GLIVENKO–CANTELLI THEOREMS

Let $X_1, X_2, \ldots$ be independent identically distributed (i.i.d.) random variables with common distribution function F, $F(x) = P(X \leqslant x)$ for $-\infty < x < \infty$, and let $\mathbb{F}_n$ denote the *empirical distribution function* of the first nX's (*see* EMPIRICAL DISTRIBUTION FUNCTION (EDF) STATISTICS) defined for $-\infty < x < \infty$ by

$$n\mathbb{F}_n(x) = [\text{number of } i \leqslant n \text{ with } X_i \leqslant x] = \sum_{i=1}^{n} 1_{(-\infty,x]}(X_i).$$

For fixed x, $n\mathbb{F}_n(x)$ has a binomial distribution* with parameters n and $F(x)$, and hence, using the *weak law of large numbers** for (3), and the classical de Moivre–Laplace *central limit theorem** for (4),

$$E\mathbb{F}_n(x) = F(x), \tag{1}$$

$$\text{var}(\mathbb{F}_n(x)) = F(x)(1 - F(x))/n, \tag{2}$$

$$\mathbb{F}_n(x) \underset{p}{\rightarrow} F(x) \quad \text{as} \quad n \to \infty, \tag{3}$$

$$n^{1/2}(\mathbb{F}_n(x) - F(x)) \underset{d}{\rightarrow} N(0, F(x)(1 - F(x))) \quad \text{as} \quad n \to \infty; \tag{4}$$

where E denotes expected value, "var" denotes the variance, "$\underset{p}{\rightarrow}$" denotes convergence in probability, and "$\underset{d}{\rightarrow}$" denotes convergence in law or in distribution (*see* CONVERGENCE OF SEQUENCES OF RANDOM VARIABLES).

The property of $\mathbb{F}_n$ that concerns us here strengthens (3) in two important ways: to uniform convergence (in x), and to convergence with probability 1 (w.p. 1) or almost sure convergence.

Theorem 1 [1,8].

$$P\left(\lim_{n\to\infty} \sup_{-\infty<x<\infty} |\mathbb{F}_n(x) - F(x)| = 0\right) = 1,$$

or, equivalently,

$$\lim_{n\to\infty} ||\mathbb{F}_n - F|| \equiv \lim_{n\to\infty} \sup_x |\mathbb{F}_n(x) - F(x)| = 0 \quad \text{w.p. 1}.$$

Theorem 1 was proved by Glivenko [8] for continuous distributions F, and by Cantelli [1] for general F (see, e.g., Loève [13] for a proof). It asserts that the empirical distribution function $\mathbb{F}_n$ estimates F to any desired degree of precision uniformly in x for sufficiently large sample size n. The true distribution function F can be "rediscovered from the data"; or the empirical distribution function $\mathbb{F}_n$ "looks like" the true distribution F for large n. The Glivenko–Cantelli theorem has been called the "central statistical theorem" by Loève [13] and the "fundamental statistical theorem" by Renyi [15].

The Glivenko–Cantelli theorem is of constant use in establishing the *consistency* of many different statistical tests and estimates. Two examples illustrate these types of applications.

Example 1. Consistency of the Kolmogorov Test. Consider testing the simple null hypotheses $H_0 : F = F_0$, where F_0 is completely specified. Kolmogorov [11] suggested that H_0 be rejected when

$$D_n \equiv \sup_x |\mathbb{F}_n(x) - F_0(x)| \equiv ||\mathbb{F}_n - F_0||$$

is large; *see* KOLMOGOROV–SMIRNOV-TYPE TESTS OF FIT. When F_0 is the true distribution function, the Glivenko–Cantelli theorem asserts that

$$P_{F_0}\left(\lim_{n\to\infty} D_n = 0\right) = 1. \tag{5}$$

Kolmogorov [11] showed in fact that the distribution of D_n does not depend on F_0 if F_0 is continuous, and that

$$\lim_{n\to\infty} P_{F_0}(n^{1/2} D_n \geqslant \lambda) = 2\sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2\lambda^2) \equiv K(\lambda)$$

