McElreath Chapter 10: Big Entropy and the Generalized Linear Model

Mike Wolfe

4/25/2021

Distributional choices in Bayesian statistical models

- Prior - for each parameter what information do we have about it and how do we describe it's uncertainty?
- Likelihood - how is the data distributed? What is the data generating process?
- So far we have been choosing Gaussian distributions for both the likelihood and priors through convential linear regression.
- But we don't have to be limited to these conventional choices
- How do we choose?
  - Part 1 - Maximum Entropy as a guiding principle
  - Part 2 - Putting together models with different distributions

Looking briefly ahead

- This chapter serves as a foundation for upcoming chapters
  - Chapter 11 - models for count variables (**SEQUENCING DATA**)
  - Chapter 12 - complicated models such as ordinal outcomes and mixtures (**Also SEQUENCING DATA**)
  - Chapter 13 - moving into multilevel models (**Where the real fun starts**)

Information entropy

- Talked about this in chapter 6 but want a measure that satisfies 3 criteria:
  - Measure should be continuous
  - It should increase if there are more possible events
  - It should be additive
- Information entropy $H(p)$ satisfies these criteria for a distribution p where p_i is the probability of event i.

$$H(p) = -\sum_i p_i \log p_i$$

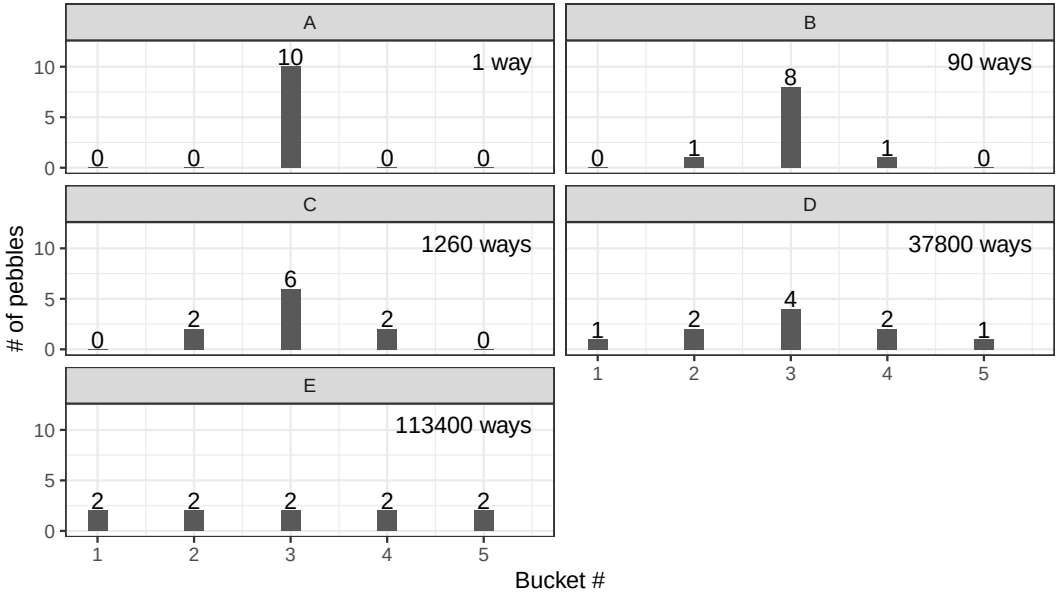
Maximum entropy

- Use this measure to choose a probability distriubtion

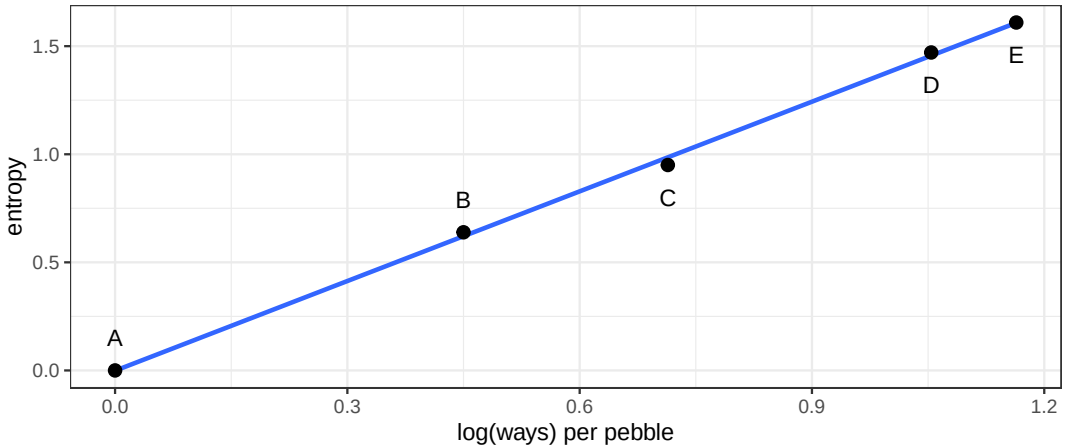
  The distribution that can happen the most ways is also the distribution with the biggest information entropy. The distribution with the biggest entropy is the most conservative distribution that obeys its constraints.

- Consider 10 pebbles randomly thrown into 5 buckets with an equal probability of a pebble landing in each bucket.

Maximum Entropy - Pebbles in buckets (Uniform Distribution)



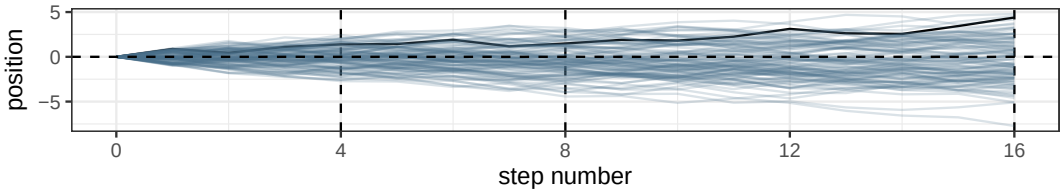
Which scenario has maximum entropy?

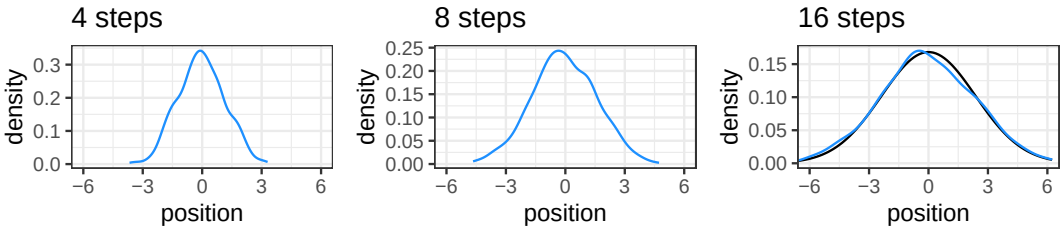


[The maximum entropy distribution's] high plausibility is conditional of our assumptions, of course.

Maximum Entropy - the Gaussian distribution

- Previously, we talked about getting a Gaussian when you add up a bunch of small fluctuations
- 100 people taking 16 steps by flipping a coin at each step





Maximum Entropy - Gaussian

- Why does the distribution show up so much?
- If all we know about some continuous values is that they have finite variance, the most likely distribution is a bell shaped one.
- The Gaussian distribution is not the only bell-shaped distribution. We can consider several different bell-shaped functions at the same known variance.

