Math 502AB - Lecture 21

Dr. Jamshidian

November 6, 2017

1 Lecture - Part 1

1.1 Factorization Theorem, Cont'd

1.1.1 Examples

1. Suppose that $X_1, ..., X_n$ is a sample from a uniform distribution on $[0, \theta]$. That is

$$f(X|\theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

Obtain a sufficient statistic for θ.

First, we will come up with the distribution,

$$\begin{aligned} f_n(x|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \left(\frac{1}{\theta} \mathcal{I}_{0<x_i<\theta} \right) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathcal{I}_{0<x_i<\theta} \end{aligned}$$

We note that,

$$\prod_{i=1}^{n} \mathcal{I}_{0<x_i<\theta} = 1 \iff 0 < x_1 < \theta, ..., 0 < x_n < \theta \iff max(x_1, ..., x_n) < \theta$$

Thus, we have,

$$f_n(x|\theta) = \frac{1}{\theta^n} \mathcal{I}_{[0,\theta]}(\max x_i)$$

And this implies that $\max X_i$ is our sufficient statistic.

2. Let $X_1, ..., X_n$ be a sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are to be estimated. Obtain a sufficient statistic for this estimation. Let's try to apply the factorization theorem again:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x_i - \mu) \right\} = \left(\frac{1}{\sqrt{2\pi}\sigma} \right)^n \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

We can factor the right hand side as:

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(x_i-\overline{x})^2+n(\overline{x}-\mu)^2\right]\right\}$$

Now, consider the exp side. What are the statistics you can see here? $\overline{x}$ is one of them. We notice that this is a function:

$$g\left(\sum(x_i-\overline{x})^2, \overline{x}|\mu,\sigma^2\right)$$

In other words, g is jointly sufficient on *both* $\overline{x}$ as well as $\sum(x_i-\overline{x})$.

1.2 Theorem: Sufficient Statistics for Exponential Family

Let $X_1, ..., X_n$ be a sample from an exponential family of distributions with

$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}$$

Where $\theta = (\theta_1, ..., \theta_d)$ with $d \leq k$. Then we have,

$$T(X) = \left[\sum_{i=1}^{n} t_1(x_i), \sum_{i=1}^{n} t_2(x_i), ..., \sum_{i=1}^{n} t_k(x_i)\right]$$

Example

1. Let $X_1, ..., X_n \sim N(\theta_1, \theta_2)$ (**iid**). Then,

$$f(x|\theta_1,\theta_2) = \exp\left\{-\frac{1}{\theta_2}x^2 + \frac{\theta_1}{\theta_2}x - \frac{\theta_1^2}{2\theta_2} - \log\sqrt{2\pi\theta_2}\right\}$$

Based on this theorem, what are the sufficient statistics? $\sum_{i=1}^{n} X_i^2$ and $\sum_{i=1}^{n} X_i$. But earlier we got two *different* sufficient statistics. How? Well, first, sufficient statistics are **not** unique. However, there is also a *bijective* function between these statistics.

1.3 Minimal Sufficient Statistics

A *minimal sufficient statistic* is a statistic that has all of the required information about θ, and can not be reduced further.

1.3.1 Definition:

$T(X)$ is a minimal sufficient statistic if, for every sufficient statistic $T'(X)$, $T'(X)$ is a function of $T(X)$.

1.3.2 Definition

Let $f_T(t|\theta)$ be a family of *pdf*'s (or *pmf*'s) for a statistic $T(X)$. A family of probability distributions is called **complete** if $E[g(T)] = 0$ implies that $P(g(T) = 0) = 1$ for all θ, and a function g.

