed. Here it has a mean of 3.9, with credible interval of (3.1, 4.8).

It is immediately clear just reporting the parameter posterior impies vastly more certainty than the predictive posteriors. We do not have just one predictive posterior, but one for every possible level of speed. Hence we also have varying levels of over-certainty. The ratio of predictive to parameter credible intervals was, 40.1 (at 1 mph), 37.8 (10 mph), 37.6 (20 mph), and 40.1 (30 mph),

The over-certainty is immense at any speed. But what is even more interesting is the enormous probability leakage at low speeds. Here we have most probability for predictive stopping distances of less than 0, a physical impossibility. The back-

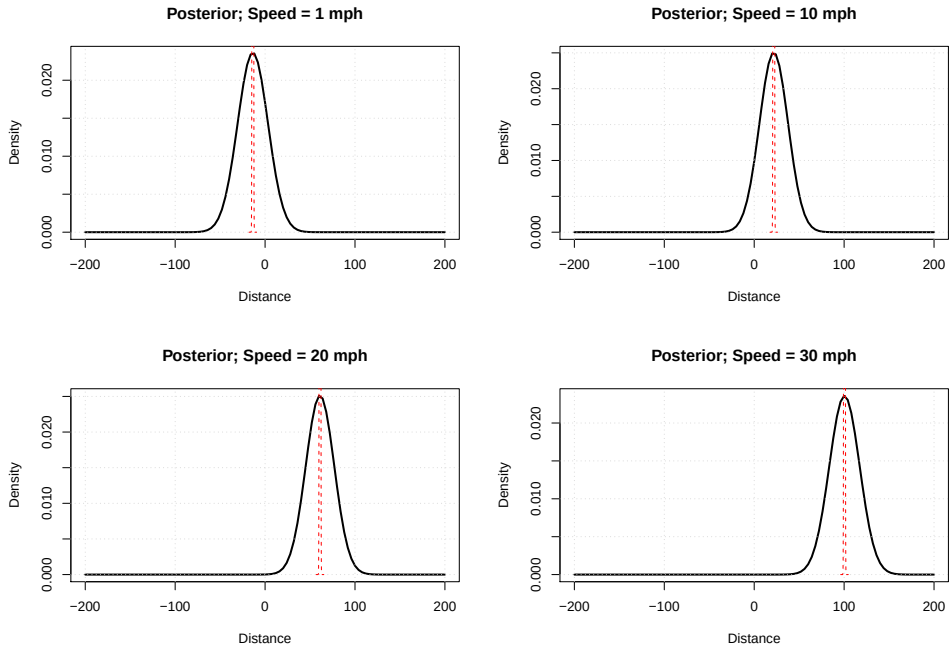

Fig. 3 In each panel, the predictive posterior is given in black, and the parameter posterior is given in dashed red. In order to highlight the comparisons, the parameter posterior, which is fixed regardless of the distance, was shifted to the peak of the predictive posterior distributions.

ground information B did not account for this impossibility, and merely said, as most do say, that a regression is a fine approximation. It is not. It stinks for low speeds.

But this astonishingly failure of the model would have gone forever unnoticed had the model not been cast in its predictive form. The parameter-centric analysis would be have missed, and almost always does miss, this glaring error.

If instead of σ^2 being known, and faked, we let it be unknown, the results for this model are much the same, only with greater uncertainties in the posteriors. We skip that and move to a more general example.

4.4 Regression 2, σ^2 Unknown

It is well to take an example which in most respects passes muster, even in a predictive sense, but which nevertheless still exaggerates certainty. We use the `oats` data in the *MASS* package. This was an experiment with three oat varieties planted at 6 blocks, and with four amounts of added nitrogen via manure: none added, 0.2cwt, 0.4cwt, and 0.6cwt per acre. The outcome is yield of oats in quarter pounds. The intent and supposition was that greater amounts of nitrogen would lead to greater yelds.

