STA 602 – Intro to Bayesian Statistics

Lecture 2

Li Ma

Probability modeling and statistical inference

- *Probability models* are assumptions (or hypotheses) that characterize the randomness that arises in the data
- *Statistical inference* goes the other direction: it incorporates the data to make *educated "guesses"* about the underlying random mechanism. Two common goals:
  - Estimating or predicting the value of some interesting quantities.
  - Verifying the assumptions and choosing among different possible hypotheses/models using data.

The *ideal* inference procedure

Let's consider the following simple situation.

- Suppose we have a *comprehensive* list of *mutually exclusive* events, which can be considered as "causes" or "states of nature"

$$E_1, E_2, \ldots, E_n.$$

  That is, $\Omega = \cup_{i=1}^n E_i$ and $E_i \cap E_j = \emptyset$ for all $i \neq j$.
- E.g., Zipcode recognition. A person handwrites a digit in $\{0,1,2,\ldots,9\}$.
- Suppose we *know* the *a priori* probabilities of these causes,

$$\mathrm{P}(E_1), \mathrm{P}(E_2), \ldots, \mathrm{P}(E_n).$$

- An expriment is performed, and we observe the outcome or an event, F. (E.g., a handwritten digit is observed.)
- Inference: What are the *a posteriori* probabilities $\mathrm{P}(E_i|F)$, that is the probability of the different causes *given* the outcome F?

- This inference is *ideal* in the sense that both our *prior* and *posterior* understanding about the underlying random mechanism is expressed *probabilistically*. Very much like how our brain works.
- So how do we go from our prior knowledge to posterior knowledge?
- How do we incorporate the information/evidence from the data that supports the different scenarios?
- Through weighing the probability of the causes $\mathrm{P}(E_1), \mathrm{P}(E_2), \ldots, \mathrm{P}(E_n)$, with the probability of the observation F under each cause,

$$\mathrm{P}(F|E_1), \mathrm{P}(F|E_2), \ldots, \mathrm{P}(F|E_n).$$

Example: COVID-19 antibody testing

A patient is given a test for detecting antibodies for the coronavirus in blood.

- Two causes:

$$E_1 = \{\text{The patient has the antibodies}\}$$
$$E_2 = \{\text{The patient doesn't have the antibodies}\}.$$

- Let $\mathrm{P}(E_1) = 1 - \mathrm{P}(E_2)$ be the prevalence of cancer in the *corresponding* population.
- The observed event is

$$F = \{\text{result of the test is positive}\}.$$

- From laboratory studies we know that

$$\mathrm{P}(F|E_1) = .9, \quad \mathrm{P}(F|E_2) = .05.$$

- Inference question: In light of F, what is the chance of actually having the antibodies, i.e. $\mathrm{P}(E_1|F)$?

Bayes' theorem

In this "ideal" situation the following theorem provides a simple recipe for inference.

Theorem (Bayes')

If $E_1, E_2, \ldots, E_n$ are comprehensive and mutually exclusive,

- *Comprehensiveness: The outcome space $\Omega = E_1 \cup E_2 \cup \ldots \cup E_n$.*
- *Mutual exclusiveness: $E_i \cap E_j = \emptyset$ for all $i \neq j$.*

then for each E_i and any event F with $\mathrm{P}(F) > 0$,

$$\mathrm{P}(E_i|F) = \frac{\mathrm{P}(E_i)\mathrm{P}(F|E_i)}{\sum_{j=1}^{n} \mathrm{P}(E_j)\mathrm{P}(F|E_j)}.$$

Proof of Bayes Theorem

This theorem is a direct consequence of the *multiplication rule*. For any two events E and F, we have

$$\mathrm{P}(E \cap F) = \mathrm{P}(F)\mathrm{P}(E|F).$$

Applying this twice we get

$$\mathrm{P}(E_i \cap F) = \mathrm{P}(F)\mathrm{P}(E_i|F) = \mathrm{P}(E_i)\mathrm{P}(F|E_i).$$

