LECTURE 1 NOTES

1. Statistical models. A *statistical model* $\mathcal{F}$ is a set of probability distributions (or a set of densities). A *parametric model* is one that is parametrized by a finite number of parameters. That is, a parametric model has the form

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}.$$

The set $\Theta \subset \mathbf{R}^p$ is the *parameter space*. For now, we focus on such parametric models. Later, we consider non-parametric models.

At a abstract level, the goal of a statistical investigation is to learn about a random variable from observations. An investigator begins by choosing a parametric model $\mathcal{F}$. The choice is usually based on domain knowledge, although it is also influenced by other (such as computational) concerns. For example, suppose an investigator who wishes to study the probability a coin comes up heads may choose the Bernoulli model:

$$\mathcal{F} = \{\text{Ber}(p) : p \in (0, 1)\},$$

where $\text{Ber}(p)$ is the Bernoulli distribution with mean p.

The investigator then collects data by observing the random variable $\mathbf{x}$ taking values in a sample space $\mathcal{X}$. Usually, the observations consists of independent observations of the same random variable: $\mathbf{x} = \{\mathbf{x}_i\}_{i\in[n]}$. In the coin study, the observations may consist of the outcomes of n tosses of the coin.

EXAMPLE 1.1 (Location-scale model). *Let* $\mathbf{x}$ *be a real-valued random variable, and let* $f(x)$ *be its density. The location-scale model of* $\mathbf{x}$ *consists of densities of the form* $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$*, where* $\mu \in \mathbf{R}$ *is the* location parameter *and* $\sigma > 0$ *is the* scale parameter*. Intuitively, the location parameter translates the density, while the scale parameter disperses or concentrates the density.*

The choice of a model is subjective. However, it is in some sense necessary. By choosing a model, the investigator introduces prior knowledge about the data generating process. There is a theorem, called the *No Free Lunch* theorem, which in essence says that without any prior knowledge of the data generating process, statistical inference is impossible. In pseudo-mathematical terms:

$$\text{Inference} = \text{data} + \text{knowledge},$$

hence the need to introduce prior knowledge by choosing a model.

A *statistic* is a function of the data that summarizes the data. When the function is not invertible (and it usually isn't), the summary is lossy. Thus statistics are a form of *data reduction*. Formally, a statistic $\mathbf{t} := \phi(\mathbf{x})$ for some function $\phi : \mathcal{X} \to \mathcal{T}$ partitions the sample space $\mathcal{X}$ into its pre-images:

$$\mathcal{X} = \bigcup_t \mathcal{A}_t, \text{ where } \mathcal{A}_t := \{x \in \mathcal{X} : \phi(\mathbf{x}) = t\}.$$

Thus observing $\{\phi(\mathbf{x}) = t\}$ informs the investigator of that $\mathbf{x} \in \mathcal{A}_t$, but not the exact value of $\mathbf{x}$. Some examples of statistics are:

1. sample mean: $\bar{\mathbf{x}} := \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i$,
2. sample variance: $\mathbf{s}^2 = \frac{1}{n-1}\sum_{i\in[n]}(\mathbf{x}_i - \bar{\mathbf{x}})$,
3. order statistics: $\mathbf{x}_{(i)} \leq \cdots \leq \mathbf{x}_{(n)}$.

In the coin tossing example, a relevant statistic may be the fraction of times the coin came up heads:

$$\bar{\mathbf{x}}_n := \frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i. \tag{1.1}$$

It is a form of data reduction because many sequences of heads and tails give the same fraction. Since $\mathbf{x}$ is stochastic, any function of $\mathbf{x}$ is also a random variable, and its distribution is called its *sampling distribution*.