The first step is computing an ordinary linear regression; here using Gibbs sampling via the `rstanarm` package version 2.18.2 in R version 3.5.2, using default priors, but here using 6 chains of 20,000 iterations each for greater resolution. Diagnostics (not shown) suggest the model converged. The results are in Fig 4.

The predictive posteriors are presented as histograms, using posterior observable draws, and the parameter posteriors as red dashed lines. We chose the first block and the Golden rain oat variety to create the predictions. There is some difference between blocks, but little between varieties. A full analysis would, of course, consider all these arrangemnets, but here these choices will be sufficient. The parameter posteriors are cut off as before so as not to lose resolution on the observables. The parameter posteriors are also re-centered for easy visual comparison by adding the mean intercept value.

The over-certainty as measured by length of the predictive to parameteric credible intervals is about 3 times for each amount of nitrogen. However, since the purpose of this was to demonstrate increasing nitrogen boosts yields, this measure may seem out of place. Rather, the mistake of assuming the uncertainties are the same is not likely to be made.

There is still lurking over-certainty, though.

The posterior probability the parameter for 0.2cwt added nitrogen is greater than 0 is very close to 1 (0.9998). Any researcher would go away with the idea that adding nitrogen guarantees boosting yield. But the posterior predictive chance that a new plot with 0.2cwt will have a greater yield than a plot with no added nitrogen is only 0.8. This implies an over-certainty of $\approx 1/0.8 = 1.25$.

Likewise, although all the parameters in the model for nitrogen have extremely high probabilities of differing from each other , the predictive probabilities of greater yeilds do not differ as much. The posterior probability the parameter for 0.4cwt is greater than 0.2cwt is 0.997, and the the parameter for 0.6cwt is greater than 0.4cwt is 0.96.

The predictive probability of greater yields in plots with 0.4cwt over 0.2cwt is 0.75, with over-certainty of 0.997/0.75 = 1.33. And the predictive probability of greater yields in plots with 0.6cwt over 0.4cwt is 0.66, with over-certainty of 0.96/0.66 = 1.45.

Notice that we were able to compute over-certainty in this case because we were comparing probabilities for “like” things, here oats yields. We cannot always make comparisons, however, as the last section details.

