Theoretical Statistics

STAT 210A

Instructor: Will Fithian

Kelvin Lee

UC Berkeley

Contents

Lecture 1

Measure Theory

1.1 Basics

1.1.1 Measures

Definition 1.1.1 (Measure)**.** Given a set $\mathcal{X}$, a **measure** μ maps subsets $A \subseteq \mathcal{X}$ to nonnegative numbers $\mu(A) \in [0, \infty]$.

Example 1.1.2. Let $\mathcal{X}$ be a countable set ($\mathcal{X} = \mathbb{Z}$ for example). Then the **counting measure** ia

$$\mu(A) = \#A = \# \text{ of points in } A.$$

Example 1.1.3. Consider $\mathcal{X} = \mathbb{R}^n$. The **Lebesgue measure** is

$$\lambda(A) = \int \cdots \int_A dx_1 \cdots dx_n = \text{Vol}(A).$$

Example 1.1.4 (Standard Gaussian Distribution)**.**

$$\mathbb{P}(A) = \mathbb{P}(Z \in A) = \int \cdots \int_A \phi(x) dx_1 \cdots dx_n$$

where $Z \sim_n (0, I_n)$ and $\phi(x) = \frac{e^{-\frac{1}{2}\sum x_i^2}}{\sqrt{(2\pi)^n}}$.

Because of pathological sets, $\lambda(A)$ is only defined for some subsets $A \subseteq \mathbb{R}^n$. In other words, it is often impossible to assign measures to all subsets A of $\mathcal{X}$. This leads to the idea of a σ-field(σ-algebra).

Definition 1.1.5 (σ-field)**.** A **σ-field** is a collection of sets on which μ is defined, satisfying certain closure properties.

In general, the domain of a measure μ is a collection of subsets $\mathcal{F} \subseteq 2^{\mathcal{X}}$ (power set), and $\mathcal{F}$ must be a σ-field.

Example 1.1.6. Let $\mathcal{X}$ be a countable set. Then $\mathcal{F} = 2^{\mathcal{X}}$. (Counting measure is defined for all subsets).

Example 1.1.7. Let $\mathcal{X} = \mathbb{R}^n$, then $\mathcal{F}$ is the **Borel σ-field** $\mathcal{B}$, the smallest σ-field containing all open rectangles $(a_1, b_1) \times \cdots \times (a_n, b_n)$ where $a_i < b_i \quad \forall i$.

Given a **measurable space** $(\mathcal{X}, \mathcal{F})$, a **measure** is any map $\mu : \mathcal{F} : [0, \infty]$ with $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ if $A_i \in \mathcal{F}$ are disjoint. If $\mu(\mathcal{X}) = 1$, then μ is a **probability measure**.

Measures let us define integrals that put weight $\mu(A)$ on $A \subseteq \mathcal{X}$.

Define

$$\int \mathbf{1}\{x \in A\} d\mu(x) = \mu(A) \qquad \text{(indicator)}$$

extend to other functions by linearity and limits:

$$\int \left(\sum c_i \mathbf{1}\{x \in A_i\}\right) d\mu(x) = \sum c_i \mu(A_i) \qquad \text{(simple function)}$$

$$\int f(x) d\mu(x) \qquad \text{(approx. by simple functions)}$$

Example 1.1.8.

- *Counting*: $\int f d\# = \sum_{x \in \mathcal{X}} f(x)$.
- *Lebesgue*: $\int f d\lambda = \int \cdots \int f(x) dx_1 \cdots dx_n$.
- *Gaussian*: $\int f dP = \int \cdots \int f(x)\phi(x) dx_1 \cdots dx_n = \mathbb{E}[f(Z)]$.

1.1.2 Densities

The λ and $\mathbb{P}$ above are closely related and we now want to make this precise.

Given $(\mathcal{X}, \mathcal{F})$, two measures $\mathbb{P}, \mu$, we say that $\mathbb{P}$ is **absolutely continuous** with respect to μ if $\mathbb{P}(A) = 0$ whenever $\mu(A) = 0$.

Notation: $P \ll \mu$ or we say μ *dominates* $\mathbb{P}$.

If $\mathbb{P} \ll \mu$, then (under mild conditions) we can always define a **density function** $p : \mathcal{X} \to [0, \infty)$ with

$$\mathbb{P}(A) = \int_A p(x) d\mu(x)$$
$$\int f(x) d\mathbb{P}(x) = \int f(x) p(x) d\mu(x).$$

The density function is also defined as

$$p(x) = \frac{d\mathbb{P}}{d\mu}(x),$$

known as *Radon-Nikodyan derivative.*

Remark. It is useful to turn $\int f d\mathbb{P}$ into $\int f p d\mu$ if we know how to calculate integrals $d\mu$.