Thus (draw a diagram)

$$\mathrm{P}(E_i|F) = \frac{\mathrm{P}(E_i \cap F)}{\mathrm{P}(F)} = \frac{\mathrm{P}(E_i)\mathrm{P}(F|E_i)}{\mathrm{P}(F)}.$$

Now $F = F \cap (\cup_{j=1}^{n} E_j) = \cup_{j=1}^{n} (F \cap E_j)$, where the events $(F \cap E_j)$ are also mutually exclusive. So

$$\mathrm{P}(F) = \sum_{j=1}^{n} \mathrm{P}(F \cap E_j) = \sum_{j=1}^{n} \mathrm{P}(E_j)\mathrm{P}(F|E_j).$$

Remark: Note that the denominator $\mathrm{P}(F)$ *plays the role of a normalizing constant to ensure that* $\sum_{i=1}^{n} \mathrm{P}(E_i|F) = 1$.

Example: COVID-19 antibody testing

We have

$$\mathrm{P}(F|E_1) = .9 \quad \mathrm{P}(F|E_2) = .05.$$

By Bayes' Theorem

$$\mathrm{P}(E_1|F) = \frac{\mathrm{P}(E_1)(.9)}{\mathrm{P}(E_1)(.9) + (1 - \mathrm{P}(E_1))(.05)}$$

and

$$\mathrm{P}(E_2|F) = 1 - \mathrm{P}(E_1|F) = \frac{(1 - \mathrm{P}(E_1))(.05)}{\mathrm{P}(E_1)(.9) + (1 - \mathrm{P}(E_1))(.05)}.$$

- If about 20% of the population has antibodies, i.e., $\mathrm{P}(E_1) = 0.2$, so $\mathrm{P}(E_1|F) = 0.2 \times 0.9/(0.2 \times 0.9 + 0.8 \times 0.05) \approx 0.82$.
- If about 2% of the local population has antibodies, i.e., $\mathrm{P}(E_1) = 0.02$, so $\mathrm{P}(E_1|F) = 0.02 \times 0.9/(0.02 \times 0.9 + 0.98 \times 0.05) \approx 0.27$.

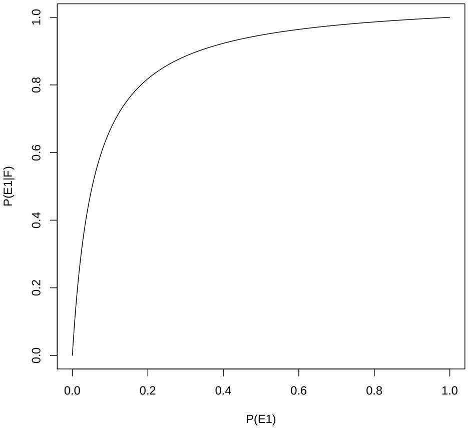
Remark: Again, note that the denominator $\mathrm{P}(F)$ is there to ensure that

$$\mathrm{P}(E_1|F) + \mathrm{P}(E_2|F) = 1.$$

Bayes factor

- The ratio of the probabilities of the outcome under the two scenarios
$$\frac{\mathrm{P}(F|E_1)}{\mathrm{P}(F|E_2)}$$
is called the *Bayes factor* (BF) between E_1 and E_2.
- Another way to express Bayes theorem is in terms of *odds* and BF:
$$\frac{\mathrm{P}(E_1|F)}{\mathrm{P}(E_2|F)} = \frac{\mathrm{P}(E_1)}{\mathrm{P}(E_2)} \cdot \frac{\mathrm{P}(F|E_1)}{\mathrm{P}(F|E_2)}.$$
That is
$$\text{Posterior odds} = \text{Prior odds} \times \text{BF}.$$
- In the above example, the BF = 18, which is typically deemed very large, but …

Relationship between $\mathrm{P}(E_1|F)$ and $\mathrm{P}(E_1)$.



