Modelling Mondays

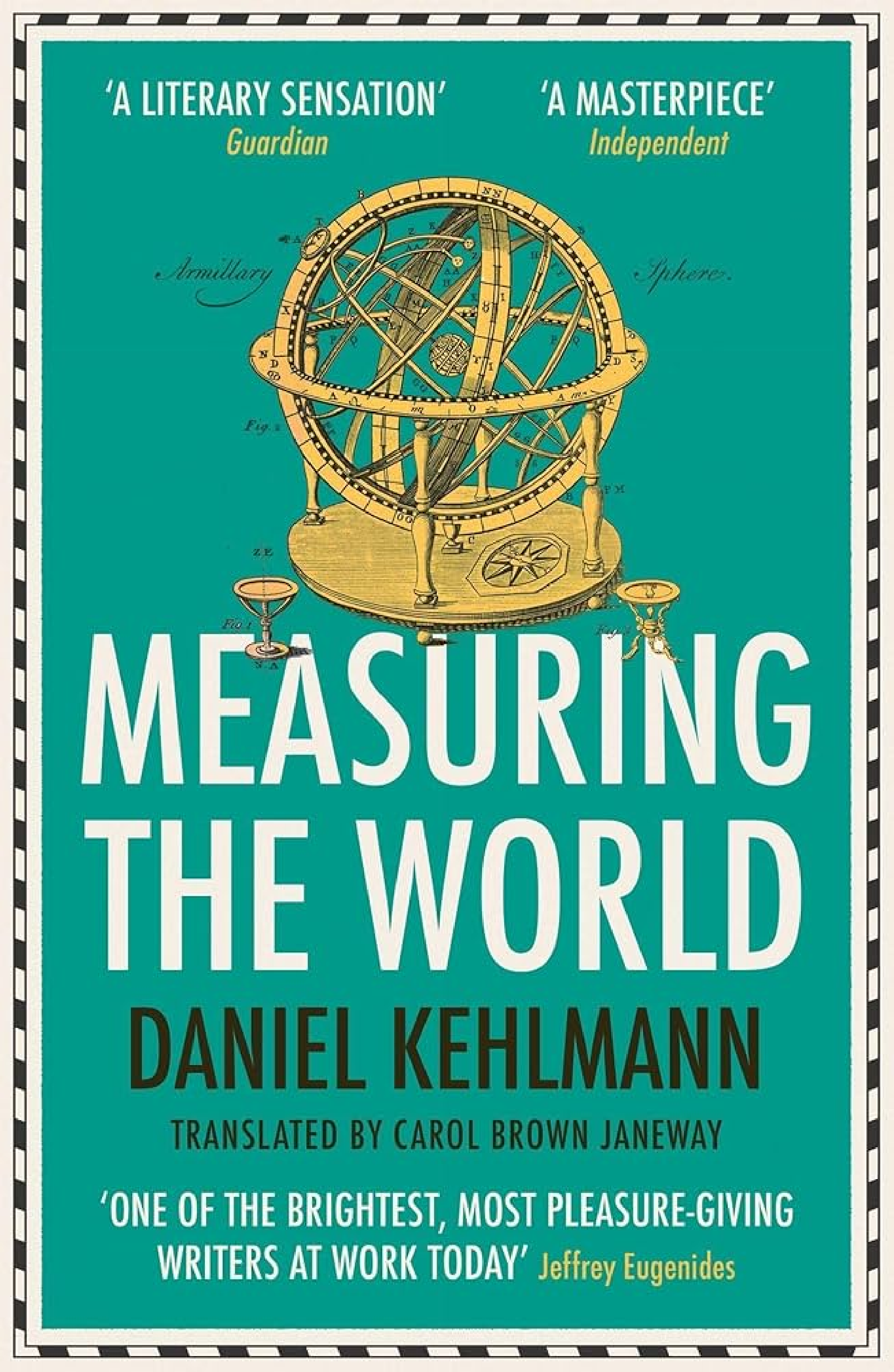
Argyris Stringaris

2024-05-06

Motivation

What is Data Generating Process (DGP) and what is a likelihood function?

Typically, we think of a DGP as a mathematical formula that gives rise to a distribution. For example, the IQ curve can be generated through the Gaussian, named after Karl Friedrich Gauß–very much worth reading about also in the novel The Measuring of the World, by Kehlmann (where the parallel lives of Gauß and Humboldt are presented).



'A LITERARY SENSATION' *Guardian*

'A MASTERPIECE' *Independent*

Armillary Sphere.