If $\mathbb{P}$ is a probability measure, μ is a Lebesgue measure, then $p(x)$ is called **probaility density function** (pdf). If μ is a counting measure, then $p(x)$ is called the **probability mass function** (pmf).

1.1.3 Probability Space and Random Variables

Typically, we set up a problem with multiple random variables having various relationships to one another. It is convenient to think of them as functions of an abstract outcome ω.

Definition 1.1.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a **probability space**. $\omega \in \Omega$ is called **outcome**. $A \in \mathcal{F}$ is called **event**. $\mathbb{P}(A)$ is called **probability of** A.

Definition 1.1.10. A **random variable** is a function $X : \Omega \to \mathcal{X}$. We say $\mathcal{X}$ has distribution Q, denoted as $X \sim Q$ if $\mathbb{P}(X \in B) = \mathbb{P}(\{\omega : X(\omega) \in B\}) = Q(B)$.

More generally, we could write events involving many random variables $X(\omega), Y(\omega), Z(\omega)$:

$$\mathbb{P}(X \geq Y + Z) = \mathbb{P}(\{\omega : X(\omega) \geq Y(\omega) + Z(\omega)\})$$

Definition 1.1.11. The **expectation** is an integral with respect to $\mathbb{P}$:

$$\mathbb{E}[f(X, Y)] = \int_\Omega f(X(\omega), Y(\omega)) d\mathbb{P}(\omega).$$

To do real calculations, we must eventaully boil $\mathbb{P}$ or $\mathbb{E}$ down to concrete integrals/sums/etc. If $\mathbb{P}(A) = 1$, we say that A occurs **almost surely**.

Lecture 2

Risk and Estimation

2.1 Estimation

Definition 2.1.1 (Statistical Model). A **statistical model** is a family of candidate probability distributions

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$$

for some random variable $X \sim P_\theta$. $X \in \mathcal{X}$ is called **data** (observed). θ is the **parameter** (unobserved).

The goal of estimation is to observe $X \sim P_\theta$ and guess value of some **estimand** $g(\theta)$.

Example 2.1.2. Suppose we flip a biased coin n times. Let $\theta \in [0,1]$ be the probability of getting a head and let X be the number of heads after n flips. Then $X \sim \text{Binom}(n,\theta)$ with $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, which is the density with respect to counting measure on $\mathcal{X} = \{0,\ldots,n\}$. A natural estimator would be $\delta_0(X) = \frac{X}{n}$.

Question. Is the natural estimator a good estimator? Is there a better one?

Definition 2.1.3 (Statistic). A **statistic** is any function $T(x)$ of data X (not of both X and θ).

Definition 2.1.4 (Estimator). An **estimator** $\delta(X)$ of $g(\theta)$ is a statistic which is intended to guess $g(\theta)$.

2.2 Loss and Risk

Definition 2.2.1 (Loss function). A **loss function** $L(\theta, d)$ measures how bad an estimate is.

Example 2.2.2. One common loss function is the *squared-error loss* $L(\theta, d) = (d - g(\theta))^2$.
Typical properties of loss functions:

(i) $L(\theta, d) \geq 0 \ \forall \theta, d$

(ii) $L(\theta, g(\theta)) = 0 \ \forall \theta$.

Definition 2.2.3 (Risk function). The **risk function** is the expected loss (risk) as a function of θ for an estimator $\delta(\cdot)$.

$$R(\theta; \delta(\cdot)) = \mathbb{E}_\theta[L(\theta, \delta(X))].$$

Remark. The subscript θ under $\mathbb{E}$ tells us which parameter value is in effect, NOT what randomness to integrate over.

Example 2.2.4 (Coin flip cont'd). We have $\delta_0(X) = \frac{X}{n}$. Then $\mathbb{E}_\theta\left[\frac{X}{n}\right] = \theta$ (unbiased). Then

$$R(\theta, \delta) = MSE(\theta; \delta_0) = \mathrm{Var}_\theta\left(\frac{X}{n}\right) = \frac{\theta(1-\theta)}{n}.$$

