Chapter 5. Consistency and Limiting Distributions

Note. The ETSU *Graduate Catalog* describes Mathematical Statistics 1 (STAT 4047/5047) as including probability distributions, random variables, distributions, and the Central Limit Theorem. The topics of confidence intervals and hypothesis testing (covered in Chapter 4) are to be covered in Mathematical Statistics 2 (STAT 4057/5057); Hogg, McKean, and Craig mention in the Preface that "the instructor would have the option of interchanging the order of Chapters 4 and 5," and we follow that plan here. In Section 5.1 we very quickly review the topics of Chapter 4. In this chapter, we introduce two kinds of convergence related to sequences of random variables: convergence in probability and convergence in distribution. These ideas will justify our use of the normal distribution in, for example, confidence intervals, as we'll see in the Central Limit Theorem (Theorem 5.3.1).

Section 5.1. Convergence in Probability

Note. We saw in Example 3.1.4 that the relative frequency of a success (in a sequence of repeated Bernoulli trials) approaches the probability p of a success (as the trials are performed more and more times) and this limit holds with probability 1. This is an example of convergence in probability. We'll formally define this idea, rove the Weak Law of large Numbers (in Theorem, 5.1.1), and prove some properties of convergence in probability. We also visit/revisit some ideas from Chapter 4 concerning sampling, statistics, and estimates.

Definition 5.1.1. Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. The sequence X_n *converges in probability* to X is, for all $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0 \text{ or equivalently } \lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1.$$

When this holds, we denote it as $X_n \xrightarrow{P} X$.

Note. Often in statistics, random variable X is a constant $X = a$, in which case we write $X \xrightarrow{P} a$. This is the case in Example 3.1.4 mentioned above in which case $X = p$.

Note. A subtlety of convergence in probability is that it deals with convergence of X_n to X where we measure distance using the function P. In Real Analysis 1 (MATH 5210), we consider the measure of sets of real numbers and we find that a set of measure 0 need not be empty (in fact, it need not be finite or even countable). A property (of a function, say) is said to hold "almost everywhere" if it holds except on a set of measure 0. Similarly, when $X_n \xrightarrow{P} X$ it is common to say that $X_n \to X$ "almost surely" (that is, with probability 1; see my online notes for Measure Theory Based Probability [not a formal ETSU class] on 4.6. Random Variables). If we are dealing with infinite probability spaces (in particular, if we have continuous random variables) then events of probability 0 are not necessarily *impossible* (i.e., such events need not be empty sets). These ideas are addressed in Graph Theory 2 (MATH 5450) where the concept of almost surely in a finite probability space is used to prove the existence of certain properties of graphs

(see my online notes for Graph Theory 2 on 13.3 Variance; notice Note 13.1.A). In Graph Theory 2, as here, Chebychev's Inequality (Theorem 1.10.3) is used to prove convergence in probability.

Theorem 5.1.1. Weak Law of Large Numbers.
Let $\{X_n\}$ be a sequence of independent and identically distributed ("iid") random variables having common mean $\mu < \infty$ and variance $\sigma^2 < \infty$. Let $\overline{X}_n = \left(\sum_{i=1}^{n} X_i\right)/n$ (this is the *sample mean*). Then $\overline{X}_n \xrightarrow{P} \mu$.

Note. Dealing with the Strong Law of Large Numbers is "beyond the level of this course!" It is normally addressed in a probability theory course which requires a background in Lebesgue measure and integration (covered in ETSU's Real Analysis 1 [MATH 5210]), abstract measure and integration (covered in ETSU's Real Analysis 2 (MATH 5220), and functional analysis (covered in ETSU's Fundamentals of Functional Analysis [MATH 5740]). I have some preliminary notes posted for Measure Theory Based Probability. Two statements of the Strong Law of Large Numbers are given in this course (here based on the statement and numbering scheme of *Probability and Measure Theory* 2nd Edition, by Robert B. Ash with contributions from Catherine Doleans-Dade, Academic Press, 2000), as follows:

Theorem 6.2.2. Kolmogorov Strong Law of Large Numbers.
Let $X_1, X_2, \ldots$ be independent random variables, each with finite mean and variance, and let $\{b_n\}$ be an increasing sequence of positive real numbers with $b_n \to \infty$. If

$$\sum_{n=1}^{\infty} \frac{\text{Var}(X_n)}{b_n^2} < \infty \text{ and } S_n = X_1 + X_2 + \cdots + X_n,$$

then we have $\dfrac{S_n - E(S_n)}{b_n} \to 0$ almost everywhere.

Theorem 6.8.3. Strong Law of Large Numbers, iid Case.

If $X_1, X_2, \ldots$ are independent and identically distributed ("iid") random variables with finite expectation m, and $S_n = X_1 + X_2 + \cdots + X_n$, then $S_n/n \to m$ almost everywhere and in L^1.

