

Likelihoods as Goodness-of-Fit Metrics

Likelihoods as Goodness-of-Fit Metrics

Overview

One way to test how well a distribution fits a set of data is to calculate the likelihood of the data given the distribution model. In Bayesian notation this is:

$$\mathcal{L}(\theta|x) = p_{\theta}(x) = p_{\theta}(X = x)$$

θ = Some model (or specific parameterization of a model)

x = Some observed data (one or more observations)

$\mathcal{L}(\theta|x)$ = Likelihood of the model, given observed data, x

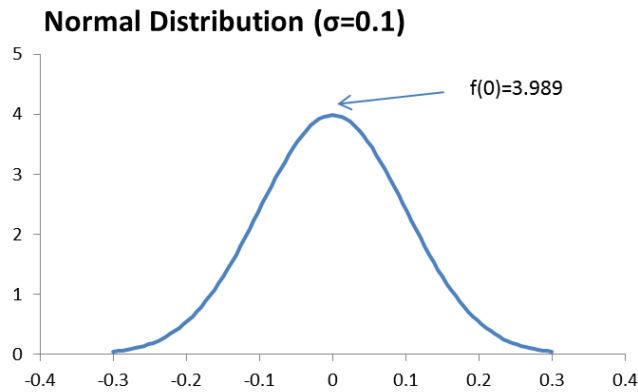
$p_{\theta}(x)$ = Probability of the observations x given the model θ

A simple example is a coin toss. The binary distribution can be measured by a single parameter, θ , indicating the probability of heads. For a fair coin, $\theta = 0.5$. If we observe the results of 4 coin-tosses as HHHT we would calculate the probability of this event as $(0.5)^4 = 0.0625$. An alternate model might assume that the coin is biased and that $\theta = 0.75$. In this case, the probability of the observed outcome is $(0.75)^3(0.25) = 0.1055$. The probability of the observed data has increased, and thus the likelihood that this model is the better one has also increased. In fact, this parameterization is the maximum likelihood estimate (MLE) as no other parameterization will fit the observed data better.

Continuous Distributions

For a continuous probability distribution the approach is identical. We simply replace the discrete probability function $p_{\theta}(x)$ with a probability mass function, $f_{\theta}(x)$. It should be noted that in this case the likelihood function can take on positive values – something that was not possible with the discrete case. This is because it is possible for the probability mass function to have positive values. The true probability of any individual event with a continuous distribution is nearly zero, but the mass function may take values approaching infinity as long as the total area under the curve is 1. The chart below shows the PDF for a normal distribution with standard deviation of 0.1. The maximum value of the PDF occurs at $x = 0$ where $f_{\theta}(x) = 3.989$. If we were to observe multiple values near zero, the likelihood function would produce results that increased towards infinity as we add more and more observations.

Likelihoods as Goodness-of-Fit Metrics



Log-Likelihood

Because the likelihood function is calculated as the product of probabilities (or values from the PDF), it will become increasingly small (or increasingly large) as we add more observations. For this reason, it is often advantageous to take the logarithm of the probability.

In the coin toss example, we only had a few observations, so the total probability was still a meaningful number. If we instead flip a fair coin 1,000 times the probability of any individual outcome becomes infinitesimally small: $(0.5)^{1000} = 9.33 \times 10^{-302}$. The log likelihood instead evaluates to:

$$1000(\ln 0.5) = 1000(-.6931) = -693.15$$

It will still increase in magnitude as we add observations but at a much slower pace. When using the log-likelihood it is still true that we want to maximize this value. Since the value above is negative, a value closer to zero would indicate a more likely outcome and a better fit.

Average Likelihood

Something that I don't see commonly done is to account for differences in observation sizes. In our coin toss example we had evaluated:

$$\begin{aligned}\mathcal{L}(\theta = 0.5|HHHT) &= (0.5)^3(0.5) = 0.0625 \\ \mathcal{L}(\theta = 0.75|HHHT) &= (0.75)^3(0.25) = 0.1055\end{aligned}$$

Suppose we instead sampled 1,000 observations, but the same 4 results repeated indefinitely: HHHT HHHT HHHT (etc.). Now, the likelihood of each parameterization would become much smaller. We can normalize for this by calculating the average probability per observation. This is simply the geometric mean of the individual probabilities. For these examples, it is calculated as:

$$\begin{aligned}\bar{\mathcal{L}}(\theta = 0.5|HHHT) &= 0.0625^{1/4} = 0.5 \\ \bar{\mathcal{L}}(\theta = 0.75|HHHT) &= 0.1055^{1/4} = 0.5699\end{aligned}$$

