Introduction to Empirical Processes and Semiparametric Inference
Lecture 02: Overview Continued

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Empirical Processes: The Main Features

A *stochastic process* is a collection of random variables $\{X(t), t \in T\}$ on the same probability space, indexed by an arbitrary index set T.

In general, an *Empirical process* is a stochastic process based on a random sample, usually of n i.i.d. random variables $X_1, \ldots, X_n$, such as $\mathbb{F}_n$, where the index set is $T = \mathbb{R}$.

More generally, an empirical processes has the following features:

- The i.i.d. sample $X_1, \ldots, X_n$ is drawn from a probability measure P on an arbitrary sample space $\mathcal{X}$.
- We define the *empirical measure* to be $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} is the measure that assigns mass 1 to X_i and mass zero elsewhere.
- For a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we define $\mathbb{P}_n f = n^{-1} \sum_{i=1}^n f(X_i)$.
- For any class $\mathcal{F}$ of functions $f : \mathcal{X} \mapsto \mathbb{R}$ we can define the empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, i.e., $\mathcal{F}$ becomes the index set.
- This approach leads to a stunningly rich array of empirical processes.

Setting $\mathcal{X} = \mathbb{R}$, we can re-express $\mathbb{F}_n$ as the empirical process $\{\mathbb{P}_n f, f \in \mathcal{F}\}$ by setting $\mathcal{F} = \{\mathbf{1}\{x \leq t\}, t \in \mathbb{R}\}$.

Recall that the law of large number yields

$$\mathbb{F}_n(t) \xrightarrow{\text{as}} F(t), \tag{1}$$

for each $t \in \mathbb{R}$.

The functional perspective invites us to consider uniform results over $t \in \mathbb{R}$, or, more generally, over $f \in \mathcal{F}$: this is also called the *sample path* perspective.

Along this line of reasoning, Glivenko (1933) and Cantelli (1933) strengthened (1) to

$$\sup_{t \in \mathbb{R}} |\mathbb{F}_n(t) - F(t)| \stackrel{\text{as}}{\to} 0. \tag{2}$$

Another way of saying this is that the sample paths of $\mathbb{F}_n$ get uniformly closer to F as $n \to \infty$, almost surely.

For the general empirical process setting, we say that a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ is P-Glivenko-Cantelli (or P-GC, or Glinvenko-Cantelli, or GC) if

$$\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \stackrel{\text{as}*}{\to} 0, \tag{3}$$

where $Pf = \int_{\mathcal{X}} f(s) P(dx)$, and $\stackrel{\text{as}*}{\to}$ is a mode of convergence slightly stronger than $\stackrel{\text{as}}{\to}$ with the following attributes:

- certain measureability problems that can arise in complicated empirical processes (including many practical survival analysis settings) are cleanly resolved;
- the modes of convergence are equivalent in the setting of (2); and
- more details to be described later.

Returning to $\mathbb{F}_n$, the central limit theorem tells us that for each $t \in \mathbb{R}$,

$$G_n(t) \equiv \sqrt{n}\left[\mathbb{F}_n(t) - F(t)\right] \rightsquigarrow G(t),$$

where

- $\rightsquigarrow$ denotes convergence in distribution and
- $G(t)$ is a mean zero normal (Gaussian) random variable with variance $F(t)(1 - F(t))$.

In fact, we know that for all t in any finite set of the form $T_k = \{t_1, \ldots, t_k\} \in \mathbb{R}$, G_n will simultaneously converge to a mean zero Gaussian vector $G = \{G(t_1), \ldots, G(t_k)\}'$, where

$$\text{cov}[G(s), G(t)] = E[G(s)G(t)] \quad = \quad F(s \wedge t) - F(s)F(t), \tag{4}$$

for all $s, t \in T_k$.

Much more can be said.

Donsker (1952) showed that the sample paths of G_n (as a function on $\mathbb{R}$) converge in distribution to a certain stochastic process G.

Weak convergence is the generalization of convergence in distribution from vectors of random variables to sample paths of stochastic processes.