EXAMPLE 1.2. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$. *Then* $\sum_{i\in[n]} \mathbf{x}_i \sim \text{bin}(n, p)$. *Indeed, the moment-generating function (MGF) of* $n\bar{\mathbf{x}}$ *is*

$$\mathbf{E}\left[e^{t\sum_{i\in[n]} \mathbf{x}_i}\right] = \mathbf{E}\left[\prod_{i\in[n]} e^{t\mathbf{x}_i}\right].$$

Since $\{\mathbf{x}_i\}_{i\in[n]}$ *are* i.i.d.*,*

$$= \prod_{i\in[n]} \mathbf{E}\left[e^{t\mathbf{x}_i}\right] = \left(\mathbf{E}\left[e^{t\mathbf{x}_1}\right]\right)^n.$$

The MGF of a $\text{Ber}(p)$ *random variable is* $1 - p + pe^t$. *Thus the MGF of* $n\bar{\mathbf{x}}_n$ *is* $\left(1 - p + pe^t\right)^n$, *which is the MGF of a* $\text{bin}(n, p)$ *random variable.*

EXAMPLE 1.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \Sigma)$. *Then* $\bar{\mathbf{x}} \sim \mathcal{N}\big(\mu, \frac{1}{n}\Sigma\big)$. *Indeed, the MGF of* $\frac{1}{n}\sum_{i\in[n]} \mathbf{x}_i$ *is*

$$\mathbf{E}\left[e^{\frac{1}{n}\sum_{i\in[n]} t^T\mathbf{x}_i}\right] = \mathbf{E}\left[\prod_{i\in[n]} e^{\frac{1}{n}t^T\mathbf{x}_i}\right] = \prod_{i\in[n]} \mathbf{E}\left[e^{\frac{1}{n}t^T\mathbf{x}_i}\right] = \left(\mathbf{E}\left[e^{\frac{1}{n}t^T\mathbf{x}_1}\right]\right)^n.$$

The MGF of a $\mathcal{N}(\mu, \Sigma)$ random variable is $e^{t^T\mu+\frac{1}{2}t^T\Sigma t}$. Thus the MGF of $\bar{\mathbf{x}}_n$ is

$$\left(e^{\frac{1}{n}t^T\mu+\frac{1}{2n^2}t^T\Sigma t}\right)^n = e^{t^T\mu+\frac{1}{2}t^T\left(\frac{1}{n}\Sigma\right)t},$$

which is the MGF of a $\mathcal{N}(\mu, \frac{1}{n}\Sigma)$ random variable.

EXAMPLE 1.4. *If $\mathbf{x} \sim \mathcal{N}(0,1)$, then $\mathbf{x}^2 \sim \chi^2_1$. The MGF of $\mathbf{x}^2$ is*

$$\begin{aligned}\mathbf{E}\left[e^{t\mathbf{x}^2}\right] &= \int_{\mathbf{R}} \frac{1}{\sqrt{2\pi}} e^{\left(t-\frac{1}{2}\right)x^2} dx \\ &= \int_{\mathbf{R}} \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{\left(t-\frac{1}{2}\right)^{-1}}} dx.\end{aligned}$$

The integrand is "almost" the density of a $\mathcal{N}\left(0, \left(t-\frac{1}{2}\right)^{-1}\right)$ random variable. To make it so, we multipy and divide by $\left(t-\frac{1}{2}\right)^{-1/2}$ to obtain

$$\begin{aligned}&= \left(t-\tfrac{1}{2}\right)^{-1/2} \int_{\mathbf{R}} \frac{1}{\sqrt{2\pi}\left(t-\frac{1}{2}\right)^{-1/2}} e^{\frac{x^2}{\left(t-\frac{1}{2}\right)^{-1}}} dx \\ &= \left(t-\tfrac{1}{2}\right)^{-1/2},\end{aligned}$$

which is the MGF of a χ^2_1 random variable. If $\mathbf{x} \sim \mathcal{N}(0, I_k)$, $\|\mathbf{x}\|_2^2 \sim \chi^2_k$:

$$\mathbf{E}\left[e^{t\|\mathbf{x}\|_2^2}\right] = \mathbf{E}\left[e^{t\sum_{i\in[k]}\mathbf{x}_i^2}\right] = \mathbf{E}\left[\prod_{i\in[k]} e^{t\mathbf{x}_i^2}\right].$$

