Chapter 1

Introduction

Danger: very rough draft from page 10 onwards

SECTION *1 describes the standard mathematical formulation of the statistical decision problem.*

SECTION *2 defines Le Cam's distance between statistical models.*

1.1 Decision theory

Introduction::decision

The standard framework for statistical decision theory consists of three parts:

(i) A ***statistical model***, which is just an indexed set $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ of probability measures all defined on the same $(\mathcal{X}, \mathcal{A})$, for some set $\mathcal{X}$ equipped with a sigma-field $\mathcal{A}$. Some authors call such a model a ***statistical experiment***. The data correspond to a point in the set $\mathcal{X}$. The elements of Θ are sometimes called the states of Nature.

(ii) A space $\mathbb{D}$ of possible actions or decisions that the Statistician can take after observing x. For example, in estimation problems we can take $\mathbb{D} = \Theta$ and for the simplest of hypothesis testing $\mathbb{D}$ could be just a two-point set. To make sense of various integrals we need $\mathbb{D}$ to be equipped with a sigma-field $\mathcal{D}$.

(iii) A loss function $L : \Theta \times \mathbb{D} \to (-\infty, \infty]$, with the interpretation that action $a \in \mathbb{D}$ incurs a loss $L(\theta, a)$ when θ is the true state of Nature.

For the most part, I will be assuming that the loss function L is nonnegative.

The Statistician's task is to choose a ***decision rule***, a map $d : \mathcal{X} \to \mathbb{D}$ that is $\mathcal{A}\backslash\mathcal{D}$–measurable. More generally, the decision can be randomized, a concept that is usually expressed by means of a Markov kernel.

Markov.kernel <1>

Definition. *A Markov kernel K from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ is a collection of probability measures $K = \{K_x : x \in \mathcal{X}\}$ on $\mathcal{B}$ for which the map $x \mapsto K_x B$ is $\mathcal{A}$-measurable for each fixed B in $\mathcal{B}$.*

Many authors take a ***randomized decision rule*** to be a Markov kernel from $(\mathcal{X}, \mathcal{A})$ to $(\mathbb{D}, \mathcal{D})$, but some parts of the decision theory require a more general concept expressible by means of Kolmogorov conditional expectations. Lucien Le Cam made a case for an even broader definition of randomization by means of linear maps between spaces of finite measures. See Chapter 2 for these generalizations and the arguments for why they lead to a cleaner mathematical theory.

Each decision rule has a ***risk function***, the function of θ defined by the expected losses for the rule,

$$R(\theta, d) = \int L(\theta, d(x))\, \mathbb{P}_\theta(dx) = \mathbb{P}_\theta^x L(\theta, d(x)).$$

The last expression is the more concise linear functional way to write the expectation. See Pollard (2001, Section 1.4) for a gentle introduction to linear functional notation. More generally, a randomized decision rule $\delta = \{\delta_x : x \in \mathcal{X}\}$ has risk function

$$R(\theta, \delta) = \iint L(\theta, a)\delta_x(da)\, \mathbb{P}_\theta(dx) = \mathbb{P}_\theta^x \delta_x^a L(\theta, a)$$

The main idea of statistical decision theory is: decision rules should be compared using only their risk functions. If $R(\theta, \delta) \le R(\theta, \delta')$ for all $\theta \in \Theta$, with strict inequality for at least one θ, then δ' is inferior to δ. Some statisticians go further, summarizing the virtues of a rule by means of its ***minimax risk***,

$$R_{\mathrm{mmax}}(\Theta, \delta) = \sup_{\theta \in \Theta} R(\theta, \delta)$$

or its ***Bayes risk*** for some prior probability distribution π on Θ,

$$R(\pi, \delta) = \int R(\theta, \delta)\, \pi(d\theta) = \pi^\theta \mathbb{P}_\theta^x \delta_x^a L(\theta, a).$$

Of course Θ should be equipped with its own sigma field $\mathcal{T}$ and the loss function should be $\mathcal{T} \otimes \mathcal{D}$-measurable if we are not to run into measure theoretic complications.

A few examples will help to prepare the ground for the definition of Le Cam's distance, as well as providing some notation. The following three models have been much studied in recent years as pieces of the prime example of successful application of Le Cam theory to nonparametric settings. In each case Θ is some (large) set of probability densities defined on (the Borel sigma-field of) $[0, 1]$. For each θ in Θ, write P_θ for the probability measure with density θ with respect to Lebesgue measure.

iid <2> **Example.** The n-fold product measure $\mathbb{P}_{\theta,n} := P_\theta^n$ defines the joint distribution of n independent observations from P_θ. Under $\mathbb{P}_{\theta,n}$ the coordinate maps $x_1, \ldots, x_n$ on $[0,1]^n$ are independent random variables, each with distribution P_θ.

□

Poisson.proc <3> **Example.** Let μ be a finite measure on the Borel sigma-field of $[0,1]$. A Poisson process with intensity measure μ is a stochastic process that produces a random finite subset of points in $[0,1]$ such that

(i) the number of points landing in a Borel set A has a Poisson distribution with mean μA

(ii) for each finite collection of disjoint Borel sets $A_1, \ldots, A_k$ the numbers of points landing in each A_j are independent random variables.

Let $\mathbb{P}_{\theta,n}$ be the distribution of the Poisson process with intensity measure $\mu = nP_\theta$. A realization of the process could be constructed by first generating a random variable N with a Poisson(n) distribution then, if $N = k$, generating k independent observations from P_θ. Many authors would consider this recipe a completely adequate way of describing a Poisson process. A more formal description would take $\mathbb{P}_{\theta,n}$ as a probability measure on the set $\mathcal{X}$ of all measures on $\mathcal{B}[0,1]$ expressible as a finite sum of point masses. The sigma-field $\mathcal{A}$ would be the smallest for which each of the maps $x \mapsto x(B)$, with $B \in \mathcal{B}[0,1]$, is an $\mathcal{A}\backslash\mathcal{B}(\mathbb{R}^+)$-measurable map from $\mathcal{X}$ into $\mathbb{R}^+$.

□

whitenoise <4> **Example.** Observe a continuous process

$$Y_t = 2\int_0^t \sqrt{\theta(s)}ds + n^{-1/2}B_t \qquad \text{for } 0 \le t \le 1,$$

where B is a standard Brownian motion (with continuous sample paths). The distribution $\mathbb{P}_{\theta,n}$ of Y is a probability measure on the Borel sigma field of the space $C[0,1]$ equipped with its uniform metric.

□

Remark. Under some conditions on Θ, I will show in a later Chapter that all three models are asymptotically equivalent in Le Cam's sense.

The concept of sufficiency plays a special role in decision theory. It allows calculations for a statistical model to be reduced to calculations with a simpler statistical model defined by the distributions of a sufficient statistic.

