ORF 524: Statistical Theory and Methods **Fall 2016**

Lecture 1: 09/15/16

Lecturer: Prof. Samory Kpotufe *Scribe: Zachary Hervieux-Moore*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

1.1 Overview

1.1.1 Satistical Problems:

We observe a random object X, n times, as $X = (X_1, X_2, \dots, X_n)$ from some unknown distribution P. What can we infer about X_{n+1}, or the rest of the unobserved "population" of X's?

Density/Distribution Estimation: Suppose we know $X \sim P \in \mathcal{P}$ (e.g. all Gaussians) where $\sim$ means "is distributed according to". We want to infer the "right" $P \in \mathcal{P}$ from the sample $X_1, \dots, X_n$.

Example 1.

- *Inferring the bias of a coin from multiple throws. From there we know* $Pr(X_{n+1} = 1)$.
- *Inferring the population of the US voting for Hillary from a survey.*

Regression/Classification: $X = (Z, Y)$ where we want to predict Y from Z.

Example 2.

- $Z \equiv$ *Netflix movies,* $Y \equiv$ *Whether you like* X
- $Z \equiv$ *Image,* $Y \equiv$ *object in image*
- $Z \equiv$ *Financial Instrument,* $Y \equiv$ *pricing*

Inference: More generally, making decisions from observations.

Example 3.

- *Is the new flu vaccine effective?*
- *Do humans cause global warming?*

1.1.2 Statistical Approach

We usually don't use the whole sample X. Rather, we computer a "statistic" from the sample.

Definition 1. *A statistic T is some quantity we compute from observations, $X_1, \dots, X_n$. We then use T to infer something about the unknown P.*

Note: All the earlier examples can be viewed as trying to estimate some functional (or characteristic) θ of P, or infer something about θ (e.g., is $\theta > \theta_0$?)

Example 4. *Bias θ of a coin: We compute $T \equiv \overline{X} = \frac{1}{n}\sum_i X_i$ and simply infer that $\theta \approx T$.*

1.1.3 Our concerns in this course

- Do we lose information about the unknown P only using T? (Is T "sufficient")?
- Why T and not T'? Is T the "best" for our problem? E.g., in computing a mean, why use $T = \overline{X} = \frac{1}{n}\sum X_i$ rather than $T' = \frac{1}{n}\sum log|X_i|$
- How do we define "good"/"best"?
- Suppose we keep all of $X = (X_1, X_2, \dots, X_n)$, how much information is there about P? How does that depend on n? That is, how "hard" is the original problem?

We'll develop various mathematical tools towards answering such questions... (called mathematical statistics).

1.2 Basic Tools

1.2.1 Probability Measures:

3 objects (Ω, Σ, P) defining a "Measure space".

Definition 2. *A sample space Ω is a non-empty set serving as an abstraction of basic events.*

Ex: $\Omega = \{HH, HT, TH, TT\}$ *for 2 coins.*

The measure "P" will serve to assign values between $[0, 1]$ to subsets of Ω (so called events when P is a "probability")

Definition 3. *σ-Algebra Σ (or σ-field)*

We want the freedom to define P over just a collection Σ of subsets of Ω, rather than over all subsets in 2^Ω (Power set of Ω). That is, some intuitive notion of measure such as "length"/Lebesgue cannot be soundly defined over all subsets of $\mathbb{R}$.

Σ will have to satisfy some basic properties for P to be sound. Namely Σ must be a σ-Algebra, it must satisfy:

- $\Sigma \neq \emptyset, \Sigma \subset 2^\Omega$
- *If $A \in \Sigma$, then $A^c \in \Sigma$*
- *If $A_1, A_2, \dots \in \Sigma$, then $\bigcap A_i \in \Sigma$. Where the intersection can be countably infinite.*

Ex: *Borel Algebra $\mathcal{B}(\mathbb{R}^d)$, is the smallest σ-algebra containing all open sets of $\mathbb{R}^d$.*

Note: (Ω, Σ) is a "measurable space". $A \in \Sigma$ is a "measurable set".

Exercise 1.

- *Show that* Σ *is closed under union.*
- *Show that it must contain* $\emptyset$ *and* Ω.
- *Suppose* $A \subset \Omega$. *What is the smallest* σ*-algebra containing* A*?*

Definition 4. $P : \Sigma \to \mathbb{R}$ *is called a measure on* (Ω, Σ) *iff:*

- $\forall A \in \Sigma$, $0 \leq P(A) \leq \infty$
- $P(\emptyset) = 0$
- $\forall$ *disjoint* $A_1, A_2, \ldots \in \Sigma$, $P(\bigcup\limits_{i=1}^{\infty} A_i) = \sum\limits_{i=1}^{\infty} P(A_i)$

Ex: *Lebesgue measures (length, volume in* $\mathbb{R}^d$*), counting measures (which counts the number of elements in* A*).*

Definition 5. *A "Probability" measure is one such that* $P(\Omega) = 1$.

Exercise 2.

- *Show that* $0 \leq P(A) \leq P(B), \forall\ A \subset B, A, B \in \Sigma$
- *Show that* $P(\cup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} P(A_i)$. *What if* $k = \infty$*?*

1.2.2 Random Element (Variable, Vector, etc.)

Definition 6. *Consider 2 measurable spaces* (Ω, Σ) *and* (Ω', Σ'). *A "measurable" map* $X : \Omega \to \Omega'$ *is a function such that* $\forall A \in \Sigma'$, $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \Sigma$

Note: If (Ω', Σ') is $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we call it a random vector. Generally, X is called a "random element" of Ω'. Usually $(\Omega', \Sigma') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then X is a random variable.