Since $\{\mathbf{x}_i\}_{i\in[n]}$ are i.i.d.*,*

$$= \prod_{i\in[n]} \mathbf{E}\left[e^{t\mathbf{x}_i^2}\right] = \left(t-\tfrac{1}{2}\right)^{-n/2}.$$

EXAMPLE 1.5. *If $\mathbf{x}_j \overset{\text{i.i.d.}}{\sim} \sim \text{Cauchy}(a,b)$, then $\bar{\mathbf{x}} \sim \text{Cauchy}(a,b)$. The Cauchy distribution is* heavy-tailed*; its MGF is undefined. However, its characteristic function is well-defined. Thus*

$$\mathbf{E}\left[e^{i\frac{t}{n}\sum_{j\in[n]}\mathbf{x}_j}\right] = \mathbf{E}\left[\prod_{j\in[n]} e^{i\frac{t}{n}\mathbf{x}_j}\right].$$

Since $\{\mathbf{x}_j\}_{j\in[n]}$ are i.i.d.*,*

$$\begin{aligned}&= \prod_{j\in[n]} \mathbf{E}\left[e^{i\frac{t}{n}\mathbf{x}_j}\right] = \left(\mathbf{E}\left[e^{\frac{t}{n}\mathbf{x}_1}\right]\right)^n \\ &= \left(e^{i\mu\frac{t}{n}-\sigma\left|\frac{t}{n}\right|}\right)^n = e^{i\mu t-\sigma|t|},\end{aligned}$$

which is the characteristic function of a Cauchy(a,b) *random variable.*

TABLE 1
The conditional distribution of three coin tosses given the number of heads $\mathbf{t} = \mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3$

$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2)$	$\mathbf{t}$	$f_p(\mathbf{x} \mid \mathbf{t})$
(0,0,0)	0	1
(1,0,0)	1	1/3
(0,1,0)	1	1/3
(0,0,1)	1	1/3
(1,1,0)	2	1/3
(0,1,1)	2	1/3
(1,0,1)	2	1/3
(1,1,1)	3	1

2. Sufficiency. In Section 1, we saw that statistics are a form of data reduction. In the rest of these notes, we formalize the notion of data reduction.

DEFINITION 2.1 (sufficient statistic). *Let* $\mathbf{x} \sim F \in \mathcal{F}$. *A statistic* $\mathbf{t} := \phi(\mathbf{x})$ *is a sufficient statistic for the parametric model* $\mathcal{F}$ *if the conditional distribution of* $\mathbf{x}$ *given* $\mathbf{t}$ *does not depend on the parameter* θ.

A sufficient statistic $\mathbf{t}$ summarizes all the information in $\mathbf{x}$ about θ. That is, as long as the goal is to learn about θ, observing a sufficient statistic $\mathbf{t}$ is as good as observing $\mathbf{x}$. Why? Consider two investigators: Alice and Bob. Alice observes $\mathbf{x}$, while Bob observes $\mathbf{t}$. Since $\mathbf{t}$ is a sufficient statistic, the conditional distribution of $\mathbf{x} \mid \mathbf{t}$ does not depend on θ, so Bob can draw a new sample $\mathbf{x}'$ from the conditional distribution. Further, $\mathbf{x}$ and $\mathbf{x}'$ are (unconditionally) identically distributed:

$$\begin{aligned} \mathbf{E}\left[g(\mathbf{x})\right] &= \mathbf{E}\left[\mathbf{E}\left[g(\mathbf{x}) \mid \mathbf{t}\right]\right] && \text{(tower property)} \\ &= \mathbf{E}\left[\mathbf{E}\left[g(\mathbf{x}') \mid \mathbf{t}\right]\right] && \text{(sufficiency)} \\ &= \mathbf{E}\left[g(\mathbf{x}')\right] \end{aligned}$$

for any g. Since Bob can draw a new sample $\mathbf{x}'$ with the same law as Alice's sample $\mathbf{x}$, Bob can use any method that Alice may use to learn about θ.

