

## Putting error bars on a cumulative distribution

DAVID J.C. MACKAY    MARCH 21, 2006

You've got a sample  $\{x_n\}_{n=1}^N$  of points assumed to come from a distribution, and you want to show what you believe the cumulative distribution looks like, given these points. We want not only a guess but also error bars. Perhaps you want to compare two cumulative distributions with each other and quantify whether they seem to be 'significantly different from each other'.

Here's a likelihood-based approach to put error bars on the graph.

Let the true unknown cumulative distribution at coordinate  $x$  be  $F_x$ . This quantity defines the probability that a random draw  $x_n$  from the distribution will fall in the range  $x_n \leq x$ . We can think about the inference of  $F_x$  given the  $N$  outcomes, each of which was either above or below  $x$ .

The probability that  $A$  outcomes fall above  $x$  and  $B$  fall below it, given  $F_x$ , is

$$P(A, B | F_x) \propto (F_x)^B (1 - F_x)^A \quad (1)$$

If  $A$  and  $B$  are both  $\geq 1$  then this likelihood function is a unimodal function that is well approximated as Gaussian in the logit basis [1]

$$a_x \equiv \ln \frac{F_x}{1 - F_x}. \quad (2)$$

$$F_x \equiv f(a_x) \equiv \frac{1}{1 + \exp(-a_x)}. \quad (3)$$

The Gaussian approximation (found by Taylor-expanding the likelihood (1)) is

$$P(A, B | F_x) \propto (F_x)^B (1 - F_x)^A \simeq \exp\left(-\frac{1}{2} \frac{(a_x - \hat{a}_x)^2}{\sigma_a^2}\right) \quad (4)$$

where  $\hat{F}_x = \frac{B}{A+B}$  (i.e.,  $\hat{a}_x = \ln B/A$ ) and

$$\sigma_a^2 = \frac{A+B}{AB} \quad (5)$$

I recommend indicating error bars on the inferred cumulative distribution using this approximation. For example, if there are  $A$  above and  $B$  below a point  $x$ , then a point estimate for  $F_x$  is

$$F_x = \frac{B}{A+B} \quad (6)$$

and upper and lower  $z$ -sigma error bars are at

$$f(\ln B/A \pm z\sigma_a). \quad (7)$$

In the case where  $A = 0$  or  $B = 0$ , the likelihood has no peak. We could nevertheless get an "error bar" on one side, by estimating the  $F_x$  at which the likelihood function falls by a factor of  $e^{1/2}$ . This little computation is left as an exercise for the reader.

## References

- [1] D. J. C. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33(1):77–86, 1998.