POL 571: Convergence of Random Variables

Kosuke Imai
Department of Politics, Princeton University

March 28, 2006

1 Random Sample and Statistics

So far we have learned about various random variables and their distributions. These concepts are, of course, all mathematical *models* rather than the real world itself. In practice, we do not know the true models of human behavior, and they may not even correspond to probability models. George Box once said that there is no true model, but there are useful models. Even if there is such a thing as "the true probability model," we can never observe it! Therefore, we must connect what we can observe with our theoretical models. The key idea here is that we use the probability model (i.e., a random variable and its distribution) to describe the *data generating process.* What we observe, then, is a particular realization (or a set of realizations) of this random variable. The goal of *statistical inference* is to figure out the true probability model given the data you have.

Definition 1 *Random variables $X_1, X_2, \ldots, X_n$ are said to be independent and identically distributed (or i.i.d.) if they are independent and share the same distribution function $F(x)$. It is also called a (an i.i.d.) random sample of size n from the population, $F(x)$.*

If we use $f(x)$ to denote the probability density (or mass) function associated with $F(x)$, then the joint probability density (or mass) function given a particular set of realizations $(x_1, x_2, \ldots, x_n)$ is given by $\prod_{i=1}^n f(x_i)$. Of course, if the random variables are not i.i.d., then the joint density (or mass) function will be much more complicated, $f(x_n \mid x_{n-1}, \ldots, x_1) \cdots f(x_2 \mid x_1) f(x_1)$.

The above definition is an example of what is sometimes called an *infinite (or super) population model* because in theory one can obtain the infinite number of random sample from the population (hence, the population size is infinite). In the social sciences, this often requires one to think about a hypothetical population from which a particular realization is drawn. For example, what does it mean to say, "the outcome of the 2000 presidential election is a particular realization from the population model"?

Another important framework is a *finite population model* where we consider the population size to be finite. This might be appropriate in a survey sampling context where a sample of respondents is drawn from the particular population, which is of course finite. The concept will also play a role in analyzing randomized experiments as well as the statistical method called bootstrap. An example of this model is given next,

Definition 2 *Consider a population of size N, $\{x_1, x_2, \ldots, x_N\}$, with $N \in \mathbf{N}$. Random variables $X_1, X_2, \ldots, X_n$ are called a simple random sample if these units are sampled with equal probability and without replacement.*

Note that $P(X_i = x) = 1/N$ for all $i = 1, 2, \ldots, n$ if x is a distinct elements of $\{x_1, x_2, \ldots, x_N\}$. This implies that the marginal distribution of X_i is the same as the case of sampling with replacement.

However, a simple random sample is no longer independent because the conditional distribution of X_2 given X_1, for example, depends on the observed value of X_1. Of course, this is one of the simplest probability sampling methods, and there are more sophisticated sampling methods available.

Given a random sample, we can define a statistic,

Definition 3 *Let $X_1, \ldots, X_n$ be a random sample of size n from a population, and Ω be the sample space of these random variables. If $T(x_1, \ldots, x_n)$ is a function where Ω is a subset of the domain of this function, then $Y = T(X_1, \ldots, X_n)$ is called a statistic, and the distribution of Y is called the sampling distribution of Y.*

What you should take away from this definition is that a statistic is simply a function of the data and that since your data set is a random sample from a population, a statistic is also a random variable and has its own distribution. Three common statistics are given below,

Definition 4 *Let $X_1, \ldots, X_n$ be a random sample from a population. Then,*

1. *The sample mean is defined by $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.*
2. *The sample variance is defined by $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ where $S = \sqrt{S^2}$ is called the sample standard deviation.*

These statistics are good “guesses” of their population counterparts as the following theorem demonstrates.

