Idea of Statistical Modeling

ISTA 410 / INFO 510 - Bayesian Modeling and Inference

University of Arizona School of Information

January 20, 2021

Outline

Outline for today:

- Goals of Bayesian analysis
- Example: kidney cancer death rates
- Understanding types of uncertainty

Motivation: Bayesian analysis

Goal: analyze and quantify uncertainty

- uncertain quantities get a probability distribution
- probability distribution is updated based on new observations

Bayesian approach:

- Named for Thomas Bayes – English minister in the 18th century
- Considered the problem of *inverse probability*
- Didn't invent the whole theory, but was one of the earliest to solve a problem with it (along with Laplace)

Generative probabilistic models

Core component: generative models

"forward" procedure

- given values of model parameters, can generate outcomes
  - given a value for the probability a coin comes up heads, we can simulate a sequence of flips
  - given a mean and standard deviation, we can simulate normally distributed values

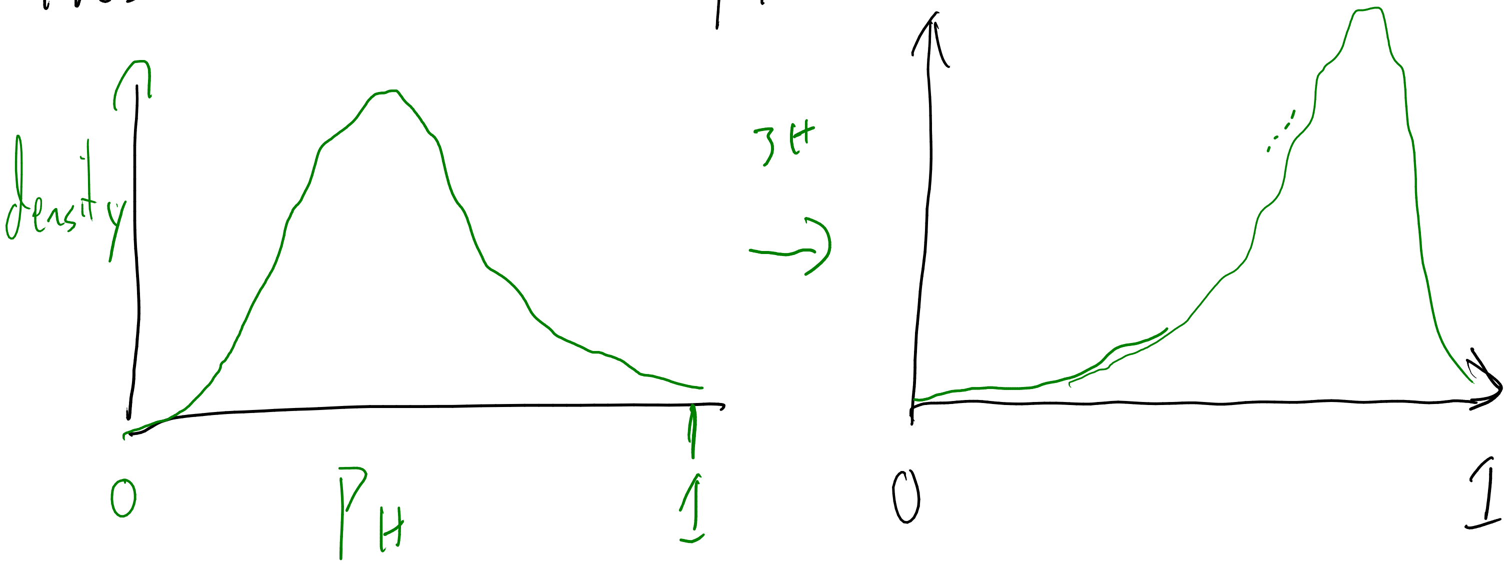
The inverse problem is: given the outcomes, infer a probability distribution for the parameters

- Both Bayes's and Laplace's early work deal with a binomial model (like the coin flip)

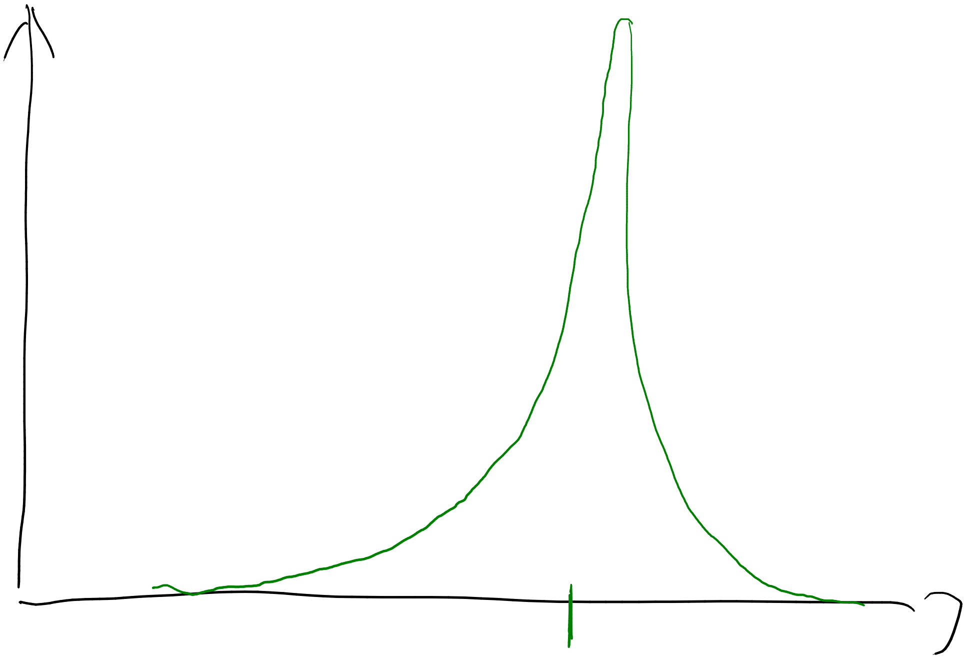
Bayes' Theorem.