In this setting, events are measurable sets and random variables are measurable functions. The term "almost everywhere" was mentioned above. The space L^1 is a Banach space (a complete normed linear space) and the claim in the second version of the Strong Law means that $S_n/n \to m$ with respect to the L^1 norm. The proof the Kolmogorov Strong Law is based (in part) on Kolmogorov's Inequality (a generalization of Chebychev's Inequality). The proof of the Strong Law, iid Case (as stated) is based on Martingale theory, though another proof is given which uses the Borel-Cantelli Lemma, the Lebesgue Dominated Convergence Theorem, and Kronecker's Lemma. A proof along these lines is presented in N. Etemadi's "An Elementary Proof of the Strong Law of Large Numbers" *Zeitschrift für Wahrscheinlichkeitstheorie* **55**, 119–122 (1981), a copy of which is online at Yao Li's webpage at the University of Massachusetts, Amherst (accessed 1/9/2021). We only need the Weak Law of Large Numbers, so we leave the Strong Law for another class.

Theorem 5.1.2. Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n + Y_n \xrightarrow{P} X + Y$.

Theorem 5.1.3. Suppose $X_n \xrightarrow{P} X$ and a is a constant. Then $aX_n \xrightarrow{P} aX$.

Note. We can combine the previous two result to deduce a "linearity" behavior of convergence in probability (under the obvious hypotheses): $aX_n + bY_n \xrightarrow{P} aX + bY$.

Theorem 5.1.4. Suppose $X_n \xrightarrow{P} a$ and the real function g is continuous at a. Then $g(X_n) \xrightarrow{P} g(a)$.

Note. We can deduce several things from Theorem 5.1.4, in particular we have for $X_n \xrightarrow{P} a$ (where a is constant) that $X_n^2 \xrightarrow{P} a^2$, $1/X_n \xrightarrow{P} 1/a$ for $a \neq 0$, and $\sqrt{X_n} \xrightarrow{P} \sqrt{a}$ for $a \geq 0$ (or, arguably, $a \geq 0$). Hogg, McKean, and Craig mention that if $X_n \xrightarrow{P} X$ and g is a continuous function then $g(X_n) \xrightarrow{P} g(X)$; they give as a reference page 104 of H. G. Tucker's *A Graduate Course in Probability*, New York: Academic Press (1967). For our purposes, we really just need the following special case of this result.

Theorem 5.1.A. Suppose $X_n \xrightarrow{P} X$. Then $X_n^2 \xrightarrow{P} X^2$.

Theorem 5.1.5. Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n Y_n \xrightarrow{P} XY$.

Note. We now turn to some of the topics from Chapter 4, "Some Elementary Statistical Inference." For a setting, we consider a random variable X with probability density (or mass) function of the form $f(x;\theta)$ for an unknown parameter $\theta \in \Omega$. For example, X might have a normal distribution, but a normal distribution has the parameters μ and σ. In this case, Ω is the set $\Omega = \{\theta - (\mu, \sigma^2) \mid \mu \in \mathbb{R}, \sigma > 0\}$ and function f gives a normal distribution and then "parameter" θ determines the specific normal distribution. The plan is to estimate θ based on a sample.

Definition. Let random variable V have probability density (or mass) function $f(x;\theta)$ where $\theta \in \Omega$. A *random sample* $X_1, X_2, \ldots, X_n$ from the distribution of X is a collection of n independent and identically distributed ("iid") random variables $X_1, X_2, \ldots, X_n$ with the common probability density/mass function $f(x;\theta)$. A function $T = T(X_1, X_2, \ldots, X_n)$ of the sample is a *statistics*. When the sample yields observed values $x_1, x_2, \ldots, x_n$, these are the *realized values* of the sample and the realized statistic $t = t(x_1, x_2, \ldots, x_n)$ is a *point estimate* of θ (that is, the sample is used to estimate the unknown parameter θ which determines $f(x;\theta)$. The *point estimator* T for θ is *unbiased* if $E[T] = \theta$.

Note. In Theorem 2.8.A, we saw that $\overline{X}$ and S^2 are unbiased estimators of μ and σ, respectively.

Definition 5.1.2. Let X be a random variable with cumulative distribution function $F(x,\theta)$ where $\theta \in \Omega$. Let $X_1, X_2, \ldots, X_n$ be a sample from the distribution of X and let T_n denote a statistic. Then T_n is a *consistent* estimator of θ if $T_n \xrightarrow{P} \theta$.

Note 5.1.A. The Weak Law of Large Numbers (Theorem 5.1.1) shows that, for $X_1, X_2, \ldots, X_n$ a random sample from a distribution with finite mean μ and finite variance σ^2, the sample mean $\overline{X}_n$ is a consistent estimator of μ. We now state the book's Example 5.1.1 as a theorem.

Theorem 5.1.B. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution of X with finite mean μ and finite variance σ^2 where $E[X^4]$ is finite, then the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

(where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$) is a consistent estimator of σ^2.

Note. We close this section by quoting the last paragraph in this section from the text (see page 326):

> "Consistency is a very important property for an estimator to have. It is a poor estimator that does not approach its target as the sample size gets large. Note that the same cannot be said for the property of unbiasedness. For example, instead of using the sample variance to estimate σ^2, suppose we use $V = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then V is consistent for σ^2, but it is biased because $E[V] = (n-1)\sigma^2/n$. Thus the bias of V is $-\sigma^2/n$, which vanishes as $n \to \infty$."