Lecture 01: Fundamentals of Statistics

Lecturer: Qiang Sun
Scribe: Yicheng Zeng

Lecture #1
January 15th, 2021

---

1 Logistics of the Course

Welcome to STAD80: **Analysis of Big Data**. In many ways this course will be similar to an introductory graduate level course. It will be split between theoretical components and computational components. Two major topics will be covered, those being:

1. Statistical Modeling,
2. Optimization - how to design effective optimization algorithms for large scale data analytics.

For most homework, you are expected to use `RStudio` and `RMarkdown`. For syllabus and office hours information, see the syllabus on Quercus. Be sure to sign up to scribe for one of the lectures!

2 Big Data and Why it is Important

Big Data is often defined as a data set with a very large sample size, typically denoted n. It might also refer to a data set that is very large in dimensionality, denoted d. We are also concerned with mixed data and unstructured data sets, for example images and text.

But why do we need to bother with this? Nowadays, we have enormous amounts of data that we never had to deal with before - for example images or geographic data. Often data is so large that it must be stored at multiple different locations. We need new techniques for dealing with this data in a practical and efficient way.

It is also much more common now to have data that is not collected for any one specific purpose. In traditional statistics, researchers would start with a plan and collect data after that. For example, in the past, a research team might conduct a medical trial, testing whether or not some new drug was effective at treating a disease. They would collect information on a group of patients that were given the drug, and on a control group. Then they would process the data and make an inference on the treatment effect. Nowadays, the plan in big data science is often more abstract, and the focus is instead on the data. We might want to make medical inferences from an enormous amount of general data collected on thousands, or even millions of people - data like MRI scans, disease histories, and demographic info.

In Figure 1, we see a typical process for addressing a scientific or business problem statistically, what we would call the "Big Data Science Loop". First, we define our scientific or business problem that is often an abstract one. Then we conduct a study and collect raw data based on both domain knowledge and statistical methodology. These collected raw data may also constitute online data repositories, like ADNI, UK Biobank, TCGA and so on. After that, data engineers will typically process the raw data into a useful form. Statisticians then design proper statistical models for the analytic data. They may also develop effective optimization algorithms for it. By this process, the statistical methods

