1. Suppose that you flip a coin many times, with the outcome of each flip being a “head” or “tail”. Suppose that each outcome is equally likely.

(a) After many independent coin tosses, approximately what will be the proportion of heads? Why?

Let $X_i$ be the outcome of the $i^{th}$ toss; i.e., $X_i = 1$ if the $i^{th}$ toss is a head, while $X_i = 0$ otherwise. The long-run average value is the expected value in one experiment (toss), which is $(0 + 1)/2 = 1/2$. The sample average in $n$ experiments converges to the expected value by virtue of the law of large numbers.

(b) In 1,000,000 coin tosses, what is the approximate probability of getting more than 500,000 heads?

$1/2$ — This can be obtained from the normal approximation, which follows from the central limit theorem. It is obvious by symmetry. It is not exactly 1/2 because there is the middle case, getting exactly 500,000 heads.

(c) In 1,000,000 coin tosses, what is the approximate probability of getting more than 505,000 heads?

0 — This also can be obtained from the normal approximation, which follows from the central limit theorem. Note that 505,000 is 5000 above the mean. However, the variance of the total number of heads is $1,000,000/4$, so that the standard deviation is $1000/2 = 500$. Thus 5000 is 10 standard deviations above the mean.

(d) In 1,000,000 coin tosses, what is the approximate probability of getting more than 500,500 heads?

0.16 — This also can be obtained from the normal approximation, which follows from the central limit theorem. Note that 500,500 is 500 above the mean, which is 1 standard deviation. The probability of being at least one standard deviation above the mean is $1 - 0.8413 \approx 0.16$. Here are more details:

\[
P(S_{1,000,000} > 500,500) = P((S_{1,000,000} - E[S_{1,000,000}])/SD(S_{1,000,000}) > (500,500 - E[S_{1,000,000}])/SD(S_{1,000,000}))
\]
\[
= P((S_{1,000,000} - 500,000)/500 > (500,500 - 500,000)/500)
\]
\[ P(N(0, 1) > 1) \approx 1 - 0.8413. \]

You might remember that \( P(N(0, 1) > 1) \approx 0.16, P(N(0, 1) > 2) \approx 0.023 \) and \( P(N(0, 1) > 3) \approx 0.0013. \) Otherwise you can look in a table of the normal distribution.

2. The CLT (central limit theorem) needed for problem 1 above can be formalized as

\[ \frac{(S_n - nm)}{\sqrt{n\sigma^2}} \Rightarrow N(0, 1) \quad \text{as} \quad n \to \infty, \]

where

\[ S_n \equiv X_1 + \cdots + X_n, \quad n \geq 1. \]

(a) What are \( m \) and \( \sigma^2 \) here?

In the statement of the CLT above, \( m \) is the mean, \( EX \), and \( \sigma^2 \) is the variance, \( Var(X) \equiv E[X^2] - (EX)^2. \)

(b) What is \( N(a, b) \)?

\( N(a, b) \) means a random variable with the normal distribution having mean \( a \) and variance \( b \).

(c) What does \( \Rightarrow \) mean?

\( \Rightarrow \) means convergence in distribution.

We say that random variables \( X_n \) with cumulative distribution functions (cdf’s) \( F_n \) (i.e., \( F_n(x) \equiv P(X_n \leq x) \)) converge in distribution to a random variable \( X \) with cdf \( F \) as \( n \to \infty \) if \( F_n(x) \to F(x) \) or, equivalently, \( P(X_n \leq x) \to P(X \leq x) \), as \( n \to \infty \) for all \( x \) that are continuity points of the function \( F(x) \equiv P(X \leq x). \) For the normal distribution, all \( x \) are continuity points of \( F(x) \equiv P(N(a, b) \leq x). \)

A reference: My book Stochastic-Process Limits is mostly beyond the scope of this course, but a few sections may be helpful. The book is available online:

http://www.columbia.edu/ ww2040/jumps.html

The relevant sections are Sections 3.2, 3.4, 11.3 and 11.4. Proofs omitted there mostly appear in P. Billingsley, Convergence of Probability Measures.