The study of sufficiency is a natural precursor to the study of a distance, between statistical models that share the same index set, defined by Lucien Le Cam. Models at zero Le Cam distance are equivalent in Blackwell's sense, a notion very closely related to the idea of sufficiency.

normal.shift <5>

Example. Suppose $X_1, \ldots, X_n$ are independent random variables, each distriuted $N(\theta, 1)$ for some unknown real θ. Suppose we wish to estimate θ using only the X_i's. That is, $\Theta = \mathbb{D} = \mathbb{R}$. It is very common for statisticians to use the squared-error loss function, $L(\theta, a) = (\theta - a)^2$.

To fit this problem into the framework described at the start of the Section, take $\mathcal{X} = \mathbb{R}^n$ equipped with its Borel sigma-field. The X_i's can then be taken as the coordinate maps: if $x = (x_1, \ldots, x_n) \in \mathbb{R}_n$ then $X_i(x) = x_i$. [Actually there seems little point in inventing a name for the observed random variables; the x_i's suffice.] The probability measure $\mathbb{P}_\theta$ is then the multivariate normal distribution, $N(\theta \mathbf{1}, I_n)$.

The sample mean, $Y(x) = \sum_{i \le n} x_i/n$, is a sufficient statistic for this problem. It has distribution $\mathbb{Q}_\theta = N(\theta, 1/n)$ under the $\mathbb{P}_\theta$ model. Observation of Y alone corresponds to a new statistical model, $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$, a set of probability measures defined on the Borel sigma-field of the real line.

There are various ways to express the sufficiency property. The essential idea is that a "probabilistic copy" $\widetilde{X}_1, \ldots, \widetilde{X}_n$, of the full sample can be created from Y without knowing θ, by means of some auxiliary randomization That is the new variables should also have joint distribution $\mathbb{P}_\theta$. Consequently, the performance of any statistical procedure based on the X_i's can be matched by an analogous procedure based on the $\widetilde{X}_i$'s. For example, for every loss function, the risk function for an estimator $T_n = T_n(X_1, \ldots, X_n)$ is exactly the same as the risk function for the randomized estimator $T_n(\widetilde{X}_1, \ldots, \widetilde{X}_n)$. The auxiliary randomization turns the deterministic function T_n on $\mathbb{R}^n$ into a randomized estimator for the $\mathcal{Q}$ model.

The $\widetilde{X}_i$'s could be constructed as follows. Independently of Y, generate independent $N(0, 1)$ distributed random variables $Z_1, \ldots, Z_n$ then define

$$\widetilde{X}_i = Y + Z_i - \overline{Z} \qquad \text{for } i = 1, 2, \ldots, n, \text{ where } \overline{Z} := \sum\nolimits_{i \le n} Z_i/n.$$

The joint distribution of $Z_1 - \overline{Z}, \ldots, Z_n - \overline{Z}$ has a $N(0, V)$ distribution, where $V = I_n - \mathbf{1}\mathbf{1}'/n$. The construction could also be described via a Markov kernel:

$$\widetilde{X}_1, \ldots, \widetilde{X}_n \mid Y = y \sim K_y := N(y\mathbf{1}, V)$$

The Z_i's provide one way to generate an observation from K_y. The relationship between the $\mathcal{P}$ and $\mathcal{Q}$ models can be expressed as

$$\mathbb{P}_\theta^a = \mathbb{Q}_\theta^y K_y^a,$$

or, if we think of K as a mapping of probability measures on $\mathcal{Y}$ to probability measures on $\mathcal{X}$,

$$\mathbb{P}_\theta = K\mathbb{Q}_\theta \qquad \text{for all } \theta \in \Theta.$$

For every randomized decision rule $\delta = \{\delta_x : x \in \mathcal{X}\}$ for $\mathcal{P}$ there is a corresponding randomized rule for $\mathcal{Q}$, which can written symbolically as

$$\widetilde{\delta}_y^a = K_y^x \delta_x^a.$$

Less formally, for the $\mathcal{P}$ model we generate a random action by

$$x \sim \mathbb{P}_\theta \qquad \text{then} \quad a \mid x \sim \delta_x$$

while for the $\mathcal{Q}$ model the scheme is

$$y \sim \mathbb{Q}_\theta \qquad \text{then} \quad x \sim \mathbb{P}_\theta \qquad \text{then} \quad a \mid x \sim \delta_x$$

For every loss function, $R(\theta, \widetilde{\delta}) = R(\theta, \delta)$ for every θ. That is, if we judge performance only by risk functions, we can always do as well with the $\mathcal{Q}$ model as with the $\mathcal{P}$ model. In the language of the next Section, the $\mathcal{P}$ and $\mathcal{Q}$ models are statistically equivalent.

□

uniform.suff <6> **Example.** Suppose X_1 and X_2 are independent observations on the unif$(0, \theta)$, for some $\theta > 0$. The joint distribution of (X_1, X_2) is $\mathbb{P}_\theta$, the uniform distribution on the square $(0, \theta)^2$. For simplicity, take the X_i's as the coordinate maps for the generic point $x = (x_1, x_2)$ of the underlying set $\mathcal{X} = \mathbb{R}_+^2$.

The random variable $Y(x) := \max(x_1, x_2)$ is a sufficient statistic. It has distribution $\mathbb{Q}_\theta$ with density $2t\{0 < t < \theta\}/\theta$ with respect to Lebesgue measure on $\mathbb{R}^+$. The conditional distribution of (x_1, x_2) given $Y = y$ is uniform on the boundary set

$$\{y\} \times (0, y) \quad \cup \quad (0, y) \times \{y\}.$$

We could construct the probabilistic copy with the help of a unif$(0, 1)$ distributed U and a Ber$(1/2)$ distributed W, with $\{Y, V, W\}$ independent, by

$$(\widetilde{X}_1, \widetilde{X}_2) = f(Y, V, W) = W(YV, Y) + (1 - W)(Y, YV).$$

The distribution of $f(y, V, W)$ defines the Markov kernel $\{K_y : y > 0\}$ for which $\mathbb{P}_\theta = K\mathbb{Q}_\theta$. Once again, the $\mathcal{P}$ and $\mathcal{Q}$ models are statistically equivalent.

□

As a small exercise, you should show that the $f(Y, V, W)$ in the previous Example can be replaced by a random variable $g(Y, U)$, with $U \sim \text{unif}(0, 1)$. That is, show how to generate (V, W) from a single uniformly distributed random variable.

In general, under mild topological assumptions on the spaces $(\mathcal{X}, \mathcal{A})$ and $(\mathcal{Y}, \mathcal{B})$, every Markov kernel from $\mathcal{Y}$ to $\mathcal{X}$ can be represented as the distribution of some (measurable) function $g(y, U)$ with $U \sim \text{unif}(0, 1)$.

