1 The Glivenko-Cantelli Theorem

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution function F on $\mathbf{R}$. The *empirical distribution function* is the function of x defined by

$$\hat{F}_n(x) = \frac{1}{n} \sum_{1 \leq i \leq n} I\{X_i \leq x\} \ .$$

For a given $x \in \mathbf{R}$, we can apply the *strong law of large numbers* to the sequence $I\{X_i \leq x\}, i = 1, \ldots n$ to assert that

$$\hat{F}_n(x) \to F(x)$$

a.s (in order to apply the strong law of large numbers we only need to show that $E[|I\{X_i \leq x\}|] < \infty$, which in this case is trivial because $|I\{X_i \leq x\}| \leq 1$). In this sense, $\hat{F}_n(x)$ is a reasonable estimate of $F(x)$ for a given $x \in \mathbf{R}$. But is $\hat{F}_n(x)$ a reasonable estimate of the $F(x)$ when both are viewed as functions of x?

The *Glivenko-Cantelli Thoerem* provides an answer to this question. It asserts the following:

Theorem 1.1 Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution function F on $\mathbf{R}$. Then,

$$\sup_{x \in \mathbf{R}} |\hat{F}_n(x) - F(x)| \to 0 \quad \text{a.s.} \tag{1}$$

This result is perhaps the oldest and most well known result in the very large field of *empirical process theory*, which is at the center of much of modern econometrics. The statistic (1) is an example of a *Kolmogorov-Smirnov* statistic.

The proof of the result will require the following lemma.

Lemma 1.1 Let F be a (nonrandom) distribution function on $\mathbf{R}$. For each $\epsilon > 0$ there exists a finite partition of the real line of the form $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that for $0 \leq j \leq k-1$

$$F(t_{j+1}^-) - F(t_j) \leq \epsilon \ .$$

PROOF: Let $\epsilon > 0$ be given. Let $t_0 = -\infty$ and for $j \geq 0$ define

$$t_{j+1} = \sup\{z : F(z) \leq F(t_j) + \epsilon\} \ .$$

Note that $F(t_{j+1}) \geq F(t_j)+\epsilon$. To see this, suppose that $F(t_{j+1}) < F(t_j)+\epsilon$. Then, by right continuity of F there would exist $\delta > 0$ so that $F(t_{j+1}+\delta) < F(t_j) + \epsilon$, which would contradict the definition of t_{j+1}. Thus, between t_j and t_{j+1}, F jumps by at least ϵ. Since this can happen at most a finite number of times, the partition is of the desired form, that is $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ with $k < \infty$. Moreover, $F(t_{j+1}^-) \leq F(t_j) + \epsilon$. To see this, note that by definition of t_{j+1} we have $F(t_{j+1} - \delta) \leq F(t_j) + \epsilon$ for all $\delta > 0$. The desired result thus follows from the definition of $F(t_{j+1}^-)$. ■

PROOF OF 1.1: It suffices to show that for any $\epsilon > 0$

$$\limsup_{n\to\infty} \sup_{x\in\mathbf{R}} |\hat{F}_n(x) - F(x)| \leq \epsilon \text{ a.s.}$$

To this send, let $\epsilon > 0$ be given and consider a partition of the real line into finitely many pieces of the form $-\infty = t_0 < t_1 \cdots < t_k = \infty$ such that for $0 \leq j \leq k-1$

$$F(t_{j+1}^-) - F(t_j) \leq \frac{\epsilon}{2} \ .$$

The existence of such a partition is ensured by the previous lemma. For any $x \in \mathbf{R}$, there exists j such that $t_j \leq x < t_{j+1}$. For such j,

$$\begin{aligned} \hat{F}_n(t_j) \leq &\quad \hat{F}_n(x) &\quad \leq \hat{F}_n(t_{j+1}^-) \\ F(t_j) \leq &\quad F(x) &\quad \leq F(t_{j+1}^-) \ , \end{aligned}$$

which implies that

$$\hat{F}_n(t_j) - F(t_{j+1}^-) \leq \hat{F}_n(x) - F(x) \leq \hat{F}_n(t_{j+1}^-) - F(t_j) \ .$$