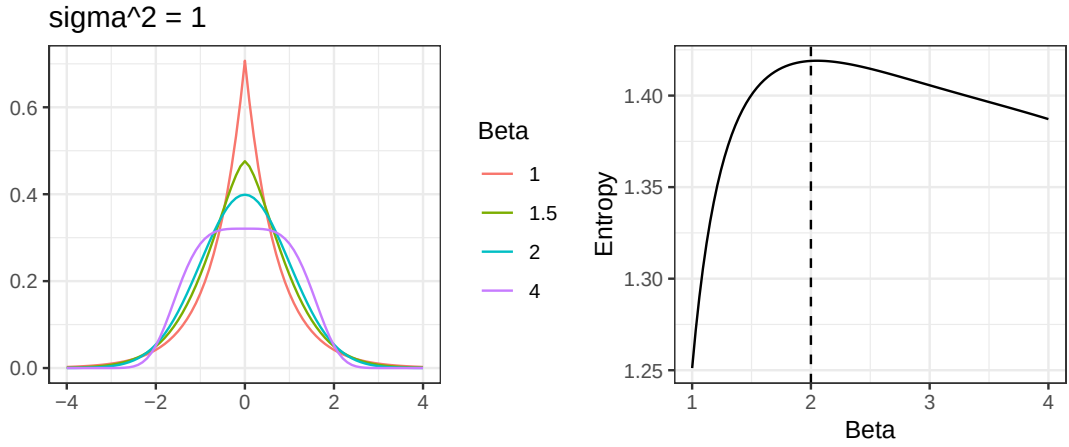
$$\Pr(y|\mu,\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} e^{-(\frac{|y-\mu|}{\mu})^{\beta}}$$

- We will consider different distributions when $\sigma^2 = 1$.
- Need to determine the relationship between σ and α, β to

$$\sigma^2 = \frac{\alpha^2\Gamma(3/\beta)}{\Gamma(1/\beta)}$$

$$\alpha = \sqrt{\frac{\sigma^2\Gamma(1/\beta)}{\Gamma(3/\beta)}}$$

Maximum Entropy - Which bell-shaped curve?



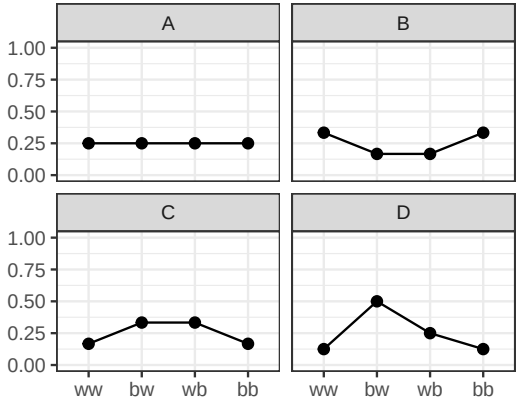
. . . the Gaussian distribution gets its shape by being as spread out as possible for a distribution with fixed variance.

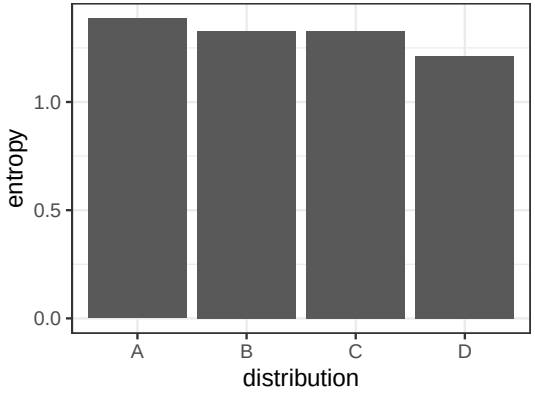
Maximum entropy- binomial distribution

- What if we have simple count data?
- Binomial distribution
  - Only two things can happen
  - There is a constant chance p across n trials of seeing event y
- This is the maximum entropy distribution for the following conditions:
  - Two unordered events
  - A constant expected value
- Consider the scenario
  - Bag with blue and white marbles of unknown quantity
  - Drawing two marbles
  - We know expected number of blue marbles of two draws is exactly 1