Definition 7. *The "Induced Measure"* P' *on* (Ω', Σ') *id one that assigns* $P'(A) = P(X^{-1}(A))$. *We write* $P'(A) = Pr(X \in A)$ *when* P *is a probability measure.*

Ex: Gaussians, Binomial, etc, are induced onto $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. **Note:** The notion of induced measure allows us to often forget about (Ω, Σ, P) and work with $(\Omega', \Sigma\, P')$. **Note:** $Pr(X \in A)$ is only defined for measurable sets $A \in \Sigma$.

1.2.3 Integration (Lebesgue):

Let (Ω, Σ) a measurable space. In all that follows, all measurable maps are assumed to be Borel.

Definition 8. *Functions* $\rho : \Omega \to \mathbb{R}$ *of the form* $\rho(\omega) = \sum\limits_{i=1}^{k} a_i 1_{\omega \in A_i}$ *for* $a_i \in \mathbb{R}$, *for some* $\{A_i\}_1^k \in \Sigma$, *are called "simple functions".*

Definition 9. *Integration is defined as follow. Let* P *be a measure on* (Ω, Σ).

- *Let ρ be simple (over $\{A_i\}_1^k$): $\int_\Omega \rho(\omega)dP(\omega) \equiv \int \rho dP \equiv \int \rho \doteq \sum_{i=1}^{k} a_i 1_{A_i}$*
- *Let $f \geq 0$: $\int f dP = \sup_{0 \leq \rho \leq f} \int \rho dP$*
- *For general f, let $f_+ = f \cdot 1_{\{f \geq 0\}}$ and $f_- = f \cdot 1_{\{f \leq 0\}}$. Then, $\int f dP = \int f_+ dP - \int f_- dP$*
- *$\forall A \in \Sigma$, $\int_A f dP \equiv \int (f \cdot 1_A) dP$*

Note: If P is a probability measure, we often write $\int f dP \doteq E[f]$, using "Expectation" notation.

Note: Let $(\Omega, \Sigma, P) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), \sigma)$. Where σ is the usual Lebesgue measure (length). Then Lebesgue integral ($\int f d\sigma$) coincides with the Riemann integral from calculus whenever the latter is well-defined.

Note: $\int f dP$ exists whenever at least one of $\int f_+ dP$, $\int f_- dP$ is $< \infty$. f is then called "integrable". This allows $\int f dP = \pm\infty$

Example 5. *Riemann is ill-defined on $\mathbb{Q}$. However, the Lebesgue integral (w.r.t. Lebesgue Measure) over $\mathbb{Q}$ is well-defined and is 0.*

Proposition 10 (Change of Measure in Integration)**.** *Consider (Ω, Σ, P), and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $f : \Omega \to \mathbb{R}$ measurable. Let $P' \doteq P(f^{-1})$ be induced by f. Then,*

$$\int_\Omega f(\omega) dP(\omega) = \int_{\mathbb{R}} \omega' dP(\omega')$$

Example 6. *X is Gaussian, X^2 is χ^2 (chi-squared), we can integrate both w.r.t. $\mathcal{N}(\mu, \sigma^2)$ and w.r.t. χ^2 measure. (Here $\Omega = \mathbb{R}$).*

1.2.4 Radon-Nikodym derivatives: (a.k.a. densities)

Definition 11. *Let μ, ν be two measures on (Ω, Σ) s.t. if $\nu(A) = 0$, then $\mu(A) = 0$ for all $A \in \Sigma$. Then we say μ is "dominated" by ν ($\mu << \nu$) or μ is "absolutely-continuous" w.r.t. ν.*

Example 7. *Any continuous/discrete μ and Lebesgue/Counting.*

Definition 12. *The measure ν is "σ-finite" iff $\exists \{A_i\} \subset \Sigma$ s.t. $\bigcup A_i = \Omega$, and $\nu(A_i) < \infty \; \forall i$.*

Exercise 3.

- *Show that Lebesgue is σ-finite (for $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$).*
- *Show that the counting measure is σ-finite iff Ω is countable.*

Theorem 13 (Radon-Nikodym Theorem)**.** *Suppose $\mu << \nu$ (both on some (Ω, Σ)), and ν is σ-finite. Then $\exists$ a Borel map f, $f \geq 0$, s.t. $\forall A \in \Sigma$, $\mu(A) = \int f \cdot 1_A d\nu \doteq \int_A f d\nu$.*

f is often denoted $\frac{d\mu}{d\nu}$ and is called the Radon-Nikodym derivative of μ w.r.t. ν (or the "density" of μ w.r.t. ν).

Example 8. *The following are all densities,*

- *The Gaussian density w.r.t. Lebesgue σ.*
- *In fact, any continuous X on $\mathbb{R}$, the normal density is w.r.t. Lebesgue.*
- *For discrete RV's X, the pmf f_X is a density w.r.t. to "a" counting measure.*

Exercise 4. *Show that for a discrete random variable $X : \Omega \to \mathbb{R}$, the pmf f is indeed the density of P_X w.r.t. the counting measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.*