(d) Is it true that

\[ P\left(\frac{(S_n - nm)}{\sqrt{n\sigma^2}} \leq x\right) \to P(N(0, 1) \leq x) \quad \text{as} \quad n \to \infty \quad \text{for all} \quad x? \]
Yes. As just explained, all \( x \) are continuity points of the limiting cdf. An example was given in the homework (involving unit point masses) for which not all \( x \) are continuity points of the limiting cdf. The correction in the broadcast email concerned that example.

(e) Is it true that
\[
P((S_n - nm)/\sqrt{n\sigma^2} \in A) \rightarrow P(N(0,1) \in A) \quad \text{as} \quad n \rightarrow \infty \quad \text{for all events} \ A?
\]

No! Consider the countably infinite set \( A \) equal to the set of numbers of the form \((k - nm)/\sqrt{n\sigma^2}\) as \( k \) ranges from 0 to \( n \), and then as \( n \) ranges from 1 upwards. For the left side the probability is 1. For the right side the probability is 0. Indeed, the probability of any countably infinite set under the normal distribution is 0: For that special \( A \),
\[
P((S_n - nm)/\sqrt{n\sigma^2} \in A) = 1, \quad \text{but} \quad P(N(0,1) \in A) = 0.
\]

We use the fact that
\[
P(N(0,1) \in A) = \int_A \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx.
\]

It turns out that \( P_n \Rightarrow P \) for probability measures (convergence in distribution) if \( P_n(A) \rightarrow P(A) \) for all events \( A \) such that the boundary of \( A \), denoted by \( \partial A \), has probability 0 under the limiting probability measure, i.e., such that \( P(\partial A) = 0 \). Here, for the special set we chose, \( \partial A \) is the entire real line, \( \mathbb{R} \). So here \( P(N(0,1) \in \partial A) = 1 \neq 0 \). The point of this is that we do not want to rule out convergence in distribution because convergence fails for a few very special bad sets. The obvious definition is \( P_n(A) \rightarrow P(A) \) for all \( A \), but that would prevent us from having convergence in distribution when we want to have it. The condition for cdf’s is a special case: For sets of the form \( A = (-\infty, x] \), \( \partial A = \{ x \} \). If \( P(X = x) > 0 \), then \( x \) is not a continuity point of \( P(X \leq x) \).


(f) Is it true that
\[
17e^{(S_n - nm)/\sqrt{n\sigma^2}} \rightarrow 17e^{N(0,1)} \quad \text{as} \quad n \rightarrow \infty ?
\]

This is true if by \( \rightarrow \) we mean \( \Rightarrow \) (convergence in distribution). This is true by the continuous mapping theorem.

THM. If \( X_n \Rightarrow X \) and if \( f \) is a continuous function (of a real variable), then \( f(X_n) \Rightarrow f(X) \). The continuous function here is of course
\[
f(x) = 17e^{(x)}.
\]

3. What is a random variable?
See top of page 7 in Ross. But it is not emphasized enough there. A random variable is a function (or mapping). The function maps an underlying probability space (S in Ross, Ω in class) into the real line \( \mathbb{R} \). The standard notation for a random variable is \( X \).

We often are only interested in the probability distribution of the random variable. Then we may never mention the underlying probability space. But the probability law of \( X \) comes from the probability in the underlying probability space. When we write \( P(X \leq x) \) we are giving the cumulative distribution function of the random variable \( X \). What does that mean?

There is an underlying probability measure, say \( P \), on \( S \). Then \( \{X \leq x\} \equiv \{s \in S : X(s) \leq x\} \) is an event in \( S \) (subset of \( S \)). The probability of this event is determined by the probability measure \( P \) originally defined on \( S \). We say it again below.

Given a probability space \( (S,P) \) and a random variable \( X \) (mapping \( S \) into \( \mathbb{R} \)), we define the probability law of \( X \), say \( P_X \), on \( \mathbb{R} \) by

\[
P_X(A) = P(X \in A) = P(\{s \in S : X(s) \in A\}) = P(X^{-1}(A)) \quad \text{for} \quad A \subseteq \mathbb{R}.
\]

The last three \( P \)'s in the chain of equalities all are the \( P \) on \( S \), i.e., \( P \) is part of the underlying probability space.