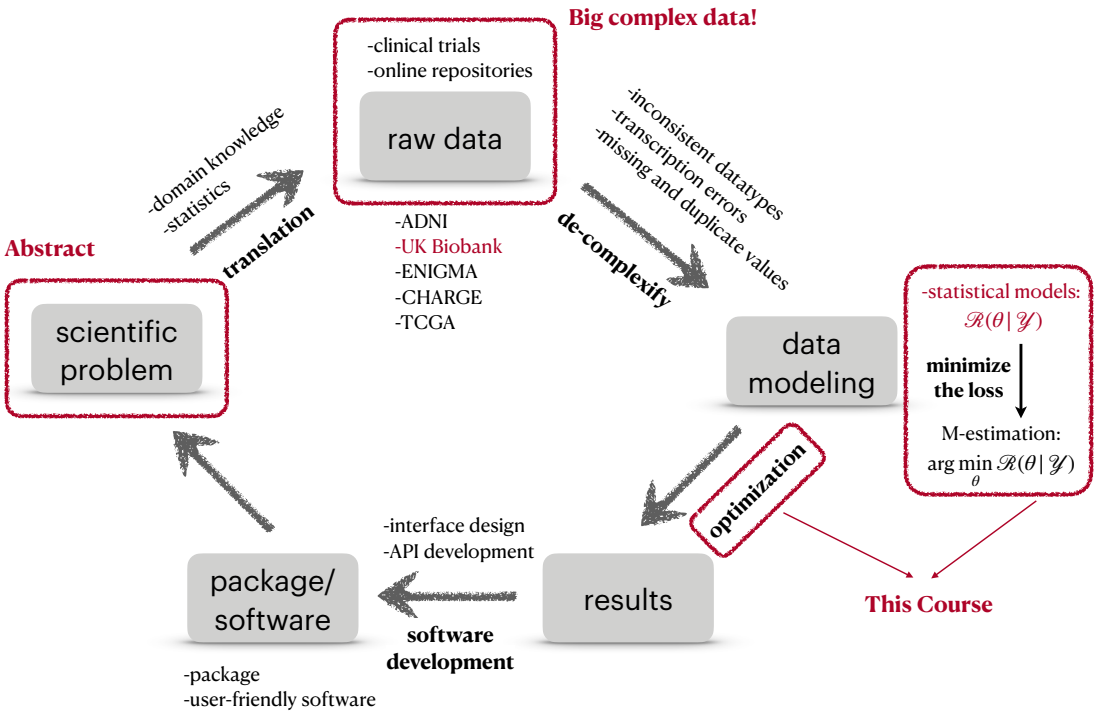

Figure 1: Big Data Science Loop

and optimization algorithms would also be translated into user-friendly packages/softwares from which other data scientists would benefit. After all of this, we may find that our data sometimes does not have the features we want, and we are not able to draw the desired inferences from it. In this case, we will either have to return to our original problem and revise it, or revise our study, collect more data, and try again. If our data does contain the features we want, we are able to move forward and make a decision or inference from it.

This course is really about data modeling, designing optimization algorithms, and drawing conclusions from the data.

3 Basic Concepts in Statistics

In this section we will recap some fundamental concepts in statistics. We start with some definitions.

Definition 3.1 (Sample Space)**.** The collection of all possible outcomes of a statistical experiment.

Definition 3.2 (Random Samples)**.** We call the random variables $X_1, \cdots, X_n$ random samples if they are independent and identically distributed (i.i.d.) according to some probability density function (pdf) $p(x)$, that is

$$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} X \sim p(x).$$

In this course, we use capital letters to denote random samples and use lower case letters for realized samples. We also denote random variables $X_1, \cdots, X_n$ by $X_{1:n}$, and random variables $x_1, \cdots, x_n$ by $x_{1:n}$.

Definition 3.3 (Realizations/observed values/observed outcomes)**.** The realized samples $x_1, \cdots, x_n$ are called realizations of the random variables $X_1, \cdots, X_n$.

Definition 3.4 (Statistic)**.** A statistic is any measurable function of random samples $X_{1:n}$.

Definition 3.5 (Cumulative Distribution Function (CDF))**.** The cumulative distribution function $F(\cdot)$ of random variable X is defined as

$$F(x) = P(X \leq x),\ x \in \mathbb{R}.$$

Definition 3.6 (Probability Density Function (PDF))**.** The probability density function $p(\cdot)$ of a random variable X is the derivative of its cumulative distribution function. That is,

$$p(x) = \frac{\partial}{\partial x} F(x),\ x \in \mathbb{R}.$$

Remark. $p(x)$ could also represent the probability mass function for a discrete random variable. When we focus on a family of probability density (mass) functions, we will use $p_\theta(x)$ to denote that the density (mass) function is parameterized by θ.

Definition 3.7 (A Statistical Model)**.** A statistical model is a family of probability distributions indexed by a parameter set Θ:

$$\mathcal{P} = \{p_\theta : \theta \in \Theta\}.$$

Definition 3.8 (A Parametric Model)**.** If there exists a finite dimensional Θ to index $\mathcal{P}$, then $\mathcal{P}$ is a parametric model.

Remark. For a parametric model, Θ must be finite dimensional, but it does not have to take a finite number of values. For example, in a Bernoulli model, there is only one parameter $\theta \in [0, 1]$ which represents the probability of success in a single trial. The parameter θ of this model is 1-dimensional, so this is a parametric model. But θ may take infinitely many possible values in the interval $[0, 1]$.

Definition 3.9 (A Nonparametric Model)**.** If there does not exist a finite dimensional Θ to index $\mathcal{P}$, we say that $\mathcal{P}$ is a nonparametric model.

Example 3.10 (A Parametric Model)**.** Consider the family of Gaussian distributions

$$\mathcal{P} = \left\{ p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} :\ \mu \in \mathbb{R},\ \sigma^2 > 0 \right\}.$$

This is a parametric model, because $p_{\mu,\sigma^2}(x)$ is indexed by the 2-dimensional parameter $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$.

