Introduction

ISTA 410 / INFO 510: Bayesian Modeling and Inference

U. of Arizona School of Information

August 24, 2020

A first exercise

An example

The following example comes from David Mackay's book:

> *Unstable particles are emitted from a source and decay at a distance* x*, which is distributed according to an exponential distribution with characteristic length* λ*. Decay events can be observed only if they occur inside a window* $1 \leq x \leq 20$*.* N *events are observed at locations* $x_1, \ldots, x_N$*. What is* λ*?*

An example

How to make an estimate?

If our observations were unconstrained (we could observe the decay event at any $x > 0$), what would our estimate be?

How to make an estimate?

If our observations were unconstrained (we could observe the decay event at any $x > 0$), what would our estimate be?

Just take the sample mean:

$$\hat{\lambda} = \bar{x}$$

How to make an estimate?

If our observations were unconstrained (we could observe the decay event at any $x > 0$), what would our estimate be?

Just take the sample mean:

$$\hat{\lambda} = \bar{x}$$

Under what circumstances is this still a reasonable estimate, given our constraints? When is it not?

No one-size-fits-all estimator

The lesson of the previous slide is that there is no single estimator that will work regardless of λ.

Instead of trying to find a function of the data that directly estimates λ, let's apply Bayes' theorem:

$$p(\lambda|x) = \frac{p(x|\lambda)p(\lambda)}{p(x)}$$

The core idea

What does this mean?

- We model the unknown parameter λ as if it were a random variable – in other words, we assign it a probability distribution $p(\lambda)$
- We apply Bayes' theorem to update this probability distribution to be conditional on the data, $p(\lambda|x)$
- In the end, this distribution is our "estimate" – it contains all that we know about the parameter

The ingredients

For a single x:

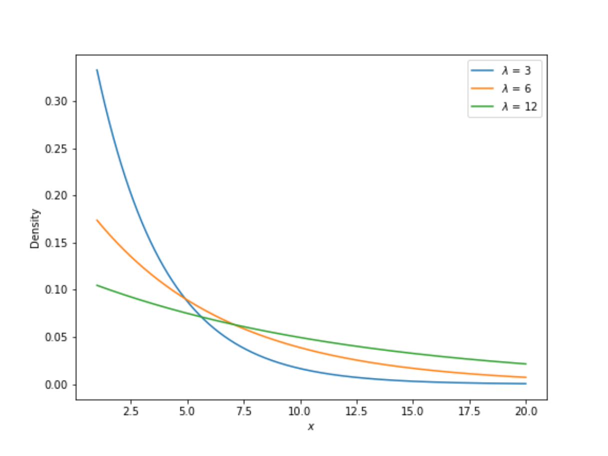
$$p(x|\lambda) = \frac{1}{Z(\lambda)} \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & 1 < x < 20 \\ 0 & \text{otherwise} \end{cases}$$

$$Z(\lambda) = \frac{1}{\lambda} \int_1^{20} e^{-x/\lambda} dx$$

Let's examine this a bit by graphing it.

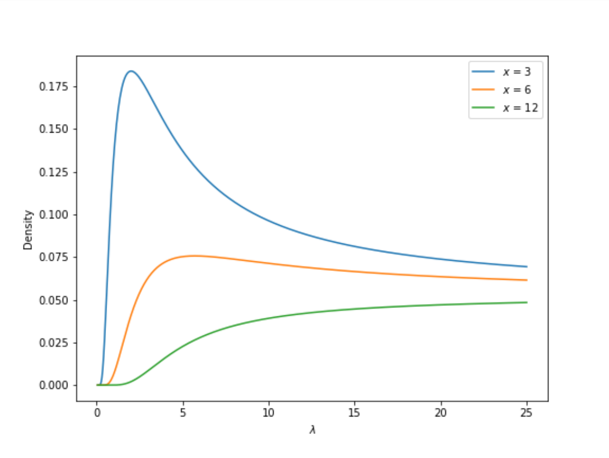
The ingredients

If we fix λ and plot $p(x|\lambda)$, we get a simple exponential curve, representing the probability of observing a decay event at various values of x:

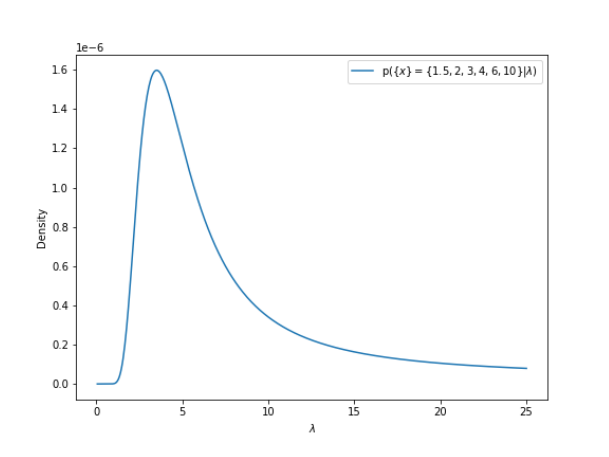


Likelihood function

If we fix x and think of $p(x|\lambda)$ as a function of λ, we get the *likelihood function*



Likelihood function



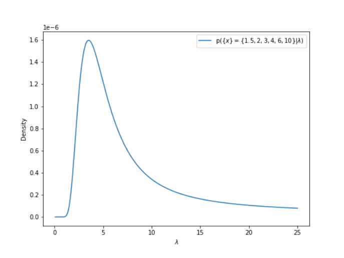
The whole picture...