4. Suppose that

\[
V_n \Rightarrow N(2,3), \quad W_n \Rightarrow N(1,5), \quad X_n \Rightarrow 7 \quad \text{and} \quad Y_n \Rightarrow 3.
\]

(a) Does \( \sqrt{n}e^{(V_n^2-12V_n^2)} \Rightarrow \sqrt{N(2,3)}e^{(N(2,3)^3-12N(2,3)^2)} \)?

Yes, by the continuous mapping theorem for convergence in distribution. The continuous function is

\[
f(x) = \sqrt{x}e^{(x^3-12x^2)}.
\]

(b) Does \( V_n + W_n \Rightarrow N(3,8) \)?

No, not without extra assumptions. If \( V_n \) and \( W_n \) are independent random variables, then we can first deduce that \( (V_n, W_n) \Rightarrow (N(2,3), N(1,5)) \) in \( \mathbb{R}^2 \). (See Section 11.4 of my book, Stochastic-Process Limits or See Chapter 4 of Billingsley.) Then we can apply the continuous mapping theorem with addition (mapping \( \mathbb{R}^2 \) into \( \mathbb{R} \)). That yields

\[
V_n + W_n \Rightarrow N(2,3) + N(1,5).
\]

But then

\[
N(2,3) + N(1,5) \overset{d}{=} N(3,8)
\]

where \( \overset{d}{=} \) means equality in distribution, because sums of normals are normals, the expectation of a sum is always the sum of the expectations, and the variance of a sum is the sum of the variances when the component random variables are independent. So, under the extra independence assumption, the answer is yes.
(c) Does \( V_n + X_n \Rightarrow N(9, 3) \)?

Yes. The fact that the limit for \( X_n \) is a constant means that we can write

\[
(V_n, X_n) \Rightarrow (N(2, 3), 7) \text{ in } \mathbb{R}^2.
\]

(Again, see Section 11.4 of my book, Stochastic-Process Limits or Chapter 4 of Billingsley.)

Then we can apply the continuous mapping theorem with addition to get

\[
V_n + X_n \Rightarrow N(2, 3) + 7 \overset{d}{=} N(9, 3).
\]

Summary: One step is to go from \( X_n \Rightarrow X \) and \( Y_n \Rightarrow Y \) in \( \mathbb{R} \) to \((X_n, Y_n) \Rightarrow (X, Y) \) in \( \mathbb{R}^2 \).

Two sufficient conditions are:

(i) \( X_n \) and \( Y_n \) are independent for all \( n \),

and

(ii) One of the limits is deterministic, e.g., \( P(Y = c) = 1 \) for some constant \( c \).

(d) Does \( V_n^4(e^{X_n^2 + Y_n^2}) - W_n V_n^{13} + 6 \Rightarrow N(2, 3)^4(e^{(49 + 27)} - N(1, 5)N(2, 3)^{13} + 6 \)?

No, not in general. But, yes, if we assume that \( V_n \) and \( W_n \) are independent for each \( n \).

Then as a first step we can write

\[
(V_n, W_n, X_n, Y_n) \Rightarrow (N(2, 3), N(1, 5), 7, 3) \text{ in } \mathbb{R}^4.
\]

Then we can apply the continuous mapping theorem for convergence in distribution for the function mapping \( \mathbb{R}^4 \) into \( \mathbb{R} \).

(e) Does \( V_n^2 \Rightarrow N(2, 3)^2 \)?

Yes, by the continuous mapping theorem.

(f) Is \( N(2, 3)^2 \overset{d}{=} N(4, 9) \), where \( \overset{d}{=} \) means equal in distribution?

No, the square of a normal random variable is not normally distributed. It has a chi-squared distribution.

(g) Answer questions (a) - (e) above under the condition that \( \Rightarrow \) in all the limits is replaced by convergence with probability one (w.p.1).

