Lecture Notes 5

1 Statistical Models

(Chapter 6.) A **statistical model** $\mathcal{P}$ is a collection of probability distributions (or a collection of densities). Examples of **nonparametric models** are

$$\mathcal{P} = \left\{ p : \int (p''(x))^2 dx < \infty \right\}, \quad \text{and} \quad \mathcal{P} = \left\{ \text{all distributions on } \mathbb{R}^d \right\}.$$

A **parametric model** has the form

$$\mathcal{P} = \left\{ p(x;\theta) : \theta \in \Theta \right\}$$

where $\Theta \subset \mathbb{R}^d$. An example is the set of Normal densities $\{p(x;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}\}$ where $\theta = (\mu, \sigma)$.

For now, we focus on parametric models. Later we consider nonparametric models.

2 Statistics

Let $X_1, \ldots, X_n \sim p(x;\theta)$. Let $X^n \equiv (X_1, \ldots, X_n)$. Any function $T = T(X_1, \ldots, X_n)$ is itself a random variable which we will call a *statistic*.

Some examples are:

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: $X_{(n)}$.

3 Sufficiency

We continue with **parametric inference**. In this section we discuss **data reduction** as a formal concept.

3.1 Sufficient Statistics

Suppose that $X_1, \ldots, X_n \sim p(x;\theta)$. T is **sufficient** for θ if the conditional distribution of $X_1, \ldots, X_n | T$ does not depend on θ. Thus, $p(x_1, \ldots, x_n|t;\theta) = p(x_1, \ldots, x_n|t)$.

Intuitively, this means that you can replace $X_1, \ldots, X_n$ with $T(X_1, \ldots, X_n)$ without losing information. (This is not quite true as we'll see later. But for now, you can think of it this way.)

Notation: In what follows we use the following notation:

$$X^n \equiv (X_1, \ldots, X_n), \quad x^n \equiv (x_1, \ldots, x_n).$$

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

$$p_{X^n|T}(x^n|t) = \mathbb{P}(X^n = x^n | T(X^n) = t) = \frac{P(X^n = x^n \text{ and } T = t)}{P(T = t)}.$$

But

$$P(X^n = x^n \text{ and } T = t) = \begin{cases} 0 & T(x_1, \ldots, x_n) \neq t \\ P(X_1 = x_1, \ldots, X_n = x_n) & T(x_1, \ldots, x_n) = t. \end{cases}$$

Hence,

$$P(X^n = x^n) = \prod_{i=1}^n \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\prod(x_i!)} = \frac{e^{-n\theta}\theta^t}{\prod(x_i!)}.$$

Now, $T(x^n) = \sum x_i = t$ *and so*

$$P(T = t) = \frac{e^{-n\theta}(n\theta)^t}{t!} \qquad \text{since } T \sim \text{Poisson}(n\theta).$$

Thus,

$$\frac{P(X^n = x^n)}{P(T = t)} = \frac{t!}{(\prod x_i)! n^t}$$

which does not depend on θ. *So* $T = \sum_i X_i$ *is a sufficient statistic for* θ. *Other sufficient statistics are:* $T = 3.7 \sum_i X_i$, $T = (\sum_i X_i, X_4)$, *and* $T(X_1, \ldots, X_n) = (X_1, \ldots, X_n)$.

3.2 Sufficient Partitions

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

Example 2 *Let* $X_1, X_2, X_3 \sim \text{Bernoulli}(\theta)$. *Let* $T = \sum X_i$.

x^n		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 3 *Let* $X_1, X_2, X_3 \sim \text{Bernoulli}(\theta)$. *Then* $T = X_1$ *is* **not** *sufficient. Look at its partition:*

x^n		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*	

3.3 The Factorization Theorem

Theorem 4 *$T(X^n)$ is sufficient for θ if the joint pdf/pmf of X^n can be factored as*

$$p(x^n;\theta) = h(x^n) \times g(t;\theta).$$

Example 5 *Let $X_1, \cdots, X_n \sim$ Poisson. Then*

$$p(x^n;\theta) = \frac{e^{-n\theta}\theta^{\sum X_i}}{\prod(x_i!)} = \frac{1}{\prod(x_i!)} \times e^{-n\theta}\theta^{\sum_i X_i}.$$

Example 6 *$X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. Then*

$$p(x^n;\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{\sum(x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right\}.$$

(a) If σ known:

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

Thus, $\overline{X}$ is sufficient for μ.