MEASURING THE WORLD

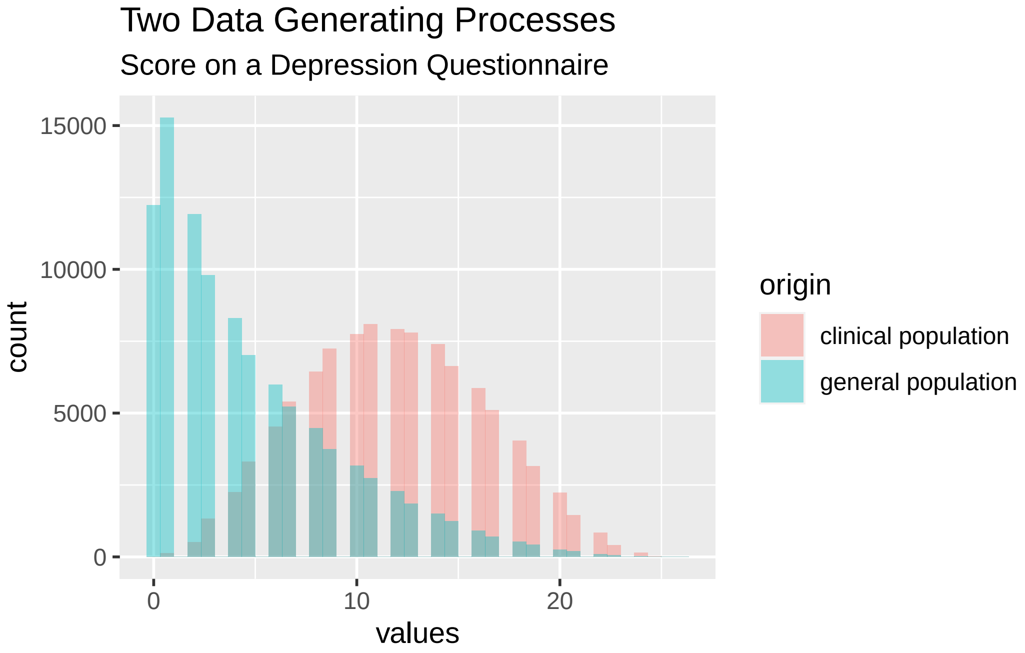
DANIEL KEHLMANN

TRANSLATED BY CAROL BROWN JANEWAY

'ONE OF THE BRIGHTEST, MOST PLEASURE-GIVING WRITERS AT WORK TODAY' Jeffrey Eugenides

But in a more abstract way, the question is, what are the mechanisms through which a set of data are generated, be it voting patterns, brain data or league games.

Consider, for example, a sample of the general population filling in a questionnaire about depression. Figure 1a. shows a typical pattern, that of a right skewed truncated distribution. The "mechanism" that gives rise to the right skew is the fact that there are far more people without many symptoms and hence many people close to the zero mark. It is also truncated because scores can't go below zero and can't go above the max of the sum of the scale. By contrast, Figure 1b, shows the



In general, we always want to consider the DGS so as to:

a) understand what gives rise to the data.

b) mathematically describe (at least) how the data arise.

c) estimate parameters (related to b)

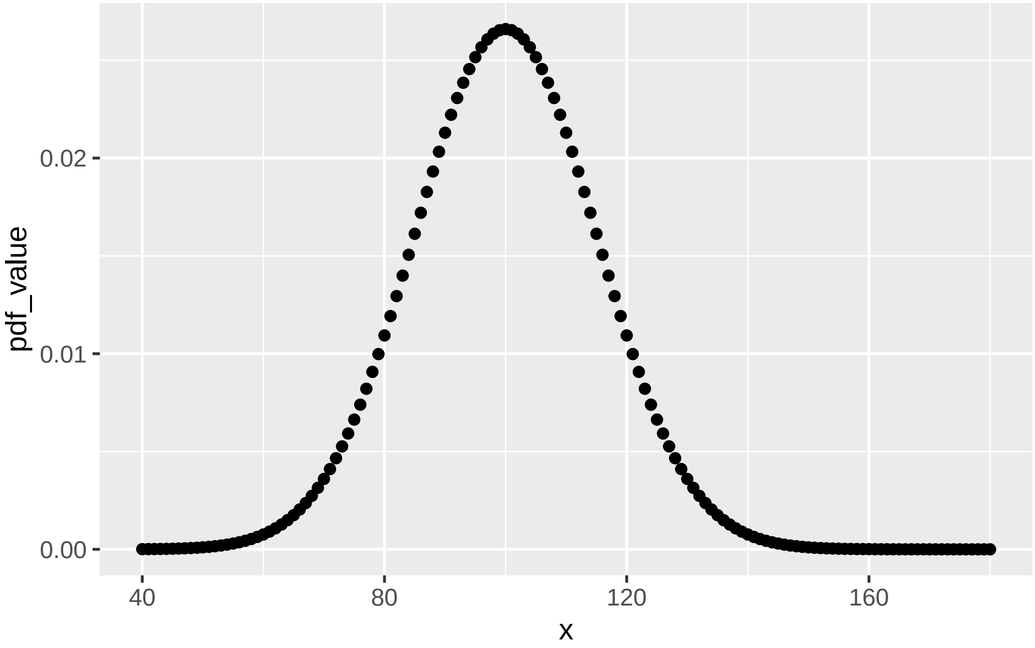
d) simulate the process to study it better.