Previously: we considered a coin that comes up heads with probability p_H

Prob. distribution for p_H:



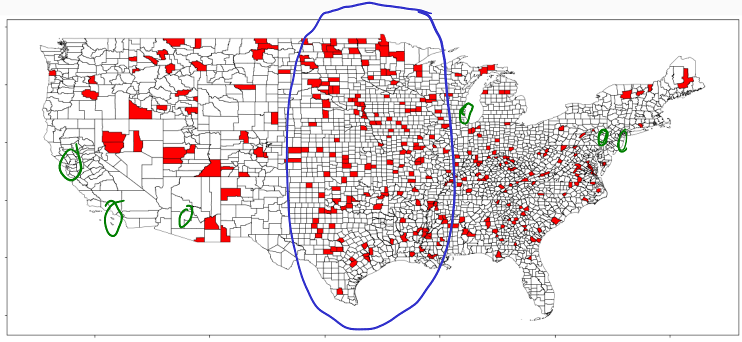
More observations:



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?

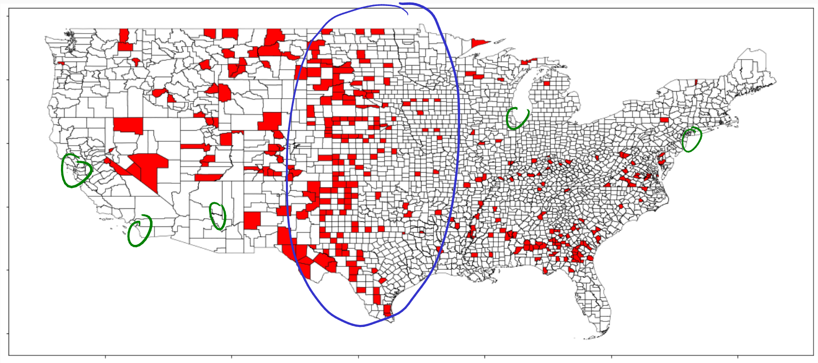
What we can notice: most counties in the middle of the country, not coasts, not around the largest cities.

Why?

- Differences in lifestyle, food, water access
- Medical resources/facilities available
- Population average age

Where is kidney cancer lowest?

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



- lower pollution -> lower cancer rates
- more time spent outdoors -> healthier lifestyle

A simple model

To understand this, let's build and run a generative model.

- This model should be able to generate simulated versions of our observed outcomes
- What are our outcomes? Deaths due to kidney cancer.
- We need to pick a probability distribution

Poisson distribution

- Defined on natural numbers $\{0, 1, 2, \ldots\}$
- Models a count of $\underbrace{\text{events}}_{\text{deaths due to kidney cancer}}$ occurring independently at a fixed rate
- Depends on a rate parameter most often written λ

Probability mass function:

$$P(X = k) = \underbrace{\frac{\lambda^k e^{-\lambda}}{k!}}_{\substack{\text{prob. of } k \text{ deaths} \\ \text{(or other event)}}}$$

Our model

We'll make the simplest possible assumption: there is no effect of geography on kidney cancer, and the only difference between counties is the population.

So, our model has one parameter, θ (the underlying death rate).

Then, the death count in each county, y_j, follows a Poisson distribution:

$$y_j \sim \text{Poisson}(\theta n_j)$$

shorthand for "y_j has Poisson dist with parameter θn_j"

where n_j is the population of the county. (In Poisson models this scaling factor is sometimes called an *exposure*)

Let's try simulating...

A simple model

Is our simple model good enough?

- Philosophically, if we are looking for differences between counties, we should allow our model to infer those

A simple model

Is our simple model good enough?

- Philosophically, if we are looking for differences between counties, we should allow our model to infer those
- Doesn't seem to have as much variability in outcomes as the real data

One explanation for the lack of variability: we ignored a form of uncertainty

Sources of uncertainty

There are two sources of uncertainty in our observations:

- aleatoric uncertainty (chance uncertainty)
  - due to random outcomes
  - identical counties may produce different numbers of deaths in the same time period
- epistemic uncertainty (knowledge uncertainty)
  - due to our lack of knowledge
  - we have uncertainty in the underlying parameter θ

Our model failed to include any epistemic uncertainty, because the estimate of θ was a constant

Adding some epistemic uncertainty

Let's allow the θ to vary between counties, and let θ also be a random variable:

$$\theta_j \sim \mathrm{Gamma}(20, 430000)$$
$$y_j \sim \mathrm{Poisson}(\theta_j n_j)$$

j - county index.

* Gamma chosen partly for convenience in inference.

* 20, 430,000 chosen by moment-matching details in § 2.7 of textbook.

Summary

Summary

What we did:

- Build and tinker with a generative model to understand how some features of a data set arise
- Identify two types of uncertainty in the data-generating process

What we didn't do:

- Any inference – we didn't make a careful attempt to infer the underlying model parameter(s) from the data

Next week

Next week:

- Further review of probability theory
- HW posted tonight – look at it over the weekend and send me questions so I know what we most need to review!