Example

1. Let $X_1, ..., X_n \sim Bernoulli(p)$ (**iid**). Then we know that $T = \sum X_i$ is a sufficient statistic for $0 < p < 1$. To show that T is the minimal sufficient statistic, we want to show that the distribution of T is **complete**. We know that
$$T \sim Binomial(n, p)$$
To show that $Binomial(n, p)$ is complete, we need to show that for a function g
$$E[g(T)] = 0 \Rightarrow P(g(T) = 0) = 1$$
So, we have,
$$\begin{aligned}
0 = E[g(T)] &= \sum_{t=0}^{n} g(t)\binom{n}{t}p^t(1-p)^{n-t} \\
&= (1-p)^n \sum_{t=0}^{n} g(t)\binom{n}{t}\left(\frac{p}{1-p}\right)^t \\
\iff \sum_{t=0}^{n} g(t)\binom{n}{t}r^t &= 0 \quad r = \left(\frac{p}{1-p}\right)
\end{aligned}$$
We know that this term is a polynomial of degree n in r. This is zero **iff** all of the coefficients are equal to zero. That is,
$$g(t)\binom{n}{t} = 0 \iff g(t) = 0 \iff P(g(t) = 0) = 1$$
And, since $g(t) = 0$ for **all** values of t, this implies that $P(g(t) = 0) = 1$.

1.3.3 Theorem (Complete Sufficient Statistics for Exponential Families)

Let $X_1, ..., X_n$ be **iid** from an exponential family of distributions with *pdf* (or *pmf*)
$$f(x|\theta) = h(x)c(\theta)\exp\left\{\sum_{i=1}^{k} w_i(\theta)t_i(x)\right\}$$
with $\theta = (\theta_1, ..., \theta_k)$. Then the statistic
$$T(X) = \left[\sum_{i=1}^{n} t_1(x_i), \sum_{i=1}^{n} t_2(x_i), ..., \sum_{i=1}^{n} t_k(x_i)\right]$$

is **complete** if

$$\{w_1(\theta), ..., w_k(\theta) : \theta \in \Theta\}$$

is an open set in $\mathbb{R}^k$

1.3.4 Theorem

If a *minimal sufficient statistic* exists, then any *complete sufficient statistic* is a *minimal sufficient statistic.*

1.3.5 Theorem

Let $f(x|\theta)$ be the *pdf* (or *pmf*) for a sample X. Suppose there exists $T(X)$ such that, for every two sample points x and y, the ratio $\frac{f(x|\theta)}{f(y|\theta)}$ is constant as a function of θ **if and only if** $T(X) = T(Y)$, then $T(X)$ is a *minimal sufficient statistic.*

Examples:

1. Let $X_1, ..., X_n \sim \textit{Bernouli}(\theta)$. $T = \sum X_i$ is a sufficient statistic. We already know that this is a minimal sufficient statistic since it is a binomial random variable. However, suppose we didn't know. By this theorem, we consider:

$$\begin{aligned}\frac{f(x|\theta)}{f(y|\theta)} &= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n y_i}}{\theta^{\sum_{i=1}^n y_i}(1-\theta)^{n-\sum_{i=1}^n y_i}} \\ &= \theta^{\sum x_i - \sum y_i}(1-\theta)^{\sum y_i - \sum x_i} \\ &= 1 \iff \sum x_i = \sum y_i\end{aligned}$$

1.4 Chapter 7: Point Estimation

Examples:

1. X = the number of accidents on the 57 freeway each year
2. X = the amount of time that it takes for a randomly selected student to get to school

A sample constitutes **iid** observations from X which we denote by $X_1, ..., X_n$. We assume that

$$X \sim f_X(x|\theta)$$

Our aim is to estimate θ based on a sample $X_1, ..., X_n$. We must understand that there are two different things we are dealing with

1. **Estimator:** $T(X_1, ..., X_n)$. This is a random variable
2. **Estimate:** $T(x_1, ..., x_n)$. This is based on an observation. In other words, once we take a sample, we plug into an estimator to obtain an estimate.