Donsker's famous result can be stated succinctly as

$$G_n \rightsquigarrow G \text{ in } \ell^\infty(\mathbb{R}), \tag{5}$$

where, for any index set T, $\ell^\infty(T)$ is the collection of all bounded functions $f : T \mapsto \mathbb{R}$.

$\ell^\infty(T)$ is a metric space with respect to the uniform metric on T and is used to remind us that we are taking the functional perspective or, more precisely, that we are thinking of distributional convergence in terms of the sample paths.

The limiting process G in (5) is a mean zero *Gaussian process* with

$$E[G(s)G(t)] = F(s \wedge t) - F(s)F(t),$$

as given in (4).

A Gaussian process is a stochastic process $\{Z(t), t \in T\}$, where

- for every finite $T_k \subset T$, $\{Z(t), t \in T_k\}$ is multivariate normal and
- all (almost all) sample path are continuous in a certain sense to be defined later.

The particular Gaussian process G in (5) can be written as $G(t) = \mathbb{B}(F(t))$, where $\mathbb{B}$ is a *Brownian bridge* process on the unit interval ($[0, 1]$).

The process $\mathbb{B}$ is a mean zero Gaussian process on $[0, 1]$,

- with covariance $s \wedge t - st$, for $s, t \in [0, 1]$,
- and has the form $\mathbb{W}(t) - t\mathbb{W}(1)$, for $t \in [0, 1]$, where $\mathbb{W}$ is a standard *Brownian motion* process.
- $\mathbb{B}$ can be generalized in a very important manner which we will discuss shortly.

The Brownian motion $\mathbb{W}$ is a mean zero Gaussian process on $[0, \infty)$ with

- continuous sample paths (almost surely);
- $\mathbb{W}(0) = 0$; and
- with covariance $s \wedge t$.

Both the Brownian bridge and Brownian motion are very important stochastic processes that arise frequently in statistics and probability.

Returning again to general empirical processes, define the random measure

$$\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P),$$

and define $\mathbb{G}$ to be a mean zero Gaussian process indexed by $\mathcal{F}$,

- with covariance $E\left[f(X)g(X)\right] - E\left[f(X)\right]E\left[g(X)\right]$, for all $f, g \in \mathcal{F}$,
- and having appropriately continuous sample paths (almost surely).

Both $\mathbb{G}_n$ and $\mathbb{G}$ can be thought of as being indexed by $\mathcal{F}$ and are completely defined once we specify $\mathcal{F}$.

We say that $\mathcal{F}$ is P-Donsker if

$$\mathbb{G}_n \rightsquigarrow \mathbb{G} \text{ in } \ell^\infty(\mathcal{F}).$$

The P or $\mathcal{F}$ may be dropped if the context is clear.

Donsker's (1952) theorem tells us that $\mathcal{F} = \{\mathbf{1}\{x \leq t\}, t \in \mathbb{R}\}$ is Donsker for all probability measures defined by a real distribution function F.

With $f(x) = \mathbf{1}\{x \leq t\}$ and $g(x) = \mathbf{1}\{x \leq s\}$,

$$E\left[f(X)g(X)\right] - E\left[f(X)\right]E\left[g(X)\right] = F(s \wedge t) - F(s)F(t).$$

For this reason, $\mathbb{G}$ is referred to as a Brownian bridge (or generalized Brownian bridge).

The range of possibilities for $\mathbb{G}_n$ and $\mathbb{G}$, as defined through classes of function $\mathcal{F}$, is stunningly vast.

Suppose we are interesting in forming simultaneous confidence bands for F over some subset $H \in \mathbb{R}$.

Since $\mathcal{F} = \{\mathbf{1}\{x \leq t\}, t \in \mathbb{R}\}$ is Glivenko-Cantelli, we can uniformly consistently estimate the covariance

$$\sigma(s,t) = F(s \wedge t) - F(s)F(t)$$

with

$$\hat{\sigma}(s,t) = \mathbb{F}_n(s \wedge t) - \mathbb{F}_n(s)\mathbb{F}_n(t).$$

Is it possible to use this to generate the desired confidence bands?

While knowledge of the covariance is sufficient to generate simultaneous confidence bands when H is finite via the chi-square distribution (for example), it is not sufficient when H is infinite (e.g., when H is a subinterval of $\mathbb{R}$).