Theorem 1 (Unbiasedness of Sample Mean and Variance) *Let $X_1, \ldots, X_n$ be an i.i.d. random sample from a population with mean $\mu < \infty$ and variance $\sigma^2 < \infty$. If $\overline{X}$ is the sample mean and S^2 is the sample variance, then*

1. *$E(\overline{X}) = \mu$, and $\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n}$.*
2. *$E(S^2) = \sigma^2$*

The theorem says that *on average* the sample mean and variances are equal to their population counterparts. That is, over repeated samples, you will get the answer right on average. This property is called *unbiasedness*. But, of course, typically we only have one random sample and so the answer you get from a particular sample you have may or may not be close to the truth. For example, each X_i is also an unbiased estimator of μ although sample mean is perhaps a better estimator because the variance is smaller. We will revisit this issue later in the course. This theorem can be also generalized to any function $g(X_i)$ provided that $E[g(X)]$ and $\mathrm{var}[g(X)]$ exist. You should be able to show $E[\sum_{i=1}^n g(X_i)/n] = E[g(X)]$ and $\mathrm{var}[\sum_{i=1}^n g(X_i)]/n] = \mathrm{var}[g(X)]/n$.

There are several useful properties of the sample mean and variance, we use later in the course, when the population distribution is normal.

Theorem 2 (Sample Mean and Variance of Normal Random Variables) *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Normal distribution with mean μ and variance σ^2. Let $\overline{X}$ and S^2 be the sample mean and variance, respectively. Then,*

1. *$\overline{X} \sim N(\mu, \sigma^2/n)$.*
2. *$(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.*
3. *$\overline{X}$ and S^2 are independent.*

4. $\sqrt{n}(\overline{X} - \mu)/S \sim t_{n-1}$.

What is important about the last result of this theorem is that the distribution of the statistic, $\sqrt{n}(\overline{X} - \mu)/S$ does not depend on the variance of X. That is, regardless of the value of σ^2, the exact distribution of the statistic is t_1. We also consider the distribution of the ratio of two sample variances.

Theorem 3 (Ratio of the Sample Variances) *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from the Normal distribution with mean μ_X and variance σ_X^2. Similarly, let $Y_1, Y_2, \ldots, Y_m$ be an i.i.d. sample from the Normal distribution with mean μ_Y and variance σ_Y^2. If S_X^2 and S_Y^2 are the sample variances, then the statistic,*

$$\frac{S_X^2/S_Y^2}{\sigma_X^2/\sigma_Y^2} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2},$$

is distributed as the F distribution with $n-1$ and $m-1$ degrees of freedom.

Finally, we give another class of statistics, which is a bit more complicated than the sample mean and variance.

Definition 5 *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from a population. The order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ can be obtained by arranging this random sample in non-decreasing order, $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$ where $(1), (2), \ldots, (n)$ is a (random) permutation of $1, 2, \ldots, n$. In particular, we define the sample median as $X_{((n+1)/2)}$ if n is odd and $(X_{(n/2)} + X_{(n/2+1)})/2$ if n is even.*

We will see later in the course that the sample median is less affected by extreme observations than the sample mean. Here, we consider the marginal distribution of the order statistics.

Theorem 4 (Order Statistics) *Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ be the order statistics from an i.i.d. random sample from a population.*

1. *If the population distribution is discrete with the probability mass function $f_X(x)$ and $x_1 < x_2 < \cdots$ are possible values of X in ascending order, then*

$$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) \left[q_i^k (1-q_i)^{n-k} - q_{i-1}^k (1-q_{i-1})^{n-k} \right],$$

where $q_i = \sum_{k=1}^{i} P(X = x_k) = P(X \leq x_i)$ and $q_0 = 0$.

2. *If the population distribution is continuous with the probability density function $f_X(x)$ and the distribution function $F_X(x)$, then*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) [F_X(x)]^{j-1} [1 - F_X(x)]^{n-j}.$$

Now, we can easily answer the following question,

Example 1 *Let $X_1, X_2, \ldots, X_n$ be an i.i.d. random sample from* Uniform$(0, 1)$. *What is the distribution of the jth order statistic?*

2 Convergence of Random Variables

The final topic of probability theory in this course is the convergence of random variables, which plays a key role in *asymptotic* statistical inference. We are interested in the behavior of a statistic as the sample size goes to infinity. That is, we ask the question of "what happens if we can collect the data of infinite size?" Of course, in practice, we never have a sample of infinite size. However, if we have a data set that is "large enough," then we might be able to use the large-sample result as a way to find a good approximation for the finite sample case. We consider four different modes of convergence for random variables,