1.2 The Le Cam distance between statistical models

Introduction::distance

Suppose $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ on $(\mathcal{X}, \mathcal{A})$ and $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$, on $(\mathcal{Y}, \mathcal{B})$, are two statistical models with the same index set Θ. Le Cam's distance $\Delta(\mathcal{P}, \mathcal{Q})$ between the two models is defined so that a small Δ implies that solutions to decision theoretic problems for $\mathcal{P}$ have similar solutions—in the sense of risk functions—to corresponding problems for $\mathcal{Q}$, and vice versa.

Let me begin with a much stronger notion of closeness of models $\mathcal{P}$ and $\mathcal{Q}$ defined on the same space $(\mathcal{X}, \mathcal{A})$. Suppose there is some (small) ϵ' for which

TV.close<7>
$$\|\mathbb{P}_\theta - \mathbb{Q}_\theta\|_{\text{TV}} \le \epsilon \qquad \text{for all } \theta \in \Theta.$$

Remark. If $\mathbb{Q}$ and $\widetilde{\mathbb{Q}}$ are two probability measures defined on the same sigma-field $\mathcal{B}$, their total variation distance is defined as

$$\left\|\mathbb{Q} - \widetilde{\mathbb{Q}}\right\|_{\text{TV}} = \sup_{B \in \mathcal{B}} |\mathbb{Q}B - \widetilde{\mathbb{Q}}B| = \frac{1}{2} \sup_{|g| \le 1} |\mathbb{Q}g - \widetilde{\mathbb{Q}}g|,$$

the second supremum running over all $\mathcal{B}$measurable functions g that are bounded in absolute value by 1. If both measures are dominated by another measure μ, with densities q and $\widetilde{q}$ then

$$2\left\|\mathbb{Q} - \widetilde{\mathbb{Q}}\right\|_{\text{TV}} = \left\|\mathbb{Q} - \widetilde{\mathbb{Q}}\right\|_1 := \int |q - \widetilde{q}|\, d\mu = \mu|q - \widetilde{q}|.$$

Suppose also that Θ is a metric space (with metric ρ) and we have some estimator $\widehat{\theta}$ for which

$$\mathbb{Q}_\theta\{\rho(\widehat{\theta}, \theta) > R_\epsilon\} \le \epsilon \qquad \text{for all } \theta \in \Theta.$$

Then we have

$$\mathbb{P}_\theta\{\rho(\widehat{\theta}, \theta) > R_\epsilon\} \le \epsilon + \epsilon' \qquad \text{for all } \theta \in \Theta.$$

If ϵ and ϵ' are both small then we have transformed a good estimator for the $\mathcal{Q}$ model into a good estimator for the $\mathcal{P}$ model. Almost the same argument

works to transform good estimators for $\mathcal{P}$ into good estimators for $\mathcal{Q}$. The bound <7> has created a strong coupling of inference problems for the two models.

bi.iid <8> **Example.** Consider the model $\mathcal{P}_n = \{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$ from Example 2, where Θ is some set of densities on $[0,1]$. That is, $\mathbb{P}_{\theta,n} = P_\theta^n$ where P_θ has density θ with respect to Lebesgue measure on $[0,1]$.

For a fixed positive integer m let $J_i := \Big(i-1)/m, i/m\Big]$ and define an approximation map A_m by

$$A_m\theta = \sum\nolimits_{i=1}^m \overline{\theta}_i\{t \in J_i\} \qquad \text{where } \overline{\theta}_i = m\int \{s \in J_i\}\theta(s)\,ds$$

Define $Q_\theta = P_{A_m\theta}$ and $\mathbb{Q}_{\theta,n} = Q_\theta^n$. Under suitable smoothness conditions on Θ, the $\mathcal{P}_n$ model will be close in the total variation sense to the model $\mathcal{Q}_n = \{\mathbb{Q}_{\theta,n} : \theta \in \Theta\}$ model.

It is usually not easy to calculate total variation distances between product measures. Instead, it is typical to work with an upper bound involving Hellinger distance, which is much friendlier to product measures:

$$\frac{1}{4}\left\|P^n - Q^n\right\|_{\mathrm{TV}}^2 \le H^2(P^n, Q_n) \le nH^2(P,Q)$$

See Pollard (2001, Section3.3, Problem 4.18) for details. If we assume that each density in Θ is bounded from below by some constant $c_0 > 0$, which is a very common simplifying assumption in the literature, then

$$\begin{aligned} H^2(P_\theta, Q_\theta) &= \int_0^1 \left(\sqrt{\theta(s)} - \sqrt{A_m\theta(s)}\right)^2 ds \\ &\le \int_0^1 \frac{\left(\theta(s) - A_m\theta(s)\right)^2}{\left(\sqrt{\theta(s)} + \sqrt{A_m\theta(s)}\right)^2}\, ds \\ &\le \frac{1}{4c_0}\sum\nolimits_{i=1}^m \int \{s \in J_i\}\left(\theta(s) - \overline{\theta}_i\right)^2 ds \end{aligned}$$

As m gets larger the approximation will usually get better, particularly so if Θ imposes a smoothness constraint. One particularly elegant way to control the bound is to take $m = 2^k$ then make some assumption about the decay of the coefficients for θ when expanded in the Haar basis. See Brown, Carter, Low, and Zhang (2004) for example.

The $\mathcal{Q}_n$ model is conceptually easier to work with because the m-dimensional vector of counts in each J_i interval is a sufficient statistic, which has a multinomial distribution.

The anaologous discretization for the Poisson process model of Example 3 is even cleaner, for then the m-vector of cell counts, which is again a sufficient statistic, contains independent Poisson($n\overline{\theta}_i/m$) variables. See Problem [3].

□

The Le Cam distance takes this total variation idea one step further. We no longer need the models defined on the same space; the coupling of the models is created by randomization. For the moment, I will think of the randomization as being defined by a Markov kernel. As you will see in Chapter 2, Le Cam allowed more general objects to be called randomizations, so that desirable theorems about randomization should become true.