Other choices:

$$\delta_1(X) = \frac{X+3}{n}$$
$$\delta_2(X) = \frac{X+3}{n+6}.$$

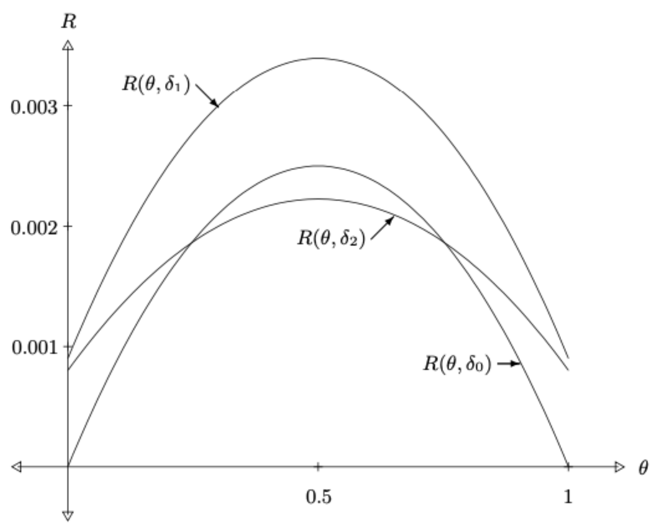

Figure 2.1: Risks for $\delta_0, \delta_1, \delta_2$.

δ_1 is bad but δ_0, δ_2 are ambiguous.

Definition 2.2.5 (Inadmissible). An estimator δ is **inadmissible** if $\exists \delta^*$ with

(i) $R(\theta; \delta^*) \le R(\theta; \delta)\ \forall \theta \in \Theta$

(ii) $R(\theta; \delta^*) < R(\theta; \delta)$ for some $\theta \in \Theta$

From the previous example, we see that δ_1 is inadmissible.

Back to the issue regarding the ambiguity of the comparison between two estimators. Here are some strategies to resolve that ambiguity:

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

   (i) Average-case risk: minimize
   $$\int_\Theta R(\theta;\delta)d\Lambda(\theta)$$
   with some measure Λ. This is called the **Bayes estimator**, and Λ is the **prior**.

   (ii) Worst-case risk: minimize
   $$\sup_{\theta\in\Theta} R(\theta,\delta).$$
   over $\delta : \mathcal{X} \to \mathbb{R}$. This is a **minimax estimator**, which is closely related to Bayes.

   Remark. We do not consider the best-case risk because the constant estimator would always ignore the data, which makes it a bad estimator.

2. Constrain the choice of estimator.

   (i) Only consider unbiased δ. $\mathbb{E}_\theta[\delta(X)] = g(\theta)\,\forall\theta\in\Theta$.

Lecture 3

Exponential Families

3.1 s-parameter Exponential Family

Definition 3.1.1 (s-parameter exponential family)**.** An **s-parameter exponential family** is a family of probability densities $\mathcal{P} = \{p_\eta : \eta \in \Xi\}$ with respect to a common measure μ on $\mathcal{X}$ of the form

$$p_\eta(x) = \exp\left[\sum_{i=1}^{s} \eta_i T_i(x) - A(\eta)\right] h(x), \quad x \in \mathcal{X}$$

where

- $T : \mathcal{X} \to \mathbb{R}^s$ is a **sufficient statistic**
- $h : \mathcal{X} \to \mathbb{R}$ is a **carrier/base density**
- $\eta \in \Xi \subseteq \mathbb{R}^s$ is a **natural parameter**
- $A : \mathbb{R}^s \to \mathbb{R}$ is a **cumulant generating function** (CGF)

Note that the CGF $A(\eta)$ is totally determined by h, T since we must have $\int_{\mathcal{X}} p_\eta d\mu = 1\, \forall \eta$. Hence,

$$A(\eta) = \log_{\mathcal{X}} \int \exp\left[\sum_{i=1}^{s} \eta_i T_i(x)\right] h(x) d\mu(x).$$

p_η is only normalizable iff $A(\eta) < \infty$.

Definition 3.1.2 (Natural parameter space)**.** The **natural parameter space** is the set of all allowable (normalizable) η:

$$\Xi_1 = \{\eta : A(\eta) < \infty\}.$$

We say $\mathcal{P}$ is in **canonical form** if $\Xi = \Xi_1$.

Remark. Note that Ξ_1 is determined by T, h, η. We could take $\Xi \subset \Xi_1$ if we wanted. $A(\eta)$ is convex $\implies$ Ξ_1 is convex (from homework).

Interpretation of Exponential Families:

- Start with a base density p_0.
- Apply an **exponential tilt**:

1. multiply by $e^{\eta^\top T}$
2. renormalize (if possible)

An exponential family in canonical form is all possible tilts of h (or any p_η) using any linear combination of T.