Definition 6 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X be a random variable.*

1. *$\{X_n\}_{n=1}^{\infty}$ is said to converge to X* **in the rth mean** *where $r \geq 1$, if*
$$\lim_{n\to\infty} E(|X_n - X|^r) = 0.$$

2. *$\{X_n\}_{n=1}^{\infty}$ is said to converge to X* **almost surely**, *if*
$$P(\lim_{n\to\infty} X_n = X) = 1.$$

3. *$\{X_n\}_{n=1}^{\infty}$ is said to converge to X* **in probability**, *if for any $\epsilon > 0$,*
$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1.$$

4. *$\{X_n\}_{n=1}^{\infty}$ is said to converge to X* **in distribution**, *if at all points x where $P(X \leq x)$ is continuous,*
$$\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x).$$

Almost sure convergence is sometimes called *convergence with probability 1* (do not confuse this with convergence in probability). Some people also say that a random variable converges *almost everywhere* to indicate almost sure convergence. The notation $X_n \overset{a.s.}{\to} X$ is often used for almost sure convergence, while the common notation for convergence in probability is $X_n \overset{p}{\to} X$ or $\text{plim}_{n\to\infty} X_n = X$. Convergence in distribution and convergence in the rth mean are the easiest to distinguish from the other two. The former says that the distribution function of X_n converges to the distribution function of X as n goes to infinity. It is often written as $X_n \overset{d}{\to} X$. Convergence in the rth mean is also easy to understand. If $r = 1$, then it implies that the mean of X_n converges to the mean of X, i.e., *convergence in the first moment* defined as $\lim_{n\to\infty} E(X_n) = E(X)$ (Why?). We are also often interested in the case of $r = 2$, which is often called *mean square convergence*.

The distinction between almost sure convergence and convergence in probability is more subtle. To understand the former, recall the definition of a random variable; it is a function $X(\omega)$ which maps the sample space Ω to a real line. Almost sure convergence says that the probability of an event, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for $\omega \in \Omega$, is one. In other words, $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for all but some $\omega \in S \subset \Omega$ with $P(S) = 0$. Alternatively, one can require $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for *all* $\omega \in \Omega$ (i.e., pointwise convergence). However, such a requirement is not necessary because it does not involve any notion of probability.

Convergence in probability is weaker than almost sure convergence because $X_n(\omega)$ itself may not converge to $X(\omega)$ for some $\omega \in S \subset \Omega$ where $P(S) > 0$. In fact, even if $X_n(\omega)$ does not converge to $X(\omega)$ for *any* $\omega \in \Omega$, X_n may converge to X in probability, as the following well-known example illustrates,

Example 2 *Let X be a uniform random variable with the sample space $\Omega = [0,1]$. Define a sequence of random variable, $\{X_n\}_{n=1}^{\infty}$, as follows: $X_1(\omega) = \omega + 1_{[0,1]}(\omega), X_2(\omega) = \omega + 1_{[0,1/2]}(\omega), X_3(\omega) = \omega + 1_{[1/2,1]}(\omega), X_4(\omega) = \omega + 1_{[0,1/3]}(\omega), X_5(\omega) = \omega + 1_{[1/3,2/3]}(\omega), X_6(\omega) = \omega + 1_{[2/3,1]}(\omega)$. Does this sequence of random variables converge to X almost surely? What about convergences in mean and in distribution?*

In addition to the relationship between almost sure convergence and convergence in probability, we may also guess that probability in distribution is the weakest form of convergence among the ones listed here. Convergence in distribution only concerns about the distribution function, and it has no reference to the sample space, for example. The following example illustrates this fact,

Example 3 *Let X be a Bernoulli random variable with equal probability, $1/2$. Now, let a sequence of "identical" (but not independent) random variables, $X_1, X_2, \ldots$, where $X_n = X$ for all n. Does this sequence of random variables converge in distribution to $Y = 1 - X$? What about convergence in other modes?*

Finally, it is easy to show an example where a sequence of random variables converge in probability, but fails to converge in mean.

