STAT 522, Lec 1 notes, Jan 3
Winter 2024, Ruibo Zhang

1 preliminaries

Syllabus

Material:

1. We will start our course by first carefully defining our objects of investigation in terms of the population (i.e, the joint distribution).

   Start with the first principle with i.i.d. examples.

2. We will then examine how to perform inference when we only observe a sample from the population, under different types of assumptions.

Computations:

1. R is **HIGHLY recommend** since it is more well-implemented than Python in terms of statistics.

2. Rstudio and VS code are highly recommended.

Participation: Lecture notes + Readings. ALWAYS due on the following day of the lecture.

Home page

Assignments/exams are listed. There is a resource page that include all available code.
Textbook, Grading: See Canvas for more information.

2 course material

Plan for the day:

- Big picture
- Review of Probability theory
- Fundamental concepts: univariate distributions

Main references: see website

- A & M Chap 1 - 2
- GC Chap 3 - 5

Big Picture

We are going to build everything from first principles, i.e. in a "population first" approach. There is a generic population problem.

- joint distribution
- Casual inference model

We are going to define the objects that we want to study in terms of the population, $\boxed{\text{nonparametrically}}$.

- "Transitional approach": We would have a model described as

  $$y_i \overset{i.i.d.}{\sim} N(\alpha + \beta x_i, \sigma^2), x_i \text{ is fixed.}$$

  And we want to figure out α, β from data. Alternatively, another way:

  $$y_i = \alpha + \beta x_i + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2).$$

  THIS IS NOT WHAT WE WANT TO DO!

We will define parameters as operations that use performance in a model. E.g.

- What is the $\boxed{\text{best}}$ linear prediction of y given x?

  What do we mean by best: we need a cost function. The best one should minimize the cost. Make sense when the model is invalid.

  "β" the partial derivative of the BLP(best linear predictor) with respect to x (we have here in the non-parametric sense.)

  $$\frac{\partial \text{BLP}(y|x)}{\partial x} = \beta.$$

Population first approach

Population: for most of time, population is the joint distribution of these observed points

$$\mathbb{P}(v), v = \{x, y, z, w, \dots\} \text{ (all variables)}$$

The statistical parameters are going to be defined in terms of $\mathbb{P}(v)$. E.g. $\mathbb{E}[Y|X]$ or $\text{BLP}(Y|X)$. (There are simpler: predictive objects and casual objects.) What we have access to is the sample. Actually sample is finite. (One can treat the population as infinite samples):

$$\mathbb{P}_n : \text{ empirical distribution }.$$

How reliable is the empirical

$$\begin{array}{ccccc}
\text{Sample}(i.i.d.) & \xrightarrow{\text{assumptions}} & \text{Population} & \overset{\text{casual assumptions}}{\rightarrow} & \begin{array}{c}\text{interventional distributions } \mathbb{P}(Y|\text{do}(X)) \\ \text{casual parameters are defined in terms of casual}\end{array} \\
\downarrow & & \downarrow & & \\
\mathbb{P}(v) & & \text{parameters} & \rightarrow & \mathbb{E}[X|Y]
\end{array}$$

Probability theory review

Definition 1. We use Ω to denote the sample space. Ω is the set of all conceivable outcomes. We use $S \subset \mathcal{P}(\Omega)$ to denote the event space (σ-algebra).

1. S is not empty.
2. S is closed under complement.
3. S is closed under countable union.

We use $\mathbb{P}$ to denote the probability measure. $\mathbb{P} : S \to [0,1]$ is a set function from S to $[0,1]$.

Definition 2. We call the triplet $(\Omega, S, \mathbb{P})$ the Probability space. Axioms:

1. For all $A \in S, \mathbb{P}(A) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. Disjoint events: $\{A_i\}_{i \in I} \in S$

$$\mathbb{P}(\cup_{i \in I} A_i) = \sum_{i \in I} \mathbb{P}(A_i).$$

where I is countable.

Conditional probability:

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Law of total probability: towering property

$$\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X].$$

Random variables

A random variable is a map from Ω to $\mathbb{R}$, i.e. $X : \Omega \to \mathbb{R}$.

Example 3. Suppose we are modeling coin flips, then

- $\Omega = \{\text{heads}, \text{tails}\}$. (These are models, not necessary the real world.).
- $S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.
- $\mathbb{P}(\emptyset) = 0, \mathbb{P}(\{H\}) = \frac{1}{2}, \mathbb{P}(\{T\}) = \frac{1}{2}, \mathbb{P}(\{T, H\}) = 1$.

In these particular case, we have that $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$.

Random variable: $X(T) = 1, X(H) = 0$.