It's not really Bayesian until we have a prior:

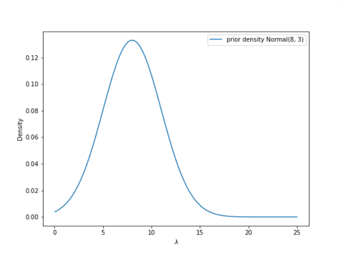
$$p(\lambda|\{x\}) = \frac{p(\{x\}|\lambda)p(\lambda)}{p(\{x\})}$$
$$\propto p(\{x\}|\lambda)p(\lambda)$$

This density function encodes what we know about λ after observing $\{x\}$.

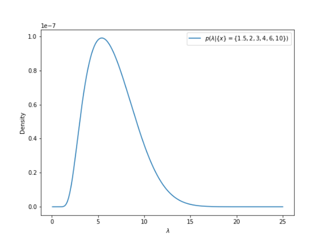
Imagine our prior knowledge is that λ should be somewhere near 8...



Likelihood



Prior



Posterior

Bayes' theorem in summary

In words, the significance of this is (attributed to Steve Gull by David Mackay):

> *what you know about λ after the data arrive is what you knew before, $p(\lambda)$, and what the data told you, $p(\{x\}|\lambda)$*

The prior

The factor $p(\lambda)$, the *prior density*, represents what we knew about λ before observing any x, and we can't compute the posterior $p(\lambda|x)$ without stating it.

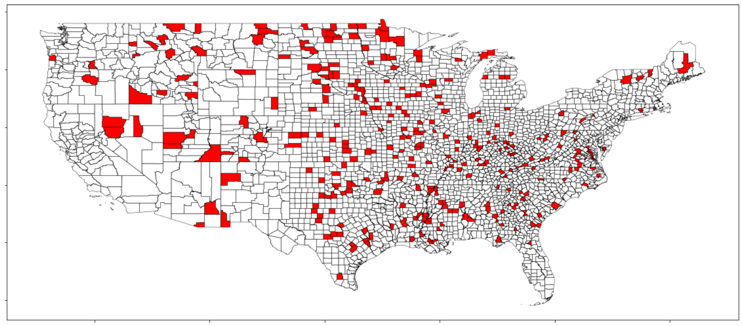
This is the same situation as we had before, in a sense – we couldn't pick an estimate without making an assumption about the likely value of λ.

This is unavoidable – can't do inference without any assumptions!

Case study: kidney cancers

Where is kidney cancer highest?

The following map shows the counties with the highest 10% death rates due to kidney cancer (1980-89).



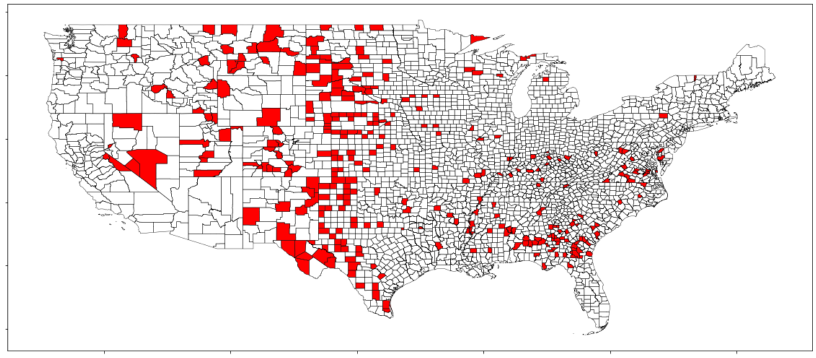
What do we notice?

What we can notice: most counties in the middle of the country, not coasts.

Why?

Where is kidney cancer lowest?

The following map shows the counties with the *lowest* 10% death rates due to kidney cancer (1980-89).



A simple model

To understand this, let's take a detour into simulation.

We'll model the number of cancer deaths per year as a Poisson random variable with parameter $\lambda = N\theta$:

$$y_i \text{ Poisson}(\mathrm{n_i}\theta) P(y_i = k|\theta) = \frac{(n_i\theta)^k e^{-n_i\theta}}{n!}$$

This encodes the assumption that all counties have the same underlying cancer rate θ (measured in deaths per capita per 5 years).

Let's go simulate!

A simple model

Key feature of Bayesian models: always *generative*

A simple model

Key feature of Bayesian models: always *generative*

- Given parameter θ, can generate data n_j
- Given data n_j, can make inferences about θ

Models encode our understanding of how the data arises

A simple model

Is our simple model good enough?

A simple model

Is our simple model good enough?

No; data generated by the model doesn't "resemble" the real world data very well (we'll formalize this more later on).

A simple model

Is our simple model good enough?

No; data generated by the model doesn't "resemble" the real world data very well (we'll formalize this more later on).

Less empirically, more theoretically: if our goal is to learn about geographic variation in cancer rates, we should have a model that allows for that variation!

A slightly more complex model

For our next pass, let's allow θ to vary – i.e., we take each county to have its own θ_j Then we set a prior distribution on θ_j:

$$\theta_j \ \mathrm{Gamma}(20, 430, 000)$$

Then the posterior distribution is

$$\theta_j | y_j \ \mathrm{Gamma}(20 + y_j, 430, 000 + n_j)$$