The omniscient person: knowing the DGS and the correct parameter.

This is someone who knows the function and its probability, is certain about the DGS. Let's say that theyknow that they are dealing with the normal distribution, which is formalised as:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x-\theta)^2}{2\sigma^2}} \tag{1}$$

where, x is the point of interest of the probability density function, θ is the mean (location parameter) of the normal distribution, and σ is the standard deviation (spread parameter).



You will all recognise this as the standard IQ curve.

Please note from Equation 1 that here the point is that the situation is phrased as:

$$f(x|\theta, \sigma)$$

i.e. we ask what the probability is of obtaining these data given the parameters θ.

The situation where you are certain about the correct parameter and only need to know the frequency of individual values or set of values is a very convenient one to be in. Often however, in the real world we may have an intuition about what the DGP might be but not know the parameter(s). That is when we ask about the likelihood.

A real person: having data, intuiting the DGS, and not knowing the parameter.

Consider having collected some data, having some intuition about the DGS and needing to find out the parameter amongst a set of parameters.

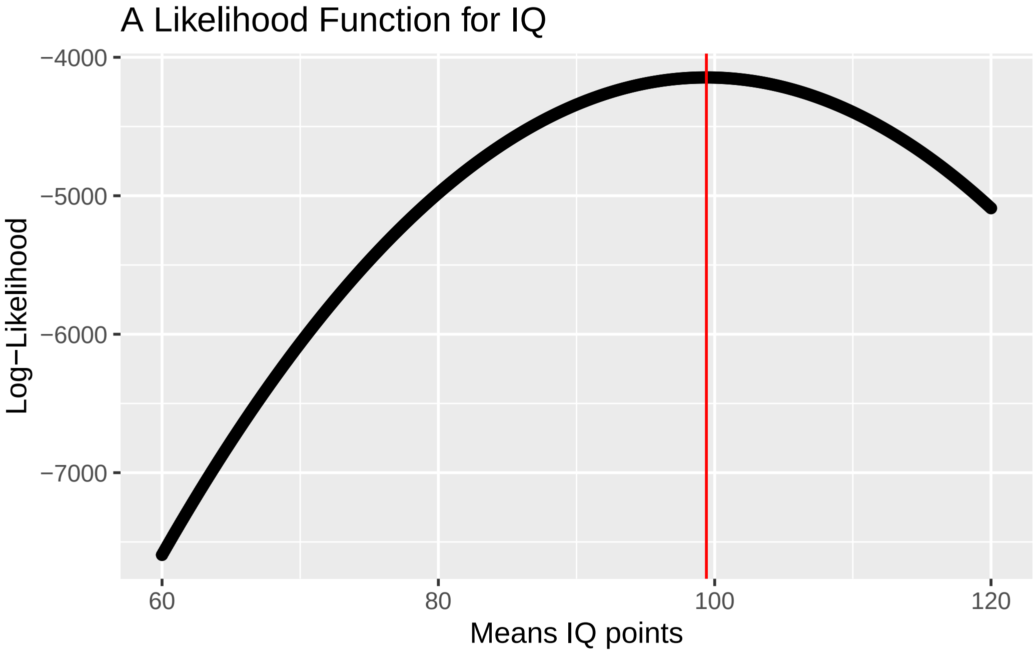
This is a more likely situation which I will illustrate here by trying to recover the mean parameter from synthetic data.

