36-705: Intermediate Statistics **Fall 2017**

Lecture 11: September 25

Lecturer: Siva Balakrishnan

The first chunk of our course has focused primarily on properties of averages of i.i.d. random variables. In particular, the question of interest roughly was "how close is an average of i.i.d. random variables to an expectation?". We developed tail bounds (non-asymptotic) and understood the limiting distribution of the average (asymptotic) as ways to attack this question. Then we started to discuss uniform laws where the question was how to argue that many (possibly related averages) converge to their respective expectations.

Now, we will switch gears and start to talk a bit more formally about statistical estimation and inference. Before we can make sense of these questions however we need to define a statistical model.

11.1 Statistical Models

The typical starting point for statistical inference should be familiar to you by now: we suppose that we obtain an i.i.d sample $\{X_1, \ldots, X_n\}$ from some distribution P.

Often, we further hypothesize restrictions on the set of possible distributions that could have generated the data, i.e. we suppose that $P \in \mathcal{P}$ where $\mathcal{P}$ is just some collection of distributions. We refer to $\mathcal{P}$ as the *statistical model*.

The common classes of distributions usually fall into one of two broad categories (with lots of fuzziness in between):

1. **Parametric models:** Here the statistical model $\mathcal{P}$ is described by a finite set of parameters. We usually write these as:

   $$\mathcal{P} = \{p(X;\theta) : \theta \in \Theta\},$$

   where $\Theta \subseteq \mathbb{R}^d$, i.e. these are a collection of distributions that we can describe using d real-valued parameters. We use the notation $p(X;\theta)$ or $p_\theta(X)$ to denote the distribution of the random variable X, where the distribution is parameterized by θ.

   A typical example would be to hypothesize that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, and in this case the parameter $\theta = (\mu, \sigma) \in \mathbb{R}^2$.

2. **Non-parametric models:** Roughly, non-parametric models are those which cannot be described by a finite set of parameters.

A few common examples are: (1) the set of all possible distributions on $\mathbb{R}$ (say):

$$\mathcal{P} = \{\text{all distributions on } \mathbb{R}\}.$$

In this case we are making no assumptions on the data generating process (beyond the i.i.d. assumption). When we studied estimating the CDF we made no assumptions and were implicitly working with this non-parametric model.

(2) Another common example is the set of distributions with smooth (say with square integrable second derivative) densities, i.e.

$$\mathcal{P} = \left\{ p : \int_{\mathbb{R}} (p''(x))^2 dx \leq C \right\}.$$

The model assumptions are the starting point for all subsequent inference, and can strongly influence our conclusions about the data. In general, non-parametric models are making much weaker assumptions about the data, while parametric models make strong (often unjustifiable assumptions).

On the other hand, as we will see later on in the course, parametric models can often be estimated using far fewer samples and can sometimes be much more interpretable. These advantages of course do not typically justify the use of parametric models.

That said, we will begin investigating parametric models and developing a comprehensive theory for estimation and inference for parametric models before turning our attention to non-parametric models.

11.2 Statistics

A statistic is simply a function of the observed sample, i.e. if we have $X_1, \ldots, X_n \sim P$ then any function $T(X_1, \ldots, X_n)$ is called a statistic. A statistic is a random variable.

Some examples of statistics include:

1. order statistics, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$
2. sample mean: $\overline{X} = \frac{1}{n} \sum_i X_i$,
3. sample variance: $S^2 = \frac{1}{n-1} \sum_i (X_i - \overline{x})^2$,
4. sample median: middle value of ordered statistics,
5. sample minimum: $X_{(1)}$
6. sample maximum: $X_{(n)}$.

11.3 Sufficient Statistics and Data Reduction

For the rest of todays lecture we will talk about a special class of statistics called sufficient statistics. There are two ways to motivate sufficient statistics: the data reduction viewpoint where we would like to discard non-informative pieces of the dataset (for storage or other benefits), and the risk reduction viewpoint where we want to construct estimators that only depend on meaningful variation in the data. We will focus on the former viewpoint today.

The goal in data reduction roughly is to find ways to reduce the size of a dataset without throwing away important information. We need to fix ideas more concretely to make progress on this abstract question.

We focus on parametric models and suppose we observe samples $\{X_1, \ldots, X_n\} \sim p(X; \theta)$. The typical statistical estimation problem is that we observe the samples as want to understand something about the unknown parameter θ. Again somewhat abstractly the goal of data reduction is to find statistics $T(X_1, \ldots, X_n)$ that contain all the information about the unknown parameter θ.

Sufficient Statistic: A statistic $T(X_1, \ldots, X_n)$ is said to be **sufficient** for the parameter θ if the conditional distribution $p(X_1, \ldots, X_n | T(X_1, \ldots, X_n) = t; \theta)$ does not depend on θ for any value of t.