for all $0 \leqslant \lambda < \infty$. Thus if $K(\lambda_\alpha) = \alpha$ and $P_{F_0}(n^{1/2} D_n \geqslant \lambda_{n,\alpha}) \equiv \alpha$, $0 < \alpha < 1$, then

$$\lim_{n\to\infty} \lambda_{n,\alpha} = \lambda_\alpha. \tag{6}$$

If, however, some $F \neq F_0$ is the true distribution function, the Glivenko–Cantelli theorem implies that

$$P_F\left(\lim_{n\to\infty} D_n = d\right) = 1, \quad (7)$$

where $d \equiv \sup_x |F(x) - F_0(x)| = ||F - F_0|| > 0$. Hence when $F \neq F_0$ is true, (6) and (7) imply that

$$\lim_{n\to\infty} P_F(n^{1/2} D_n \geqslant \lambda_{n,\alpha}) = 1. \quad (8)$$

In other words, the probability of rejecting the null hypothesis $F = F_0$ when $F \neq F_0$ is the true distribution increases to 1 as the sample size becomes large. The Kolmogorov test is *consistent*.

Example 2. Consistency of the Mann–Whitney Estimator of $P(X \leqslant Y)$ in a Two-Sample Problem. Suppose that $Y_1, Y_2, \ldots$ are i.i.d. with common distribution function G, independent of the X's above, and let $\mathbb{G}_n$ denote the empirical distribution function of the first nY's. Consider estimating $P(X \leqslant Y) = \int F dG$ based on the first mX's and first nY's. The Mann–Whitney estimator of this probability is

$$W_{mn} \equiv \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1_{[X_i \leqslant Y_j]} = \int \mathbb{F}_m d\mathbb{G}_n.$$

To show that $W_{mn} \to P(X \leqslant Y) = \int F dG$ w.p. 1, add and subtract $\int F d\mathbb{G}_n$ and integrate the second term by parts to obtain

$$\begin{aligned}
&\left|\int \mathbb{F}_m d\mathbb{G}_n - \int F dG\right| \\
&\quad = \left|\int (\mathbb{F}_m - F) d\mathbb{G}_n + \int F d(\mathbb{G}_n - G)\right| \\
&\quad \leqslant ||\mathbb{F}_m - F|| + \left|\int (\mathbb{G}_n - G) dF\right| \\
&\quad \leqslant ||\mathbb{F}_m - F|| + ||\mathbb{G}_n - G|| \\
&\quad \to 0 + 0 = 0 \qquad \text{w.p. 1}
\end{aligned}$$

as $m \to \infty, n \to \infty$, by the Glivenko–Cantelli theorem. Thus W_{mn} is a (strongly) consistent estimator of $P(X \leqslant Y)$. (*See also* MANN–WHITNEY–WILCOXON STATISTIC for further information concerning W_{mn}. Another proof of the consistency of W_{mn} is based on the fact that W_{mn} is a *U-statistic** and hence a reverse martingale*.)

Before leaving the classical case, two important related results should be mentioned: an exponential inequality for the random variable $||\mathbb{F}_n - F||$, and a law of the iterated logarithm*.

The *inequality* of Dvoretzky et al. [4] asserts that

$$P(||\mathbb{F}_n - F|| \geqslant \lambda) \leqslant C \exp(-2n\lambda^2) \quad (9)$$

for all $\lambda > 0$ where C is an absolute constant. [$C = 58$ works; the smallest C for which (9) holds is still unknown.] The factor of 2 appearing in this inequality is best possible; note that the lead term in the distribution $K(\lambda)$ is $2\exp(-2\lambda^2)$. For example,

$$P(||\mathbb{F}_n - F|| \geqslant 0.04) \leqslant 0.10$$

if $n \geqslant \frac{1}{2} \cdot 625 \cdot \log(580) \cong 1989$.

