Chapter 4

Randomization

SECTION 1 : reminder of about role of randomization in Le Cam dist
SECTION 2 : facts about linear fnals
SECTION 3 : introduce M-, K-, and L-randomizations
SECTION 4 : express conditioning via randomizations
SECTION 5 : prove K-sufficiency of LR; canonical measures
SECTION 6 : advantages of Le Cam randomization
SECTION 7 : choice of sample space

Chapter incomplete. First draft. Many remarks intended only as reminders to DP.

1. Le Cam distance

A probability model consists of a probability measure on a sigma-field $\mathcal{A}$ of some set $\mathcal{X}$. A statistical model is an indexed family of probability measures, $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{A})$. Let $\mathcal{Q} := \{\mathbb{Q}_\theta : \theta \in \Theta\}$ be another statistical model, with the same index set Θ, but with the $\mathbb{Q}_\theta$ measures living on some $(\mathcal{Y}, \mathcal{B})$. The Le Cam distance $\delta(\mathcal{Q}, \mathcal{P})$ from $\mathcal{Q}$ to $\mathcal{P}$ is defined as the infimum of all those constants ϵ for which there exists a randomization K for which

<1> $$\tfrac{1}{2}\|K\mathbb{Q}_\theta - \mathbb{P}_\theta\|_1 \le \epsilon \qquad \text{for every } \theta \text{ in } \Theta.$$

The symmetric form of the Le Cam distance, $\Delta(\mathcal{P}, \mathcal{Q})$, is defined as the maximum of $\delta(\mathcal{P}, \mathcal{Q})$ and $\delta(\mathcal{Q}, \mathcal{P})$. If $\Delta(\mathcal{P}, \mathcal{Q}) = 0$ then the experiments are said to be equivalent (in Le Cam's sense).

comment on whether inf achieved; note advantage of L-randn

The subtlety of the definition lies in the meaning given to the word *randomization*. In this Chapter, I shall temporarily distinguish between three types of construction that capture the idea of randomization, to which I shall give slightly artificial names: M-randomization, K-randomization, and L-randomization. Here the M is intended as a reminder of the role played by Markov kernels, the K refers to the similarity to Kolmogorov's abstract conditional expectation, and the L refers to Le Cam.

For the moment I shall ignore the question of whether the infima in the definition of $\Delta(\mathcal{P}, \mathcal{Q})$ are achieved, and refer to statistical models $\mathcal{P}$ and $\mathcal{Q}$ as being M-equivalent if there are M-randomizations $\mathbb{K}$ and $\mathbb{K}'$ for which $\mathbb{K}\mathbb{Q}_\theta = \mathbb{P}_\theta$ and $\mathbb{K}'\mathbb{P}_\theta = \mathbb{Q}_\theta$ for all θ in Θ. The definitions for K- and L-equivalence are analogous. In fact, one of the advantages of the L-randomizations will be that the infima in the definitions are achieved, whereas extra regularity conditions are needed before the same assertion can be made for the other two concepts.

K-equiv subtle, because of negligible sets

My main purpose is to explain the value of the abstract definition introduced by Le Cam (1964)—expanding on ideas of Bohneblust, Shapley & Sherman (1949), Blackwell (1951), and Blackwell (1953)—as a way to sidestep a host of "regularity conditions" that Le Cam regarded as superfluous to the statistical interpretation.

Check 64 citation

REMARK. Over the years I have slowly come to agree with much of Le Cam's argument, although I cannot claim complete comfort with some of the abstractions. As I explain in Section 7, I believe some of my difficulties are related to an unjustifiable belief in the idea that probability models should start with a choice of sample space.

2. Linear functionals and measures

For a set $\mathcal{X}$ equipped with a sigma-field $\mathcal{A}$, define

$$\mathcal{M}^+(\mathcal{X}, \mathcal{A}) := \text{all } \mathcal{A}\text{-measurable functions from } \mathcal{X} \text{ into } \overline{\mathbb{R}}^+ := \mathbb{R}_+ \cup \{\infty\},$$
$$\mathbb{M}^+(\mathcal{X}, \mathcal{A}) := \text{all bounded functions in } \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

Remember that countably additive, nonnegative measures on $\mathcal{A}$ can be identified with those increasing linear functionals on $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$ that have the Monotone Convergence property. For finite measures, it is more convenient to make an identification with increasing linear functionals on $\mathbb{M}^+(\mathcal{X}, \mathcal{A})$.

mention interpretation as finitely additive measures?

Define $\mathbb{L}^+(\mathcal{X}, \mathcal{A})$ to consist of all maps $\lambda : \mathbb{M}^+(\mathcal{X}, \mathcal{A}) \to \mathbb{R}^+$ for which

$$\lambda(\alpha_1 f_1 + \alpha_2 f_2) = \alpha_1 \lambda(f_1) + \alpha_2 \lambda(f_2) \qquad \text{for all } \alpha_i \in \mathbb{R}^+ \text{ and } f_i \in \mathbb{M}^+(\mathcal{X}, \mathcal{A}).$$

For functionals in $\mathbb{L}^+(\mathcal{X}, \mathcal{A})$, write $\lambda_1 \geq \lambda_2$ to mean that $\lambda_1 f \geq \lambda_2 f$ for all f in $\mathbb{M}^+(\mathcal{X}, \mathcal{A})$.