Roughly, once we know the value of the sufficient statistic, the joint distribution no longer has any more information about the parameter θ. One could imagine keeping only $T(X_1, \ldots, X_n)$ and throwing away all the data. Take this loose interpretation with a grain of salt - we will return to it.

There is another more subtle question about what it means to condition on the sufficient statistic when it comes from a continuous distribution (since typically the conditioning set will have zero probability). It turns out that one can salvage this using something called the factorization theorem (which we will come to shortly). For now however, you should think about the discrete case and suppose that the probability that $T(X_1, \ldots, X_n) = t$ is strictly positive.

In order to check if a given statistic is sufficient we simply need to compute the conditional distribution $p(X_1, \ldots, X_n | T(X_1, \ldots, X_n) = t; \theta)$ and verify that it does not depend on t.

Poisson sufficient statistics: $X_1, \cdots, X_n \sim \text{Poisson}(\theta)$. Let $T = \sum_{i=1}^n X_i$.

Then,

$$p(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{p(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p(T = t)}.$$

There are two possibilities, that the sum of the x_is is equal to t and that the sum is not

equal to t. In the latter case the probability is zero so we obtain,

$$\frac{p(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p(T = t)} = \frac{p(X_1 = x_1, \ldots, X_n = x_n)\mathbb{I}(T = t)}{p(T = t)}$$

The sum of n independent Poissons with parameter θ is Poisson with parameter $n\theta$ so we obtain,

$$\begin{aligned}\frac{p(X_1 = x_1, \ldots, X_n = x_n)\mathbb{I}(T = t)}{p(T = t)} &= \frac{\exp(-n\theta)\theta^{\sum_{i=1}^n x_i}\mathbb{I}(T = t)(\sum_{i=1}^n x_i)!}{[\prod_{i=1}^n (x_i)!]\exp(-n\theta)(n\theta)^{\sum_{i=1}^n x_i}} \\ &= \frac{\mathbb{I}(T = t)t!}{[\prod_{i=1}^n (x_i)!]\, n^t},\end{aligned}$$

which does not depend on θ. So we can conclude that $T(X_1, \ldots, X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic. You can similarly verify that the average is sufficient, and that $3.7\sum_{i=1}^n X_i$ is sufficient. Furthermore, you can always condition on more things without destroying sufficiency so that $(\sum_{i=1}^n X_i, X_1, X_{17})$ is also sufficient.

In some sense we might believe that the sum is a better sufficient statistic than $(\sum_{i=1}^n X_i, X_1, X_{17})$. We will return to this idea of "minimal sufficient statistics" in the next class.

Binomial sufficient statistics: Suppose we observe $X_1, \ldots, X_n \sim \text{Ber}(p)$, then once again we can verify that the sum is sufficient. This proceeds in exactly the same way, i.e.

$$\begin{aligned}p(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{p(X_1 = x_1, \ldots, X_n = x_n, T = t)}{p(T = t)} \\ &= \frac{\mathbb{I}(T = t)\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\mathbb{I}(T = t)}{\binom{n}{t}},\end{aligned}$$

which does not depend on θ.

11.3.1 Sufficient Statistics - The Partition Viewpoint

It is better to describe sufficiency in terms of partitions of the sample space.

Example 11.1 *Let $X_1, X_2, X_3 \sim$ Bernoulli(θ). Let $T = \sum X_i$.*

(x_1, x_2, x_3)		t	$p(x\|t)$
(0, 0, 0)	$\to$	$t = 0$	*1*
(0, 0, 1)	$\to$	$t = 1$	*1/3*
(0, 1, 0)	$\to$	$t = 1$	*1/3*
(1, 0, 0)	$\to$	$t = 1$	*1/3*
(0, 1, 1)	$\to$	$t = 2$	*1/3*
(1, 0, 1)	$\to$	$t = 2$	*1/3*
(1, 1, 0)	$\to$	$t = 2$	*1/3*
(1, 1, 1)	$\to$	$t = 3$	*1*
8 elements	$\to$	*4 elements*	

1. A partition $B_1, \ldots, B_k$ is sufficient if $f(x|X \in B)$ does not depend on θ.
2. A statistic T induces a partition. For each t, $\{x : \ T(x) = t\}$ is one element of the partition. T is sufficient if and only if the partition is sufficient.
3. Two statistics can generate the same partition: example: $\sum_i X_i$ and $3\sum_i X_i$.
4. If we split any element B_i of a sufficient partition into smaller pieces, we get another sufficient partition.