In terms of partitions, a sufficient statistic partitions the sample space into regions on which the distribution of $\mathbf{x} \mid \mathbf{t}$ does not depend on θ. Going back to the coin tossing example, let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2 \overset{i.i.d.}{\sim} \text{Ber}(p)$. Tables 1 and 2 gives the conditional distributions of $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \mid \mathbf{x}_1+\mathbf{x}_2+\mathbf{x}_3$ and $(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \mid \mathbf{x}_1$. We observe that the former (conditional) distribution does not depend on p, but the latter depends on p. Thus $\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3$ is a sufficient statistic, but $\mathbf{x}_1$ is not.

TABLE 2
The conditional distribution of three coin tosses given the outcome of the first toss $\mathbf{x}_1$

$(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2)$	$\mathbf{x}_1$	$f_p(\mathbf{x} \mid \mathbf{x}_1)$
(0,0,0)	0	$(1-p)^2$
(0,1,0)	0	$p(1-p)$
(0,0,1)	0	$p(1-p)$
(0,1,1)	0	p^2
(1,0,0)	1	$(1-p)^2$
(1,1,0)	1	$p(1-p)$
(1,0,1)	1	$p(1-p)$
(1,1,1)	1	p^2

Definition 2.1 is intuitive, but not very useful for finding sufficient statistics. If a statistical model has densities, sufficient statistics can be found by factorizing the densities.

THEOREM 2.2. *A statistic* $\mathbf{t} := \phi(\mathbf{x})$ *is sufficient for a parametric model* $\mathcal{F}$ *if and only if there are functions* $g_\theta : \mathcal{X} \to \mathbf{R}$ *and* $h : \mathcal{X} \to \mathbf{R}$ *such that any density* $f_\theta(x) \in \mathcal{F}$ *has the form*

$$f_\theta(x) = g_\theta(\phi(x))h(x).$$

PROOF. A proof of the factorization theorem for continuous densities depends on measure theoretic arguments. We focus on the discrete case.

Assume $\mathbf{t} := \phi(\mathbf{x})$ is a sufficient statistic. It is possible to check that

$$g_\theta(t) = \mathbf{P}_\theta(\phi(\mathbf{x}) = t) \text{ and } h(x) = \mathbf{P}(\mathbf{x} = x \mid \phi(\mathbf{x}) = \phi(x))$$

is a valid factorization. Indeed,

$$\begin{aligned} g_\theta(\phi(x))h(x) &= \mathbf{P}_\theta(\phi(\mathbf{x}) = \phi(x))\mathbf{P}(\mathbf{x} = x \mid \phi(\mathbf{x}) = \phi(x)) \\ &= \mathbf{P}_\theta(\mathbf{x} = x, \phi(\mathbf{x}) = \phi(x)) \\ &= \mathbf{P}_\theta(\mathbf{x} = x), \end{aligned}$$

Assume $\mathbf{P}_\theta(\mathbf{x} = x) = g_\theta(\phi(x))h(x)$. It is possible to check that the conditional law of $\mathbf{x}$ given $\phi(\mathbf{x})$ does not depend on θ:

$$\begin{aligned} \mathbf{P}_\theta(\mathbf{x} = x \mid \phi(\mathbf{x}) = t) &= \frac{\mathbf{P}_\theta(\mathbf{x} = x, \phi(\mathbf{x}) = t)}{\mathbf{P}_\theta(\phi(\mathbf{x}) = t)} \\ &= \frac{\mathbf{P}_\theta(\mathbf{x} = x, \phi(\mathbf{x}) = t)}{\sum_{x \in \mathcal{X}} \mathbf{P}_\theta(\mathbf{x} = x)\mathbf{1}\{\phi(\mathbf{x}) = t\}}, \end{aligned}$$

which, by assumption, is

$$\begin{aligned}
&= \frac{g_\theta(t)h(x)}{\sum_{x\in\mathcal{X}} g_\theta(\phi(x))h(x)\mathbf{1}\{\phi(x)=t\}} \\
&= \frac{g_\theta(t)h(x)}{\sum_{x\in\mathcal{X}} g_\theta(t)h(x)\mathbf{1}\{\phi(x)=t\}} \\
&= \frac{h(x)}{\sum_{x\in\mathcal{X}} h(x)\mathbf{1}\{\phi(x)=t\}}.
\end{aligned}$$