The finite (countably additive, nonnegative) measures on $\mathcal{A}$ correspond to the subset $\mathbb{L}^+_\sigma(\mathcal{X}, \mathcal{A})$ of such functionals that are ***σ-smooth at*** 0:

$$\text{if } \{f_n : n \in \mathbb{N}\} \subseteq \mathbb{M}^+(\mathcal{X}, \mathcal{A}) \text{ and } f_n \downarrow 0 \text{ pointwise then } \lambda f_n \downarrow 0.$$

If $\mathbb{Q} \in \mathbb{L}^+_\sigma(\mathcal{X}, \mathcal{A})$ then $\mathbb{L}^+_\mathbb{Q}(\mathcal{X}, \mathcal{A})$ will denote the subset of $\mathbb{L}^+_\sigma(\mathcal{X}, \mathcal{A})$ consisting of the measures dominated by $\mathbb{Q}$.

density needn't be bdd; still need results for $\mathcal{M}^+$

When the choice of $\mathcal{X}$ or of $\mathcal{A}$ is clear, I will abbreviate $\mathbb{M}^+(\mathcal{X}, \mathcal{A})$ to $\mathbb{M}^+(\mathcal{X})$, or $\mathbb{M}^+(\mathcal{A})$, or even just $\mathbb{M}^+$. And so on.

REMARK. The sets $\mathbb{M}^+$, $\mathbb{L}^+$, and $\mathbb{L}^+_\sigma$ are the positive cones of various vector lattices. Everything to be discussed in this Section is actually just a special case of general results for abstract vector lattices. See Appendix D for the general case.

There is a small collection of results about linear functionals that explains some of the reasons for the success of Le Cam's abstract approach to decision theory.

<2> **Lemma.** *Let λ_1 and λ_2 be functionals in $\mathbb{L}^+(\mathcal{X},\mathcal{A})$. The functionals defined by*

$$\mu f := \sup\{\lambda_1 f_1 + \lambda_2 f_2 : f = f_1 + f_2 \text{ with } f_i \in \mathbb{M}^+(\mathcal{X},\mathcal{A})\}$$
$$\nu f := \inf\{\lambda_1 f_1 + \lambda_2 f_2 : f = f_1 + f_2 \text{ with } f_i \in \mathbb{M}^+(\mathcal{X},\mathcal{A})\}$$

both belong to $\mathbb{L}^+(\mathcal{X},\mathcal{A})$. The linear functional μ is the smallest for which $\mu \geq \lambda_i$ for $i = 1, 2$, and ν is the largest for which $\nu \leq \lambda_i$ for $i = 1, 2$.

REMARK. The functional μ is usually denoted by $\lambda_1 \vee \lambda_2$, and is called the lattice-theoretic maximum of λ_1 and λ_2. Similarly ν is denoted by $\lambda_1 \wedge \lambda_2$, and is called their lattice-theoretic minimum.

Proof. It is easy to see that $\mu(g+h) \geq \mu(g) + \mu(h)$ for all $g, h \in \mathbb{M}^+$. For if $g_1 + g_2 = g$ and $h_1 + h_2 = h$ then

$$\lambda_1 g_1 + \lambda_2 g_2 + \lambda_1 h_1 + \lambda_2 h_2 = \lambda_1(g_1 + h_1) + \lambda_2(g_2 + h_2) \leq \mu(g+h).$$

Take the supremum over all such g_i pairs and h_i pairs to get the stated inequality. For the reverse inequality, suppose $f_1 + f_2 = g + h$. It is possible to split each f_i into $g_i + h_i$ in such a way that $g = g_1 + g_2$ and $h = h_1 + h_2$:

$g_1 := f_1 \wedge g$	$g_2 := (g - f_1)^+ = (f_2 - h)^+$	g
$h_1 := (h - f_2)^+ = (f_1 - g)^+$	$h_2 := f_2 \wedge h$	h
f_1	f_2	$f_1 + f_2 = g + h$

We then have

$$\mu g + \mu h \geq (\lambda_1 g_1 + \lambda_2 g_2) + (\lambda_1 h_1 + \lambda_2 h_2) = \lambda_1 f_1 + \lambda_2 f_2.$$

Take the supremum over all such f_i pairs to get $\mu(g+h) \leq \mu(g) + \mu(h)$. An even easier argument shows that $\mu(\alpha f) = \alpha \mu(f)$ for $\alpha \in \mathbb{R}^+$ and $f \in \mathbb{M}^+$. Thus $\mu \in \mathbb{L}^+$.

If γ is another functional for which $\gamma \geq \lambda_i$ for $i = 1, 2$, and if $f_1 + f_2 = f$, then $\lambda_1 f_1 + \lambda_2 f_2 \leq \gamma f_1 + \gamma f_2 = \gamma f$, which implies that $\mu \leq \gamma$.

The arguments for $\nu \in \mathbb{L}^+$ are similar. □

<3> **Corollary.** *If $\lambda_1 \in \mathbb{L}^+(\mathcal{X},\mathcal{A})$ and $\lambda_2 \in \mathbb{L}^+_\sigma(\mathcal{X},\mathcal{A})$ then $\lambda_1 \wedge \lambda_2 \in \mathbb{L}^+_\sigma(\mathcal{X},\mathcal{A})$.*

? mention interpretation: $\mathbb{L}_\sigma$ is a band in $\mathbb{L}$

Proof. If $f_n \downarrow 0$ pointwise, then $(\lambda_1 \wedge \lambda_2) f_n \leq \lambda_2 f_n \downarrow 0$. □

<4> **Theorem.** *For each λ in $\mathbb{L}^+(\mathcal{X},\mathcal{A})$ the collection $B_\lambda := \{\mu \in \mathbb{L}^+_\sigma(\mathcal{X},\mathcal{A}) : \mu \leq \lambda\}$ contains a largest member, denoted by $\pi_\sigma\lambda$. The map $\lambda \mapsto \pi_\sigma\lambda$ is linear and ...*

check projection argument for what more is needed

Proof. Define $\delta := \sup\{\mu 1 : \mu \in B_\lambda\}$. Note that $\delta \leq \lambda 1 < \infty$. Find $\{\mu_i : i \in \mathbb{N}\}$ with $\sup_i \mu_i 1 = \delta$. Let γ be a finite measure, such as $\sum_i 2^{-i}\mu_i$, that dominates each μ_i. Write m_i for the density $d\mu_i/d\gamma$. Define μ_∞ as the measure with density $m_\infty := \sup_i m_i$ with respect to γ.