Suppose K is a Markov kernel from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$. Randomization via K serves to turn an observation x from some unknown $\mathbb{P}_\theta$ into an observations y from a corresponding probability measure on $\mathcal{Y}$: first $x \sim \mathbb{P}_\theta$ then $y \mid x \sim K_x$. These two random operations create a pair of random variables (x, y) with a joint distribution Γ_θ,

$$\Gamma_\theta(A \times B) = \int_A (K_x B)\, \mathbb{P}_\theta(dx) = \mathbb{P}_\theta^x K_x^y \{x \in A, y \in B\}.$$

More generally, at least for every nonnegative, $\mathcal{A} \otimes \mathcal{B}$-measurable function f on $\mathcal{X} \times \mathcal{Y}$,

$$\Gamma_\theta^{x,y} f(x, y) = \mathbb{P}_\theta^x K_x^y f(x, y)$$

The probability measure Γ_θ has marginals $\mathbb{P}_\theta$ and $\widetilde{\mathbb{Q}}_\theta$, where

$$\widetilde{\mathbb{Q}}_\theta g = \mathbb{P}_\theta^x K_x^y g(y)$$

at least for every nonnegative, $\mathcal{B}$-measurable function g on $\mathcal{Y}$.

Under the joint distribution Γ_θ, the kernel K becomes the conditional distribution of y given x. For the model $\{\Gamma_\theta : \theta \in \Theta\}$ on $(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\otimes\mathcal{B})$, the x coordinate is a sufficient statistic. To indicate the dependence of both $\widetilde{\mathbb{Q}}_\theta$ and Γ_θ in a linear way on both $\mathbb{P}_\theta$ and the Markov kernel K, I like to write $\widetilde{\mathbb{Q}}_\theta = K\mathbb{P}_\theta$ and $\Gamma_\theta = \mathbb{P}_\theta \otimes K$. (See Pollard (2001, Section 4.3) for a more detailed discussion.)

Le Cam's deficiency distance measures how well $\mathcal{Q}$ can be approximated in a total variation sense by $\{\widetilde{\mathbb{Q}}_\theta : \theta \in \Theta\}$ obtained by a randomization of $\mathcal{P}$.

deficieny.def <9> **Definition.** *The **deficiency** of the $\mathcal{P}$ model with respect to the $\mathcal{Q}$ model is defined as*

deficiency<10>
$$d_{\text{LeCam}}(\mathcal{P}, \mathcal{Q}) = \inf_K \sup_{\theta\in\Theta} \|\mathbb{Q}_\theta - K\mathbb{P}_\theta\|_{\text{TV}}\,,$$

the infimum runnning over all randomizations K. Note well: *the K is not allowed to depend on θ.*

*The **Le Cam distance** between $\mathcal{P}$ and $\mathcal{Q}$ is defined as*

$$\Delta_{\text{LeCam}}(\mathcal{P},\mathcal{Q}) = \max\left(d_{\text{LeCam}}(\mathcal{P},\mathcal{Q}), d_{\text{LeCam}}(\mathcal{Q},\mathcal{P})\right)$$

If $\Delta_{\text{LeCam}}(\mathcal{P},\mathcal{Q}) = 0$ then models are said to be equivalent in the Blackwell sense.

Remark. For the moment you should think of the randomizations as being defined by Markov kernels. In Chapter 2 the class of K's will be made larger, which, amongst other benefits, will ensure that the infimum over K in the definition of $\delta(\mathcal{P},\mathcal{Q})$ is achieved some some randomization.

Often we have not just one pair of models but a whole family of pairs of models, $\mathcal{P}_n = \{\mathbb{P}_{\theta,n} : \theta \in \Theta_n\}$ and $\mathcal{Q}_n = \{\mathbb{Q}_{\theta,n} : \theta \in \Theta_n\}$, indexed by another parameter (such as sample size n). If $\Delta_{\text{LeCam}}(\mathcal{P}_n,\mathcal{Q}_n) \to 0$ as $n \to \infty$ then the models $\mathcal{P}_n$ and $\mathcal{Q}_n$ are said to be ***asymptotically equivalent*** in the Le Cam sense.

The choice of the total variation distance in the definition of $d_{\text{LeCam}}(\mathcal{P},\mathcal{Q})$ once again leads to a strong coupling of inference problems. Suppose, for example, that Θ is a metric space and we have some estimator $\widehat{\theta}$ for which

$$\mathbb{Q}_\theta\{d(\widehat{\theta},\theta) > R_\epsilon\} \le \epsilon \qquad \text{for all } \theta \in \Theta.$$

Suppose also that there is a Markov kernel K such that $\left\|\mathbb{Q}_\theta - \widetilde{\mathbb{Q}}_\theta\right\|_{\text{TV}} \le \epsilon'$ for all θ, where $\widetilde{\mathbb{Q}}_\theta = K\mathbb{P}_\theta$. The Markov kernel creates a ***randomized estimator***, that is, a Markov kernel $\widetilde{\theta} = \{\widetilde{\theta}_x : x \in \mathcal{X}\}$ from $\mathcal{X}$ to Θ, defined by $\widetilde{\theta}_x =$ the distribution of $\widehat{\theta}(y)$ where $y \mid x \sim K_x$. Then

$$\begin{aligned}\mathbb{P}_\theta^x\widetilde{\theta}_x^t\{d(t,\theta) > R_\epsilon\} &= \mathbb{P}_\theta^x K_x^y\{d(\widehat{\theta}(y),\theta) > R_\epsilon\} \\ &= \widetilde{\mathbb{Q}}_\theta\{d(\widehat{\theta}(y),\theta) > R_\epsilon\} \le \epsilon + \epsilon' \qquad \text{for all } \theta \in \Theta.\end{aligned}$$

The final inequality uses the fact that

$$\widetilde{\mathbb{Q}}_\theta A - \mathbb{Q}_\theta A \le \left\|\mathbb{Q}_\theta - \widetilde{\mathbb{Q}}_\theta\right\|_{\text{TV}} \le \epsilon'$$

with $A = \{y \in \mathcal{Y} : d(\widehat{\theta}(y),\theta) > R_\epsilon\}$.

Remark. If you are having trouble deciphering these calculations it might help you to rewrite them using $L(\theta,t) = \{d(t,\theta) > R_\epsilon\}$, a loss function taking values in $\{0,1\}$.

The Chapter from here on is just a series of sketched examples to illustrate some cases where useful bounds can be calculated. Some of these bounds will be needed to establish equivalance with the white noise model. I am not sure whether they belong in this Chapter. After the lectures I will probably have a better idea idea of what should go where.

poisson.bayes <11>

Example. Suppose $\mathbb{P}_{\theta,n} := \text{Poisson}(n + \sqrt{n}\theta)$. (I chose this parametrization so that it is easy to distinguish between two fixed θ values that are widely separated but hard if they are very close together.) The set $\mathcal{X}$ equals the set $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. Expressed in the fanciest way possible, the probability measure $\mathbb{P}_{\theta,n}$ has density

$$p_{\theta,n}(x) := \mathbb{P}_{\theta,n}\{x\} = \exp(-n - \sqrt{n}\theta)\frac{(n + \sqrt{n}\theta)^x}{x!} \qquad \text{for } x \in \mathbb{N}_0$$

with respect to counting measure on $\mathbb{N}_0$.