The *iterated logarithm law* of Smirnov [17] and Chung [2] gives a rate of convergence for the Glivenko–Cantelli theorem: it asserts that

$$\begin{aligned}
&\limsup_{n\to\infty} \frac{n^{1/2}||\mathbb{F}_n - F||}{(2\log\log n)^{1/2}} \\
&\quad = \sup_x [F(x)\{1 - F(x)\}]^{1/2} \leqslant \tfrac{1}{2} \quad \text{w.p. 1.} \quad (10)
\end{aligned}$$

Thus

$$||\mathbb{F}_n - F|| = O(n^{-1/2}(\log\log n)^{1/2}) \quad \text{w.p. 1;}$$

the supremum distance between $\mathbb{F}_n$ and F goes to zero only a little more slowly than $n^{-1/2}$ w.p. 1.

Since 1960 the Glivenko–Cantelli theorem has been extended and generalized in several directions: to random vectors and to observations X with values in more general metric spaces; to empirical probability measures indexed by families of sets; to observations that may be dependent or nonidentically distributed; and to metrics other than the supremum metric. Here we briefly summarize some of this work. More detailed information and further references can be found in the survey by Gaenssler and Stute [7].

Let $X_1, X_2, \ldots$ be i.i.d. random variables with values in a (measurable) space $(\mathbb{X}, \mathscr{B})$ and common probability measure P on $\mathbb{X}$;

for many important applications in statistics $(\mathbb{X}, \mathscr{B}) = (R^k, \mathscr{B}^k)$, k-dimensional Euclidean space with its usual Borel sigma field. The *empirical measure* $\mathbb{P}_n$ of the first nX's is the probability measure that puts mass $1/n$ at each of $X_1, \ldots, X_n$:

$$\mathbb{P}_n = (\delta_{X_1} + \cdots + \delta_{X_n})/n, \tag{11}$$

where $\delta_x(A) = 1$ if $x \in A$; 0 if $x \notin A$, for $A \in \mathscr{B}$.

Many of the generalizations referred to above assert that, in some sense, "$\mathbb{P}_n$ looks like P" for large n. It has become common practice to refer to any such theorem as a "Glivenko–Cantelli theorem."

For $(\mathbb{X}, d)$ a separable metric space, the convergence of $\mathbb{P}_n$ to P was first investigated by Fortet and Mourier [6] and Varadarajan [22], who proved that $\beta(\mathbb{P}_n, P) \to 0$ w.p. 1, where β is the dual-bounded-Lipschitz metric (see Dudley [3]) and $\mathbb{P}_n \to P$ weakly w.p. 1, respectively.

Let $\mathscr{C} \subset \mathscr{B}$ be some specified subclass of sets and set

$$D_n(\mathscr{C}, P) = \sup_{C \in \mathscr{C}} |\mathbb{P}_n(C) - P(C)|. \tag{12}$$

A number of results assert that $D_n(\mathscr{C}, P) \to 0$ w.p. 1 for specific spaces $\mathbb{X}$ and classes of sets $\mathscr{C}$. For example, when $\mathbb{X} = \mathbb{R}^k$ and $\mathscr{C}$ = all intervals in $\mathbb{R}^k$, or all half-spaces in $\mathbb{R}^k$, or all closed balls in $\mathbb{R}^k$, then $D_n(\mathscr{C}, P) \to 0$ for any probability measure P [5,6,10]. For a general class of sets $\mathscr{C}$, however, some restriction on P may be necessary: If $\mathbb{X} = \mathbb{R}^k$ and $\mathscr{C}$ = all convex sets in $\mathbb{R}^k$, then $D_n(\mathscr{C}, P) \to 0$ w.p. 1 if $P_c(\partial C) = 0$ for all $C \in \mathscr{C}$ where P_c is the nonatomic part of P [14]. For a discussion of more results of this type and further references, see Gaenssler and Stute [7].