When H is infinite, and more generally, it is useful to make use of the Donsker result for G_n.

In particular, let $U_n = \sup_{t \in H} |G_n(t)|$, and note that the distribution of U_n can be used to construct the desired confidence bands.

The *continuous mapping theorem* tells us that whenever a process $\{Z_n(t), t \in H\}$ converges weakly to a tight limiting process $\{Z(t), t \in H\}$ in $\ell^\infty(H)$, then

$$h(Z_n) \rightsquigarrow h(Z) \text{ in } h(\ell^\infty(H))$$

for any continuous map h.

In our setting, $U_n = h(G_n)$, where $g \mapsto h(g) = \sup_{t \in H} |g(t)|$, for any $g \in \ell^\infty(\mathbb{R})$, is a continuous real valued function.

Thus $U_n \rightsquigarrow U = \sup_{t \in H} |G(t)|$ by the continuous mapping theorem.

Note that

$$U = \sup_{t \in H} |G(t)| = \sup_{t \in H} |\mathbb{B}(F(t))| = \sup_{u \in [0,1]} |\mathbb{B}(u)|$$

is distribution free and has a known distribution.

In particular, for $s > 0$,

$$P[U \leq s] = 1 + 2 \sum_{k=1}^{\infty} (-1)^k e^{-2k^2 s^2}$$

(Billingsly, 1968, Page 85); e.g., a 95% confidence band is

$$\mathbb{F}_n(\cdot) \pm \frac{1.92}{\sqrt{n}}.$$

An alternative approach is to use bootstraps of $\mathbb{F}$.

These bootstraps have the form

$$\hat{\mathbb{F}}_n(t) = n^{-1} \sum_{i=1}^{n} W_{ni} \mathbf{1}\{X_i \leq t\},$$

where $(W_{n1}, \ldots, W_{nn})$ is a multinomial random n-vector with

- probabilities $(1/n, \ldots, 1/n)$,
- number of trials n,
- independence from the data $X_1, \ldots, X_n$.

More general weights $(W_{n1}, \ldots, W_{nn})$ are possible (and useful), and it can be shown that

$$\hat{G}_n = \sqrt{n}\left(\hat{\mathbb{F}}_n - \mathbb{F}_n\right)$$

converges weakly, conditional on the data $X_1, \ldots, X_n$, to G in $\ell^\infty(\mathbb{R})$.

Thus the bootstrap provides valid confidence bands for F.

Consider now the more general setting, where $\mathcal{F}$ is a Donsker class and we wish to construct confidence band for $Ef(X)$ for all $f \in \mathcal{H} \subset \mathcal{F}$.

Under minimal (moment) conditions on $\mathcal{F}$,

$$\hat{\sigma}(f,g) = \mathbb{P}_n\left[f(X)g(X)\right] - \mathbb{P}_n f(X)\mathbb{P}_n g(X)$$

is uniformly consistent for

$$\sigma(f,g) = P[f(X)g(X)] - Pf(X)Pg(X).$$

Whenever $\mathcal{H}$ is infinite, knowledge of σ is not enough to form confidence bands.

Fortunately, the bootstrap is always valid when $\mathcal{F}$ is Donsker and can thus be used for arbitrary and/or infinite $\mathcal{H}$.

If $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, where

$$\hat{\mathbb{P}}_n f = n^{-1} \sum_{i=1}^{n} W_{ni} f(X_i),$$

then the conditional distribution of $\hat{\mathbb{G}}_n$ given the data $X_1, \ldots, X_n$ converges weakly to $\mathbb{G}$ in $\ell^\infty(\mathcal{F})$, and $\mathcal{F}$ can be replaced with any $\mathcal{H} \subset \mathcal{F}$.

The bootstrap for $\mathbb{F}_n$ is a special case of this.

Although many statistics do not have the form $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, many can be written as $\phi(\mathbb{P}_n)$, where $\phi : \ell^\infty(\mathcal{F}) \mapsto B$ is continuous and B is a set (possibly infinite-dimensional).

Consider for example, $\xi_n(p) = \mathbb{F}_n^{-1}(p)$ for $p \in [a, b] \subset [0, 1]$.