Example 3.11 (A Nonparametric Model)**.** Consider

$$\mathcal{P} = \{\text{all cumulative distribution functions}\}$$

This is nonparametric model as we cannot find a finite dimensional parameterization for it.

Remark. A nonparametric model also has parameters, it is just that the parameters are actually infinitely dimensional.

4 Asymptotic Theory

In this course, we will also touch on asymptotic theory, which is a field of statistics that focuses on the behaviour of a model when the sample size approaches infinity. It answers the question, "What happens when we get more and more data?" Though an inference should hold no matter what the sample size is, asymptotic theory is historically significant in statistics, and so will be covered in this course.

We first look at some definitions on convergence of random variables as follows.

Definition 4.1 (Convergence in Probability)**.** Given a random variable X and a sequence of random variables $\{X_n\}_{n\geq 1}$. If it holds that

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0$$

for any $\epsilon > 0$, we say X_n converges to X in probability, denoted as $X_n \xrightarrow{P} X$.

Definition 4.2 (Convergence in Distribution)**.** Given a random variable X and a sequence of random variables $\{X_n\}_{n\geq 1}$. Let

$$F_n(x) = P(X_n \leq x), \quad F(x) = P(X \leq x).$$

If it holds that

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for any $x \in \mathbb{R}$, we say X_n converges to X in distribution, denoted as $X_n \xrightarrow{D} X$.

Definition 4.3 (Sample Mean/Empirical Average)**.** The sample mean (or empirical average) of random samples $X_1, \cdots, X_n$ is defined as

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Next, we display the two most fundamental results in asymptotic theory .

Theorem 4.4 (The Law of Large Numbers)**.** Given random samples $X_1, \cdots, X_n$ with expectation $E(X_i) = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then the Law of Large Numbers says that $\bar{X}_n \xrightarrow{P} \mu$ as $n \to \infty$.

Theorem 4.5 (The Central Limit Theorem)**.** Assume the same situation as in Theorem 4.4. The Central Limit Theorem says that

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1), \text{ as } n \to \infty.$$

Remark. Note that the Central Limit Theorem implies the Law of Large Numbers. LLN is more useful for estimation, whereas CLT is more useful for inference, for example creating a confidence interval, or finding a p-value.

5 Statistical Estimators

Now we briefly introduce some statistical techniques.

Definition 5.1 (Point Estimation)**.** Assume $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$. A point estimation refers to making a single best guess on θ.

Definition 5.2 (Estimator)**.** An estimator is a rule for calculating an estimate of a parameter based on the data. Given random samples $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$, an estimator $\widehat{\theta}_n$ has a form of

$$\widehat{\theta}_n = g(X_1, \cdots, X_n),$$

where $g(\cdot)$ is some measurable function from $\mathbb{R}^n$ to Θ.

Definition 5.3 (Estimate)**.** An estimate refers to the value we use to guess the true parameter. Given realizations $x_1, \cdots, x_n$ of the random samples $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$, $g(x_1, \cdots, x_n)$ is an estimate of θ, where $g(\cdot)$ is the same function as in Definition 5.2.

Remark. When we apply the estimator $\widehat{\theta}_n(X_1, \cdots, X_n) = g(X_1, \cdots, X_n)$ to realizations $x_1, \cdots, x_n$, we get an estimate $\widehat{\theta}_n(x_1, \cdots, x_n) = g(x_1, \cdots, x_n)$. Note that an estimator is an random quantity, while an estimate is fixed.

For an estimator $\widehat{\theta}_n$, we hope it converges to θ in probability, which is called consistency. We introduce the formal definition below.

Definition 5.4 (Consistent Estimator)**.** If $\widehat{\theta}_n$ converges to θ in probability as n goes to infinity, then $\widehat{\theta}_n$ is called a consistent estimator for θ.

Another quantity that will be used to evaluate an estimator is the bias.