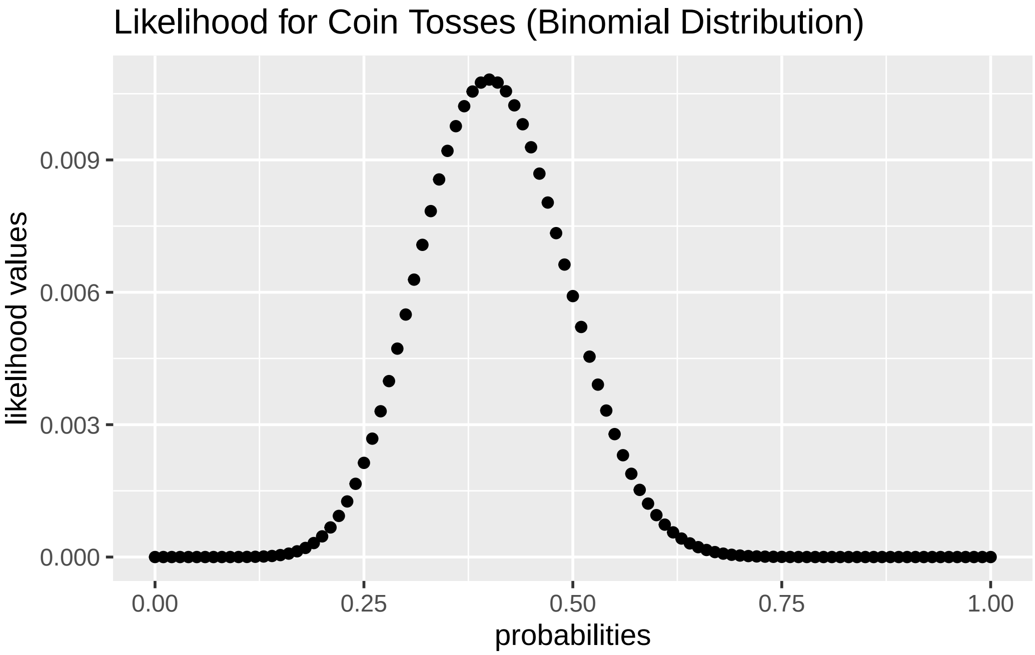
For the moment it is safe to say that what were trying to do here is to invert the process above, i.e. what you do with the probability density function. Instead of asking what data are likely to occur given a parameter (such as the mean and sd) that you *already* know about, here you ask, what is the most likely parameter that has given rise to the data I have.

$$L(\mu, \sigma^2 | x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{2}$$

Equation 2 states precisely that: what is the likelihood of this mean and variance, given all these data points? Equation 2 on the right hand side contains the PDF, as above, but what it says is that it takes the probability at each step and multiplies them altogether, this is what that giant Greek Π stands for, the product.

Notice that when I tried this with fewer data points, I was able to get the likelihood, but when I increased them, I needed the natural log. Try it for yourself.





Now let's turn to the simple linear regression model. Let's start by asking how to think formally of the data generating mechanism of any linear model. It should be a

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3}$$

where, ϵ_i follows a normal distribution with mean zero and variance σ^2

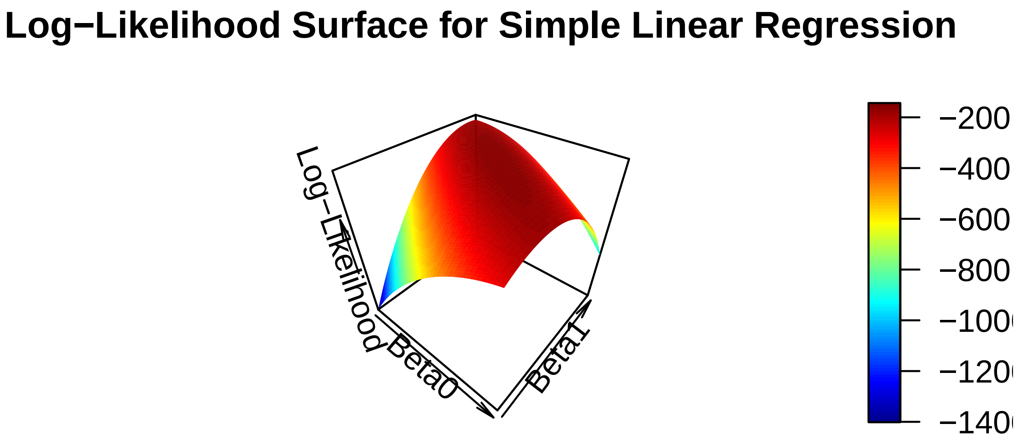
$$L(\beta_0, \beta_1 | x_1, y_1, x_2, y_2, \dots, x_n, y_n) = \prod_{i=1}^{n} f(y_i | \beta_0 + \beta_1 x_i) \tag{4}$$

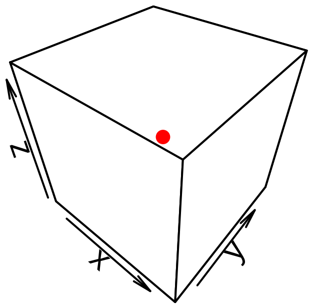
where

$$f(y_i | \beta_0 + \beta_1 x_i)$$

is the probability density function (PDF) of the normal distribution with mean $(\mu_i = \beta_0 + \beta_1 x_i)$ and constant variance \$ \sigma^2 \$

To demonstrate this, I will first create synthetic data





In the code chunk below, I explain how the outer product and vectorisation works.