- The prior can impact the inference substantially!
- It is usually worthwhile to study the robustness of a Bayesian analysis with respect to prior choices through a *sensitivity analysis*.

Example: The Monty Hall game

You are on a TV show in which you are presented with three doors.

- Behind one of them there is a Porsche.
- Behind each of the other two there is a goat (or a problem set)!

You get to choose a door to open and whatever is behind the door is yours to take home.

- You feel lucky and pick Door 1.
- Just before you open Door 1, the host opens Door 2 and you see a goat behind it.
- The host then asks "Are you sure you want to open Door 1?"

What should you do? You can try the game online at *http://math.ucsd.edu/~crypto/Monty/monty.html*.

What would Bayes do?

- Three possible causes E_1, E_2 and E_3.

$$E_i = \{\text{The car is behind Door } i\}.$$

- The effect

$$F = \{\text{The host opens Door 2}\}.$$

- $\mathrm{P}(E_1) = \mathrm{P}(E_2) = \mathrm{P}(E_3) = 1/3$.
- $\mathrm{P}(F|E_1) = 1/2$, $\mathrm{P}(F|E_2) = 0$ and $\mathrm{P}(F|E_3) = 1$.

By Bayes' Theorem

$$\mathrm{P}(E_1|F) = \frac{(1/3)(1/2)}{(1/3)(1/2)+(1/3)(0)+(1/3)(1)} = 1/3$$

$$\mathrm{P}(E_2|F) = \frac{(1/3)(0)}{(1/3)(1/2)+(1/3)(0)+(1/3)(1)} = 0$$

$$\mathrm{P}(E_3|F) = \frac{(1/3)(1)}{(1/3)(1/2)+(1/3)(0)+(1/3)(1)} = 2/3.$$

Yes. You should switch!

- We can in fact know this answer without doing all the calculation. How?

- In proving Bayes' theorem, we have used nothing but multiplication rule.
- There is no disagreement in the truth of this theorem.
- The inference feels very "natural".
- Why isn't every statistical inference problem solved in this manner?

More general versions of Bayes' Theorem

A similar argument as our proof can be used to extend Bayes' Theorem to more general cases.

For example, suppose

- X and Θ are two continuous random variables.
  - Θ corresponds to the unobserved effects E_i's.
  - X corresponds to the observed outcome F (or data).
- We have available two pieces of information:
  1. The *marginal p.d.f.* of Θ, $p(\theta)$, corresponding to $\mathrm{P}(E_i)$.
  2. The *conditional p.d.f.* of X given Θ, $p(x|\theta)$, corresponding to $\mathrm{P}(F|E_i)$.
- What is the conditional distribution of Θ given X, $p(\theta|x)$?

The joint p.d.f. of X and Θ

$$p(x, \theta) = p(x|\theta)p(\theta) = p(\theta|x)p(x).$$

From this we get that for x such that $p(x) > 0$,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

Since this is a p.d.f., it must integrate to 1. That is,

$$\int_{-\infty}^{\infty} \frac{p(x|\theta)p(\theta)d\theta}{p(x)} = 1.$$

Therefore we must have

$$p(x) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta)d\theta.$$

Note that this is consistent with the definition of *marginal* p.d.f.

There we have

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

for all x such that $p(x) = \int_{-\infty}^{\infty} p(x|\theta)p(\theta)d\theta > 0$.

- For each fixed x, this gives a p.d.f. of Θ.
- The denominator depends only on the fixed x, not on Θ.
- The denominator is only a normalizing constant so that the density in θ integrates to 1.
- To emphasize this, we will often write

  $$p(\theta|x) \propto p(x|\theta)p(\theta),$$

  meaning $p(\theta|x)$ is proportional to $p(x|\theta)p(\theta)$ *as a function in θ*.