Maximum entropy - binomial marbles simple

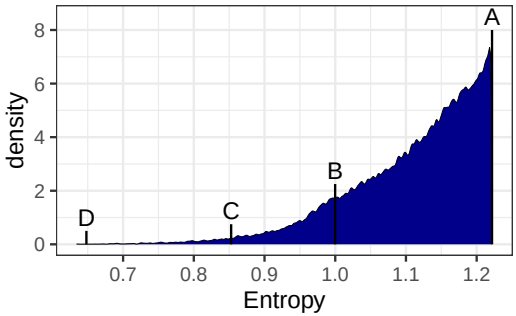
- We want to find the distribution with the highest entropy over four possible outcomes: ww, bw, wb, bb.
- Consider the following four distributions with the same expected value of 1 blue marble in two draws

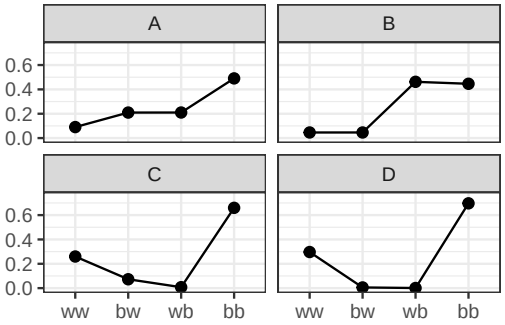




Maximum Entropy - binomial marbles complex

- What if instead we made the constraint that the expected value be 1.4 blue marbles in 2 draws? (i.e. p = 0.7 or 7 blue marbles and 3 white marbles)





If only two un-ordered outcomes are possible and you think the process generating them is invariant in time, then the distribution that is the most conservative is the binomial.

Generalized Linear Models

- So far we have been focusing on linear regression using Gaussian distributions
- Linear regression will make ridiculous predictions on bounded variables
- What if we want to use some of these other distributions we have been talking about?

Generalized Linear Models

- The key to GLMs is that you have models like this:

$$y_i \sim \text{Binomial}(n, p_i)$$
$$f(p_i) = \alpha + \beta(x_i - \bar{x})$$

* Two changes to note: * Likelihood is now a binomial instead of Gaussian * Now contains a link function, $f(p_i)$ in this example. * Need a link function since the parameter p is not unbounded in both directions

Exponential family of likelihood functions

- Every member is a maximum likelihood function for some constraints
- We have already talked about
  - Gaussian distribution (linear regression)
  - Binomial distribution (logisitic regression - next chapter)
- However, we have some new distributions:
  - **Exponential** - Used for distance and duration. If prob of an event is constant in time or space. Maximum entropy among all non-negative distributions with the same average displacement
  - **Gamma** - Similar to exponential but can have peak above zero. Waiting times between to exponentially distributed events. Maximum entropy among all distributions with the same mean and same average logarithm
  - **Poisson** - Binomial when n is large and p is small. Maximum entropy under the same constraints as binomial

Linking linear models and distributions using the link function

- Link functions allow you to convert from linear space to the non-linear space of a parameter
- Example, the binomial distribution can use a logit function

$$y_i \sim \text{Binomial}(n, p_i)$$
$$\text{logit}(p_i) = \alpha + \beta x_i$$

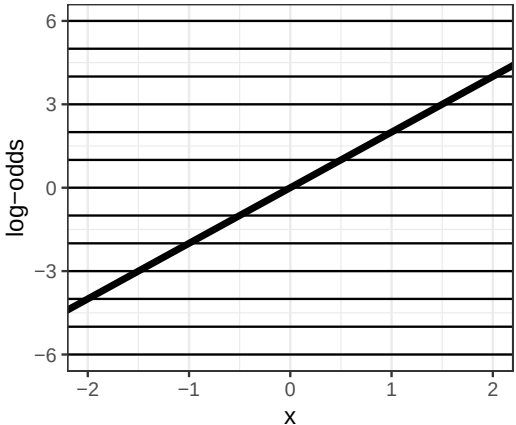
- Where logit function defines the log-odds