Furthermore,

$$\begin{aligned} \hat{F}_n(t_j) - F(t_j) + F(t_j) - F(t_{j+1}^-) &\leq \hat{F}_n(x) - F(x) \\ \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + F(t_{j+1}^-) - F(t_j) &\geq \hat{F}_n(x) - F(x) \ . \end{aligned}$$

By construction of the partition, we have that

$$\begin{aligned} \hat{F}_n(t_j) - F(t_j) - \frac{\epsilon}{2} &\leq \hat{F}_n(x) - F(x) \\ \hat{F}_n(t_{j+1}^-) - F(t_{j+1}^-) + \frac{\epsilon}{2} &\geq \hat{F}_n(x) - F(x) \ . \end{aligned}$$

The desired result now follows from the Strong Law of Large Numbers. ■

2 The Sample Median

We now give a brief application of the Glivenko-Cantelli Theorem. Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution F. Suppose one is interested in the median of F. Concretely, we will define

$$\text{Med}(F) = \inf\{x : F(x) \geq \frac{1}{2}\} \ .$$

A natural estimator of $\text{Med}(F)$ is the sample analog, $\text{Med}(\hat{F}_n)$. Under what conditions is $\text{Med}(\hat{F}_n)$ a reasonable estimate of $\text{Med}(F)$?

Let $m = \text{Med}(F)$ and suppose that F is well behaved at m in the sense that $F(t) > \frac{1}{2}$ whenever $t > m$. Under this condition, we can show using the Glivenko-Cantelli Theorem that $\text{Med}(\hat{F}_n) \to \text{Med}(F)$ a.s. We will now prove this result.

Suppose F_n is a (nonrandom) sequence of distribution functions such that

$$\sup_{x \in \mathbf{R}} |F_n(x) - F(x)| \to 0 \ .$$

Let $\epsilon > 0$ be given. We wish to show that there exists $N = N(\epsilon)$ such that for all $n > N$

$$|\text{Med}(F_n) - \text{Med}(F)| < \epsilon \ .$$

Choose $\delta > 0$ so that

$$\begin{aligned} \delta &< \frac{1}{2} - F(m - \epsilon) \\ \delta &< F(m + \epsilon) - \frac{1}{2} \ , \end{aligned}$$

which in turn implies that

$$\begin{aligned} F(m-\epsilon) &< \frac{1}{2}-\delta \\ F(m+\epsilon) &> \frac{1}{2}+\delta \ . \end{aligned}$$

(It might help to draw a picture to see why we should pick δ in this way.) Next choose N so that for all $n > N$,

$$\sup_{x\in\mathbf{R}} |F_n(x) - F(x)| < \delta \ .$$

Let $m_n = \text{Med}(F_n)$. For such n, $m_n > m - \epsilon$, for if $m_n \leq m - \epsilon$, then

$$F(m-\epsilon) > F_n(m-\epsilon) - \delta \geq \frac{1}{2} - \delta \ ,$$

which contradicts the choice of δ. We also have that $m_n < m + \epsilon$, for if $m_n \geq m + \epsilon$, then

$$F(m+\epsilon) < F_n(m+\epsilon) + \delta \leq \frac{1}{2} + \delta \ ,$$

which again contradicts the choice of δ. Thus, for $n > N$, $|m_n - m| < \epsilon$, as desired.

By the Glivenko-Cantelli Theorem, it follows immediately that $\text{Med}(\hat{F}_n) \to \text{Med}(F)$ a.s.

