

# Statistics 210A Lecture 2 Notes

Daniel Raban

August 31, 2021

## 1 Estimation and Introduction to Exponential Families

### 1.1 Review of measure theory

Last time, we introduced some ideas from measure theory. Let's review:

A **measure**  $\mu$  assigns a “weight” to subsets  $A \subseteq \mathcal{X}$  (for  $A \in \mathcal{F}$ ).

**Example 1.1.** The **counting measure** is  $\#(A) = \text{card}(A)$ .

**Example 1.2.** **Lebesgue measure** gives  $\lambda(A) = \text{vol}(A)$  (in  $\mathbb{R}^n$ ).

**Example 1.3.** The **Gaussian distribution** gives  $P(A) = \int_A \phi(x) dx$ .

Measures give rise to integrals:

$$\int f(x) d\mu(x) = \begin{cases} \mu(A) & f(x) = \mathbb{1}_{\{x \in A\}} \\ \sum_i c_i \mu(A_i) & f(x) = \sum_i c_i \mathbb{1}_{\{x \in A_i\}} \\ \text{limit} & f(x) \text{ nice enough.} \end{cases}$$

If  $P \ll \mu$  (meaning  $\mu(A) = 0 \implies P(A) = 0$ ), there is a **density**  $p(x)$  with  $p : \mathcal{X} \rightarrow [0, \infty)$  such that  $\int f dP = \int f p d\mu$  for all (nice)  $f$ .

The **outcome space**  $\Omega$  containing outcomes  $\omega$  is equipped with a measure  $\mathbb{P}$ . Random variables are functions with  $X(\omega) \in \mathcal{X}$  (e.g.  $\mathcal{X} = \mathbb{R}$ ). You can think of  $X$  “decoding” the randomness  $\omega$  to tell you what the value in our nicer space  $\mathcal{X}$  is. We write  $X \sim Q$  if  $\mathbb{P}(X \in B) = Q(B)$ .

### 1.2 Estimation

In statistics, there are multiple possible distributions that could have generated the data.

**Definition 1.1.** A **statistical model** is a family of candidate probability distributions  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  for a random variable  $X \sim P_\theta$ .  $X$  is called the **data**, and  $\theta$  is called the **parameter**.

The data  $X$  is observed by the statistical analyst, whereas  $\theta$  is unobserved by the analyst. For now,  $\theta$  is fixed and unknown.<sup>1</sup> The goal of estimation is to observe  $X \sim P_\theta$  and guess the value of some estimand  $g(\theta)$ .

**Example 1.4.** Flip a biased coin  $n$  times. The parameter  $\theta \in [0, 1]$  is the probability of heads, and  $X \sim \text{Binom}(n, \theta)$  is the number of heads after  $n$  flips.  $X$  has a density  $p_\theta(x) = \theta^x(1 - \theta)^{n-x} \binom{n}{x}$  for  $x = 0, 1, \dots, n$  (this is a density with respect to counting measure on  $\{0, 1, \dots, n\}$ ).

**Definition 1.2.** A **statistic** is any function  $T(X)$  of  $X$ .

In particular, a statistic is not a function of  $\theta$ . It is something the statistical analyst can calculate.

**Definition 1.3.** An **estimator**  $\delta(X)$  of  $g(\theta)$  is a statistic intended to guess  $g(\theta)$ .

**Example 1.5.** In our coin flipping example, the natural estimator for  $\theta$  is  $\delta_0(X) = X/n$ .

### 1.3 Loss and risk

How can we tell if an estimator is good?

**Definition 1.4.** The **loss function**  $L(\theta, d)$  measures how badly an estimate is.

**Example 1.6.** One important loss function is the **squared error loss**  $L(\theta, d) = (d - g(\theta))^2$ .

Usually,  $L(\theta, d) \geq 0$  for all  $\theta, d$  with  $L(\theta, g(\theta)) = 0$ .

**Definition 1.5.** The **risk function**  $R(\theta; \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(X))]$  is the expected loss as a function of  $\theta$ .

**Remark 1.1.** The  $\mathbb{E}_\theta$  notation refers to the expectation with respect to  $X$ , where  $\theta$  is the true parameter. This is in contrast to other disciplines which use the notation  $\mathbb{E}_X$  to denote what variables we are conditioning on in the expectation. We will use the notation  $\mathbb{E}[f(X, X') \mid X']$  when we want to only integrate over certain random variables.

**Example 1.7.** The **mean squared error** is the risk function  $\text{MSE}(\theta, \delta_0(\cdot)) = \mathbb{E}_\theta[(\delta(x) - \theta)^2]$ .

