STA 602 – Intro to Bayesian Statistics

Lecture 1

Li Ma

Course logistics

- Sakai, Gradescope, and Course Website.
- Grading structure: weekly labs (10%), weekly homework (10%), 2 quizzes (15%), midterm (30%), and final (35%).
- Zoom—*use your NetID to sign in at* `http://duke.zoom.us`. Only Zoom accounts associated with Duke emails can access the Zoom links. Make sure you sign into Zoom with the correct email.
- Office hours (Zoom)—link provided on Sakai. Poll.
- TAs: Aihua Li, Zhuoqun (Carol) Wang, Haoliang Zheng.
- Questions?

Statistics, Bayesian analysis, and decision analysis

- Statistics is the *prime* information science.
- It deals with the *extraction* of useful *information* from *data*, accounting for the appropriate *uncertainty*.
- Such information can help us
  - *understand the data generative mechanism*;
  - *make predictions about future observations*;
  - *make decisions* with proper evaluation of potential risk.
- *Bayesian* inference provides a principled approach to summarize our understanding of the data generative mechanism with fully probabilistic quantification of uncertainty.
- *Decision analysis* utilizes such summary of knowledge to evaluate possible losses and risks to guide decision making.

Probability modeling and statistical inference

- Statistics is the science of *extracting information from data*, while accounting for the *uncertainty*.
- Probability is the tool for formulating *uncertainty* in a mathematical fashion.
  - What is a model? assumptions.
  - What is a probability model?

} exactly Bayesian.

Probability modeling and statistical inference

- In undergraduate probability we have learned several families of probability distributions.
  - E.g. normal, binomial, negative binomial, geometric, exponential, Poisson, gamma, etc. (You need to be very familiar with these!)
  - These are the basic building blocks for more complex data generative mechnisms.
- Probability modeling is the art of making the appropriate *assumptions* about the underlying randomness in the data.
  - For example: What family of distributions to use? What additional assumptions—such as independence, stationarity, etc?
  - "All models are wrong, but some are useful." —George Box.
- Based on these assumptions, statistics aims at "completing the picture"— drawing inference ("highly educated guess") about various aspects of the underlying random mechanisms.

Common statistical inference problems

- Estimation: What is the value of certain quantities of interest?
- Prediction: What may be the value of a future observation?
- Hypothesis testing/Model selection: Are certain assumptions reasonable, or not (in comparison to some alternatives)?

In real problems, statistical inference and probability modeling typically form a *two-way* process.

- Model construction $\longleftrightarrow$ model application/validation.
- Some modern statistical methods try to be "model-free". (Most of them still involve assumptions, albeit very weak and general ones.)
- Bayesian inference generally requires the model construction part *though in modern literature (last 40 years) the model can be very flexible — Bayesian nonparametrics*.

Example: Coin tossing

- I've got a coin which I tossed 10 times.
- My data: *HHTHHTHTHH*.
- Q: What can be said about the chance for getting H?
  - Estimation: What's the chance of getting a head on each toss?
  - Prediction: What is the chance for two new tosses to both be H?
  - Hypothesis testing: Is this a fair coin? Are the tosses independent?

What *probability model* can we use to represent this *experiment*?

Statistical experiment and random variables

- *Experiment*: A process, planned or unplanned, with an (observable or unobservable) outcome, ω.
- *Outcome space*: The collection of all possible outcomes, denoted by Ω.
- *Event*: A subset of the outcome space. (In some cases not all subsets are events, but we don't have to worry about that in this course. See Section 1.4 of textbook.)
- *Random variable*: A real-valued function defined on Ω.
  - $X : \Omega \to \mathbb{R}$
  - The set of values X can take, i.e.,

$$\{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}$$

is called the *sample space* of the random variable X. That is, the space on which your data sits.

Example: A coin flipping experiment

- An experiment that consists of a series of n *Bernoulli* trials.
- The outcome of each single trial is either 1, called a success, or 0, called a failure.