$\xi_n(p)$ is the sample quantile process and can be shown to be expressible, under reasonable regularity conditions, as $\phi(\mathbb{F}_n)$, where ϕ is a "Hadamard differentiable" functional of $\mathbb{F}_n$.

In this example, and in general, the "functional delta method" states that $\sqrt{n}\left[\phi(\mathbb{P}_n) - \phi(P)\right]$ converges weakly in B to $\phi'(\mathbb{G})$, whenever $\mathcal{F}$ is Donsker and ϕ is Hadamard differentiable.

Moreover, the empirical process bootstrap discussed previously is also automatically valid, i.e., $\sqrt{n}\left[\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)\right]$, conditional on the data, converges weakly to $\phi'(\mathbb{G})$.

Many other statistics can be written as zeros, or approximate-zeros, of estimating equations based on empirical processes: these are called "Z-estimators".

An example is $\hat{\beta}$ from linear regression which can be written as a zero of $U_n(\beta) = \mathbb{P}_n \left[X(Y - X'\beta) \right]$.

Yet other statistics can be written as maxima or minima of objective functions based on empirical processes: these are called "M-estimators".

Examples include least-squares, maximum likelihood and minimum penalized likelihoods such as $\tilde{L}(\beta, \eta)$ from partly-linear logistic regression.

Thus many important statistics can be viewed as involving an empirical process.

A key asymptotic issue is studying the limiting behavior of these empirical processes in terms of their samples paths (the functional perspective).

Primary achievements in this direction include:

- Glivenko-Cantelli results which extend the law of large numbers,
- Donsker results which extend the central limit theorem,
- the validity of the bootstrap for Donsker classes, and
- the functional delta method.

Stochastic Convergence

There is always a metric space $(\mathbb{D}, d)$ involved, where $\mathbb{D}$ is the set of points and d is a metric satisfying, for all $x, y, z \in \mathbb{D}$,

- $d(x, y) \geq 0$,
- $d(x, y) = d(y, x)$,
- $d(x, z) \leq d(x, y) + d(y, z)$, and
- $d(x, y) = 0$ iff $x = y$.

Often $\mathbb{D} = \ell^\infty(T)$, the set of bounded functions $f : T \mapsto \mathbb{R}$, where T is an index set of interest and $d(x, y) = \sup_{t \in T} |x(t) - y(t)|$ is the *uniform distance*.

When we speak of X_n converging to X in $\mathbb{D}$, we mean that the samples paths of X_n behave more and more like the sample paths of X (as points in $\mathbb{D}$).

When X_n and X are *Borel measurable*, weak convergence, denoted $X_n \rightsquigarrow X$, is equivalent to $Ef(X_n) \to Ef(X)$ for every bounded, continuous $f : \mathbb{D} \mapsto \mathbb{R}$, which set of functions we denote $C_b(\mathbb{D})$, and where continuity is in terms of d, i.e., $|f(x) - f(y)| \to 0$ as $d(x, y) \to 0$.

We will define Borel measurability of X_n later, but it basically means that there are certain important subsets $A \subset \mathbb{D}$ for which $P(X_n \in A)$ is not defined.

What about when X_n is not Borel measurable?

We need to introduce *outer expectation* for arbitrary maps (not necessarily random variables) $T : \Omega \mapsto \mathbb{R}\big[-\infty, \infty\big]$, where Ω is a sample space.

The outer expectation of T, denoted E^*T is the infemum of all EU, where

- $U : \Omega \mapsto \mathbb{R}$ is measurable,
- $U \geq T$, and
- EU exists.

We analogously define $E_*T = -E^*\big(-T\big)$ as the *inner expectation*.

We can show that there exists a *minimal measurable majorant* T^* such that T^* is measurable and $ET^* = E^*T$.

Conversely, the *maximal measurable minorant* T_* also exists and satisfies $T_* = -(-T)^*$ and $E_*T = ET_*$.

We can also define *outer probability* $P^*(A)$ as the infemum over all $P(B)$, where $A \subset B \subset \Omega$ and B is measurable.

$P_*(A) = 1 - P^*(\Omega - A)$ is the *inner probability*.