Sometimes it is more convenient to use a different parameterization:

$$p_\theta(x) = \exp\left\{\eta(\theta)^\top T(x) - B(\theta)\right\} h(x), \quad \text{where } B(\theta) = A(\eta(\theta)).$$

Example 3.1.3 (Gaussian Family)**.** Consider $X \sim (\mu, \sigma^2)$ where $\mu \in \mathbb{R}, \sigma^2 > 0$. $\theta = (\mu, \sigma^2)$.

$$\begin{aligned} p_\theta(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \exp\left\{ \underbrace{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2}_{\eta(\theta)^\top T(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{B(\theta)} \right\} \cdot \underbrace{1}_{h(x)} \end{aligned}$$

This is a two-parameter exponential family with $\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$ and $T(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$, $h(x) = 1$, and $B(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$.

Remark. $h(x)$ can also be $1/\sqrt{2\pi}$ if we did not include the factor $1/(\sqrt{2\pi}\sigma)$ into the exponentiation. In that case, $B(\theta) = \mu^2/(2\sigma^2) + \log\sigma$.

In canonical form,

$$\begin{aligned} p_\eta(x) &= \exp\left\{\eta^\top \begin{pmatrix} x \\ x^2 \end{pmatrix} - A(\eta)\right\} \\ A(\eta) &= -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \end{aligned}$$

Example 3.1.4. Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} (\mu, \sigma^2)$. Then their joint density is

$$\begin{aligned} p_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^n p_\theta^{(1)}(x_i) \\ &= \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x_i-\mu)^2}{(2\sigma^2)}\right\}\right] \\ &= \exp\left\{\sum_{i=1}^n \left[\frac{\mu}{\sigma^2}x_i - \frac{1}{2\sigma^2}x_i - \left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right]\right\} \\ &= \exp\left\{\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 - n\left(\frac{\mu}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)\right\}. \end{aligned}$$

These densities also form a two-parameter exponential family with $\eta(\theta) = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}, T(x) = \begin{pmatrix} \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i^2 \end{pmatrix}$, $B(\theta) = nB^{(1)}(\theta)$.

Generally, suppose $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\eta^{(1)}(x) = \exp\{\eta^\top T(x) - A(\eta)\}h(x)$. Then

$$\begin{aligned}
X \sim \prod_{i=1}^n p_\eta^{(1)}(x_i) &= \prod_{i=1}^n \exp\left\{\eta^\top T(x_i) - A(\eta)\right\} h(x_i) \\
&= \exp\left\{\eta^\top \underbrace{\sum_{i=1}^n T(x_i)}_{\text{sufficient statistic}} - \underbrace{nA(\eta)}_{\text{cgf}}\right\} \underbrace{\prod_{i=1}^n h(x_i)}_{\text{carrier density}} .
\end{aligned}$$

Suppose $X \in \mathcal{X}$ follows an exponential family. Then $T(X)$ also follows a closely related exponential family. $T(X) \in \mathcal{T} \subseteq \mathbb{R}^s$. If $X \sim p_\eta(x) = \exp\left\{\eta^\top T(x) - A(\eta)\right\}$ (WLOG assume $h(x) = 1$) with respect to μ.

For a set $B \subseteq \mathcal{T}$, define $\nu(B) = \mu(T^{-1}(B))$. Then $T(X) \sim q_\eta(t) = \exp\left\{\eta^\top t - A(\eta)\right\}$ with respect to ν.

Discrete case:

$$\begin{aligned}
\mathbb{P}_\eta(T(X) \in B) &= \sum_{x : T(x) \in B} \exp\left\{\eta^\top T(x) - A(\eta)\right\} \mu(\{x\}) \\
&= \sum_{t \in B} \sum_{x : T(x) = t} \exp\{\eta^\top t\} \mu(\{x\}) \\
&= \sum_{t \in B} \exp\left\{\eta^\top t - A(\eta)\right\} \mu(T^{-1}(\{t\})) \\
&= \sum_{t \in B} \exp\left\{\eta^\top t - A(\eta)\right\} \nu(\{t\}).
\end{aligned}$$

Thus, $T \sim \exp\{\eta^\top t - A(\eta)\}$ with respect to ν.

Example 3.1.5 (Binomial). Let $X \sim \text{Binom}(n, \theta)$. n is fixed and so the parameter is $\theta \in [0, 1]$. Then

$$\begin{aligned}
p_\theta(x) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \\
&= \binom{n}{x} \left(\frac{\theta}{1-\theta}\right)^x (1-\theta)^n \\
&= \binom{n}{x} \exp\left\{x \log\left(\frac{\theta}{1-\theta}\right) + n \log(1-\theta)\right\}
\end{aligned}$$

with natural parameter $\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ called the *log odds ratio*.