3 A Generalized Glivenko-Cantelli Theorem

Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P. We no longer require that the random variables are real-valued. The *empirical measure* is the measure defined by

$$\hat{P}_n = \frac{1}{n} \sum_{1\leq i\leq n} \delta_{X_i} \ ,$$

where δ_x is the measure that puts mass one at x. In this language, the empirical distribution function is simply $\hat{P}_n(-\infty, x]$ and the uniform convergence over $x \in \mathbf{R}$ asserted in (1) can instead be understood as uniform

convergence of the empirical measure to the true measure over sets of the form $(-\infty, x]$ for $x \in \mathbf{R}$. Can this result be extended to other classes of sets?

It turns out that the convergence of the empirical measure to the true measure is uniform for large classes of sets provided that the sets do not "pick out" too many subsets of any set of n points in the space.

Definition 3.1 *We say that a class* $\mathbf{V}$ *of subsets of a space* S *has polynomial discrimination (of degree* v*) if there exists a polynomial* $\rho(\cdot)$ *(of degree* v*) so that from every set of* n *points in* S *the class picks out at most* $\rho(n)$ *distinct subsets. Formally, if* $S_0 \subseteq S$ *consists of* n *points, then there are at most* $\rho(n)$ *distinct sets of the form* $V \cap S_0$ *with* $V \in \mathbf{V}$*.*

It is easy to see that half-intervals of the form $(-\infty, x]$ with $x \in \mathbf{R}$ have polynomial discrimination of degree 1. Unfortunately, it is difficult to come up with more classes of sets have polynomial discrimination. For that purpose, the following observation is useful. A class of sets $\mathbf{V}$ is said to "shatter" a set of points S_0 if it can "pick out" every possible subset of S_0, i.e., if each subset of S_0 can be written as $V \cap S_0$ for some $V \in \mathbf{V}$. A lemma due to Sauer asserts that if a class of sets shatters no set of v points, then it has polynomial discrimination of degree (no greater than) $v - 1$. For example, on the real line, the half-intervals shatter no set of two points, which is consistent with our earlier observation. More interestingly, the class of all closed discs in $\mathbf{R}^2$ shatters no set of four points. This follows from the following result, which allows us to generate many classes of sets that have polynomial discrimination.

Lemma 3.1 *Let* $\mathbf{G}$ *be a vector space of real-valued functions on* S *with dimension* $v - 1$*. The class of sets of the form* $\{g \geq 0\}$ *as* g *varies over* $\mathbf{G}$ *shatters no set of* v *points in* S*.*

PROOF: Choose $s_1, \ldots, s_v$ distinct points from S. We must show that there is some subset of these points that is not "picked out" by the sets of the

form $\{g \geq 0\}$ as g varies over $\mathbf{G}$. Define the map $L : \mathbf{G} \to \mathbf{R}^v$ by the rule

$$L(g) = (g(s_1), \ldots, g(s_v)) \ .$$

L is a linear map. Hence, $L(\mathbf{G})$ is a linear subspace of $\mathbf{R}^v$ of dimension no greater than $v - 1$. As a result, there exists a nonzero $\gamma \in \mathbf{R}^v$ that is orthogonal to $L(\mathbf{G})$, i.e.,

$$\sum_{1 \leq i \leq v} \gamma_i g(s_i) = 0 \text{ for all } g \in \mathbf{G} \ .$$

Equivalently,

$$\sum_{i:\gamma_i \geq 0} \gamma_i g(s_i) = \sum_{i:\gamma_i < 0} -\gamma_i g(s_i) \ . \tag{2}$$

Assume without loss of generality that $\{i : \gamma_i < 0\} \neq 0$ (otherwise replace γ by $-\gamma$). Consider the set $\{s_i : \gamma_i \geq 0\}$. Suppose by way of contradiction that there exists a $g \in \mathbf{G}$ such that $\{g \geq 0\} \cap \{s_1, \ldots, s_v\} = \{s_i : \gamma_i \geq 0\}$. For such a g, (**??**) cannot hold. ■

The following theorem generalizes the classical Glivenko-Cantelli Theorem to all classes of sets with polynomial discrimination, including those that can be generated using the previous lemma.