To prove that $\mu \in B_\lambda$, fix an n, then define sets $A_i \in \mathcal{A}$ such that $f_i = \max_{j \leq n} f_j$ on A_i. For $f \in \mathbb{M}^+$, we have

$$\gamma\left(f \max_{j\leq n} m_j\right) = \sum\nolimits_{j\leq n} \gamma\left(f m_j A_j\right) = \sum\nolimits_{j\leq n} \mu_j\left(f A_j\right) \leq \sum\nolimits_{j\leq n} \lambda\left(f A_j\right) = \lambda f.$$

Let n increase to infinity to deduce that $\mu f \leq \lambda f$.

Clearly $\mu_\infty \geq \mu_i$ for each i, and hence $\mu_\infty 1 = \delta$. If μ_0 is any other member of B_λ, with no loss of generality we may suppose it has a density m_0 with respect

to γ. A small variation on the argument in the preceding paragraph shows that $\mu_\infty \vee \mu_0 \in B_\lambda$, and hence $\delta \geq (\mu \vee \mu_0)1 = \gamma(m_\infty \vee m_0) \geq \gamma(m_\infty) = \mu 1 = \delta$. It follows that $m_\infty \vee m_0 = m_\infty$ a.e. $[\gamma]$, and $\mu_0 \leq \mu_\infty$. The measure μ_∞ defines $\pi_\sigma(\lambda)$, the largest member of B_λ.

It is easy to see that $\pi_\sigma(\lambda_1 + \lambda_2) \geq \pi_\sigma(\lambda_1) + \pi_\sigma(\lambda_2)$ and $\pi_\sigma(\alpha\lambda) = \alpha\pi_\sigma(\lambda)$ for $\alpha \in \mathbb{R}^+$, because $\mathbb{L}_\sigma^+$ is stable under sums and multiplication by positive constants. Write μ_i for $\pi_\sigma(\lambda_i)$. To establish the reverse inequality, suppose $\mu_0 \in B_{\lambda_1+\lambda_2}$. Then $\lambda_1 + \lambda_2 = \mu_0 + \gamma_0$ for some γ_0 in $\mathbb{L}^+$. There is a decomposition $\mu_0 = \mu_1 + \mu_2$ with $\mu_i \leq \lambda_i$ for $i = 1, 2$:

$\mu_1 := \mu_0 \wedge \lambda_1$	$\gamma_1 := (\lambda_1 - \mu_0)^+ = (\gamma_0 - \lambda_1)^+$	λ_1
$\mu_2 := (\lambda_2 - \gamma_0)^+ = (\mu_0 - \lambda_1)^+$	$\gamma_2 := \gamma_0 \wedge \lambda_2$	λ_2
μ_0	γ_0	$\lambda_1 + \lambda_2$

Here $(\gamma_0 - \lambda_1)^+$ equals $\gamma_0 - \gamma_0 \wedge \lambda_1$, the smallest γ in $\mathbb{L}^+$ for which $\gamma \geq \gamma_0 - \lambda_1$, and so on. We therefore have $\mu_0 = \mu_1 + \mu_2 \leq \pi_\sigma(\lambda_1) + \pi_\sigma(\lambda_2)$, for every μ_0 in $B_{\lambda_1+\lambda_2}$. The linearity of π_σ follows. □

REMARK. It is no accident that the proofs of Lemma <2> and Theorem <4> both involved the decomposition shown in the tabular displays. In fact, both constructions correspond to the same fact for vector lattices. Perhaps it would be better to argue directly from the more general constructions in Appendix D.

3. Randomization: three possibilities

To understand the meaning given by Le Cam to the K in <1>, you should think of a randomization as a mechanism to convert observations from one distribution into observations from another distribution: sample a y from $\mathbb{Q}$, crank up the randomizer, then out pops an observation x from $\mathbb{P}$. The randomization might be provided by a Markov kernel, $\mathbb{K} := \{\mathbb{K}_y : y \in \mathcal{Y}\}$, a family of probability measures on $\mathcal{A}$ for which $\mathbb{K}_y^x f(x) \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B})$ for each f in $\mathcal{M}^+(\mathcal{X}, \mathcal{A})$. In that case, $\mathbb{P}f$ is given by

$$\mathbb{P}f = \mathbb{Q}^y \mathbb{K}_y^x f(x) \qquad \text{for } f \in \mathcal{M}^+(\mathcal{X}, \mathcal{A}).$$

More generally, we can use the kernel to define a linear map from $\mathbb{L}_\sigma^+(\mathcal{Y}, \mathcal{B})$ to $\mathbb{L}_\sigma^+(\mathcal{X}, \mathcal{A})$, by $(\mathbb{K}\mu)f := \mu^y \mathbb{K}_y^x f(x)$. If μ is a probability measure then so is $\mathbb{K}\mu$. I will call a randomization defined in this way an ***M-randomization***.

<5> **Example.** The simplest example of a M-randomization is given by a measurable map, $S : \mathcal{Y} \to \mathcal{X}$. Take $\mathbb{K}_y$ as the probability measure degenerate at the point $S(y)$ then $\mathbb{K}\mathbb{Q}$ is just the image measure $S\mathbb{Q}$, defined by $(S\mathbb{Q})f = \mathbb{Q}(f \circ S)$. □

If we focus not on the mechanism by which the x is generated from the y, but on the correspondence between the $\mathbb{Q}$ that goes in and the $\mathbb{P}$ that comes out, then it is natural to think of a randomization as a map from measures to measures that preserves total mass. This idea underlies Le Cam's definition of randomizations (which he called transitions). Define an ***L-randomization*** from $\mathcal{Y}$ to $\mathcal{X}$ to be a map K from $\mathbb{L}_\sigma^+(\mathcal{Y}, \mathcal{B})$ to $\mathbb{L}_\sigma^+(\mathcal{X}, \mathcal{A})$ for which

(i) $K(\alpha_1\mu_1 + \alpha_2\mu_2) = \alpha_1 K\mu_1 + \alpha_2 K\mu_2$ for all $\mu_i \in \mathbb{L}_\sigma^+(\mathcal{Y}, \mathcal{B})$ and $\alpha_i \in \mathbb{R}^+$,

(ii) $(K\mu)(\mathcal{X}) = \mu(\mathcal{Y})$.