(b) If (μ,σ^2) unknown then $T = (\overline{X}, S^2)$ is sufficient. So is $T = (\sum X_i, \sum X_i^2)$.

3.4 Minimal Sufficient Statistics (MSS)

We want the greatest reduction in dimension.

Example 7 *$X_1, \cdots, X_n \sim N(0, \sigma^2)$. Some sufficient statistics are:*

$$\begin{aligned}
T(X_1,\cdots,X_n) &= (X_1,\cdots,X_n) \\
T(X_1,\cdots,X_n) &= (X_1^2,\cdots,X_n^2) \\
T(X_1,\cdots,X_n) &= \left(\sum_{i=1}^{m} X_i^2, \sum_{i=m+1}^{n} X_i^2\right) \\
T(X_1,\cdots,X_n) &= \sum X_i^2.
\end{aligned}$$

> T is a **Minimal Sufficient Statistic** if the following two statements are true:
>
> 1. T is sufficient and
> 2. If U is any other sufficient statistic then $T = g(U)$ for some function g.

In other words, T generates the **coarsest sufficient partition**.

Suppose U is sufficient. Suppose $T = H(U)$ is also sufficient. T provides greater reduction than U unless H is a $1-1$ transformation, in which case T and U are equivalent.

Example 8 $X \sim N(0, \sigma^2)$. *X is sufficient. $|X|$ is sufficient. $|X|$ is MSS. So are* X^2, X^4, e^{X^2}.

Example 9 *Let* $X_1, X_2, X_3 \sim \text{Bernoulli}(\theta)$. *Let* $T = \sum X_i$.

x^n		t	$p(x\|t)$	u	$p(x\|u)$
(0, 0, 0)	$\to$	$t=0$	*1*	$u=0$	*1*
(0, 0, 1)	$\to$	$t=1$	*1/3*	$u=1$	*1/3*
(0, 1, 0)	$\to$	$t=1$	*1/3*	$u=1$	*1/3*
(1, 0, 0)	$\to$	$t=1$	*1/3*	$u=1$	*1/3*
(0, 1, 1)	$\to$	$t=2$	*1/3*	$u=73$	*1/2*
(1, 0, 1)	$\to$	$t=2$	*1/3*	$u=73$	*1/2*
(1, 1, 0)	$\to$	$t=2$	*1/3*	$u=91$	*1*
(1, 1, 1)	$\to$	$t=3$	*1*	$u=103$	*1*

Note that U and T are both sufficient but U is not minimal.

3.5 How to find a Minimal Sufficient Statistic

Theorem 10 *Define*

$$R(x^n, y^n; \theta) = \frac{p(y^n; \theta)}{p(x^n; \theta)}.$$

Suppose that T has the following property:

> $R(x^n, y^n; \theta)$ **does not depend on θ if and only if** $T(y^n) = T(x^n)$**.**

Then T is a MSS.

Example 11 *$Y_1, \cdots, Y_n$ iid Poisson (θ).*

$$p(y^n;\theta) = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i}, \quad \frac{p(y^n;\theta)}{p(x^n;\theta)} = \frac{\theta^{\sum y_i - \sum x_i}}{\prod y_i!/\prod x_i!}$$

which is independent of θ iff $\sum y_i = \sum x_i$. This implies that $T(Y^n) = \sum Y_i$ is a minimal sufficient statistic for θ.

The minimal sufficient statistic is not unique. But, the minimal sufficient partition is unique.

Example 12 *Cauchy.*

$$p(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

Then

$$\frac{p(y^n;\theta)}{p(x^n;\theta)} = \frac{\prod_{i=1}^{n}\{1+(x_i-\theta)^2\}}{\prod_{j=1}^{n}\{1+(y_j-\theta)^2\}}.$$

The ratio is a constant function of θ if

$$T(Y^n) = (Y_{(1)}, \cdots, Y_{(n)}).$$

It is technically harder to show that this is true only if T is the order statistics, but it could be done using theorems about polynomials. Having shown this, one can conclude that the order statistics are the minimal sufficient statistics for θ.

4 What Sufficiency Really Means

If T is sufficient, then T contains all the information you need from the data to compute the **likelihood function**. It does not contain all the information in the data. We will define the likelihood function shortly.