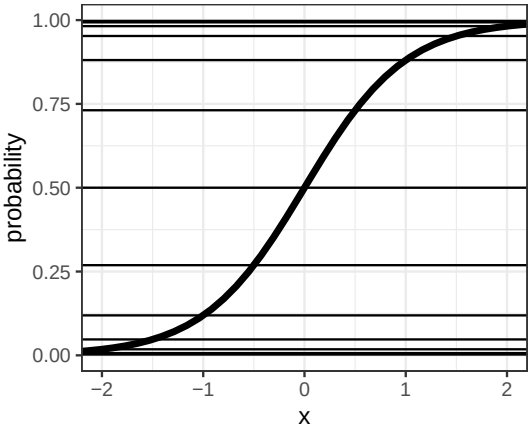
$$\text{logit}(p_i) = \log \frac{p_i}{1 - p_i}$$

- And p_i is therefore described with the logistic function or inverse-logit

$$p_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}$$

Logit link function





The key lesson now is just that no regression coefficient, such as beta from a GLM ever produces a constant change on the outcome scale.

Log link function

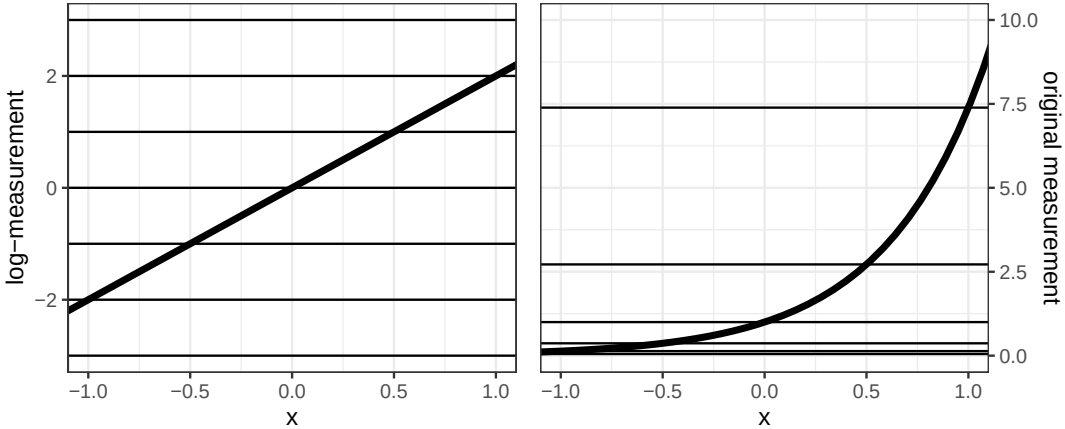
- An example, parameter that only maps onto positive reals:

$$y_i \sim \text{Normal}(\mu, \sigma_i)$$
$$\log(\sigma_i) = \alpha + \beta x_i$$

- The inverse link is the following

$$\sigma_i = \exp(\alpha + \beta x_i)$$

Log link function



Using a log link implies an exponential scaling of the outcome with the predictor variable

GLMs odds and ends

- Omitted variable bias can be a huge issue with GLMs due to the relative effects of the estimates
- Big coefficients don't mean big changes on the outcome scale. Coefficients are giving you relative changes
- Information criteria can only be used on models with the same likelihood function. Don't compare a logistic regression to a linear regression using information criteria
- You can use maximum entropy principles to choose priors as well

Summary and looking ahead

- Maximum entropy principles can help you choose distributions.
- Most commonly used distributions are maximimum entropy for some known constraint
- GLMs let you used distributions other than Gaussian distributions for your likelihood function
- GLMs require a link function that results in the outcome not scaling linearly with the predictors
- Caution must be used when interpreting coefficients as they live on a relative and not absolute scale
- Next McElreath Chapter will focus on models for count data led by Jeremy
- Next Neidhardt Chapter will be Chapter 14 led by Yulduz.