In the just stated results the classes $\mathscr{C}$ were formed by subsets of $\mathbb{R}^k$ which have a common geometric structure; the methods of proof of the corresponding Glivenko–Cantelli theorems rely heavily on this fact. For arbitrary sample spaces $(\mathbb{X}, \mathscr{B})$ where geometrical arguments are not available, the most appealing approach to obtain Glivenko–Cantelli theorems for classes $\mathscr{C} \subset \mathscr{B}$ was given by Vapnik and Chervonenkis [21]. Based on combinatorial arguments they showed that given a class $\mathscr{C} \subset \mathscr{B}$ such that for some finite n, "$\mathscr{C}$ does not cut all subsets of any $E \subset \mathbb{X}$ with $\text{card}(E) = n$" [i.e., for any $E \subset \mathbb{X}$ with $\text{card}(E) = n$ there is a subset of E which is not of the form $E \cap C$ for some $C \in \mathscr{C}$], then (under some measurability assumptions) $D_n(\mathscr{C}, P) \to 0$ w.p. 1 for any probability measure P.

Dependent Observations. When $\mathbb{X} = \mathbb{R}^1$, $\mathscr{C} = \{(-\infty, x] : x \in \mathbb{R}^1\}$, and

$$\mathbb{F}_n(x) = \mathbb{P}_n(-\infty, x],$$

Tucker [20] generalized the classical Glivenko–Cantelli theorem to *strictly stationary** sequences:

$$||\mathbb{F}_n - F_\omega|| \to 0 \qquad \text{w.p. 1}, \tag{13}$$

where F_ω is a (possibly random) distribution function; when the X's are also *ergodic**, F_ω is simply the common one-dimensional marginal law of the X's. Tucker's Glivenko–Cantelli theorem applies to sequences of random variables satisfying a wide range of *mixing conditions*; it has been generalized to higher-dimensional spaces and more general index sets by Stute and Schumann [19] (see also Steele [18] and Kazakos and Gray [9]).

Nonidentically Distributed Observations. If the X's are independent but not identically distributed, there is no common probability measure P to be recovered from the data. Nevertheless, letting P_i denote the probability law of X_i, $i = 1, 2, \ldots$, we still have

$$\begin{aligned} E\mathbb{P}_n(C) &= n^{-1}(P_1 + \cdots + P_n)(C) \\ &\equiv \overline{P}_n(C). \end{aligned}$$

Thus it is still reasonable to expect that the empirical measure $\mathbb{P}_n$ "looks like" the average measure $\overline{P}_n$. When $\mathbb{X} = \mathbb{R}^1$, $\mathscr{C} = \{(-\infty, x]; x \in \mathbb{R}^1\}$, $\mathbb{F}_n(x) = \mathbb{P}_n(-\infty, x]$, and $\overline{F}_n(x) = \overline{P}_n(-\infty, x]$, Koul [12] and Shorack [16] have shown that

$$||\mathbb{F}_n - \overline{F}_n|| \equiv \sup_x |\mathbb{F}_n(x) - \overline{F}_n(x)| \to 0 \quad \text{w.p. 1}$$

always. When $(\mathbb{X}, d)$ is a separable metric space, Wellner [23] has shown that if $\{\overline{P}_n\}$ is tight, then $\beta(\mathbb{P}_n, \overline{P}_n) \to 0$ and $\rho(\mathbb{P}_n, \overline{P}_n) \to 0$ w.p. 1, where β and ρ are the dual-bounded Lipschitz and Prohorov metrics, respectively.