Between the two extremes of M- and L-randomizations lies a concept suggested by the definition of the Kolmogorov conditional expectation from classical probability theory.

<6> **Definition.** *Let $\mathbb{Q}$ be measure on $\mathcal{B}$, and let κ be a map from $\mathbb{M}^+(\mathcal{X}, \mathcal{A})$ to $\mathbb{M}^+(\mathcal{Y}, \mathcal{B})$, with the value of κf at the point y being denoted by $\kappa(y, f)$. Call κ a* ***K-randomization*** *modulo $\mathbb{Q}$ if*

(i) $\kappa(x, 1) = 1$ a.e. $[\mathbb{Q}]$.

(ii) for each $f_1, f_2 \in \mathcal{M}_+(\mathcal{X}, \mathcal{A})$ *and each* $\alpha_1, \alpha_2 \in \mathbb{R}^+$,

$$\kappa(x, \alpha_1 f_1 + \alpha_2 f_2) = \alpha_1\kappa(y, f_1) + \alpha_2(x)\kappa(y, f_2) \qquad \text{a.e. } [\mathbb{Q}]$$

(iii) if $\{f_n\}$ *is a sequence in* $\mathbb{M}^+(\mathcal{X}, \mathcal{A})$ *that decreases pointwise to* 0 *then* $\kappa(y, f_n) \downarrow 0$ a.e. $[\mathbb{Q}]$.

REMARK. Ignore this remark. [In what sense could we think of a K-randomization $\kappa(y, \cdot)$ as a recipe for generating a x from an observation y on $\mathbb{Q}$? There need be no probability measure $\mathbb{K}_y$ for which $\mathbb{K}_y f = \kappa(y, f)$; there is no probability distribution from which to generate x. However, if we are only interested in some particular random variable $Z := Z(x)$ it is possible to define a Markov kernel $\mathbb{H}_y(\cdot)$ from $\mathcal{Y}$ to $\mathbb{R}$, in such a way that $\kappa(y, h(Z)) = \mathbb{H}_y^z h(z)$, for each h in $\mathbb{M}^+(\mathbb{R})$. Indeed, $\kappa(y, h(Z))$ defines a K-randomization modulo $\mathbb{Q}$ from $\mathcal{Y}$ into the locally compact metric space $\mathbb{R}$, so Corollary <9> applies.]

Every M-randomization defines a K-randomization, in the obvious way: $\kappa(y, f) := \mathbb{K}_y^x f(x)$. The precautions about $\mathbb{Q}$-negligible sets are not needed in this case. Indeed, it is largely the accumulation of uncountably many negligible sets that prevents us, in general, from finding a single $\mathbb{Q}$-negligible set $\mathcal{N}$ such that $\kappa(y, \cdot)$ is a probability measure for each y in $\mathcal{N}^c$.

Every K-randomization defines a L-randomization, but the construction must accommodate the fact that $\kappa(y, \cdot)$ might behave badly on $\mathbb{Q}$-neglible sets. If $\mu \in \mathbb{L}_\mathbb{Q}^+(\mathcal{Y}, \mathcal{B})$, that is, if $\mu \in \mathbb{L}_\sigma^+(\mathcal{Y}, \mathcal{B})$ and $\mu \ll \mathbb{Q}$, then $(\kappa\mu) f := \mu^y \kappa(y, f)$ defines a linear functional in $\mathbb{L}_\sigma^+(\mathcal{X}, \mathcal{A})$. (Countable additivity, in the form of σ-smoothness at 0, comes from property (iii) of κ.) In fact, $\kappa\mu \in \mathbb{L}_\mathbb{P}^+(\mathcal{X}, \mathcal{A})$, where $\mathbb{P} := \kappa\mathbb{Q}$. For if $0 = \mathbb{P} f = \mathbb{Q}\kappa(y, f)$, with $f \in \mathbb{M}^+(\mathcal{X}, \mathcal{A})$, then $\kappa(y, f) = 0$ a.e. $[\mathbb{Q}]$, which implies $\kappa(y, f) = 0$ a.e. $[\mu]$ and $(\kappa\mu) f = 0$.

If μ does not belong to $\mathbb{L}_\mathbb{Q}^+(\mathcal{Y}, \mathcal{B})$, write it as $\tilde{\mu} + \mu^\perp$, with $\tilde{\mu} \ll \mathbb{Q}$ and $\mu^\perp$ singular, that is, it concentrates on a $\mathbb{Q}$-negligible set. For fixed probability measure P_0 on $\mathcal{A}$ (the same for every μ), define a measure $\kappa\mu$ in $\mathbb{L}_\sigma^+(\mathcal{X}, \mathcal{A})$ by

$$(\kappa\mu) f := \tilde{\mu}^y \kappa(y, f) + \mu^\perp(\mathcal{Y}) P_0 f \qquad \text{for } f \in \mathbb{M}^+(\mathcal{X}, \mathcal{A}).$$

Linearity of the map $\mu \mapsto \tilde{\mu}$ ensures that κ is linear in μ. The construction also ensures that $(\kappa\mu)(\mathcal{X}) = \tilde{\mu}\mathcal{Y} + \mu^\perp\mathcal{Y} = \mu\mathcal{Y}$. That is, κ is an L-randomization.

REMARK. Ignore this remark. [The proof works because $\mathbb{L}_\mathbb{Q}$ is a band in $\mathbb{L}_\sigma$. The map $\mu \mapsto \tilde{\mu}$ is the projection onto the band. See Appendix D.]

The construction from the last two paragraphs can be run in reverse, to build many K-randomizations from any L-randomization K.