Consider the Bayes test between two hypotheses, $\theta = \theta_0 = 0$ versus $\theta = \theta_1$, with prior probability $\pi\{\theta_0\} = \pi\{\theta_1\} = 1/2$. Use a zero-one loss function,

$$L(\theta, t) = \{\theta \neq t\} \qquad \text{for } t,\ \theta \in \Theta_0 := \{\theta_0, \theta_1\}.$$

That is, the task is to guess which model ($\theta = \theta_0$ or $\theta = \theta_1$) generates the observation x, with a loss of one unit for an incorrect guess and a zero loss for a correct guess and $\mathbb{D} = \Theta_0$.

As a heuristic, note that $W_n := (x - n)/\sqrt{n}$ is approximately distributed as $\mathbb{Q}_\theta := N(\theta, 1)$ under the $\mathbb{P}_{\theta,n}$ model. It might seem that inferences about θ for the $\mathcal{P}_n$ model should be approximately like inferences about θ under the $\mathcal{Q} = \{\mathbb{Q}_t h : \theta \in \mathbb{R}\}$ model. In various senses, if we restrict the range of θ, it is true that $\mathcal{P}_n$ and $\mathcal{Q}$ are close in the Le Cam sense but it is not for the heuristic reason just mentioned. (See the next Example to understand why closeness of data distributions is not enough.) The Le Cam distance involves much more than just closeness in distribution for the observed data, for each θ; it depends on the way different members of the model are related to other, as the present example will show. In fact, it is the behavior of the likelihood ratio,

<12>

$$\frac{p_{\theta,n}(x)}{p_{0,n}(x)} = \exp\left(-\theta\sqrt{n} + x\log(1 + \theta/\sqrt{n})\right) = \psi_n(W_n, \theta)$$

where $\Psi_n(z,\theta) = \exp(\psi_n(z,\theta))$ and

$$\begin{aligned}\psi_n(z,\theta) &= n\log(1+\theta/\sqrt{n}) - \theta\sqrt{n} + z\sqrt{n}\log(1+\theta/\sqrt{n}) \\ &= -\frac{1}{2}\theta^2 + O(|\theta|^3/\sqrt{n}) + z\left(\theta + O(\theta^2/\sqrt{n})\right)\end{aligned} \tag{<13>}$$

In what follows I will drop n from indexing, but you should remember that the statistical problems only get close as n goes off to infinity.

A randomized test is just a Markov kernel $\delta = \{\delta_x : x \in \mathbb{N}_0\}$ from $\mathbb{N}_0$ to Θ_0. The Bayes risk of δ is

$$\begin{aligned}R(\pi,\delta) &= \frac{1}{2}\mathbb{P}_0^x\delta_x\{\theta_1\} + \frac{1}{2}\mathbb{P}_{\theta_1}^x\delta_x\{\theta_0\} \\ &= \frac{1}{2}\sum_{x\in\mathbb{N}_0} p_0(x)\delta_x\{\theta_1\} + p_{\theta_1}(x)\delta_x\{\theta_0\}\end{aligned}$$

The Bayes rule minimizes the Bayes risk by choosing

$$\delta_x\{0\} = \begin{cases} 1 & \text{if } p_0(x) \le p_{\theta_1}(x) \\ 0 & \text{otherwise}\end{cases}$$

That is, the minimum Bayes risk is

$$\begin{aligned}\frac{1}{2}\sum_{x\in\mathbb{N}_0}\min\left(p_0(x), p_{\theta_1}(x)\right) &= \frac{1}{2}\sum_{x\in\mathbb{N}_0} p_0(x)\min\left(1, p_{\theta_1}(x)/p_0(x)\right) \\ &= \frac{1}{2}\mathbb{P}_0\min\left(1,\Psi_n(W_n,\theta_1)\right) \qquad \text{by <12>.}\end{aligned}$$

Under $\mathbb{P}_0$, the random variables converge in distribution to $\mathbb{Q}_0$. Also $\psi_n(z,\theta_1)$ converges uniformly on compact sets of z to the function $\psi(z,\theta) := z\theta - \frac{1}{2}\theta^2$. These two facts together imply (Why?) that the minimum Bayes risk converges to

$$\frac{1}{2}\mathbb{Q}^y\min\left(1,\Psi(y,\theta_1)\right) \qquad \text{where } \Psi(z,\theta) := \exp\left(\psi(z,\theta)\right)$$

The last expression is, in fact, the minimum Bayes risk for testing $\mathbb{Q}_0$ against $\mathbb{Q}_{\theta_1}$, for the same prior and loss function. You can establish this fact by proving that the Bayes estimator for this problem is

$$\delta_y\{0\} = \begin{cases} 1 & \text{if } \phi(y) \le \phi(y-\theta_1) \\ 0 & \text{otherwise}\end{cases}$$

and then noting that

$$\frac{d\mathbb{Q}_\theta}{d\mathbb{Q}_0}(y) = \frac{\phi(y-\theta)}{\phi(y)} = \exp(\theta y - \theta^2/2) = \Psi(y,\theta).$$

It is the convergence in distribution of the likelihood ratios that gave the asymptotic equality for the Bayes risks.

□

bad.poisson <14>

Example. Choose $\mathbb{P}_{0,n} = \text{Poisson}(n)$ as in the last Example, but take $\mathbb{P}_{\theta_1,n}$ as the distribution of $X + e^{-n}$ where $X \sim \text{Poisson}\,(n + \sqrt{n}\theta_1)$. Once again note that $W_n := (x - n)/\sqrt{n}$ is approximately distributed as $\mathbb{Q}_\theta := N(\theta, 1)$ under the $\mathbb{P}_{\theta,n}$ model. But this time there is a perfect test (What?) between $\theta = 0$ and $\theta = \theta_1$, with zero Bayes risk. The minimum Bayes risk for the $\mathcal{Q}$ model is the same as before.

The great difference occurs because $\mathbb{P}_{0,n}$ and $\mathbb{P}_{\theta_1,n}$ are mutually singular, which has a great effect on the likelihood ratio but no effect on the limiting distribution of W_n under each alternative.

□

local.Poisson.normal <15>

Example. It might not be obvious to you that the result in Example 11 is actually due to asymptotic equivalence in Le Cam's sense. Let me construct the "randomizations" that proves that $d_{\text{LeCam}}(\mathcal{Q}, \mathcal{P}_n) \to 0$ for the models $\mathcal{P}_n = \{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$ and $\mathcal{Q} = \{\mathbb{Q}_{\theta,n} : \theta \in \Theta\}$ for any fixed, bounded subset Θ of the real line.