A heuristic proof

$$\begin{aligned}
p(\theta|x)\Delta\theta &\approx \mathrm{P}(\theta \le \Theta < \theta + \Delta\theta \,|\, x \le X < x + \Delta x) \\
&= \frac{\mathrm{P}(\theta \le \Theta < \theta + \Delta\theta, x \le X < x + \Delta x)}{\mathrm{P}(x \le X < x + \Delta x)} \\
&\approx \frac{p(\theta,x)\Delta\theta\Delta x}{p(x)\Delta x} = \frac{p(\theta,x)\Delta\theta}{p(x)}.
\end{aligned}$$

Hence

$$p(\theta|x) = \frac{p(\theta,x)}{p(x)}$$

Similarly, flipping the places of θ and x, we have

$$p(x|\theta) = \frac{p(\theta,x)}{p(\theta)}.$$

Hence,

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

Even more generally

- Θ and X can each be either *discrete or continuous*. Still we have

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

  or simply

$$p(\theta|x) \propto p(x|\theta)p(\theta).$$

- $p(x|\theta)$ is the pdf or pmf of X given θ.
- $p(\theta)$ is the pdf or pmf of θ.
- $p(x)$ is the pdf or pmf of the *marginal distribution* of X, integrating out θ, i.e.,

$$p(x) = \int p(x|\theta)p(\theta)d\theta \quad \text{or} \quad \sum_{\theta} p(x|\theta)p(\theta)d\theta$$

  depending on whether θ is continuous or discrete.

Example: A political poll

A polling organization wishes to determine the fraction of Democrats in favor of the incumbent governer of North Carolina.

- They *randomly* select $n = 100$ names from the list of all registered Democrats to be interviewed.
- Assuming that all n are interviewed and expressed an opinion.
- The poll results in a count X for the governor and a count $n - X$ against.

Let θ be the actual proportion of Democrats who support the governor in the population. (People often don't differentiate the notation of the random variable Θ and its value θ.)

- After observing the data, what can we learn about θ?

If the sample is truly random, it is reasonable to model this poll as a Binomial experiment.

- Given θ, we know the distribution of X is Binomial(n, θ):

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

- We are uncertain about the true value of θ, so may treat it as a random variable as well.
    - We express our uncertainty using a probability distribution $p(\theta)$.
    - Suppose we “have no idea” about θ, and choose $p(\theta)$ to be Uniform$(0, 1)$.
    - This represents our *prior* (i.e. before observing data) knowledge about the value of θ.

Now we have the two pieces needed in Bayes' Theorem. Inference becomes a simple application of the theorem.

- Suppose we observe $X = 40$. The theorem says

$$\begin{aligned} p(\theta|X=40) &\propto p(X=40|\theta)p(\theta) \\ &= \binom{100}{40}\theta^{40}(1-\theta)^{60}\cdot 1 \quad \text{for } 0<\theta<1, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The corresponding normalizing constant is

$$\int_0^1 \binom{100}{40}\theta^{40}(1-\theta)^{60}d\theta.$$

We recognize that the portion of the density that involves θ, namely $\theta^{40}(1-\theta)^{60}$, is exactly the variable part of a Beta(41,61) distribution, so the two distributions *must agree* as both integrate to 1.

$$\begin{aligned} p(\theta|X=40) &= \frac{\Gamma(102)}{\Gamma(41)\Gamma(61)}\theta^{40}(1-\theta)^{60} \quad \text{for } 0<\theta<1, \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

What is our model?