<7> **Theorem.** *For each L-randomization K and each $\mathbb{Q}$ in $\mathbb{L}^+_\sigma(\mathcal{Y},\mathcal{B})$ there exists a K-randomization κ modulo $\mathbb{Q}$ such that $\kappa\mu = K\mu$ for each μ in $\mathbb{L}^+_{\mathbb{Q}}(\mathcal{Y},\mathcal{B})$.*

better to use $\mathbb{M}^+$ then pass to limit?

Proof. For $g \in \mathcal{M}^+(\mathcal{Y})$ with $\mathbb{Q}g < \infty$, write $\mathbb{Q}_g$ for the finite measure with density g with respect to $\mathbb{Q}$. Then define $\nu_f(g) := (K\mathbb{Q}_g)(f)$ for $f \in \mathbb{M}^+(\mathcal{X})$. For fixed f, the map $g \mapsto \nu_f(g)$ is linear. If $0 \le f \le C$, with C constant, then

$$\nu_f(g) \le C\left(K\mathbb{Q}_g\right)(1) = C\mathbb{Q}_g 1 = C\mathbb{Q}g.$$

Thus $\nu_f \in \mathbb{L}^+_{\mathbb{Q}}(\mathcal{Y})$, with density $\kappa(y, f)$ with respect to $\mathbb{Q}$ bounded above by C:

$$(K\mathbb{Q}_g)(f) = \nu_f(g) = \mathbb{Q}^y\left(g(y)\kappa(y, f)\right) \qquad \text{for } f \in \mathbb{M}^+(\mathcal{X}) \text{ and } g \in \mathbb{M}^+(\mathcal{Y}).$$

We need to check properties (i), (ii), and (ii) of Definition <6>. For $f \equiv 1$ we have $\nu_1 = \mathbb{Q}$, which lets us choose $\kappa(y, 1) \equiv 1$. For fixed g, the map $f \mapsto \nu_f(g)$ is linear. Thus, for all $\alpha_i \in \mathbb{R}^+$ and all $g \in \mathbb{M}^+(\mathcal{Y})$,

$$\mathbb{Q}\left(g(y)\kappa(y, \alpha_1 f_1 + \alpha_2 f_2)\right) = \alpha_1\mathbb{Q}\left(g(y)\kappa(y, f_1)\right) + \alpha_2\mathbb{Q}\left(g(y)\kappa(y, f_2)\right).$$

With strategic choices for g we then recover the linearity property (ii). (In consequence, if $f_1 \le f_2$ then $\kappa(y, f_1) \le \kappa(y, f_2)$ a.e. $[\mathbb{Q}]$.)

Finally, if $\{f_i : i \in \mathbb{N}\} \subseteq \mathbb{M}^+(\mathcal{X})$ and $f_i \downarrow 0$ pointwise, then

$$\mathbb{Q}^y\kappa(y, f_n) = (K\mathbb{Q})f_n \downarrow 0,$$

from which property (iii) follows. Thus κ is a K-randomization. The definining property for κ becomes $(K\mu)(f) = \mu^y\kappa(y, f) = (\kappa\mu)(f)$ if we write μ for $\mathbb{Q}_g$. □

> REMARK. Not surprisingly, the method of proof for the Theorem is based on the same idea as the method for proving existence of conditional expectations in Kolmogorov's sense.

<8> **Example.** Let $\mathcal{X} = \mathcal{Y} = [0, 1]$ equipped with its Borel sigma-field $\mathcal{A} = \mathcal{B}$. Define an -randomization K by $K\delta_y := \delta_{1-y}$ for point masses, and $K\nu := \nu$ for nonatomic ν. Extend by linearity. Note that the K cannot be represented by a Markov kernel. □

> REMARK. Ignore this remark. [The Example will be worth a revisit when we consider the Kakutani representation, whereby every L-randomization is represented by a Markov kernel. Something interesting happens to the sample space where the measures live, so that point masses live on a part of the space disjoint from the support of nonatomic measures.]

Under topological regularity conditions, the conclusion of Theorem <7> can be strengthened, replacing the K-randomization by an M-randomization.

<9> **Corollary.** *If $\mathcal{X}$ is a locally compact metric space, with $\mathcal{A}$ its Borel sigma-field, then the conclusion of Theorem <7> holds with κ replaced by a Markov kernel.*

<10> **Corollary.** *If $\mathcal{X}$ is a compact Hausdorff space, with $\mathcal{A}$ its Borel sigma-field, and if $\mathcal{B}$ contains all $\mathbb{Q}$-negligible sets, then the conclusion of Theorem <7> holds with κ replaced by a Markov kernel.*

Is $\mathbb{Q}$ completion really needed?

Use separability argument for Corollary <9>, as in construction of disintegration. Use linear lifting for Corollary <10>. In both cases, build the kernel as a linear functional on the continuous functions in $\mathbb{M}^+(\mathcal{X})$ with compact support.

The bottom line: the distinctions between the three forms of randomizations are largely a matter of regularity conditions to manage negligible sets, and choice of the state space to get σ-smoothness automatically for continuous functions with compact support.

4. Conditioning

In classical probability theory, randomizations also appear in the study of conditioning. Consider the case where T is a measurable map from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$, with $\mathbb{Q} = T\mathbb{P}$, the image of some probability measure $\mathbb{P}$ on $\mathcal{A}$.

The strongest concept is that of a disintegration, or regular conditional distribution. The conditional distribution for $\mathbb{P}$ is then a Markov kernel $\mathbb{K}$ from $(\mathcal{Y}, \mathcal{B})$ to $(\mathcal{X}, \mathcal{A})$, for which $\mathbb{K}\mathbb{Q} = \mathbb{P}$ and $\mathbb{K}_y\{T \neq y\} = 0$ for $\mathbb{Q}$-almost all y. Sometimes the integral $\mathbb{K}_y^x f(x)$ is written $\mathbb{P}(f \mid T = y)$, which, unfortunately, is also used to denote the related Kolmogorov conditional expectation.

