ENGG2430A Probability and Statistics for Engineers

Chapter 5: Limit Theorems

Instructor: Shengyu Zhang

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

Background

- We will discuss fundamental issues related to the asymptotic behavior of sequences of random variables.

- Our principal context involves a sequence $X_1, X_2, \ldots$ of independent identically distributed (i.i.d.) random variables with mean μ and variance σ^2.

Background

- Let

$$S_n = X_1 + \cdots + X_n$$

be the sum of the first n of them.

- *Limit theorems* are mostly concerned with the properties of S_n and related random variables as n becomes very large.

Background

- Because of independence, we have

$$\mathrm{var}(S_n) = \mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n) = n\sigma^2$$

- The distribution of S_n spreads out as n increases
- Thus S_n cannot have a meaningful limit.
- But the situation is different if we consider the sample mean

$$M_n = \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n}.$$

Background

- A quick calculation yields,

$$\mathbf{E}[M_n] = \mu, \ \ \mathrm{var}(M_n) = \frac{\sigma^2}{n}.$$

- The variance of M_n decreases to zero as n increases.
- Thus the bulk of the distribution of M_n must be very close to the mean μ.
- This phenomenon is the subject of certain laws of large numbers

Background

- The laws generally assert that *the sample mean M_n converges to the true mean μ*.
- These laws provide a mathematical basis for the loose interpretation of an expectation $\mathbf{E}[X] = \mu$ …
- … as the average of a large number of independent samples drawn from the distribution of X.

Background

- We will also consider a quantity which is intermediate between S_n and M_n.
- Z_n is defined as follows.

1. subtract $n\mu$ from S_n, to obtain the zero-mean random variable $S_n - n\mu$
2. then divide by $\sigma\sqrt{n}$, to form the random variable

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Background

- It can be seen that

$$\mathbf{E}[Z_n] = 0, \qquad \mathrm{var}[Z_n] = 1$$

- Since the mean/variance of Z_n remain unchanged as n increases, its distribution neither spreads, nor shrinks to a point.
- The central limit theorem is concerned with
  - the asymptotic shape of the distribution of Z_n
  - and asserts that Z_n becomes the standard normal distribution.

Application

- Limit theorems are useful for several reasons:

- (a) Conceptually. They provided an interpretation of expectations/probabilities in terms of a long sequence of identical independent experiments.

Application

- (b) They allow for an approximate analysis of the properties of random variables such as S_n.
  - This is to be contrasted with an exact analysis which requires a formula for the PMF or PDF of S_n, a complicated and tedious task when n is large.

- (c) They play a major role in inference and statistics, in the presence of large data sets.

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

Markov and Chebyshev Inequalities

- These inequalities use the mean and possibly the variance of a random variable to draw conclusions on the probabilities of certain events.
- primarily useful in situations
  - exact values or bounds for the mean and variance of a random variable X are easily computable.
  - but the distribution of X is either unavailable or hard to calculate.

Markov inequality

- Theorem (Markov Inequality). If a random variable X can only take nonnegative values, then

$$P(X \geq a) \leq \frac{E[X]}{a}, \quad \text{for all } a > 0$$

- If a nonnegative random variable has a small mean, then the probability that it takes a large value must also be small.

Markov inequality (proof)

- To justify the Markov inequality, let us fix a positive number a and consider the random variable Y_a defined by

$$Y_a = \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a. \end{cases}$$

- It is seen that the relation

$$Y_a \leq X$$

always holds and therefore

$$\mathbf{E}[Y_a] \leq \mathbf{E}[X].$$

- On the other hand,

$$\mathbf{E}[Y_a] = a\mathrm{P}(Y_a = a) = a\mathrm{P}(\mathrm{X} \geq \mathrm{a})$$

- From which we obtain

$$a\mathrm{P}(X \geq a) \leq \mathbf{E}(X)$$

Figure 1: Illustration of the derivation.

- Part (a): the PDF of a nonnegative random variable X.
- Part (b): the PMF of a related random variable Y_a.