Even if X_n is not Borel measurable, we can have weak convergence: specifically, if $E^* f(X_n) \to Ef(X)$, for all $f \in C_b(\mathbb{D})$, then we say $X_n \rightsquigarrow X$.

Such weak convergence carries with it the following implicit measurability requirements:

- X is Borel measurable and
- $E^* f(X_n) - E_* f(X_n) \to 0$, for every $f \in C_b(\mathbb{D})$.

A sequence X_n satisfying the second requirement above is called *asymptotically measurable*.

We can define outer measurable forms of weak and strong convergence in probability:

convergence in probability: $X_n \xrightarrow{\text{P}} X$ if $P\{d(X_n, X)^* > \epsilon\} \to 0$ for every $\epsilon > 0$.

outer almost sure convergence: $X_n \xrightarrow{\text{as*}} X$ if there exists a sequence Δ_n of measurable random variables such that

- $d(X_n, X) \leq \Delta_n$ and
- $P\{\limsup_{n\to\infty} \Delta_n = 0\} = 1$.

While these modes of convergence are slightly different than the usual ones, they are identical when all the random quantities involved are measurable.

For most settings we will study, the limiting quantity X will be *tight* (related to smoothness) in addition to being Borel measurable.

Moreover, the most common choice for $\mathbb{D}$ will be $\ell^\infty(T)$ with the uniform metric d.

Let $\rho(s,t)$ be a *semimetric* on T: a semimetric ρ satisfies all of the requirements of a metric except that $\rho(s,t) = 0$ does not necessarily imply $s = t$.

We say that the semimetric space (T, ρ) is *totally bounded* if, for every $\epsilon > 0$, there exists a finite subset $T_k = \{t_1, \ldots, t_k\} \subset T$ such that for all $t \in T$, we have $\rho(s, t) \leq \epsilon$ for some $s \in T_k$.

Define $UC(T, \rho)$ to be the subset of $\ell^\infty(T)$ consisting of functions x for which

$$\lim_{\delta \downarrow 0} \sup_{s,t \in T \text{ with } \rho(s,t) \leq \delta} |x(t) - x(s)| = 0.$$

The functions of $UC(T, \rho)$ are also called the ρ-equicontinuous functions.

The "UC" refers to uniform continuity, and $UC(T, \rho)$ is a very important metric space.

A Borel measurable stochastic process X in $\ell^\infty(T)$ is tight if $X \in UC(T, \rho)$ almost surely for some ρ making T totally bounded.

If X is a Gaussian process, then ρ can be chosen without loss of generality to be $\rho(s,t) = (\text{var}[X(s) - X(t)])^{1/2}$.

Tight Gaussian processes will be the most important limiting processes we will consider.

Theorem 2.1. X_n *converges weakly to a tight* X *in* $\ell^\infty(T)$ *if and only if:*

(i) For all finite $\{t_1, \ldots, t_k\} \subset T$*, the multivariate distribution of* $\{X_n(t_1), \ldots, X_n(t_k)\}$ *converges to that of* $\{X(t_1), \ldots, X(t_k)\}$*.*

(ii) There exists a semimetric ρ *for which* T *is totally bounded and*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^* \left\{ \sup_{s,t \in T \text{ with } \rho(s,t) < \delta} |X_n(s) - X_n(t)| > \epsilon \right\} = 0,$$

for all $\epsilon > 0$*.*

Usually, establishing Condition (i) is easy, while establishing Condition (ii) is hard.

For empirical processes from i.i.d. data, we want to show $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $\ell^\infty(\mathcal{F})$, where

- $\mathcal{F}$ is a class of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$,
- $\mathcal{X}$ is the sample space, and
- $Ef^2(X) < \infty$ for all $f \in \mathcal{F}$.

Thus Condition (i) is automatic by the standard central limit theorem.

Establishing Condition (ii) is much harder.

When $\mathcal{F}$ is Donsker:

- The limiting process $\mathbb{G}$ is always a tight Gaussian process,
- $\mathcal{F}$ is totally bounded w.r.t. $\rho(f,g) = \{\text{var}[f(X) - g(X)]\}^{1/2}$, and
- Condition (ii) is satisfied with $T = \mathcal{F}$, $X_n(f) = \mathbb{G}_n f$, and $X(f) = \mathbb{G}f$, for all $f \in \mathcal{F}$.