REMARK. Unfortunately, existence of regular conditional distributions does not come for free. If we want conditional distributions represented by Markov kernels we have to impose regularity conditions on the map T and on the probability measures. Le Cam would probably have pointed to these regularity conditions as unwanted intrusions on the idea of randomization.

reference?

The Kolmogorov conditional expectation can be defined in any of three equivalent ways. It is a K-randomization modulo $\mathbb{Q}$ for which either

(i) $\mathbb{P}^x(f(x)g(Tx)) = \mathbb{Q}^y(g(y)\kappa(y, f))$ for all $g \in \mathcal{M}^+(\mathcal{Y})$ and $f \in \mathbb{M}^+(\mathcal{X})$

Use $\mathcal{M}^+$ or $\mathbb{M}^+$ for g?

(ii) $\kappa\mathbb{Q} = \mathbb{P}$ and $\kappa(y, (g \circ T) f) = g(y)\kappa(y, f)$ a.e. $[\mathbb{Q}]$, for all $f \in \mathbb{M}^+(\mathcal{X})$ and $g \in \mathbb{M}^+(\mathcal{Y})$

(iii) for each μ in $\mathbb{L}^+_{\mathbb{Q}}(\mathcal{Y}, \mathcal{B})$, if $d\mu/d\mathbb{Q} = g$ then $d(\kappa\mu)/d\mathbb{P} = g \circ T$.

even if g is not bdd?

The form (iii) suggests that we might regard a Le Cam randomization as representing conditioning on T if $d(K\mu)/d\mathbb{P} = g \circ T$ when $d\mu/d\mathbb{Q} = g$. Under conditions such as those of Corollary <9> or <10>, and a mild assumption such as product measurability of $\{(x, Tx) : x \in \mathcal{X}\}$, this definition would lead to a regular conditional distribution.

details?

5. Sufficiency and canonical measures

Let $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ be a statistical model, with all the measures living on $(\mathcal{X}, \mathcal{A})$, and let T be a measurable map from $(\mathcal{X}, \mathcal{A})$ into $(\mathcal{Y}, \mathcal{B})$. Roughly speaking, if it is possible to recover $\mathbb{P}_\theta$ from $\mathbb{Q}_\theta := T\mathbb{P}_\theta$ by a randomization *which does not depend on* θ, then the measurable map T is said to be ***sufficient*** for $\mathcal{P}$. More formally, say

that T is ***M-sufficient*** if there is a M-randomization for which $\mathbb{K}\mathbb{Q}_\theta = \mathbb{P}_\theta$ for all θ, with analogous definitions for ***K–sufficient*** and ***L–sufficient***.

REMARK. The K-sufficiency requires a κ that is a K-randomization modulo $\mathbb{Q}_\theta$ for every θ. If $\mathcal{P}$ is not dominated by a sigma-finite measure, the construction of such a κ would require extremely delicate handling of negligible sets.

Statistical folklore asserts that, if $\mathcal{P}$ is dominated by some $\mathbb{P}_{\theta_0}$ then the likelihood ratios $\rho_\theta(x) := d\mathbb{P}_\theta/d\mathbb{P}_{\theta_0}$ are sufficient. More formally, the map $x \mapsto \{\rho_\theta(x) : \theta \in \Theta\}$ from $\mathcal{X}$ into $\mathbb{R}^\Theta$ is supposed to be sufficient, in some sense that is seldom made explicit. It is possible to make the idea more precise in several cases. To avoid difficulties related to the the choice of versions of densities and the handling of infinite families of negligible sets, let us first consider the case where the index set Θ is finite, dominated by a sigma-finite measure that need not necessarily be a member of $\mathcal{P}$.

<11> **Theorem.** *Let the statistical model $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$, with Θ finite, be dominated by a measure λ, for which $d\mathbb{P}_\theta/d\lambda = p_\theta$. Define $T : \mathcal{X} \to \mathbb{R}_+^\Theta$ as the map taking x to the point with coordinates $p_\theta(x)$. Then T is K-sufficient for $\mathcal{P}$.*

Proof. Without loss of generality suppose $\Theta = \{1, 2, \ldots, k\}$. Write $\mathbb{Q}_i$ for $T\mathbb{P}_i$, for $i = 1, \ldots, k$. Define $\mathbb{P} := (\mathbb{P}_1 + \ldots + \mathbb{P}_k)/k$ and $\mathbb{Q} := T\mathbb{P} = (\mathbb{Q}_1 + \ldots + \mathbb{Q}_k)/k$. Let κ be the Kolmogorov conditional expectation for $\mathbb{P}$ given T. That is, κ is a K-randomization modulo $\mathbb{Q}$ (and hence also modulo $\mathbb{Q}_i$ for each i) such that: if μ is a finite measure with density $d\mu/d\mathbb{Q} = g$ then $d(\kappa\mu)/d\mathbb{P} = g(Tx)$. It is enough to show that $\kappa\mathbb{Q}_i = \mathbb{P}_i$ for each i.

For each i and $y \in \mathbb{R}_+^k$, define $q_i(y) := ky_i\{y_1 + \ldots + y_k > 0\}/(y_1 + \ldots + y_k)$. The measure $\mathbb{P}_i$ has density

$$q_i(Tx) = \frac{kp_i(x)}{p_1(x) + \ldots + p_k(x)}\{p_1(x) + \ldots + p_k(x) > 0\}$$

with respect to $\mathbb{P}$. (More precisely, the ratio is one possible choice for the density.) It follows that $\mathbb{Q}_i$ has density q_i with respect to $\mathbb{Q}$, because

$$\mathbb{Q}_i g(y) = \mathbb{P}_i^x g(Tx) = \mathbb{P}^x q_i(Tx) g(Tx) = \mathbb{Q}^y \left(q_i(y) g(y)\right).$$

By definition of κ, the measure $\kappa\mathbb{Q}_i$ has density $q_i(Tx)$ with respect to $\mathbb{P}$. That is, $\kappa\mathbb{Q}_i = \mathbb{P}_i$, as required. □

To do: Develop general form of the factorization theorem for sufficiency.