Since the conditional law has no dependence on θ, we deduce $\mathbf{t} := \phi(\mathbf{x})$ is a sufficient statistic. □

Returning to the coin tossing example, we check that $\sum_{i\in[n]} x_i$ is a sufficient statistic for the model

$$\mathcal{F} = \{\mathrm{Ber}(p)^n = \mathrm{Ber}(p) \times \cdots \times \mathrm{Ber}(p) : p \in (0,1)\}$$

by factorizing the density:

$$\mathbf{P}_p(\mathbf{x}=x) = \prod_{i\in[n]} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i\in[n]} x_i}(1-p)^{1-\sum_{i\in[n]} x_i}.$$

In light of the factorization theorem, let

$$g_p(t) = \mathbf{P}_\theta(\phi(\mathbf{x})=t) = \tbinom{n}{t}p^t(1-p)^{1-t}$$

and observe

$$\frac{\mathbf{P}_p(\mathbf{x}=x)}{g_p\big(\sum_{i\in[n]} x_i\big)} = \binom{n}{\sum_{i\in[n]} x_i}^{-1},$$

which does not depend on p.

EXAMPLE 2.3. *Let* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathrm{Poi}(\theta)$. *Then the distribution of* $\mathbf{x}$ *is*

$$f_\theta(x) = \prod_{i\in[n]} \frac{\theta^{x_i}}{x_i!}e^{-\theta} = \frac{\theta^{\sum_{i\in[n]} x_i}e^{n\theta}}{\prod_{i\in[n]} x_i!},$$

which factorizes as

$$= \underbrace{\Big(\prod_{i\in[n]} x_i!\Big)^{-1}}_{h(x)} \underbrace{\big(\theta^{\sum_{i\in[n]} x_i}e^{n\theta}\big)}_{g_\theta(x)}.$$

Thus $\mathbf{t} := \sum_{i\in[n]} \mathbf{x}_i$ *is a sufficient statistic.*

Two other examples of sufficient statistics are

1. **max of uniform:** Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be *i.i.d.* according to the uniform law on $[0, \theta]$. The statistic $\mathbf{t} := \max_{i \in [n]}(\mathbf{x}_i)$ is sufficient for the model

$$\mathcal{F} = \{\text{unif}(0, \theta) : \theta > 0\}.$$

   Intuitively, given the max is t, the other $n-1$ random variables are *i.i.d.* according to the uniform law on $[0, t]$. Since the conditional law does not depend on θ, the max is sufficient.
2. **order statistics:** The order statistics $\mathbf{t} = \begin{bmatrix}\mathbf{x}_{(1)} & \ldots & \mathbf{x}_{(n)}\end{bmatrix}$ are sufficient for any model of *i.i.d.* random variables:

$$\mathcal{F} = \{F_\theta^n : \theta \in \Theta\}.$$

   Given the order statistics, the possible values of $\mathbf{x}$ are the $n!$ permutations of $\mathbf{t}$. Since $\{\mathbf{x}_i\}_{i \in [n]}$ are *i.i.d.*, each permutation occurs with probability $\frac{1}{n!}$. Thus the conditional law of $\mathbf{x} \mid \mathbf{t}$ is independent of θ.

To warp up, we observe that sufficient statistics are not unique. If $\mathbf{t}$ is a sufficient statistic, $g(\mathbf{t})$ for any invertible g is also a sufficient statistic. Since the events $\{\mathbf{t} = t\}$ and $\{g(\mathbf{t}) = g(t)\}$ are equivalent, the laws conditioned on $\{\mathbf{t} = t\}$ and $\{g(\mathbf{t}) = g(t)\}$ are identical.

Yuekai Sun
Berkeley, California
December 3, 2015