Likelihoods as Goodness-of-Fit Metrics

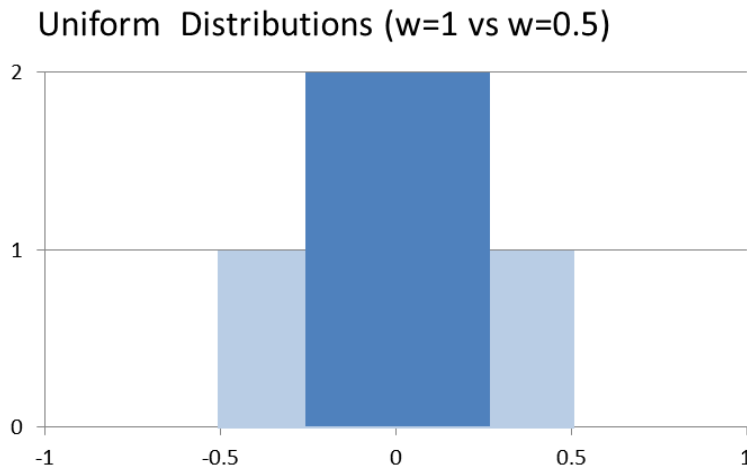
We would get these same values even if we produced 1 million observations with the same distribution of heads and tails results.

If we have log-likelihoods, we can do this same process. We first calculate the log-likelihood of the observation and then normalize it by dividing by the number of observations. This gives us the average log-likelihood for each observation. This calculation is usually preferred since it avoids the intermediate results becoming increasingly small. We can also exponentiate the result if we'd like to get it back to the average probability per observation, but this is optional.

Log-likelihood is frequently used in goodness-of-fit metrics like AIC and BIC without normalizing for sample size. This can cause problems if you are trying to make comparisons of models across different samples. Normalizing these metrics so that they are invariant to observation size is helpful in these cases.

Interpreting Likelihood Goodness-of-Fit Metrics

One difficulty in using the likelihood-based metrics for goodness-of-fit is that they can be difficult to interpret. In the coin toss example one model increased the normalized likelihood from 0.5 to 0.5999. Is this good? Is it a significant improvement? One way to think about this is to imagine a uniform distribution and its average probability level. If the distributions spans the domain -0.5 to 0.5 it will have a width of 1 and an average probability level of 1 (remember: the total area under the curve must be 1). If we cut its domain in half so that it instead ranges from -0.25 to .25, its average probability will double to 2.



In general, we find the following relations holds:

$$w_1 h_1 = w_2 h_2 = 1$$

$$\frac{h_2}{h_1} = \frac{w_1}{w_2}$$

Likelihoods as Goodness-of-Fit Metrics

where h and w are the heights and widths of the two uniform distributions.

Using this model, if we increase the average probability from 0.5 to 0.5999, this is similar to shrinking the domain of a uniform distribution function from a width of 2 down to 1.755, a decrease of 12.3%. The intuitive model feels similar to measuring a forecast or predictive model where the goal is to minimize the residual error (or the residual squared error). If the standard error of a forecast reduced by 12.3%, that would indeed be a significant accomplishment.

Relation to AIC and BIC

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are commonly used in selecting and evaluating models. Both of these are based on the likelihood goodness-of-fit metric. Their innovation is that they introduce a penalty for the number of parameters in a model.

AIC is defined as:

$$AIC = 2k - 2 \ln \hat{L}$$

k = number of parameters

\hat{L} = estimated likelihood

The log-likelihood - and thus the AIC - for a model will change as the number of observations changes. To normalize for this we can use the average log-likelihood. We will also need to normalize the parameter penalty in the same way through:

$$AIC' = \frac{k}{n} - \frac{1}{n} \ln \hat{L}$$

This modification will produce AIC values that are comparable to each other regardless of sample size. The first term in the equation will still change, however. As we add more observations, the penalty for parameters will decrease toward zero. With a large enough sample size, there will be no parameter penalty and the AIC will be identical to the average log-likelihood.

Some impose a correction to the AIC to account for statistical effects of limited observation sizes. The corrected AIC is defined as:

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

The AIC component can be normalized in the manner shown above. The second term can also be normalized by dividing by n . Once again we have a term that will decrease towards zero as we add more and more observations, resulting in a metric that is identical to the log-likelihood.

The BIC is defined as:

Likelihoods as Goodness-of-Fit Metrics

$$\text{BIC} = \ln(n) k - 2 \ln \hat{L}$$

Once again, we can normalize the second term so that it is the average log-likelihood and is invariant with regard to the number of observations. This will result in the following adjusted metric:

$$\text{BIC} = \frac{\ln(n) k}{2n} - \frac{1}{n} \ln \hat{L}$$

As with the AIC, the first term will approach zero as we add more observations. However, for those who want to penalize model parameters in the manner prescribed by the AIC and BIC, the corrections above provide ways to do that while keeping the magnitude of the metric relatively stable with regard to sample size.