Example 4 *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables where X_n is defined by: $X_n = n^3$ with probability $1/n^2$ and $X_n = 0$ with probability $1 - 1/n^2$. Does $\{X_n\}_{n=1}^{\infty}$ converge in probability to 0? What about mean convergence?*

After building up the intuition, we prove the following results,

Theorem 5 (Convergence of Random Variables) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X be another random variable.*

1. *If $\{X_n\}_{n=1}^{\infty}$ converges in probability to X, then $\{X_n\}_{n=1}^{\infty}$ converges to X in distribution.*
2. *If $\{X_n\}_{n=1}^{\infty}$ converges in mean to X, then $\{X_n\}_{n=1}^{\infty}$ also converges to X in probability.*
3. *If $\{X_n\}_{n=1}^{\infty}$ converges almost surely to X, then $\{X_n\}_{n=1}^{\infty}$ also converges to X in probability.*

Note that the converse is not generally true and that there is no general ordering between almost sure convergence and convergence in mean. An exception is that If $\{X_n\}_{n=1}^{\infty}$ converges in distribution to $c \in \mathbf{R}$ where c is a constant, then $\{X_n\}_{n=1}^{\infty}$ also converges to c in probability. Before we consider the applications of these convergence concepts, we collect some important results about the convergence of random variables.

Theorem 6 (Slutzky's Theorem) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables which converges in distribution to X. Let $\{Y_n\}_{n=1}^{\infty}$ be a sequence of random variables which converges in probability to c where $c \in \mathbf{R}$ is a constant.*

1. *The sequence, $\{X_nY_n\}_{n=1}^{\infty}$, converges in distribution to cX.*
2. *The sequence, $\{X_n + Y_n\}_{n=1}^{\infty}$, converges in distribution to $X + c$.*

An important special case of this theorem is that if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$, then $X_n + Y_n \xrightarrow{d} X$. In this case, we say that $Z_n = X_n + Y_n$ and X_n are *asymptotically equivalent* because $Z_n - X_n \xrightarrow{p} 0$.

The Slutzky's theorem as well as the convergence concepts can be readily extended to random vectors and random matrices. For example, let X_n be a sequence of random vectors and A_n be a sequence of random matrices. If $X_n \xrightarrow{d} X$ and $A_n \xrightarrow{p} A$ where X is a random vector and A is a constant matrix, then $A_n X_n \xrightarrow{d} AX$ provided that relevant vectors and matrices are conformable. In addition, if A_n and A are nonsingular, we have $X_n^\top A_n^1 X_n \xrightarrow{d} X^\top AX$. These are useful theorems, which we will often employ when learning about statistical theory.

Another important result is that convergence in distribution and in probability as well as almost sure convergence preserve continuous transformation. We skip the proof of the result concerning convergence in probability, which is beyond the scope of this class. To prove the result concerning convergence in distribution, we use the *Skorokhod's representation theorem*, which states that if $X_n \xrightarrow{d} X$, (although X_n may not converge to X in any other mode) there exist a sequence of random variable $\{Y_n\}_{n=1}^\infty$, which is distributed identically to X_n, and converge almost surely to a random variable Y, which is distributed identically to X.

Theorem 7 (Continuous Mapping Theorem) *Let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables, $f : \mathbf{R} \mapsto \mathbf{R}$ be a continuous function, and X be a random variable.*

1. *If $X_n \xrightarrow{a.s.} X$, then $f(X_n) \xrightarrow{a.s.} f(X)$.*
2. *If $X_n \xrightarrow{p} X$, then $f(X_n) \xrightarrow{p} f(X)$.*
3. *If $X_n \xrightarrow{d} X$, then $f(X_n) \xrightarrow{d} f(X)$.*

These results hold equivalently for a sequence of random vectors and matrices. An important special case here is that $X = c$ where $c \in \mathbf{R}$. Then, if $X_n \xrightarrow{a.s.} c$, then $f(X_n) \xrightarrow{a.s.} f(c)$. Similarly, if $X_n \xrightarrow{p} c$, then $f(X_n) \xrightarrow{p} f(c)$. Now answer the following question.