2 Lecture - Part 2

2.1 Section 7.2.1 - Method of Moments

Let $X_1, ..., X_n$ be a sample from a population with *pdf* (or *pmf*) $f(X|\theta_1, ..., \theta_k)$. Method of moment estimators are obtained by equating k sample moments (usually the first k) to their corresponding population moments, and solving for the parameters $\theta_1, ..., \theta_k$. Specifically, the k^{th} sample moment is

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_1^k$$

and the k^{th} population moment is a function of θ, as follows:

$$\mu'_k = E[X_1^k]$$

2.1.1 Examples

1. Let $X_1, ..., X_n \sim Poisson(\lambda)$. We then have $E[X_1] = \lambda$. One way to estimate this moment is to look at

$$\mu'_1 = m_1$$
$$\lambda = \frac{1}{n}\sum_{i=1}^{n} x_1$$

   Another way is to consider $E[X_1^2] = \lambda + \lambda^2$. This yields,

$$\lambda^2 + \lambda = \frac{1}{n}\sum_{i=1}^{n} x_1^2$$

   Thus, we can see that these estimators are *not unique*.

2. Suppose that $X_1, ..., X_n \sim N(\theta, \sigma^2)$ (**iid**). Estimate θ and σ^2 by *method of moments.*

$$\mu'_1 = E[X_1] = \theta \quad \mu'_2 = E[X_1^2] = \theta^2 + \sigma^2$$
$$m_1 = \frac{1}{n}\sum x_i = \overline{x} \quad m_2 = \frac{1}{n}\sum x_i^2$$

   Thus we have,

$$\begin{cases} \widetilde{\theta} = \overline{X} \\ \widetilde{\theta}^2 + \widetilde{\sigma}^2 = \frac{1}{n}\sum x_i^2 \Rightarrow \widetilde{\sigma}^2 = \frac{1}{n}\sum x_i^2 - (\overline{x})^2 \end{cases}$$

   Note: Here, we have introduced the notation of $\widetilde{\theta}$ as a *method of moments* estimator.

3. Suppose that $X_1, ..., X_n \sim Binomial(k, p)$. Use *method of moments* to estimate both k and p. This is a strange problem, since in the past we always knew k making p easy to figure out.

$$E[X] = kp$$
$$E[X^2] = kp(1-p) + k^2p^2$$

This gives us

$$\begin{cases} \widetilde{k}\widetilde{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} \\ \widetilde{k}\widetilde{p}(1-\widetilde{p}) + \widetilde{k}^2\widetilde{p}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

Combining, we have:

$$\begin{aligned} \widetilde{k}\widetilde{p} - \widetilde{k}\widetilde{p}^2 + \widetilde{k}^2\widetilde{p}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \overline{X} - \frac{\overline{X}^2}{\widetilde{k}} + \overline{X}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\ \widetilde{k}\left(\overline{X} + \overline{X}^2 - \frac{1}{n}\sum_{i=1}^{n} X_i^2\right) &= \overline{X}^2 \\ \widetilde{k} = \frac{\overline{X}^2}{\overline{X} + \overline{X}^2 - \frac{1}{n}\sum X_i^2} &= \frac{\overline{X}^2}{\overline{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{aligned}$$

This is an example of where a *method of moments* estimator might not be the best option since the denominator might be negative giving a nonsensical estimate.

4. (Estimating Population Size) Consider a population, labeled 1 to θ (thus, the population has θ members). A sample of size n with replacement is taken from this population and $X_1, ..., X_n$ is the lable for the members in the sample. Obtain an estimate of θ using *method of moments*.

First, we must talk about the distribution of X_i. This is a *discrete uniform distribution*:

$$f_X(x) = \begin{cases} \frac{1}{\theta} & X = 1, 2, ..., \theta \\ 0 & \text{otherwise} \end{cases}$$

We now have:

$$E[X_1] = \sum_{x=1}^{\theta} \frac{1}{\theta} X = \frac{1}{\theta} \cdot \frac{\theta(\theta+1)}{2} = \frac{\theta+1}{2}$$

This yields

$$\frac{\widetilde{\theta}+1}{2} = \overline{X} \Rightarrow \widetilde{\theta} = 2\overline{X} - 1$$

Where might this estimator go wrong? Well consider the situation where $X_1 = 1$, $X_2 = 2$, $X_3 = 9$. This yields $\overline{x} = 4$ and $\widetilde{\theta} = 7$. But it can't be 7; we observed someone with 9!