Of course, the hard part is showing $\mathcal{F}$ is Donsker.

Theorem (continuous mapping). Suppose $g : \mathbb{D} \mapsto \mathbb{E}$ *is continuous at every point of* $\mathbb{D}_0 \subset \mathbb{D}$ *and* $X_n \rightsquigarrow X$*, where* X *takes its values almost surely on* $\mathbb{D}_0$*. Then* $g(X_n) \rightsquigarrow g(X)$*.*

Example. Let $g : \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$ be $g(x) = \|x(f)\|_\mathcal{F} \equiv \sup_{f\in\mathcal{F}} |x(f)|$:

- The continuous mapping theorem now implies that $\|\mathbb{G}_n\|_\mathcal{F} \rightsquigarrow \|\mathbb{G}\|_\mathcal{F}$.
- This can be used for confidence bands for Pf.

Entropy for Glivenko-Cantelli and Donsker Theorems

In order to obtain Glivenko-Cantelli and Donsker results for $\mathcal{F}$, we need to evaluate the complexity (or entropy) of $\mathcal{F}$.

The easiest way is with *entropy with bracketing* (also called *bracketing entropy*), which we now introduce.

First, for $1 \leq r < \infty$, define $L_r(P)$ to be the space of measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{P,r} \equiv [Pf^r(X)]^{1/r} < \infty$.

An ϵ-bracket in $L_r(P)$ is a pair of functions $l, u \in L_r(P)$ with $l(X) \leq u(X)$ P-almost surely and $\|u - l\|_{P,r} \leq \epsilon$.

A function $f \in \mathcal{F}$ lies in the bracket (l, u) if $l(X) \leq f(X) \leq u(X)$ P-almost surely.

The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of ϵ-brackets in $L_r(P)$ needed in order to ensure that every $f \in \mathcal{F}$ is contained in at least one bracket.

The logarithm of the bracketing number is the entropy with bracketing.

Theorem 2.2. Let $\mathcal{F}$ be a class of measurable functions and suppose $N_{[]}(\epsilon, \mathcal{F}, L_r(P)) < \infty$ for every $\epsilon > 0$. Then $\mathcal{F}$ is P-Glivenko-Cantelli.

Example. Let $\mathcal{X} = \mathbb{R}$, $\mathcal{F} = \{\mathbf{1}\{x \leq t\}, t \in \mathbb{R}\}$, and P be the distribution F on $\mathcal{X}$.

For each $\epsilon > 0$, there exists a finite set of real numbers with $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $F(t_j-) - F(t_{j-1}) \leq \epsilon$, all $1 \leq j \leq k$, with $F(t_0) = 0$ and $F(t_k) = 1$.

Now consider the brackets $\{(l_j, u_j), 1 \leq j \leq k\}$, with $l_j(x) = \mathbf{1}\{x \leq t_{j-1}\}$ and $u_j(x) = \mathbf{1}\{x < t_j\}$ (note that u_j is not in $\mathcal{F}$).

Clearly, each $f \in \mathcal{F}$ is contained in one of these brackets and the number of such brackets is finite.

This implies that $\mathbb{F}_n$ is uniformly consistent for F almost surely.

For Donsker results, a more refined assessment of entropy is needed.

The *bracketing integral* (or *bracketing entropy integral*) is

$$J_{[]}(\delta, \mathcal{F}, L_r(P)) \equiv \int_0^\delta \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_r(P))}\, d\epsilon.$$

This allows the bracketing entropy to go to ∞, as $\epsilon \downarrow 0$, but keeps this from happening too fast.

Theorem 2.3. Let $\mathcal{F}$ be a class of measurable functions with $J_{[]}(\infty, \mathcal{F}, L_2(P)) < \infty$. Then $\mathcal{F}$ is P-Donsker.

Example, continued. For each $\epsilon > 0$, choose the brackets (l_j, u_j) as before and note that $\|u_j - l_j\|_{P,2} = (\|u_j - l_j\|_{P,1})^{1/2} \leq \epsilon^{1/2}$.