Example 11.2 *Let* $X_1, X_2, X_3 \sim$ Bernoulli(θ). *Then* $T = X_1$ *is* **not** *sufficient. Look at its partition:*

(x_1, x_2, x_3)		t	$p(x\|t)$
(0, 0, 0)	$\to$	$t = 0$	$(1-\theta)^2$
(0, 0, 1)	$\to$	$t = 0$	$\theta(1-\theta)$
(0, 1, 0)	$\to$	$t = 0$	$\theta(1-\theta)$
(0, 1, 1)	$\to$	$t = 0$	θ^2
(1, 0, 0)	$\to$	$t = 1$	$(1-\theta)^2$
(1, 0, 1)	$\to$	$t = 1$	$\theta(1-\theta)$
(1, 1, 0)	$\to$	$t = 1$	$\theta(1-\theta)$
(1, 1, 1)	$\to$	$t = 1$	θ^2
8 elements	$\to$	*2 elements*	

11.4 The Factorization Theorem

Checking the definition of sufficiency directly is often a tedious exercise since it involves computing the conditional distribution. A much simpler characterization of sufficiency comes from what is called the Neyman-Fisher factorization criterion.

Theorem 11.3 *$T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint pdf/pmf of $(X_1, \ldots, X_n)$ can be factored as*

$$p(x_1, \ldots, x_n; \theta) = h(x_1, \ldots, x_n) \times g(T(x_1, \ldots, x_n); \theta).$$

This version does not involve conditioning and thus typically makes sense even when X has a continuous distribution. Let us consider an example of this.

Example 11.4 *$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Then*

$$p(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}.$$

(a) If (μ, σ^2) unknown then we can write the joint as:

$$p(x_1, \ldots, x_n; \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} x_i + n\mu^2\right)\right\},$$

so that $T = (\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$ is sufficient using the factorization theorem.

(b) If σ is known: we can add and subtract the sample mean $\overline{x}$ in the exponent and use the fact that, $\sum_{i=1}^n (x_i - \overline{x}) = 0$ to see that,

$$p(x_1, \ldots, x_n; \mu) = \underbrace{\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{\frac{-\sum(x_i - \overline{x})^2}{2\sigma^2}\right\}}_{h(x^n)} \quad \underbrace{\exp\left\{\frac{-n(\overline{x} - \mu)^2}{2\sigma^2}\right\}}_{g(T(x^n)|\mu)}.$$

Thus, using the factorization theorem $\overline{X}$ is sufficient for μ.

The factorization theorem is relatively straightforward to prove, at least in the discrete case.

Proof: Factorization $\implies$ sufficiency:

$$\begin{aligned} p(x_1, \ldots, x_n | T = t; \theta) &= \frac{p(x_1, \ldots, x_n, T = t; \theta)}{p(T = t; \theta)} \\ &= \frac{\mathbb{I}(T(x_1, \ldots, x_n) = t) h(x_1, \ldots, x_n) \times g(t; \theta)}{\sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} h(x_1, \ldots, x_n) \times g(t; \theta)} \\ &= \frac{\mathbb{I}(T(x_1, \ldots, x_n) = t) h(x_1, \ldots, x_n)}{\sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} h(x_1, \ldots, x_n)}, \end{aligned}$$

which does not depend on θ.

Sufficiency $\implies$ factorization: We simply define $g(t; \theta) = p(T = t; \theta)$, and $h(x_1, \ldots, x_n) = p(x_1, \ldots, x_n | T = t; \theta)$, where by sufficiency we note that the latter function does not depend on θ. Now, it is straightforward to verify that factorization theorem holds.

11.5 The likelihood and its relationship to sufficiency

The likelihood function arises from viewing the joint density as a function of the unknown parameter θ, i.e.

$$\mathcal{L}(\theta) = \mathcal{L}(\theta; x_1, \ldots, x_n) = p(x_1, \ldots, x_n; \theta).$$

The likelihood is central in computing point estimates (maximum likelihood estimation) and Bayesian inference. We will return to these.

Some important points to remember:

1. The likelihood is a function of θ, it is not a probability.
2. Typically we ignore constants that do not depend on θ when computing the likelihood, i.e. we care only about the relative value of the likelihood as a function of θ. More formally, the likelihood is only defined upto a constant of proportionality, i.e. it is an equivalence class of functions.
3. The likelihood in the i.i.d. case has the form:

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} p(x_i; \theta),$$

and in this case we often will find it convenient to work with the log-likelihood:

$$\mathcal{LL}(\theta) = \sum_{i=1}^{n} \log p(x_i; \theta).$$

Relationship to sufficiency: Using the factorization theorem we can see that for any sufficient statistic T, we can write the likelihood as:

$$\mathcal{L}(\theta) = g(T(x_1, \ldots, x_n); \theta) h(x_1, \ldots, x_n),$$

and noting once again that the likelihood is only defined upto constants that do not depend on θ, we can simply ignore the $h(x_1, \ldots, x_n)$ term, i.e. we can define the likelihood as:

$$\mathcal{L}(\theta) = g(T(x_1, \ldots, x_n); \theta).$$

Thus, once we have any sufficient statistic, we have everything we need to compute the likelihood function. If we were to base our subsequent data analysis only on the likelihood function then we have lost nothing. We will see a further connection between sufficiency and the likelihood function once we define the notion of minimal sufficiency.