Theorem 3.1 *Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P on S. For every (suitably measurable) class $\mathbf{V}$ of subsets of S with polynomial discrimination,*

$$\sup_{V \in \mathbf{V}} |\hat{P}_n V - PV| \to 0 \ \textit{a.s.} \ .$$

We will break the proof of this result up into several steps.

Lemma 3.2 *Let $\{Z(t) : t \in T\}$ and $\{Z'(t) : t \in T\}$ be independent stochastic processes. Suppose there exists $\alpha > 0$ and $\beta > 0$ such that $P\{|Z'(t)| \leq \alpha\} \geq \beta$ for every $t \in T$. Then,*

$$P\{\sup_{t \in T} |Z(t)| > \epsilon\} \leq \frac{1}{\beta} P\{\sup_{t \in T} |Z(t) - Z'(t)| > \epsilon - \alpha\} \ .$$

PROOF: Let τ be a random variable such that

$$\sup_{t\in T}|Z(t)| > \epsilon \Longleftrightarrow |Z(\tau)| > \epsilon \ .$$

Note that

$$\begin{aligned} P\{|Z(\tau)| > \epsilon, |Z'(\tau)| \leq \alpha\} &= P\{|Z'(\tau)| \leq \alpha || Z(\tau)| > \epsilon\} P\{|Z(\tau)| > \epsilon\} \\ &\geq \beta P\{|Z(\tau)| > \epsilon\} \ , \end{aligned}$$

where the inequality follows because

$$\begin{aligned} P\{|Z'(\tau)| \leq \alpha || Z(\tau)| > \epsilon\} &= E[P\{|Z'(\tau)| \leq \alpha | Z\} || Z(\tau)| > \epsilon] \\ P\{|Z'(\tau)| \leq \alpha | Z\} &\geq \beta \ . \end{aligned}$$

Furthermore,

$$\begin{aligned} P\{|Z(\tau)| > \epsilon, |Z'(\tau)| \leq \alpha\} &\leq P\{|Z(\tau) - Z'(\tau)| > \epsilon - \alpha\} \\ &\leq P\{\sup_{t\in T}|Z(t) - Z'(t)| > \epsilon - \alpha\} \ . \end{aligned}$$

The desired result thus follows. ■

Lemma 3.3 *Let $X_i, i = 1, \ldots, n$ and $X'_i, i = 1, \ldots, n$ be independent i.i.d. sequences of random variables with distribution P on S. For every (suitably measurable) class $\mathbf{V}$ of subsets of S,*

$$P\{\sup_{V\in\mathbf{V}}|\hat{P}_n V - PV| > \epsilon\} \leq 2P\{\sup_{V\in\mathbf{V}}|\hat{P}_n V - \hat{P}'_n V| > \frac{\epsilon}{2}\}$$

whenever $n \geq 8/\epsilon^2$.

PROOF: By Chebychev's Inequality,

$$P\{|\hat{P}'_n V - PV| > \frac{\epsilon}{2}\} \leq \frac{4}{n\epsilon^2} \ .$$

Hence,

$$P\{|\hat{P}'_n V - PV| \leq \frac{\epsilon}{2}\} \geq \frac{1}{2}$$

whenever $n \geq 8/\epsilon^2$. We may therefore apply the preceding lemma to $Z(V) = \hat{P}_n V - PV$ and $Z'(V) = \hat{P}'_n V - PV$ with $\alpha = \epsilon/2$ and $\beta = 1/2$ to reach the desired conclusion. ■

Lemma 3.4 *Let $X_i, i = 1, \ldots, n$ and $X'_i, i = 1, \ldots, n$ be independent i.i.d. sequences of random variables with distribution P on S. Independently, let $\sigma_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution putting equal mass on -1 and 1. Let*

$$\hat{P}_n^o = \frac{1}{n} \sum_{1 \le i \le n} \sigma_i \delta_{X_i} \ .$$

For every (suitably measurable) class $\mathbf{V}$ of subsets of S,

$$P\{\sup_{V \in \mathbf{V}} |\hat{P}_n V - \hat{P}'_n V| > \frac{\epsilon}{2}\} \le 2P\{\sup_{V \in \mathbf{V}} |\hat{P}_n^o V| > \frac{\epsilon}{4}\} \ .$$