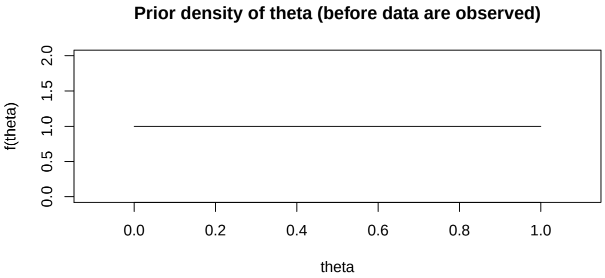
Our model (i.e., assumptions/hypotheses) is

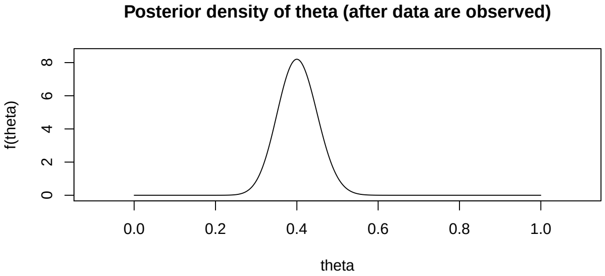
(1) θ is Uniform (0,1) *a priori*—representing our knowledge before data is observed. *(The prior.)*

(2) Given θ, X is Binomial(100,θ). *(The sampling model.)*

By combining the model with data, Bayes' Theorem allows us to reach the conclusion

(3) θ is Beta(41,61) *a posteriori*—representing our updated knowledge after data is observed. *(The posterior.)*


Prior density of theta (before data are observed)
f(theta)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
theta


Posterior density of theta (after data are observed)
f(theta)
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
theta

Summary

If we treat the state of nature (or the parameter θ) as a *random* quantity, and specify its *prior* probability distribution as a representation of our *a priori* knowledge, then Bayes' Theorem provides a simple recipe to incorporate information from data and produce the *posterior* distribution of the state of nature.

Based on this posterior distribution, we can make probabilistic statement about the state of nature or the parameters. For example,

- What is the chance that the support rate of the governer is over 45% given the data?
  - $\mathrm{P}(\theta > 0.45|x) = \int_{0.45}^{\infty} p(\theta|x)d\theta \approx 0.16.$
  - What is the posterior mean/median/mode of θ?
- This is an example of *Bayesian* inference.

Exchangeability

- Let $X_1, X_2, \ldots$ be an *infinitely exchangeable sequence* of random variables. That is, for any finite n, $(X_1, X_2, \ldots, X_n)$ are exchangeable (distribution invariant with respect to permutation):

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = P(X_{\sigma(1)} \in A_1, X_{\sigma(2)} \in A_2, \ldots, X_{\sigma(n)} \in A_n)$$

  for any permutation σ on $\{1, 2, \ldots, n\}$. In terms of pdfs, that is

$$p(x_1, x_2, \ldots, x_n) = p(x_{\sigma(1)}, x_{\sigma(2)}, \ldots, x_{\sigma(n)}).$$

- This is a generalization of an i.i.d. sequence. “The order of the data points doesn’t matter."
- Check: All i.i.d. sequences are exchangeable.

de Finetti's Theorem

- Then there exists a class of distribution $p(\cdot|\boldsymbol{\theta})$ and a probability measure p on $\boldsymbol{\theta}$, such that

$$P(X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n) = \int \left(\prod_{i=1}^{n} p(A_i \,|\, \boldsymbol{\theta}) \right) p(d\boldsymbol{\theta})$$

  and in terms of pdfs

$$p(x_1, x_2, \ldots, x_n) = \int \left(\prod_{i=1}^{n} p(x_i \,|\, \boldsymbol{\theta}) \right) p(d\boldsymbol{\theta}).$$

- All exchangeable sequences can be modeled as an i.i.d. draws from some distribution $p(\cdot \,|\, \boldsymbol{\theta})$ with some prior p on $\boldsymbol{\theta}$.
- Require flexible priors on the space of "all" distributions—$\boldsymbol{\theta}$ can be "infinite-dimensional".

Next

- More examples of inference using Bayes' Theorem.
- Why doesn't everyone use this simple scheme to solve all statistical problems?
- Contrast with the sampling approach.