Thus an L_2 ϵ-bracket is an L_1 ϵ^2-bracket, and hance the minimum number of L_2 ϵ-brackets needed to cover $\mathcal{F}$ is $1 + 1/\epsilon^2$.

Since $\log(1+a) \leq 1 + \log(a)$ for all $a \geq 1$, we now have that

$$\log N_{[]}(\epsilon, \mathcal{F}, L_2(P)) \leq \left[1 + \log(1/\epsilon^2)\right] \mathbf{1}\{\epsilon \leq 1\},$$

and thus

$$J_{[]}(\infty, \mathcal{F}, L_2(P)) \leq \int_0^\infty u^{1/2} e^{-u/2} du = \sqrt{2\pi} < \infty,$$

where we used the variable substitution $u = 1 + \log(1/\epsilon^2)$.

Many other classes of function have bounded bracketing entropy integral, including many parametric classes of functions and the class $\mathcal{F}$ of all monotone functions $f : \mathcal{X} = \mathbb{R} \mapsto [0, 1]$, for all $1 \leq r < \infty$ and any P on $\mathcal{X} = \mathbb{R}$.

There are many classes $\mathcal{F}$ for which bracketing entropy does not work, but a different kind of entropy, *uniform entropy* (or just *entropy*) works.

For a probability measure Q, the *covering number* $N(\epsilon, \mathcal{F}, L_r(Q))$ is the minimum number of $L_r(Q)$ ϵ-balls needed to cover $\mathcal{F}$, where an ϵ-ball around a function $g \in L_r(Q)$ is the set

$$\{h \in L_r(Q) : \|h - g\|_{Q,r} < \epsilon\} :$$

- For a collection of ϵ-balls to cover $\mathcal{F}$, all elements of $\mathcal{F}$ must be contained in at least one of the ϵ-balls.
- It is not necessary that the centers of the balls in the collection be in $\mathcal{F}$.

The logarithm of the covering number is the *entropy*.

What we really need, though, is the *uniform covering number*:

$$\sup_{Q} N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)), \tag{6}$$

where

- $F : \mathcal{X} \mapsto \mathbb{R}$ is an *envelope* for $\mathcal{F}$, meaning that $|f(x)| \leq F(x)$ for all $x \in \mathcal{X}$ and all $f \in \mathcal{F}$, and where
- the supremum is taken over all finitely discrete probability measures Q with $\|F\|_{Q,r} > 0$.

The logarithm of (6) is the *uniform entropy*: note that it does not depend on the probability measure P for the observed data.

The *uniform entropy integral* is

$$J(\delta, \mathcal{F}, L_r) = \int_0^\delta \sqrt{\log \sup_Q N(\epsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q))}\, d\epsilon,$$

where the supremum is taken over the same set used in (6).

The following two theorems are the Glivenko-Cantelli and Donsker theorems for uniform entropy:

Theorem 2.4. Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $\sup_Q N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty$ for every $\epsilon > 0$, where the supremum is taken over the same set used in (6). If $P^* F < \infty$, then $\mathcal{F}$ is P-Glivenko-Cantelli.

Theorem 2.5. Let $\mathcal{F}$ be an appropriately measurable class of measurable functions with $J(1, \mathcal{F}, L_2) < \infty$. If $P^* F^2 < \infty$, then $\mathcal{F}$ is P-Donsker.

An important class of functions $\mathcal{F}$ for which $J(1, \mathcal{F}, L_2) < \infty$ are the Vapnik-Červonenkis (VC) classes:

- these include the indicator function class example from above;
- classes that can be expressed as finite-dimensional vector spaces; and
- many other classes.

There are preservation theorems for building VC classes from other VC classes as well as preservation theorems for Donsker, Glivenko-Cantelli, and other classes: we will go into this in depth in Chapter 9.

Bootstrapping Empirical Processes

We mentioned earlier that measurability in weak convergence can be tricky.

This is even more the case with the bootstrap, since there are two sources of randomness:

- the data and
- the random weights (resampling) used in the bootstrap.