PROOF: Note that the distribution of the random variables $I\{X_i \in V\} - I\{X'_i \in V\}$ for $i = 1, \ldots, n$ and $V \in \mathbf{V}$ and $\sigma_i[I\{X_i \in V\} - I\{X'_i \in V\}]$ for $i = 1, \ldots, n$ and $V \in \mathbf{V}$ are the same. To see this, simply consider the distribution conditional on $\sigma_i, i = 1, \ldots, n$. Then,

$$\begin{aligned} & P\{\sup_{V \in \mathbf{V}} |\hat{P}_n V - \hat{P}'_n V| > \frac{\epsilon}{2}\} \\ & \qquad = P\{\sup_{V \in \mathbf{V}} |\frac{1}{n} \sum_{1 \le i \le n} \sigma_i [I\{X_i \in V\} - I\{X'_i \in V\}]| > \frac{\epsilon}{2}\} \\ & \qquad \le P\{\sup_{V \in \mathbf{V}} |\frac{1}{n} \sum_{1 \le i \le n} \sigma_i I\{X_i \in V\}| > \frac{\epsilon}{4}\} \\ & \qquad\qquad + P\{\sup_{V \in \mathbf{V}} |\frac{1}{n} \sum_{1 \le i \le n} \sigma_i I\{X'_i \in V\}| > \frac{\epsilon}{4}\} \ , \end{aligned}$$

from which the desired result follows. ■

We are now almost prepared to prove the desired result, but to do so we will require the following inequality due to Hoeffding for tail probabilities of sums of independent, bounded random variables.

Lemma 3.5 *Let $Y_i, i = 1, \ldots, n$ be independent random variables with zero mean and satisfying $a_i \le Y_i \le b_i$. For each $\eta > 0$,*

$$P\{|\sum_{1 \le i \le n} Y_i| \ge \eta\} \le 2 \exp(-2\eta^2 / (\sum_{1 \le i \le n} (b_i - a_i)^2) \ .$$

PROOF OF THEOREM **??**: Let $\mathbf{X}_n = \{X_1, \ldots, X_n\}$ and consider

$$P\{\sup_{V \in \mathbf{V}} |\hat{P}_n^o V| > \frac{\epsilon}{4} | \mathbf{X}_n\} \ .$$

Since $\mathbf{V}$ has polynomial discrimination, we may replace $\mathbf{V}$ with $\mathbf{V}^*$, which is a collection of at most $\rho(n)$ sets from $\mathbf{V}$. Note that $\mathbf{V}^*$ may depend on $\mathbf{X}_n$. Thus,

$$P\{\sup_{V \in \mathbf{V}} |\hat{P}_n^o V| > \frac{\epsilon}{4} | \mathbf{X}_n\} \leq \rho(n) \max_{V \in \mathbf{V}^*} P\{|\hat{P}_n^o V| > \frac{\epsilon}{4} | \mathbf{X}_n\} \ .$$

Apply Hoeffding's inequality to $Y_i = \sigma_i I\{X_i \in V\}$ to conclude that

$$P\{|\hat{P}_n^o V| > \frac{\epsilon}{4} | \mathbf{X}_n\} \leq 2 \exp(-n\epsilon^2/32) \ .$$

Integrating out with respect to $\mathbf{X}_n$ and using the previous lemmas, we have that

$$P\{\sup_{V \in \mathbf{V}} |\hat{P}_n V - PV| > \epsilon\} \leq 8\rho(n) \exp(-n\epsilon^2/32) \ .$$

It follows that

$$\sum_{1 \leq n < \infty} P\{\sup_{V \in \mathbf{V}} |\hat{P}_n V - PV| > \epsilon\} < \infty \ ,$$

so the desired result follows from the Borel-Cantelli lemma. ■