Example 5 *Show that if $X_n \xrightarrow{d} c$ and $f : \mathbf{R} \mapsto \mathbf{R}$ is a continuous function, then $f(X_n) \xrightarrow{p} f(c)$.*

Armed with good understanding of convergence, we prove two most important theorems in probability theory. First, consider a sequence of i.i.d. random variables $\{X_n\}_{n=1}^\infty$. One can also define a sequence of sample mean, $\{\bar{X}_n\}_{n=1}^\infty$, by $\bar{X}_n = \sum_{i=1}^n X_n/n$. The Strong Law of Large Numbers says that a sequence of sample mean converges to the true mean as n goes to infinity.

Theorem 8 (Strong Law of Large Numbers) *Let $\{X_n\}_{n=1}^\infty$ be a sequence of i.i.d. random variables with $\mu = E(X_i)$. Define a sequence, $\{\bar{X}_n\}_{n=1}^\infty$ where $\bar{X}_n = \sum_{i=1}^n X_n/n$. If $E(|X_i|) < \infty$, then $\bar{X}_n \xrightarrow{a.s.} \mu$.*

The proof of this theorem is beyond the scope of this course. Since almost sure convergence implies convergence in probability, $\bar{X}_n \xrightarrow{p} \mu$ also holds, which is called *the weak law of large numbers*. If we invoke an additional assumption that the variance is also finite, i.e., $\text{var}(X_i) < \infty$, then we can prove the weak law. This is not necessary for both strong and weak laws to hold, but makes the proof much easier. In statistics, if a sequence of statistics converge in probability to the population value as the sample size goes to infinity according to the weak law of large numbers, the statistic is called *consistent*.

Example 6 *What is a sufficient condition under which the sample variance $S_n^2 = \sum_{i=1}^n (X_i - \mu)^2/(n-1)$ is a consistent estimator of the population variance $\sigma^2 = \text{var}(X_i)$?*

Under this condition, the sample standard deviation S_n is a consistent estimator of the population standard deviation σ (Why?) However, S_n is not necessarily an unbiased estimator of σ (Why?) In this case the bias disappears as n goes to infinity.

The next theorem is one of the most amazing results in probability theory. Unfortunately, the proof of the theorem is beyond the scope of this course.

Theorem 9 (Central Limit Theorem) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with finite mean μ and finite non-zero variance σ^2. If we let $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

What is remarkable about this theorem is that the distribution of X_i does not matter so long as its mean and variance exist. The Central Limit theorem as well as Law of Large Numbers also holds for a sequence of random vectors, in which case the limiting distribution is the multivariate normal distribution. In statistical theory, if we regard $\bar{X}_n$ as an *estimator* and μ as an *estimand*, then $\bar{X}_n$ is called $\sqrt{n}$*-consistent* (because $\bar{X}_n \xrightarrow{p} \mu$ by the Weak Law of Large Numbers) and σ^2 is called *asymptotic variance*.

Example 7 *Pick your favorite random variable and apply the Central Limit Theorem.*

In statistics, we are interested in the limiting distribution of a function of random variable. The following theorem provides a way to calculate this. We prove a general case in terms of a random vector,

Theorem 10 (Delta Method) *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of k dimensional random vector such that $X_n \xrightarrow{p} \mu$ and $\sqrt{n}(X_n - \mu) \xrightarrow{d} X$. If $f : \mathbf{R}^k \to \mathbf{R}^r$ has continuous first derivative. Then,*

$$\sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} J(\mu)X,$$

where $J(\mu) = \frac{\partial f(\mu)}{\partial \mu^\top}$ is the $r \times k$ Jacobian matrix.

An important special case if where X is a normal random variable with the variance-covariance matrix Σ. In that case, we have $\sqrt{n}(f(X_n) - f(\mu)) \xrightarrow{d} N(0, J(\mu)\Sigma J(\mu)^\top)$.

Example 8 *What is the limiting distribution of the two variance parameters, σ_1^2 and σ_2^2, and the correlation ρ if the asymptotic distribution of $(\log(\sigma_1^2), \log(\sigma_2^2), 0.5\log[(1+\rho)/(1-\rho)])$ is $N(\mu, \Sigma)$?*

That's it. We are done with the probability theory!