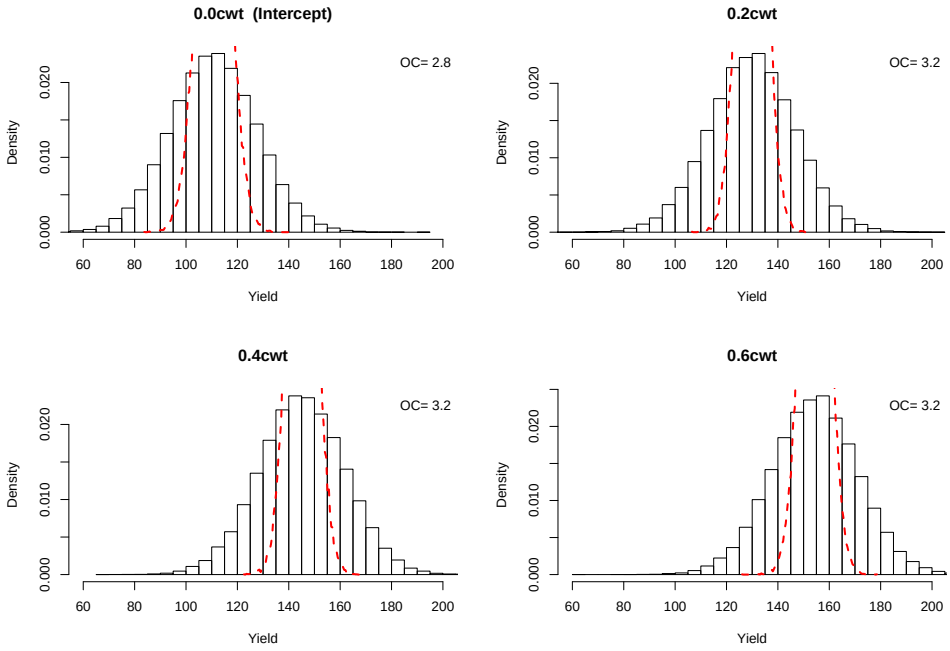

Fig. 4 The predictive posteriors are presented as histograms, using posterior observable draws, and the parameter posteriors as red dashed lines. These are cut off as before so as not to lose resolution on the observables. The parameter posteriors are also re-centered for easy visual comparison by adding the mean intercept value. The length of the predictive to parameteric credible intervals are given in the top corner (as OC = Over-certainty).

5 Comparing Apples & Kumquats

There are many more probability models beside regression and the like, but the techniques described in this paper only work where the parameters represent the same kind of thing as the observable. A normal model for the observable, say, weight, and the central parameter for that model are both in pounds (or kilograms), making comparison natural. But in other models, no such comparison can be made.

Consider the simplest logistic model with $p = \Pr(y|\mathrm{DB})$ and

$$\log\left(\frac{p}{1-p}\right) = \beta, \tag{3}$$

with B specifying β and the model form, and with D (here) null.

Obviously, the odds of y are equal to e^{β}. If β is known, there is still uncertainty in the observable, which has probability $p = e^{\beta}/(1+e^{\beta})$. Any uncertainty interval around β would be 0, but what does it mean to have an "interval" around p? The probability p is also certain: it is equal to p! It is the observable which is uncertain. It will happen or not, with probability p. There is no interval.

One might guess entropy might rescue the notion of uncertainty, but it is not so. The entropy for our uncertainty of β is 0, while for the observable it is $p\log(p) + (1-p)\log(1-p) > 0$. At first glance it thus appears entropy will help. But consider that B instead specifies we have β or $\beta+\varepsilon$ where $\varepsilon \ll 1$, both with a probability of 0.5. The entropy of the probability distribution of the parameter is $\log(2)$. But the enropy of the observable is based on the distribution $p = e^{\beta}/(1+e^{\beta}) \times 0.5 + e^{\beta+\varepsilon}/(1+e^{\beta+\varepsilon}) \times 0.5$. Since ε is small, $p \approx e^{\beta}/(1+e^{\beta})$ and the entropy is also approximately the same: $p\log(p) + (1-p)\log(1-p)$.

If β is at all large, the enropy of the observable will be near 0. Thus, according to entropy with "large" β, there is more uncertainty in the parameter than in the observable! There is now the idea of an interval around the parameter, but not around y.

The reason entropy doesn't work, nor credible intervals, nor variance which is here similar to entropy, is because the parameter is of a different kind than the observable, of a different nature. In logistic models the "βs are usually multipliers of odds, and multipliers of odds aren't anything "like" observables, which are "successes" or "failures".

All that one can do in these instances, then, is to emphasize that uncertainty of parameters just isn't anything like uncertainty in the observable. The two do not translate. It's like comparing apples to kumquats. When we see authors make the mistake of swapping uncertainties, all we can do is tell them to cut it out.

References

1. W.M. Briggs, *Uncertainty: The Soul of Probability, Modeling & Statistics* (Springer, New York, 2016)
2. T. Ando, Biometrika **94**, 443 (2007)
3. E. Arjas, A. Andreev, Lifetime Data Analysis **6**, 187 (2000)
4. J. Berkhof, I. van Mechelen, Computational Statistics **15**, 337 (2000)
5. B.S. Clarke, J.L. Clarke, *Predictive Statistics* (Cambridge University Press, Cambridge, 2018)
6. R.J. Hyndman, The American Statistician **50**, 120 (2012)
7. J.M. Bernardo, A.F.M. Smith, *Bayesian Theory* (Wiley, New York, 2000)
8. W.M. Briggs, arxiv.org/abs/1201.3611 (2013)