Example 3.1.6 (Poisson). Let $X \sim \text{Poisson}(\lambda)$. Then

$$\begin{aligned}
p_\lambda(x) &= \frac{\lambda^x e^{-\lambda}}{x!} \qquad i = 0, 1, 2, \dots \\
&= \exp\{x \log(\lambda) - \lambda\} \frac{1}{x!}
\end{aligned}$$

with natural parameter $\eta(\lambda) = \log(\lambda)$.

3.2 Differential Identities

Write

$$e^{A(\eta)} = \int \exp\left\{\eta^\top T(x)\right\} h(x) d\mu(x).$$

Theorem 3.2.1. *For $f : \mathcal{X} \to \mathbb{R}$, let*

$$\Xi_f = \left\{\eta \in \mathbb{R}^s : \int |f| \exp\{\eta^\top T\} h d\mu < \infty\right\}.$$

Then the function

$$g(\eta) = \int f \exp\{\eta^\top T\} h d\mu$$

has continuous partial derivatives of all orders for $\mu \in \Xi_f^\circ$ (interior of Ξ_f), which can be computed by differentiating under the integral.

Differentiating $e^{A(\eta)}$ once:

$$\begin{aligned}
\frac{\partial}{\partial \eta_j} e^{A(\eta)} &= \frac{\partial}{\partial \eta_j} \int e^{\eta^\top T(x)} h(x) d\mu(x) \\
\cancel{e^{A(\eta)}} \frac{\partial A}{\partial \eta_j}(\eta) &= \int T_j(x) e^{\eta^\top T(x) - A(\eta)} h(x) d\mu(x) \\
\frac{\partial A}{\partial \eta_j}(\eta) &= \mathbb{E}_\eta[T_j(X)].
\end{aligned}$$

Thus, we have

$$\boxed{\nabla A(\eta) = \mathbb{E}_\eta[T(X)]}.$$

Differentiating it again:

$$\begin{aligned}
\frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{A(\eta)} &= \int \frac{\partial^2}{\partial \eta_j \partial \eta_k} e^{\eta^\top T(x)} h(x) d\mu(x) \\
\cancel{e^{A(\eta)}} \left(\frac{\partial^2 A}{\partial_i \partial \eta_k} + \underbrace{\frac{\partial A}{\partial \eta_j}}_{\mathbb{E}_\eta[T_j]} \cdot \underbrace{\frac{\partial A}{\partial \eta_k}}_{\mathbb{E}_\eta[T_k]} \right) &= \underbrace{\int T_j T_k e^{\eta^\top T - A(\eta)} h d\mu}_{\mathbb{E}_\eta[T_j T_k]} \\
\frac{\partial^2 A}{\partial \eta_j \partial \eta_k}(\eta) &= \mathbb{E}_\eta[T_j T_k] - \mathbb{E}_\eta[T_j]\mathbb{E}_\eta[T_k] \\
&= \mathrm{Cov}_\eta(T_j, T_k).
\end{aligned}$$

Thus, we have

$$\boxed{\nabla^2 A(\eta) = \mathrm{Var}_\eta(T(X))}.$$

3.2.1 Moment Generating Functions

Differentiating $e^{A(\eta)}$ repeatedly, we get

$$\mathrm{e}^{-A(\eta)} \left(\frac{\partial^{k_1 + \cdots + k_s}}{\partial \eta_1^{k_1} \cdots \partial \eta_s^{k_s}} \mathrm{e}^{A(\eta)} \right) = \mathbb{E}_\eta[T_1^{k_1} \cdots T_s^{k_s}]$$

since $M_{T(X)}(u) = e^{A(\eta+u)-A(\eta)}$ is the **moment generating function** (**MGF**) of $T(X)$ when $X \sim p_\eta$.

$$\begin{aligned} M_{T(X)}(u) &= \mathbb{E}_\eta[\mathrm{e}^{u^\top T(X)}] \\ &= \int \mathrm{e}^{u^\top T + \eta^\top T - A(\eta)} h d\mu \\ &= \mathrm{e}^{A(\eta+u)-A(\eta)} \int \mathrm{e}^{(\eta+u)^\top T - A(\eta+u)} h d\mu \\ &= \mathrm{e}^{A(\eta+u)-A(\eta)}. \end{aligned}$$

The cumulant generating function is $K_{T(X)}(u) = \log M_{T(X)}(u) = A(\eta+u) - A(\eta)$.