5. (Satterthwaite Approximation) Let $Y_i \sim \chi^2_{(r_i)}$ for $i = 1, ..., k$. Then, $\sum Y_i \sim \chi^2_{(\sum r_i)}$. Let $a_1, ..., a_k$ be given constants. Then the distribution of $\sum a_i Y_i$ is *not* tractable.

   Satterthwaite was interested in approximating a degrees of freedom ν such that
   $$\sum_{i=1}^{k} a_i Y_i \sim \frac{\chi^2(\nu)}{\nu}$$
   First, let's try equating first moments
   $$E\left[\sum_{i=1}^{k} a_i Y_i\right] = \sum_{i=1}^{k} a_i E(Y_i) = \sum_{i=1}^{k} a_i r_i$$
   $$E\left(\frac{\chi^2_\nu}{\nu}\right) = \frac{1}{\nu} E[\chi^2_{(\nu)}] = 1 \Rightarrow \sum_{i=1}^{k} a_i r_i = 1$$
   This gives us no information about ν, so this doesn't help us at all. So, let's try the second moment:
   $$E\left(\sum_{i=1}^{k} a_i Y_i\right)^2 = E\left(\frac{\chi^2_{(\nu)}}{\nu}\right)^2$$
   $$E\left(\frac{\chi^2_{(\nu)}}{\nu}\right)^2 = \frac{1}{\nu^2}\left[var\left(\chi^2_{(\nu)}\right) + E\left[\chi^2_{(\nu)}\right]^2\right]$$
   $$= \frac{1}{\nu^2}\left[2\nu + \nu^2\right] = \frac{2}{\nu} + 1$$
   $$\Rightarrow \left(\sum a_i Y_i\right)^2 = \frac{2}{\nu} + 1$$
   $$\hat{\nu} = \frac{2}{\left(\sum (a_i Y_i)\right)^2 - 1}$$

2.2 Method of Maximum Likelihood

Suppose that random variables $X_1, ..., X_n$ have a joint density $f(x_1, ..., x_n|\theta)$. Given the observed values $X_i = x_i$ for $i = 1, ..., n$, the *likelihood* of θ as a function of $x_1, ..., x_n$ is defined by

$$\mathcal{L}(\theta) = f(x_1, ..., x_n|\theta)$$

2.2.1 Examples:

1. Let θ be the probability that a coin comes up heads. Let X_1 and X_2 be the number of heads in two independent trials of 3 flips of a coin, respectively. Here, our sample size is 2.

Suppose we observed $X_1 = 2$ and $X_2 = 0$. The likelihood for this event is

$$\begin{aligned} P(X_1 = 2, X_2 = 2) &= P(X_1 = 2)P(X_2 = 0), \text{ by independence} \\ &= \binom{3}{2}\theta^2(1-\theta)\binom{3}{0}\theta^0(1-\theta)^3 \\ &= 3\theta^2(1-\theta)^4 \end{aligned}$$

To obtain *MLE* for θ, we maximize

$$\begin{aligned} \mathcal{L}(\theta) &= 3\theta^2(1-\theta)^4 \\ \log\mathcal{L}(\theta) &= \log 3 + 2\log\theta + 4\log(1-\theta) \\ \frac{d}{d\theta}\log\mathcal{L}(\theta) &= \frac{2}{\hat{\theta}} - \frac{4}{1-\hat{\theta}} = 0 \Rightarrow \hat{\theta} = \frac{1}{3} \end{aligned}$$

If $X_1, ..., X_n \sim f_{X_1}(x_1|\theta)$, then the *likelihood* is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta)$$

MLE is obtained by maximizing $\mathcal{L}(\theta)$, provided that a maximum exists. Often, it will be easier to maximize the log likelihood

$$\ell(\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i|\theta)$$