REFERENCES

1. Cantelli, F. P. (1933). *G. Ist. Ital. Attuari*, **4**, 421–424. (One of the original works; Glivenko's result for continuous distribution functions is extended to arbitrary df's.)
2. Chung, K. L. (1949). *Trans. Amer. Math. Soc.*, **67**, 36–50. [Contains a proof of the law of the iterated logarithm (10) for $||\mathbb{F}_n - F||$.]
3. Dudley, R. M. (1969). *Ann. Math. Statist.*, **40**, 40–50. (Bounds for expected Prohorov and dualbounded-Lipschitz distances between the empirical measure $\mathbb{P}_n$ and true measure P are given using metric entropy methods.)
4. Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). *Ann. Math. Statist.*, **27**, 642–669. [The exponential bound (9) is proved and used in a study of the asymptotic minimax properties of the empirical distribution function $\mathbb{F}_n$ as an estimator of F.]
5. Elker, J., Pollard, D., and Stute, W. (1979). *Adv. Appl. Prob.*, **11**, 820–833. (Contains Glivenko–Cantelli theorems for the empirical measure $\mathbb{P}_n$ indexed by convex sets in k-dimensional Euclidean space.)
6. Fortet, R. M. and Mourier, E. (1953). *Ann. Sci. École Norm. Sup.*, **70**, 266–285. (Convergence of the dual-bounded-Lipschitz distance from the empirical measure $\mathbb{P}_n$ to the true measure P is established.)
7. Gaenssler, P. and Stute, W. (1979). *Ann. Prob.*, **7**, 193–243. (A survey of results for empirical distribution functions and empirical processes with an extensive bibliography.)
8. Glivenko, V. (1933). *G. Ist. Ital. Attuari*, **4**, 92–99. (Here Theorem 1 was first established for continuous distribution functions F.)
9. Kazakos, P. P. and Gray, R. M. (1979). *Ann. Prob.*, **7**, 989–1002. (Glivenko–Cantelli theorems for finite-dimensional distributions of stationary processes.)
10. Kiefer, J. and Wolfowitz, J. (1958). *Trans. Amer. Math. Soc.*, **87**, 173–186. [The Glivenko–Cantelli Theorem 1 and the law of the iterated logarithm (10) are extended to empirical distribution functions of random vectors in k-dimensional Euclidean space.]
11. Kolmogorov, A. N. (1933). *G. Ist. Ital. Attuari*, **4**, 83–91. (One of the original and most important papers.)
12. Koul, H. L. (1970). *Ann. Math. Statist.*, **41**, 1768–1773. (Contains a Glivenko–Cantelli theorem for the case of independent nonidentically distributed random variables.)
13. Loève, M. (1977). *Probability Theory*, 4th ed. Springer, New York. (Graduate-level textbook; excellent reference for basic probability theory.)
14. Ranga Rao, R. (1962). *Ann. Math. Statist.*, **33**, 659–680.
15. Renyi, A. (1962). *Wahrscheinlichkeitsrechnung*, VEB, Deutscher Verlag der Wissenschaften, Berlin.
16. Shorack, G. R. (1979). *Statist. Neerlandica*, **33**, 169–189. (The Glivenko–Cantelli theorem for independent but nonidentically distributed random variables is proved, and weak convergence of empirical processes of such variables studied.)
17. Smirnov, N. V. (1944). *Uspehi Mat. Nauk*, **10**, 179–206. [The law of the iterated logarithm (10) for $||\mathbb{F}_n - F||$ is proved.]
18. Steele, M. (1978). *Ann. Prob.*, **6**, 118–127. (General Glivenko–Cantelli theorems for empirical measures indexed by sets are obtained by using the combinatorial methods of Vapnik and Chervonenkis in combination with subadditive ergodic theory.)
19. Stute, W. and Schumann, G. (1980). *Scand. J. Statist.*, **7**, 102–104. (Generalization of the result of Tucker [20] for stationary processes to empirical measures indexed by sets.)
20. Tucker, H. G. (1959). *Ann. Math. Statist.*, **30**, 828–830. (Here the Glivenko–Cantelli theorem is proved for strictly stationary processes.)
21. Vapnik, V. N. and Chervonenkis, A. Ya. (1971). *Theory Prob. Appl.*, **16**, 264–280. (Combinatorial methods are introduced and used to prove Glivenko–Cantelli theorems for general sample spaces and classes of index sets.)
22. Varadarajan, V. S. (1958). *Sankhyā*, **19**, 23–26. (Classical paper showing that the empirical measure $\mathbb{P}_n$ converges weakly to P with probability 1 when the sample space is a separable metric space.)
23. Wellner, J. A. (1981). *Stoch. Proc. Appl.*, **11**, 309–312.

See also CONVERGENCE OF SEQUENCES OF RANDOM VARIABLES; EMPIRICAL DISTRIBUTION FUNCTION (EDF) STATISTICS; LAW OF THE ITERATED LOGARITHM; and LAWS OF LARGE NUMBERS.

PETER GAENSSLER
JON A. WELLNER