REMARK. Ignore this remark. [Compare with Le Cam (1986, Sections 3.1 and 3.2) for general Θ. Maybe state as L-equivalence for general Θ.]

As the proof of the Theorem <11> shows, it is convenient to take $\lambda := \sum_\theta \mathbb{P}_\theta$ as the dominating measure when Θ is finite. The densities $p_\theta(x) := d\mathbb{P}_\theta/d\lambda$ then sum to one. The vector of densities $T(x) := \{p_\theta(x) : \theta \in \Theta\}$ maps $\mathcal{X}$ into the simplex $\mathcal{S}_\Theta := \{\mathbf{y} \in \mathbb{R}_+^\Theta : \sum_\theta y_\theta = 1\}$. The image measure $\mu := T\lambda$ concentrates on the simplex, with total mass equal to the cardinality of Θ. The measure μ is called the ***canonical measure*** of $\mathcal{P}$. The image measure $\mathbb{Q}_\theta := T\mathbb{P}_\theta$ has μ-density y_θ for

each θ. The statistical model $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ is K-equivalent to $\mathcal{P}$. If two statistical models have the same canonical measure then they must be equivalent, because the canonical measures generate the same $\mathcal{Q}$ on $\mathbb{R}_+^\Theta$. The converse proposition is slightly less obvious.

<12> **Theorem.** *If $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ and $\widetilde{\mathcal{P}} := \{\widetilde{\mathbb{P}}_\theta : \theta \in \Theta\}$ are L-equivalent, and Θ is finite, then both models have the same canonical measure.*

Proof. With no loss of generality, assume $\Theta := \{1, 2, \ldots, k\}$. Write $k\mathbb{Q}$ and $k\widetilde{\mathbb{Q}}$ for the two canonical measures, living on separate copies $\mathcal{X}$ and $\mathcal{Y}$ of the simplex $\mathcal{S}_k$. They generate statistical models $\mathcal{Q} = \{\mathbb{Q}_i : i \in \Theta\}$ and $\widetilde{\mathcal{Q}} = \{\widetilde{\mathbb{Q}}_\theta : \theta \in \Theta\}$, with $d\mathbb{Q}_i/d\mathbb{Q} = kx_i$ and $d\widetilde{\mathbb{Q}}_i/d\widetilde{\mathbb{Q}} = ky_i$. Notice that $k\mathbb{Q} = \sum_i \mathbb{Q}_i$ and $k\widetilde{\mathbb{Q}} = \sum_i \widetilde{\mathbb{Q}}_i$. We are given that $\mathcal{Q}$ and $\widetilde{\mathcal{Q}}$ are L-equivalent. We need to deduce that $\mathbb{Q} = \widetilde{\mathbb{Q}}$.

As $\mathcal{S}_k$ is a compact metric space, L-equivalence implies M-equivalence. Thus there exist Markov kernels $\mathbb{K}$ (from $\mathcal{X}$ to $\mathcal{Y}$) and $\widetilde{\mathbb{K}}$ (from $\mathcal{Y}$ to $\mathcal{X}$) for which

$$\mathbb{K}\mathbb{Q}_i = \widetilde{\mathbb{Q}}_i \qquad \text{and} \qquad \widetilde{\mathbb{K}}\widetilde{\mathbb{Q}}_i = \mathbb{Q}_i \qquad \text{for each } i.$$

Sum over i to deduce that $\mathbb{K}\mathbb{Q} = \widetilde{\mathbb{Q}}$ and $\widetilde{\mathbb{K}}\widetilde{\mathbb{Q}} = \mathbb{Q}$.

The kernel $\mathbb{K}$ and the measure $\mathbb{Q}$ define a probability measure $\Gamma := \mathbb{Q} \otimes \mathbb{K}$ on $\mathcal{X} \times \mathcal{Y}$, with marginals $\mathbb{Q}$ and $\mathbb{K}\mathbb{Q} = \widetilde{\mathbb{Q}}$. We can also write Γ as $\widetilde{\mathbb{Q}} \otimes \mathbb{H}$, with $\mathbb{H}$ a Markov kernel from $\mathcal{Y}$ to $\mathcal{X}$.

REMARK. At the moment we know nothing about the relationship between the two kernels $\widetilde{\mathbb{K}}$ and $\mathbb{H}$, except that $\widetilde{\mathbb{K}}\widetilde{\mathbb{Q}} = \mathbb{Q} = \mathbb{H}\widetilde{\mathbb{Q}}$. In particular, we cannot assert that $\widetilde{\mathbb{Q}} \otimes \widetilde{\mathbb{K}} = \Gamma$.

The 'conditional mean' $\mathbb{H}_y^x x$ has a striking property. For each g in $\mathbb{M}^+(\mathcal{Y})$,

$$\begin{aligned}
\widetilde{\mathbb{Q}}\left(y_i g(y)\right) &= \widetilde{\mathbb{Q}}_i^y g(y) \\
&= \left(\mathbb{K}\mathbb{Q}_i\right)^y g(y) \\
&= \mathbb{Q}^x \left(x_i \mathbb{K}_x^y g(y)\right) \\
&= \Gamma^{x,y}\left(x_i g(y)\right) \\
&= \widetilde{\mathbb{Q}}^y \mathbb{H}_y^x \left(x_i g(y)\right) = \widetilde{\mathbb{Q}}^y \left(g(y) \mathbb{H}_y^x x_i\right).
\end{aligned}$$