What is the outcome space?

$$\Omega = \{0,1\} \times \{0,1\} \times \ldots \times \{0,1\} = \{0,1\}^n.$$

For example, for a fair coin,

- $n = 10$, "1"=H.
- Each possible outcome ω is of the form 1001100101.
- The event that we get six heads is the set of all strings in which there are six 1's.

Random variables defined on this sample space

- For example,

$$X = \#\text{ heads} = \#\text{ 1's}$$

  is a *discrete* random variable.
- More explicitly, $X : \Omega \to \mathbb{R}$, with

$$X(\omega) = \#\text{ 1's in the outcome } \omega.$$

  What is the sample space of X?
- Another example,

$$Y = \#\text{ tails that follow another tail.}$$

A probability model: the Binomial model

- Two *assumptions* regarding the randomness of this experiment:
  - Each trial is independent.
  - The chance of getting a 1 in each trial is the same.
- Based on this model, by the *multiplication rule*, the probability of each observable outcome given the parameter value θ is

$$P(1101101011|\theta) = \theta \times \theta \times (1-\theta) \times \ldots \times \theta \times \theta = \theta^7(1-\theta)^3.$$

- What is the *probability mass function* (pmf) of X, the number of heads?

The Binomial(n, θ) distribution

- The pmf is

$$P(X = k \,|\, \theta) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$
$$= 0 \quad \text{otherwise.}$$

- The *parameter* $\theta \in [0, 1]$ determine this distribution.
- The Binomial distribution is an example of a *parametric family* of probability distributions.
  - A family (or collection) of distributions indexed by a finite number of parameters.
- *Parametric*, *semi-parametric*, and *nonparametric* statistical inference.

finite collection of parameters

$f(x-u)$: f is flexible and complex while u is used to parameterize mean.

Statistical inference (that you have already learned)

- Possible *inference* question: What is the head probability θ?
- As we will learn in this course, based on this model, a common *estimator* for the θ is the value of θ that maximizes this probability

$$\hat{\theta} = \#\ 1\text{'s}/n = X/n,$$

good under square loss.

This treats θ as a fixed unknown quantity.

- But just having a point guess is not sufficient. How certain are we about the guess? (Uncertainty quantification is needed.)
- An estimate for the *standard deviation* of X is

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\hat{\theta}(1-\hat{\theta})/n}.$$

This quantifies uncertainty *under repeated experiments with the same fixed parameter*.

Some questions of interest

- Is $\hat{\theta}$ any good?
  - What does "good" mean? Decision theory is required—need notions of loss or reward. optimization, hard.
  - How to optimize: finite n or $n \to \infty$? With finite n, what if $X = 0$ and so $\hat{\theta} = 0$? 532, 732
- Can we have an interval estimate? E.g., $\hat{\theta} \pm 1.96 \cdot \mathrm{SE}(\hat{\theta})$.
  - What is the meaning of this interval?
  - Assuming $n \to \infty$.

Another inference question

- Prediction regarding future observations.
- Examples:
  - What is the chance for the next toss to be head?
  - What is the chance for the next three tosses to have 2 heads and 1 tail?
- How might you address these questions?
- Is my prediction good? Decision theory is required.
- What about the uncertainty in your prediction?