All answers are yes. With w.p.1 convergence, we can ignore probability. All steps work directly. For the set with probability one, we are dealing with ordinary functions.
(h) Answer questions (a) - (e) above under the condition that ⇒ in all the limits is replaced by convergence in probability.

As with ⇒, you need to first get the joint convergence in parts (b) and (d) in order to apply the continuous mapping theorem. As before, that is implied by the extra independence assumption. There is a continuous mapping theorem for convergence in probability, just like the continuous mapping theorem for convergence in distribution. (See the discussion about w.p.1 convergent representations below.)

5. What is the difference between the weak law of large numbers and the strong law of large numbers?

The classical law of large numbers (LLN) concerns the average or sample mean of \( n \) independent and identically distributed (IID) random variables. Given a sequence \( \{X_n : n \geq 1\} \) of IID random variables, the sample mean is \( \bar{X}_n \equiv n^{-1}S_n \), where \( S_n \equiv X_1 + \cdots + X_n \) is the \( n^{th} \) partial sum. The LLN states that \( \bar{X}_n \) converges to the mean \( EX \). The weak LLN (WLLN) is convergence in probability; the strong LLN (SLLN) is convergence with probability one (w.p.1). Generalizations of the classical LLN draw the same conclusion with the IID conditions weakened. A version of the classical SLLN is proved in the Appendix to Chapter 1 in Ross. It requires that \( E[X_1^4] < \infty \). (See question 7 below.) The SLLN actually holds under the weaker condition that \( E[|X_1|] < \infty \).

6. Consider the following limits (as \( n \to \infty \)):

(i) \( X_n \Rightarrow X \) (convergence in distribution)
(ii) \( EX_n \to EX \)
(iii) \( E[|X_n - X|] \to 0 \)
(iv) \( P(X_n \to X) = 1 \)
(v) \( X_n \to X \) in probability

What are the implications among these limits? For example, does limit (i) imply limit (ii)?

First we give the standard definitions:

(i) \( X_n \Rightarrow X \) (convergence in distribution) if \( P(X_n \leq x) \to P(X \leq x) \) for all points \( x \) that are continuity points of the function (cdf) \( P(X \leq x) \). (There are other equivalent definitions.)

(ii) Straightforward, since expectations are just real numbers.

(iii) Straightforward, since expectations are just real numbers.

(iv) \( P(X_n \to X) = 1 \) means convergence with probability one, i.e., that, for each \( s \in A \), where \( A \) is an event in the underlying sample space \( S \) (with a probability measure, making the space \( (S, P) \)) with \( P(A) = 1 \), that \( X_n(s) \to X(s) \). To appreciate this definition, you must realize that random variables are functions mapping the underlying probability space \( (S, P) \) into the real line.
(v) \( X_n \to X \) in probability means convergence in probability, i.e., for each \( \epsilon > 0 \) and each \( \eta > 0 \) there is an \( n_0 \) such that

\[
P(|X_n - X| > \eta) < \epsilon \quad \text{for all} \quad n > n_0.
\]

Convergence w.p.1 is equivalent to the stronger property: for each \( \epsilon > 0 \) and each \( \eta > 0 \) there is an \( n_0 \) such that

\[
P(\sup_{m \geq n} |X_m - X| > \eta) < \epsilon \quad \text{for all} \quad n > n_0.
\]

almost-surely convergent representations

Convergence in distribution: \( X_n \Rightarrow X \) if and only if there exist random variables \( \tilde{X}_n, n \geq 1 \), and \( \tilde{X} \) such that \( \tilde{X}_n \) has the same probability law (distribution or measure) as \( X_n \) for all \( n \) and \( \tilde{X} \) has the same probability law as \( X \), and

\[
P(\tilde{X}_n \to \tilde{X} \quad \text{as} \quad n \to \infty) = 1.
\]

Proof. Easy way: Given such random variables \( \tilde{X}_n \) and \( \tilde{X} \) with \( P(\tilde{X}_n \to \tilde{X} \quad \text{as} \quad n \to \infty) = 1 \), we obtain

\( \tilde{X}_n \Rightarrow \tilde{X} \)

because convergence w.p.1 implies convergence in distribution. However, since \( \tilde{X}_n \) has the same probability law (distribution or measure) as \( X_n \) for all \( n \) and \( \tilde{X} \) has the same probability law as \( X \), we necessarily also have

\( X_n \Rightarrow X \).