I put quotes around the word randomizations in the last paragraph because they are of a degenerate kind: deterministic functions. It would be more exciting to prove that $d_{\text{LeCam}}(\mathcal{P}_n, \mathcal{Q}) \to 0$, for then we would need (nondegenerate) Markov kernels. I leave that exercise to you.

Once again I omit various n subscripts when I do not need to emphasize the dependence on n.

The deterministic functions are given by the ***quantile transformation*** (Pollard 2001, Example 2.35). Let Φ be the $N(0, 1)$ distribution function and F_n be the distribution function of W_n under $\mathbb{P}_{0,n}$. Define $\gamma_n(y) := F_n^{-1}(\Phi(y))$. If $Y \sim N(0, 1)$ then $\gamma_n(Y)$ is a random variable with distribution function F_n. Equivalently, under $\mathbb{Q}_0$ the random variable $n + \sqrt{n}\gamma_n(y)$ has distribution $\mathbb{P}_{0,n}$.

The distribution function F_n converges uniformly to Φ because W_n converges in distribution (under $\mathbb{P}_{0,n}$) to $N(0, 1)$. It follows that

gamn.cgce<16>

$$\sup_{|y|\le C} |\gamma_n(y) - y| \to 0 \qquad \text{for each finite } C.$$

Define $\widetilde{\mathbb{P}}_\theta$ to be the distribution of $n+\sqrt{n}\gamma_n(y)$ under the $\mathbb{Q}_\theta$. Note that $\gamma_n(y)$ under $\mathbb{Q}_0$ has the same distribution as W_n under $\mathbb{P}_0$. Calculate $\left\|\widetilde{\mathbb{P}}_\theta - \mathbb{P}_\theta\right\|_{\mathrm{TV}}$ by taking a supremum of $|\widetilde{\mathbb{P}}_\theta g - \mathbb{P}_\theta g|$ over all functions g with $|g| \le 1$. For a fixed such g temporarily write $G_n(y)$ for $g(n+\sqrt{n}\gamma_n(y))$. Then

$$\begin{aligned}
|\widetilde{\mathbb{P}}_\theta g - \mathbb{P}_\theta g| &= |\mathbb{Q}_\theta^y G_n(y) - \mathbb{P}_\theta^x g(x)| \\
&= |\mathbb{Q}_0^y G_n(y)\Psi(y,\theta) - \mathbb{P}_0^x g(x)\Psi_n(W_n,\theta)| \\
&= |\mathbb{Q}_0^y G_n(y)\Psi(y,\theta) - \mathbb{Q}_0^y G_n(y)\Psi_n(\gamma_n(y),\theta)| \\
&\le \mathbb{Q}_0^y |\Psi(y,\theta) - \Psi_n(\gamma_n(y),\theta)| \qquad \text{because } |G_n(y)| \le 1.
\end{aligned}$$

By construction, the function $\Psi(y,\theta) = d\mathbb{Q}_\theta/d\mathbb{Q}_0$ is nonnegative and it integrates to one. Similarly $\Psi_n(\gamma_n(y),\theta)$ is nonnegative and

$$\mathbb{Q}_0^y \Psi_n(\gamma_n(y),\theta) = \mathbb{P}_0^x \Psi_n(W_n,\theta) = \mathbb{P}_0 \frac{d\mathbb{P}_\theta}{d\mathbb{P}_0} = 1.$$

By Scheffé's lemma (Pollard 2001, Exercise 3.6),

$$\mathbb{Q}_0^y |\Psi(y,\theta) - \Psi_n(\gamma_n(y),\theta)| = 2\mathbb{Q}_0^y \left(\Psi(y,\theta) - \Psi_n(\gamma_n(y),\theta)\right)^+,$$

which converges uniformly in θ to zero because of <13> and <16>. [Is this sketch too terse? Would more details be helpful?]

□

1.3 Facts about the Poisson distribution

Introduction::poisson,facts

The convex function

<17>
$$\begin{aligned}
h(t) &= (1+t)\log(1+t) - t \qquad \text{for } -1 \le t \\
&= \frac{t^2}{2} - \frac{t^3}{6} + O(t^4) \qquad \text{as } t \to 0
\end{aligned}$$

achieves its minimum value of zero at $t = 0$.

Poisson.dist<18> **Lemma.** *Suppose X has a Poisson(λ) distribution, with $\lambda \ge 1$.*

(i) If $\ell = \lambda + x \in \mathbb{N}$ then

$$\log\left(\sqrt{2\pi\lambda}\mathbb{P}\{W = \ell\}\right) = -\lambda h(x/\lambda) - \tfrac{1}{2}\log(1 + x/\lambda) + O(1/\ell)$$

(ii) $\mathbb{P}\{W = \ell\} \le \exp(-\lambda h(x/\lambda))$ *for all* $\ell = \lambda + x \in \mathbb{N}_0$.

(iii) For all $x \ge 0$,

$$\mathbb{P}\{|W - \lambda| \ge x\} \le 2\exp(-\lambda h(x/\lambda))$$

PROOF By Stirling's formula,

$$\log(\ell!/\sqrt{2\pi}) = (\ell + \tfrac{1}{2})\log(\ell) - \ell + r_\ell \qquad \text{where } \frac{1}{12\ell+1} \le r_\ell \le \frac{1}{12\ell}.$$

Thus

$$\begin{aligned}\log\bigl(\sqrt{2\pi\lambda}\mathbb{P}\{W=\ell\}\bigr) &= -\lambda + \ell\log(\lambda) - \log(\ell!/\sqrt{2\pi}) + \tfrac{1}{2}\log(\lambda)\\ &= -\lambda h(u) - \tfrac{1}{2}\log(1+u) + O(\ell^{-1}),\end{aligned}$$

which gives (i).

For (ii), first note that $\mathbb{P}\{W=0\} = e^{-\lambda} = \exp(-\lambda h(-1))$. For $\ell \ge 1$ we have

$$\begin{aligned}\log\Bigl(\sqrt{2\pi}\mathbb{P}\{W=\ell\}\Bigr) &= -\lambda + \ell\log(\lambda) - (\ell+\tfrac{1}{2})\log(\ell) + \ell - r_\ell\\ &\le -\lambda + \ell\log(\lambda) - \ell\log(\ell) + \ell = \lambda h(u).\end{aligned}$$

Inequality (iii) comes from two appeals to the usual trick with the moment generating function $\mathbb{P}e^{tW} = \exp(\lambda(e^t - 1))$. For $x \ge 0$,

$$\mathbb{P}\{W \ge \lambda + x\} \le \inf_{t\ge 0} \mathbb{P}e^{t(W-\lambda-x)} = \inf_{t\ge 0} \exp\bigl(-t(\lambda+x) + \lambda(e^t-1)\bigr)$$

The infimum is achieved at $t = \log(1+x/\lambda)$, giving the bound $\exp(-\lambda h(x/\lambda))$. Similarly

$$\mathbb{P}\{W \le \lambda - x\} \le \inf_{t\ge 0} \mathbb{P}e^{t(\lambda-x-W)} = \inf_{t\ge 0} \exp\bigl(t(\lambda-x) + \lambda(e^{-t}-1)\bigr)$$

with the infimum achieved at $t = -\log(1-x/\lambda)$ if $0 \le x < \lambda$ or as $t \to \infty$ if $x = \lambda$. The inequality is trivial for $x > \lambda$.