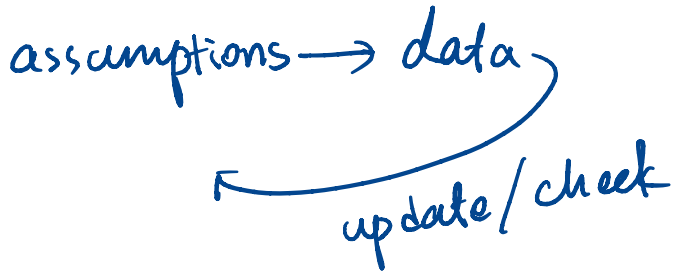
Another common inference question

- Suppose instead of guessing the value of θ, we have some *hypothesis* about its value, e.g., $\theta = 0.5$.
- Does the data support/contradict the hypothesis?
- For example, suppose in our observed data, $\hat{\theta} = 0.7$, what should we conclude?
- Apparently uncertainty quantification is needed to make such a conclusion.
- Now suppose $\text{SE}(\hat{\theta}) \approx 0.14$. What can we conclude?
- Is our acceptance/rejection good? Decision theory is required. Finite n or $n \to \infty$.
- For example, when n is finite, what if it so happens that $\hat{\theta} = 0$?

$0.7 \pm 1.96 \times 0.14$

The two-way process

- The previous inference was carried out assuming that the binomial model is correct.
- If our observed data is *HHHHHHHHTTT*, the binomial model may not be appropriate. Why? Independent
- Recall the two assumptions.
- Can this stickiness be purely due to randomness?
- One can checking such assumptions based on diagnostic statistics. (More on this later).



Sampling and Bayesian approaches

$\mathbb{P}_\theta\{(\hat{\theta}_l(\underline{x}), \hat{\theta}_u(\underline{x})) \ni \theta\} = 0.95$ (✓)

$\mathbb{P}_\theta\{\theta \in (0.2, 0.8)\} = 95\%$ (✗).

- Two schools of thoughts on how inference should proceed.
- The sampling (frequentist) approach:
  - One assumes that the unknown parameters are *fixed*. Inference is made based *only* on the probability of the data *given* the "true" but unknown parameter θ under repeated sampling.

$X|\theta \sim F_{X|\theta}$

- The Bayesian approach:
  - One *treats the parameters as random variables as well*, place modeling assumptions on them, and draw inference based on the *joint* distribution of the data and the parameters. In particular, the conditional distribution of the unknowns given the observed quantities are key to inference.

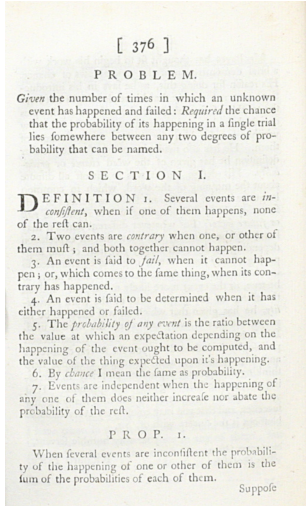
$\theta \sim F_\theta$

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

- In this class, we will mainly focus on the Bayesian approach but comparison to the sampling approach will be helpful.
- Which approach do you think is more natural?

{ How to spell out $p(\theta)$
How to get $p(\theta|x)$ from $p(x,\theta)$.

A little history of how it all begain



[376]

PROBLEM.

Given the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.

SECTION I.

DEFINITION 1. Several events are *inconsistent*, when if one of them happens, none of the reſt can.

2. Two events are *contrary* when one, or other of them muſt; and both together cannot happen.

3. An event is ſaid to *fail*, when it cannot happen; or, which comes to the ſame thing, when its contrary has happened.

4. An event is ſaid to be determined when it has either happened or failed.

5. The *probability of any event* is the ratio between the value at which an expectation depending on the happening of the event ought to be computed, and the value of the thing expected upon it's happening.

6. By *chance* I mean the ſame as probability.

7. Events are independent when the happening of any one of them does neither increaſe nor abate the probability of the reſt.

PROP. 1.

When ſeveral events are inconſiſtent the probability of the happening of one or other of them is the ſum of the probabilities of each of them.

Suppoſe

- "An essay towards solving a problem in the doctrine of chances". Rev. Thomas Bayes and Richard Price (1764).
- In response to David Hume's 1748 essay "On Miracles".
- Temporary loss of popularity during early 20th century.
- Big come back in the later part of the 20th century.
  - Computational advances.
  - Algorithmic advances.
- Today: Arguably the most common approach by field scientists.