Hard way: Given \( X_n \Rightarrow X \), let \( F_n \) be the cdf of \( X_n \) and let \( F \) be the cdf of \( X \). Let \( U \) be a fixed random variable uniformly distributed on the interval \([0, 1]\). Let

\[
\tilde{X}_n \equiv F_n^-(U), \quad n \geq 1, \quad \text{and} \quad \tilde{X} \equiv F^-(U),
\]

where \( F^- \) is the left-continuous inverse defined in homework 1. The rest of the details are in homework 1.

Convergence in probability: \( X_n \to X \) in probability if and only if for every subsequence \( \{X_m : m \geq 1\} \) of the original sequence \( \{X_n : n \geq 1\} \) there is a further subsequence \( \{X_{m_k} : m_k \geq 1\} \) such that \( X_{m_k} \) converges w.p.1 to \( X \) as \( m_k \to \infty \). (This property is somewhat obscure, so we will not pursue it.)

See Chapter 4 of Chung, *A Course in Probability Theory* for a discussion of convergence concepts. The following implications hold:

\( (iv) \to (v) \to (i), \quad (iii) \to (ii) \quad \text{and} \quad (iii) \to (v) \)

To see that \( (iv) \) does not imply \( (ii) \) or \( (iii) \), let the underlying probability space be the unit interval \([0, 1]\) with the uniform distribution (which coincides with Lebesgue measure). Let \( X = 0 \) w.p.1 and let \( X_n = 2^n \) on the interval \((a_n, a_n + 2^{-n})\) where \( a_n = 2^{-1} + 2^{-2} + \ldots + 2^{-(n-1)} \) with \( a_1 = 0 \), and let \( X_n = 0 \) otherwise. Then \( P(X_n \to X \equiv 0) = 1 \), but \( E[|X_n - X|] = EX_n = 1 \) for all \( n \), but \( EX = 0 \). (To see that indeed \( P(X_n \to X \equiv 0) = 1 \), note that the interval on which \( X_k \) is positive for any \( k > n \) has probability going to 0 as \( n \to \infty \).)

From the example above, it follows that \( (v) \) does not imply \( (ii) \) and that \( (i) \) does not imply \( (ii) \). However, \( (i) \) does imply \( (ii) \) under regularity conditions, namely, under uniform

To see that (iii) does not imply (iv), again let the underlying probability space be the unit interval $[0,1]$ with the uniform distribution (which coincides with Lebesgue measure). Let $X = 0$ w.p.1. Somewhat like before, let $X_n = 1$ on the interval $(a_n, a_n + n^{-1})$ where $a_n = a_{n-1} + (n - 1)^{-1} \mod 1$, with $a_1 = 0$, and let $X_n = 0$ otherwise. (The mod1 means that there is “wrap around” from 1 back to 0.) (To see that indeed $P(X_n \to X \equiv 0) = 0$, note that the $X_k = 1$ infinitely often for each sample point. On the other hand, $E[|X_n - X|] = EX_n = 1/n \to 0$ as $n \to \infty$.

Additional remark: Convergence in probability is equivalent to convergence in distribution when the limiting random variable $X$ is a constant, i.e., when there is a constant $c$ such that $P(X = c) = 1$.

---

7. Prove the SLLN under the assumption that $E[X^4] < \infty$.

See pages 56–58 in Ross. See pages 1–6, especially page 4, for relevant background.

---

8. Identify the following people:

(a) Borel Cantelli (His (or more properly their) theorem is applied to prove the SLLN.)
(b) Radon Nikodym
(c) Pollaczek Khintchine (big stars of queueing theory)
(d) Cauchy Schwartz
(e) Riemann Lebesgue
(f) Rao Blackwell
(g) Glivenko Cantelli
(h) Goldfarb Iyengar (fine fellows too)

Part (h) explains all: Each name is a combination of two last names, sharing in the fame.