Definition 5.5 (Bias)**.** The bias of an estimator $\widehat{\theta}_n$ is defined as

$$\text{Bias}(\widehat{\theta}_n) = E(\widehat{\theta}_n) - \theta,$$

where the expectation $E(\cdot)$ is taken with respect to the joint probability distribution of $(X_1, \cdots, X_n)$, namely $p_\theta(x_1) \cdots p_\theta(x_n)$.

Definition 5.6 (Unbiased Estimator)**.** $\widehat{\theta}_n$ is an unbiased estimator if $\text{bias}(\widehat{\theta}_n) = 0$.

Unbiasedness and consistency are not necessarily related to each other. We use the following examples to show the difference.

Example 5.7. Suppose $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$.

(a). The estimator $\widehat{\mu}^{(1)} = X_1$ is unbiased but not consistent, because it only ever takes one sample X_1.

(b). The estimator $\widehat{\mu}_n^{(2)} = \frac{1}{n}\sum_{i=1}^n X_i$ is consistent and unbiased, because $E(\widehat{\mu}_n^{(2)}) = \mu$ and $\widehat{\mu}_n^{(2)} \xrightarrow{P} \mu$.

(c). The estimator $\widehat{\mu}_n^{(3)} = \frac{1}{n+1}\sum_{i=1}^n X_i$ consistent but biased. Note that $E(\widehat{\mu}_n^{(3)}) = \frac{n}{n+1}\mu \neq \mu$. By Slutsky's Theorem and the fact that $\widehat{\mu}_n^{(2)} \xrightarrow{P} \mu$, it holds that $\widehat{\mu}_n^{(3)} \xrightarrow{P} \mu$.

6 The Maximum Likelihood Estimator

Next, we briefly introduce the Maximum Likelihood Estimator (MLE).

Definition 6.1 (Likelihood)**.** The likelihood function related to the random sample X_i is

$$L(\theta, X_i) = p_\theta(X_i).$$

Remark. $L(\theta, X_i)$ is a random quantity, since X_i is a random variable.

Definition 6.2 (Joint Likelihood)**.** The joint likelihood of θ w.r.t the entire data set $\{X_1, \cdots, X_n\}$ is defined as

$$L_n(\theta) = L(\theta; X_1, \cdots, X_n) = p_\theta(X_1, \cdots, X_n).$$

Definition 6.3 (Joint Log Likelihood)**.** The joint log likelihood of θ w.r.t the entire data set $\{X_1, \cdots, X_n\}$ is defined as

$$l_n(\theta) = \log(L_n(\theta)).$$

Now we come to our key definition.

Definition 6.4 (Maximum Likelihood Estimator)**.** The estimator $\widehat{\theta}_n$ is the MLE of θ if

$$L_n(\widehat{\theta}_n) \geq L_n(\theta)$$

for all $\theta \in \Theta$.

Remark. If $L_n(\theta)$ has a unique maximum, then we have

$$\widehat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L_n(\theta)$$

We will use an example to further clarify what the Maximum Likelihood Estimator is.

Example 6.5 (Gaussian Distribution)**.** Assume $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$. Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{1}{n}(\sum_{i=1}^n X_i)^2$. It can be shown that they are the MLE's of μ and σ^2, respectively.

But why are we even interested in the MLE of an unknown parameter? It provides a "unified framework" under which we can work towards a model that is, in some sense, "optimal". However, it does assume we know the model, and are just looking for the parameter. What do we mean by "optimal"? This is the central focus of asymptotic statistics. The following lemma is on the asymptotics of MLE.

Theorem 6.6 (MLE is Asymptotically Normal and Efficient)**.** Under certain regularity conditions, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{D} N(0, I_\theta^{-1}),$$

where $I_\theta = -E_\theta\left\{\frac{\partial^2}{\partial\theta^2}\log(p_\theta(x))\right\}$ is called the Fisher information matrix. Here we use the expectation with respect to underlying truth. Moreover, we have $\text{var}(\sqrt{n}\widehat{\theta}_n) \leq \text{var}(\sqrt{n}\widehat{\theta})$ for any $\widehat{\theta}$ that is asymptotically unbiased and locally regular.