Thus $\mathbb{H}_y^x x_i = y_i$ a.e. $[\widetilde{\mathbb{Q}}]$. Consequently, the cross-product term in

$$\mathbb{H}_y^x (x_i - 1)^2 = \mathbb{H}_y^x \left((x_i - y_i)^2 + 2(x_i - y_i)(y_i - 1) + (y_i - 1)^2\right)$$

vanishes, leaving an identity that integrates to

<13> $$\mathbb{Q}^x (x_i - 1)^2 = \widetilde{\mathbb{Q}}^y \mathbb{H}_y^x (x_i - 1)^2 = \Gamma^{x,y}(x_i - y_i)^2 + \widetilde{\mathbb{Q}}^y (y_i - 1)^2 \geq \widetilde{\mathbb{Q}}^y (y_i - 1)^2.$$

If we reverse the roles of $\mathcal{Q}$ and $\widetilde{\mathcal{Q}}$, and repeat the whole argument leading to <13> (with a joint distribution different from Γ), we arrive at a similar identity with the roles of x and y reversed, which implies that $\widetilde{\mathbb{Q}}^y (y_i - 1)^2 \geq \mathbb{Q}(x_i - 1)^2$. We must therefore conclude that $\widetilde{\mathbb{Q}}^y (y_i - 1)^2 = \mathbb{Q}(x_i - 1)^2$, and hence $\Gamma^{x,y}(x_i - y_i)^2 = 0$. That is, $x_i = y_i$ a.e. $[\Gamma]$. The joint distribution Γ concentrates on the diagonal $\{(x, y) \in \mathcal{X} \times \mathcal{Y} : x = y\}$. Its marginals, $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$, are therefore equal. □

6. Advantages of Le Cam randomizations

Expand the following into a real proof.

Not yet edited from here onwards. Old notation

<14> **Theorem.** *Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ and $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ be statistical models with the property that, for each finite subset S of Θ there exists an L-randomization ρ_S such that $K_S\mathbb{P}_\theta = \mathbb{Q}_\theta$ for each $\theta \in S$. Then there exists an L-randomization K such that $K\mathbb{P}_\theta = \mathbb{Q}_\theta$ for every θ in Θ.*

Proof. Regard the collection $\mathcal{S}$ of all finite subsets of Θ as a directed set, ordered by inclusion. Let $\{S_i : i \in \mathcal{I}\}$ be a universal subnet of $\{K_S : S \in \mathcal{S}\}$. For each $g \in \mathbb{M}(\mathcal{Y})$ and each λ in $\mathrm{ca}(\mathcal{X})$, the net $\{\lambda K_S g : S \in \mathcal{S}\}$ takes value in a bdd interval. Hence

$$R(\lambda, g) = \lim_i \lambda K_{S_i} g$$

is well defined and finite. As a functional on $\mathbb{M}(\mathcal{Y})$ the map $g \mapsto R(\lambda, g)$ is increasing and linear. Identify R with a linear map from $\mathrm{ca}_+(\mathcal{X})$ into the Banach lattice $\mathbb{M}_+(\mathcal{Y})$. It defines an element of the dual space $\mathbb{M}_+(\mathcal{Y})$, which contains the space $\mathrm{ca}(\mathcal{Y})$ as a band. Write τ for the projection operator onto that band. Check that

$$K\lambda = \tau \circ R + (1 - \|\tau \circ R\|)\mathbb{Q}_0$$

for a fixed but arbitrary probability measure $\mathbb{Q}_0$ on $\mathcal{Y}$, is a Le Cam-randomization that maps each $\mathbb{P}_\theta$ onto the corresponding $\mathbb{Q}_\theta$. □

7. Choice of sample space

Meaning of equivalence?

Le Cam: the space $(\mathcal{X}, \mathcal{A})$ is not uniquely determined. Effectively he regarded experiments as equivalent if they generate isomorphic Banach lattices.

8. Problems

[1] Use a linear lifting (Appendix AppLifting) to select a "linear version" of the Kolmogorov conditional expectation.

[2] Show how to get Markov-randomization for a function taking values in a locally compact Hausdorff space, for dominated experiments.

[3] Let $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ be dominated by a sigma-finite measure λ. Show that $\mathcal{P}$ is also dominated by a probability measure of the form $\mathbb{P} = \sum_{i=1}^{\infty} 2^{-i}\mathbb{P}_{\theta_i}$ for some choice of $\{\theta_i\}$.

9. Notes

See Schaefer (1986) for the theory of topological vector spaces, including the separation theorems for locally convex, topological vector spaces..

See Schaefer (1974) for facts about L-spaces and M-spaces.

Torgersen (1991), Millar (1983), Strasser (1985), van der Vaart (1984). Le Cam (1986, Section 2.3) for characterization of δ-distance via comparison of risks. Le Cam (1964).

Blackwell for canonical measure.

Following Bohneblust et al. (1949) Blackwell (1953) and Blackwell (1951), Le Cam defined a notion of equivalence, between two experiments with the same index set, using an extension of the idea of randomization...

Blackwell for canonical measure. Explain connection with dilation. Check Strassen paper.

References

Blackwell, D. (1951), Comparison of experiments, *in* 'Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability', pp. 93–102.

Blackwell, D. (1953), 'Equivalent comparisons of experiments', *Annals of Mathematical Statistics* **24**, 265–272.

Bohneblust, H. F., Shapley, L. S. & Sherman, S. (1949), Reconnaisance in game theory, Technical report, Rand Corporation.

Kolmogorov, A. N. (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin. (Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.).

Le Cam, L. (1964), 'Sufficiency and approximate sufficiency', *Annals of Mathematical Statistics* **35**, 1419–1455.

Le Cam, L. (1986), *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Millar, P. W. (1983), 'The minimax principle in asymptotic statistical theory', *Springer Lecture Notes in Mathematics* **976**, 75–265.

Schaefer, H. H. (1974), *Banach Lattices and Positive Operators*, Springer-Verlag.

Schaefer, H. H. (1986), *Topological Vector Spaces*, Springer-Verlag.

Strasser, H. (1985), *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*, De Gruyter, Berlin.

Torgersen, E. (1991), *Comparison of Statistical Experiments*, Cambridge University Press.

van der Vaart, A. (1984), Limits of experiments, Lecture notes, December 1994. Part of a book to be published by Cambridge University Press.