1.4 Normal versus Poisson

ntroduction::normal.Poisson

Compare with Brown, Carter, Low, and Zhang (2004).

Suppose $X \sim \text{Poisson}(\lambda)$ is independent of $U \sim \text{Unif}(-1/2, 1/2)$, and suppose $Y \sim N(\lambda, \lambda)$. Show that

$$H^2(X+U, Y) = O(1/\lambda) \qquad \text{as } \lambda \to \infty.$$

Equivalently, show that $H^2(X+U-\lambda, Y-\lambda) = O(1/\lambda)$.

By Problems [2] and [4], there is no loss of generality in assuming that λ is a positive integer.

The distribution of $X+\lambda-U$ has density

$$p(y) = \sum\nolimits_{k\ge -\lambda} \{|y-k|<1/2\}\mathbb{P}\{X=\lambda+k\}$$

From Lemma 18,

$$\log\left(\sqrt{2\pi\lambda}\mathbb{P}\{X=\lambda+k\}\right) = -\lambda h(k/\lambda) - \tfrac{1}{2}\log(1+k/\lambda) + r_{\lambda+k}$$

where $r_n = O(1/n)$ and

$$h(t) = (1+t)\log(1+t) - t = \sum_{k\ge 2}\frac{(-1)^k t^k}{(k-1)k} \qquad \text{for } |t|<1.$$

Write Q for the $N(0,\lambda)$ with density q. Let K be the smallest integer for which $K+\frac{1}{2} \ge \sqrt{2\lambda\log\lambda}$ and define $A = (K-1/2, K+1/2)$, so that $QA^c \le 2/\lambda$. Then

$$\begin{aligned}
&H^2(X+U-\lambda, Y-\lambda)\\
&\quad\le 2QA^c + \int_A q\log(q/p) \qquad \text{by Problem [5](i)}\\
&\quad= O(\lambda^{-1}) + \sum_{k=-K}^{K}\int_{J_k} q\log(q/p) \qquad \text{where } J_k = (k-1/2, k+1/2)
\end{aligned}$$

By symmetry of q, the last sum equals

$$\begin{aligned}
&\int_{J_0} q(y)\left(-\frac{y^2}{2\lambda} - \log\left(\sqrt{2\pi\lambda}\mathbb{P}\{X=\lambda\}\right)\right)dy\\
&\quad+\sum_{k=1}^{K}\int_{J_k} q(y)\left(-\frac{2y^2}{2\lambda} - \log\left[2\pi\lambda\,\mathbb{P}\{X=\lambda+k\}\mathbb{P}\{X=\lambda-k\}\right]\right)dy
\end{aligned}$$

By Lemma 18(i), the J_0 contribution is at most $O(1/\lambda)$ and the J_k contribution is at most

$$\int_{J_k} q(y)\left(-\frac{y^2}{\lambda} + \lambda\left[h(k/\lambda)+h(-k/\lambda)\right] + \frac{1}{2}\log(1-k^2/\lambda^2) - r_{\lambda+k} - r_{\lambda-k}\right)dy$$

How to finish the calculation:

(i) Show that $\frac{1}{2}\log(1-k^2/\lambda^2) - r_{\lambda+k} - r_{\lambda-k}$ contributes only a $O(1/\lambda)$ term.

(ii) Use the series <17> to kill many terms in the expansions of $h(\pm k/\lambda)$, leaving

$$\lambda\,[h(k/\lambda) + h(-k/\lambda)] = k^2/\lambda + O(k^4/\lambda^3)$$

(iii) Bound $\int_{J_k} q(y)k^4/\lambda^3$ by a constant multiple of $\int_{J_k} q(y)y^4/\lambda^3$. Sum over k, bounding by $C\int q(y)y^4/\lambda^3 = O(1/\lambda)$.

(iv) The tricky part is the contribution from

$$(k^2-y^2)/\lambda \le -2k(y-k)/\lambda + \frac{1}{4\lambda}$$

Use

$$\Big|\int_{J_k} q(y)(-2k)(y-k)\Big| \le 2kQJ_k|Q(y-k \mid y\in J_k)|$$

Use Problem [7] to bound the contribution from the conditional expectation by $2k/(4\lambda)$. Sum over k, bound k by $2y$, then bound the whole sum by a multiple of $Qy^2/\lambda^2 = O(1/\lambda)$.

Alternatively: Replace the appeal to Problem [5](i) by an appeal to Problem [5](ii) (Harry's method). That way we don't need to use the symmetry to get cancellations. This method is probably better.

1.5 Variance stabilization

Introduction::rootnormal

Let $P = N(\lambda,\lambda)$ and $Q = N(2\sqrt{\lambda},1)$. Let $\psi:\mathbb{R}\to\mathbb{R}$ be continuous and strictly increasing, with $\psi(x) = 2\sqrt{x}$ for $x \ge 1$. Show that

$$H^2(\psi(P),Q) = O(1/\lambda) \qquad \text{as } \lambda\to\infty.$$

Define $\ell = 2\sqrt{\lambda}$.

Note that $\psi(P)$ has density $\gamma(y)$ with respect to Lebesgue measure, with

$$\psi(y) = \frac{y}{\ell}\phi\left(\frac{y^2/4-\lambda}{\ell/2}\right) \qquad \text{for } y\ge 2.$$

Also Q has density

$$q(y) = \phi\,(y-\ell)$$

Define $A = \{y \in \mathbb{R} : |y - \ell| \le r\}$ with $r = \sqrt{2 \log \lambda}$. Note that $QA^c \le 2/\lambda$. Use Problem [5].

$$\int \{y \in A\} q(y) \log\left(\frac{q(y)}{\gamma(y)}\right) dy = \int_{-r}^{r} \phi(x) \log\left(\frac{\phi(x)}{\gamma(x+\ell)}\right) dx$$

Note that

$$(x+\ell)^2/4 - \lambda = \tfrac{1}{4}x^2 + \tfrac{1}{2}x\ell$$

The log term equals

$$\begin{aligned} &-\tfrac{1}{2}x^2 - \log(1 + x/\ell) + \frac{1}{2\lambda}\left((x+\ell)^2/4 - \lambda\right)^2 \\ &\quad = -\frac{x^2}{2} - \frac{x}{\ell} + O\left(\frac{x^2}{\ell^2}\right) + \frac{1}{32\lambda}(x^4 + 4x^3\ell + 4x^2\ell^2) \\ &\quad = O\left(\frac{x^2 + x^4}{\lambda}\right) + \text{terms in } x \text{ and } x^3 \end{aligned}$$

By symmetry, the terms in x and x^3 integrate out to zero, leaving a quantity of order $1/\lambda$.