**Example 1.8.** In our coin flipping example, we have the estimator  $\delta_0(X) = X/n$  with  $\mathbb{E}_\theta[X/n] = \theta$  (this is an **unbiased estimator**). The loss is

$$\text{MSE}(\theta, \delta_0(\cdot)) = \mathbb{E}_\theta[(\delta(x) - \theta)^2]$$

---

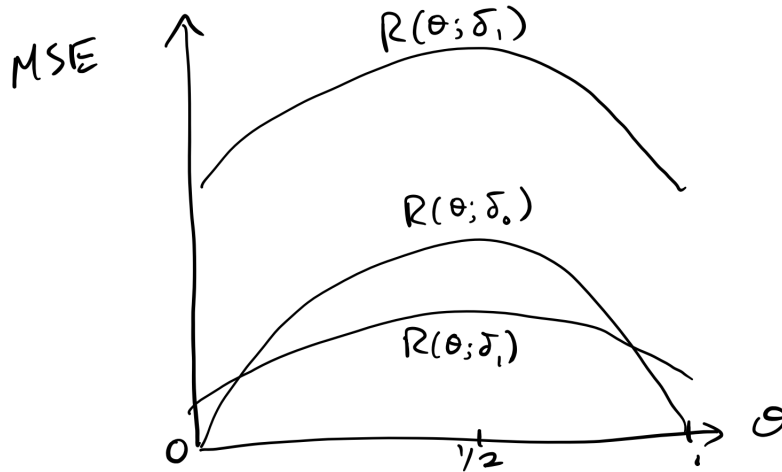
<sup>1</sup>This is a frequentist perspective. With a Bayesian perspective, we may assume that  $\theta$  follows some distribution.

$$\begin{aligned}
&= \text{Var}_\theta(X/n) \\
&= \frac{\theta(1-\theta)}{n}.
\end{aligned}$$

Here are other choices of estimators. We could take

$$\delta_1(X) = \frac{X+3}{n}.$$

$$\delta_2(X) = \frac{X+3}{n+6}$$



There is no estimator which is always the best; if  $\theta = 3/4$ , then the constant estimator  $\delta(X) = 3/4$  would be better than any estimator which has a chance of suggesting anything other than  $3/4$ .

## 1.4 Comparing estimators

**Definition 1.6.** An estimator  $\delta(X)$  is **inadmissible** if there exists another estimator  $\delta^*(X)$  such that

- (a)  $R(\theta; \delta^*) \leq R(\theta; \delta)$  for all  $\theta$ ,
- (b)  $R(\theta; \delta^*) < R(\theta; \delta)$  for some  $\theta$ .

In our previous example,  $\delta_0$  rendered  $\delta_1$  inadmissible.  
Here are some strategies to resolve the ambiguity:

1. Summarize  $R(\theta)$  by a scalar:

- (a) **Average-case risk:** Minimize  $\int_{\Theta} R(\theta; \delta) d\Lambda(\theta)$ . The minimizer  $\delta$  is called the **Bayes estimator**.
- (b) **Worse-case risk:** Minimize  $\sup_{\theta \in \Theta} R(\theta; \delta)$ . The minimizer  $\delta$  is called the **minimax estimator**.

2. Constrain the choice of estimator:

- (a) Only consider **unbiased**  $\delta(X)$  ( $\mathbb{E}_{\theta}[\delta(X)] = g(\theta)$ ).

## 1.5 Exponential families

**Definition 1.7.** An  $s$ -parameter exponential family is a family  $\mathcal{P} = \{P_{\eta} : \eta \in \Xi\}$  with densities  $p_{\eta}(x)$  with respect to a common dominating measure  $\mu$  on  $\mathcal{X}$  of the form

$$p_{\eta}(x) = e^{\eta^{\top} T(x) - A(\eta)} h(x),$$

where

- $T : \mathcal{X} \rightarrow \mathbb{R}^s$  is called the **sufficient statistic**,
- $h : \mathcal{X} \rightarrow [0, \infty)$  is called the **carrier/base density**,
- $\eta \in \Xi \subseteq \mathbb{R}^s$  is called the **natural parameter**,
- $A : \mathbb{R}^s \rightarrow \mathbb{R}$  is called the **cumulant generating function** (or the **normalizing constant**).

**Remark 1.2.**  $A(\eta)$  is totally determined by  $h, T$ , since we always must have  $\int_{\mathcal{X}} p_{\eta} d\mu = 1$  for all  $\eta$ . So we can solve

$$A(\eta) = \log \left[ \int_{\mathcal{X}} e^{\eta^{\top} T(x)} h(x) d\mu(x) \right] \leq \infty.$$

**Definition 1.8.**  $p_{\eta}$  is **normalizable** if  $A(\eta) < \infty$ . The **natural parameter space** is  $\Xi_1 = \{\eta : A(\eta) < \infty\}$ . We say  $\mathcal{P}$  is in **canonical form** if  $\Xi = \Xi_1$ .

**Remark 1.3.**  $A(\eta)$  is a convex function, so  $\Xi_1$  is a convex set.

In general, you can think of an  $s$ -parameter exponential family as describing an  $s$ -dimensional hyperplane in the space of log-densities.