For this reason, we have to assess convergence of conditional laws (the bootstrap given the data) differently than regular weak convergence.

An important result is that $X_n \rightsquigarrow X$ in the metric space $(\mathbb{D}, d)$ if and only if

$$\sup_{f \in BL_1} |E^* f(X_n) - Ef(X)| \to 0, \tag{7}$$

where BL_1 is the space of functions $f : \mathbb{D} \mapsto \mathbb{R}$ with Lipschitz norm bounded by 1, i.e.

- $\|f\|_{\mathbb{D}} \leq 1$ and
- $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{D}$.

We can now define bootstrap weak convergence.

Let $\hat{X}_n$ be a sequence of bootstrapped processes in $\mathbb{D}$ with random weights which we denote by M.

For some tight process X in $\mathbb{D}$, we use the notation $\hat{X}_n \underset{M}{\overset{\mathsf{P}}{\rightsquigarrow}} X$ to:

- mean that $\sup_{h \in BL_1} \left| E_M h(\hat{X}_n) - Eh(X) \right| \overset{\mathsf{P}}{\to} 0$ and
- $E_M h(\hat{X}_n)^* - E_M h(\hat{X}_n)_* \overset{\mathsf{P}}{\to} 0$, for all $h \in BL_1$,
- where subscript M denotes conditional expectation over M given the data,
- and where $h(\hat{X}_n)^*$ and $h(\hat{X}_n)_*$ are the measurable majorants and minorants with respect to the joint data (include M).

In addition to using the multinomial weights mentioned previously which correspond to the nonparametric bootstrap, we have a useful alternative that performs better in some settings.

Let $\vec{\xi} = (\xi_1, \xi_2, \ldots)$ be an infinite sequence of nonnegative i.i.d. random variables, also independent of the data $\vec{X}$, which have

- mean $0 < \mu < \infty$ and
- variance $0 < \tau^2 < \infty$,
- and which satisfy $\|\xi\|_{2,1} < \infty$, where

$$\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(\|\xi\| > x)}\, dx.$$

We can now define the *multiplier bootstrap* empirical measure $\tilde{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n (\xi_i / \bar{\xi}_n) f(X_i)$, where $\tilde{\mathbb{P}}_n$ is defined to be zero if $\bar{\xi}_n = 0$.

Note that the weights involved add up to n for both the multiplier and nonparametric bootstraps.

When ξ has a standard exponential distribution, for example, the moment conditions are easily satisfied and the resulting multiplier bootstrap has Dirichlet weights.

Let $\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$, $\tilde{\mathbb{G}}_n = \sqrt{n}(\mu/\tau)(\tilde{\mathbb{P}}_n - \mathbb{P}_n)$, and $\mathbb{G}$ be the Brownian bridge in $\ell^\infty(\mathcal{F})$.

Theorem 2.6. The following are equivalent:

1. $\mathcal{F}$ *is* P*-Donsker.*
2. $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathsf{P}}{\rightsquigarrow}} \mathbb{G}$ *and the sequence* $\hat{\mathbb{G}}_n$ *is asymptotically measurable.*
3. $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{\mathsf{P}}{\rightsquigarrow}} \mathbb{G}$ *and the sequence* $\tilde{\mathbb{G}}_n$ *is asymptotically measurable.*

Theorem 2.7. The following are equivalent:

1. $\mathcal{F}$ *is* P*-Donsker and* $P^* \left[\sup_{f\in\mathcal{F}} \left(f(X) - Pf\right)^2\right] < \infty$.
2. $\hat{\mathbb{G}}_n \underset{W}{\overset{\mathsf{as}*}{\rightsquigarrow}} \mathbb{G}$.
3. $\tilde{\mathbb{G}}_n \underset{\xi}{\overset{\mathsf{P}}{\rightsquigarrow}} \mathbb{G}$.

A few comments:

- Note that most other bootstrap results use conditional almost sure consistency; however, convergence in probability is sufficient for almost all statistical applications.
- We can use these reasults for many complicated inference settings.
- There are continuous mapping results for the bootstrap.
- There are also Glivenko-Cantelli type results which are needed for some applications.
- More on the bootstrap will be discussed in Chapter 10.