1.6 Many Poissons

Introduction::many

Combine the results from the last two sections to get randomizations for $\mathbb{P}_\lambda = \otimes_{i \le m} \text{Poisson}(\lambda_i)$ with $\lambda = (\lambda_1, \ldots, \lambda_m) \in \Lambda$ and $\mathbb{Q}_\lambda = \otimes_{i \le m} N(2\sqrt{\lambda_i}, 1)$. Work with independent random variables $Y_i \sim N(2\sqrt{\lambda_i}, 1)$ and $X_i \sim \text{Poisson}(\lambda_i)$.

Define

$$\widetilde{Y}_i = \psi(X_i + U_i) \qquad \text{for independent } U_i \sim \text{Unif}(-1/2, 1/2)$$

and

$$\widetilde{X}_i = \text{closest integer to } \psi^{-1}(Y_i)$$

Work directly with Hellinger distances or use affinities?

$$\prod_{i \le m} \alpha_2(\widetilde{X}_i, X_i) = O(\ldots)$$

or

$$\sum_{i \le m} H^2(\widetilde{X}_i, X_i) = O\left(\sum\nolimits_i 1/\lambda_i\right)$$

Use idea from Brown, Carter, Low, and Zhang (2004) to avoid calculations for the Y_is and $\widetilde{Y}_i$'s.

Should I also comment on conditional quantile transformations with Binomials at this point?

1.7 Problems

Introduction::problems

- If $P, Q \ll \mu$ with densities p and q, Hellinger affinity is $\alpha_2(P,Q) = \mu\sqrt{pq}$. Note that $H^2(P,Q) = 2 - 2\alpha_2(P,Q)$.
- Include facts about total variation versus Hellinger.
- Comment on convenient notation $H(X,Y)$ for $H(P,Q)$ if $X \sim P$ and $Y \sim Q$.

smooth [1] Suppose X, Y, U are independent random variables with $U \sim \text{Unif}(-1/2, 1/2)$ and both X and Y taking only integer values. Show that $H^2(X,Y) = H^2(X+U, Y+U)$.

hell.poisson [2] Show that $\alpha_2(\text{Poisson}(\lambda), \text{Poisson}(\mu)) = \exp(-\frac{1}{2}(\sqrt{\lambda} - \sqrt{\mu})^2)$. Deduce that $H^2(\text{Poisson}(\lambda), \text{Poisson}(\mu)) = O(\sqrt{\lambda} - \sqrt{\mu})^2$.

Poisson.proc [3] Let $\mathbb{P}_\mu$ denote the distribution of a Poisson process with intensity measure μ on $[0,1]$. Show that

$$H^2(\mathbb{P}_\lambda, \mathbb{P}_\mu) = 2 - 2\exp\left(-\frac{1}{2}H^2(\lambda,\mu)\right)$$

[Is this correct? Try to reduce to an affinity calculation for independent Poissons from a nested family of partitions of $[0,1]$.]

hell.normal [4] Hellinger distance between normals.

(i) Show that

$$H^2\left(N(\mu_1,\sigma^2), N(\mu_1,\sigma^2)\right) = 2 - 2\exp\left(-(\mu_1-\mu_2)^2/8\sigma^2\right).$$

(ii) Show that

$$H^2\left(N(\mu,\sigma_1^2), N(\mu,\sigma_2^2)\right) = 2 - 2\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}.$$

(iii) Deduce that

$$H^2\left(N(\mu_1,\sigma_1^2), N(\mu_2,\sigma_2^2)\right) \le \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{4|\sigma_1^2-\sigma_2^2|^2}{\sigma_2^4}$$

hell.J.KL

[5] Let P and Q be probability measures on $(\mathcal{X}, \mathcal{A})$ with densities p and q with respect to some dominating measure μ. Suppose P and Q are mutually absolutely continuous on some set $A \in \mathcal{A}$, with $p = qe^{2\eta}$.

(i) Show that $2q - 2\sqrt{pq} = 2q(1 - e^{\eta}) \le -2q\eta$. Deduce that

$$H^2(P, Q) \le 2QA^c + \int_A q \log(q/p)$$

(ii) On the subset of A where $\eta \ge 0$, show that

$$(\sqrt{q} - \sqrt{p})^2 = q(e^{\eta} - 1)^2 \le q\eta^2 e^{2\eta} = p\eta^2.$$

When $\eta < 0$ interchange the roles of p and q to get an analogous bound $q(-\eta)^2$. Deduce that

$$H^2(P, Q) \le PA^c + QA^c + \int_A (p + q) \left(\log(p/q)\right)^2$$

conditional

[6] Suppose X is a real valued random variable such that $\mathbb{P}\{X \in I\} > 0$ for every nondegenerate interval I. Show that $g(a, b) := \mathbb{P}(X \mid a \le X < b)$ is an increasing function of both a and b. Hint: For $t \in (a, b)$, define $\alpha(t) = \mathbb{P}\{a \le X < t \mid a \le X < b\} - 1 - \overline{\alpha}(t)$. Show that

$$g(a, b) = \alpha(t)g(a, t) + \overline{\alpha}(t)g(t, b) \qquad \text{and} \qquad g(a, t) \le t \le g(t, b).$$

condit.normal

[7] Suppose W has a $N(\mu, \sigma^2)$ distribution. For each $h > 0$ and each $x \in \mathbb{R}$, show that

$$|\mathbb{P}(W - x \mid x - h \le W < x + h)| \le \frac{2|x - \mu|h^2}{\sigma^2}.$$

Hint: Reduce to the case where $\mu = 0$ and $\sigma = 1$. For that case define $F(h) := \int_{z-h}^{z+h} (t - z)\phi(t)\, dt$ and $G(h) := \int_{z-h}^{z+h} \phi(t)\, dt$. Show that

$$\begin{aligned} \mathbb{P}(W - x \mid x - h \le W < x + h) &= |F(h)/G(h)| \\ &= |F'(s)/G'(s)| \qquad \text{for some } 0 < s < h \\ &= s\,|\tanh(xs)| \le h\,(1 \wedge 2|xh|) \end{aligned}$$

References

Brown, L. D., A. V. Carter, M. G. Low, and C.-H. Zhang (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. *Annals of Statistics 32*(5), 2074–2097.

Pollard, D. (2001). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.