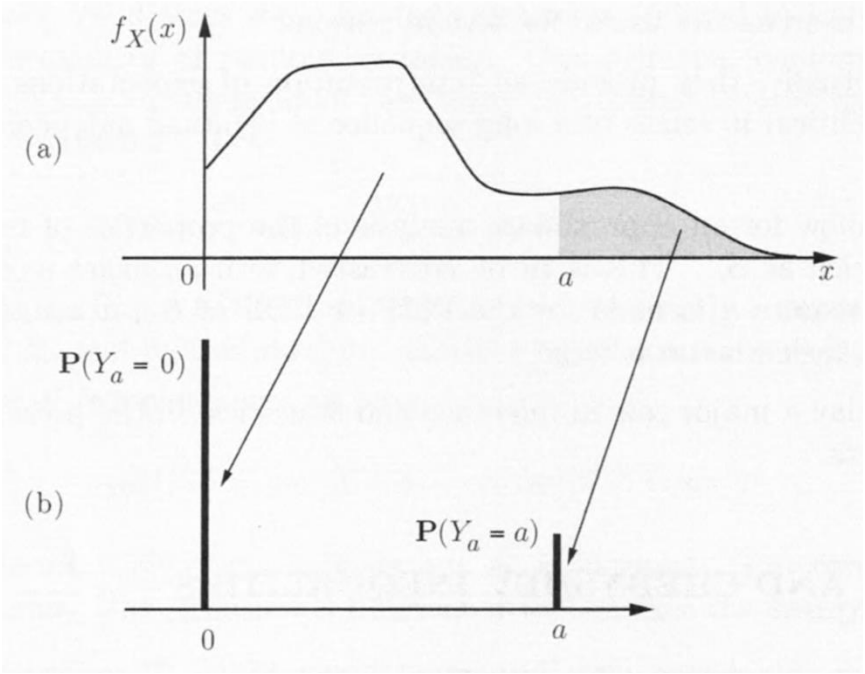
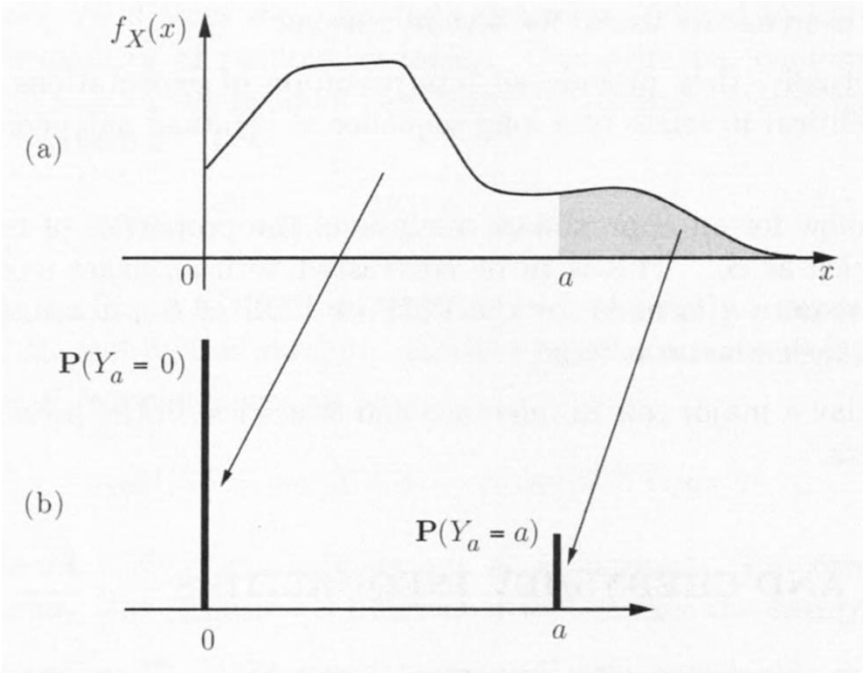


Figure 1: Illustration of the derivation.

- All of the mass in the PDF of X that lies between 0 and a is assigned to 0,
- All of the mass that lies above a is to a.
- Since mass is shifted to the left, the expectation can only decrease and therefore

$$\begin{aligned}\mathbf{E}[X] &\geq \mathbf{E}[Y_a] \\ &= a\mathrm{P}(Y_a = a) \\ &= a\mathrm{P}(X \geq a).\end{aligned}$$



Example 1.

- Let X be uniformly distributed in $[0,4]$.
- Note that $\mathbf{E}[X] = 2$.
- Then, the Markov inequality asserts that

$$\mathrm{P}(X \geq 2) \leq \frac{2}{2} = 1.$$

$$\mathrm{P}(X \geq 3) \leq \frac{2}{3} = 0.67.$$

$$\mathrm{P}(X \geq 4) \leq \frac{2}{4} = 0.5.$$

Example 1.

- By comparing with the exact probabilities

$$\mathrm{P}(X \geq 2) \leq \frac{2}{2} = 1. \qquad \mathrm{P}(X \geq 2) = 0.5.$$

$$\mathrm{P}(X \geq 3) \leq \frac{2}{3} = 0.67. \quad \mathrm{P}(X \geq 3) = 0.25.$$

$$\mathrm{P}(X \geq 4) \leq \frac{2}{4} = 0.5. \quad \mathrm{P}(X \geq 4) = 0.$$

- We see that the bounds provided by the Markov inequality can be quite loose.

Chebyshev inequality

- Theorem (Chebyshev Inequality). If X is a random variable with mean μ and variance σ^2, then

$$\mathrm{P}(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2}, \qquad \text{for all } c > 0$$

- If a random variable has small variance, then the probability that it takes a value far from its mean is also small.
  - Note: does not require X to be nonnegative.

Chebyshev inequality (Proof)

- Let's prove the Chebyshev inequality

$$\mathrm{P}(|X-\mu| \geq c) \leq \sigma^2/c^2$$

- Consider the nonnegative random variable $(X-\mu)^2$ and apply the Markov inequality:

$$\mathrm{P}\big((X-\mu)^2 \geq c^2\big) \leq \frac{\mathbf{E}[(X-\mu)^2]}{c^2} = \frac{\sigma^2}{c^2}$$

- Finally note that the event $(X-\mu)^2 \geq c^2$ is identical to the event $|X-\mu| \geq c$. Thus

$$\mathrm{P}(|X-\mu| \geq c) = \mathrm{P}\big((X-\mu)^2 \geq c^2\big) \leq \sigma^2/c^2$$

An alternative form

- An alternative form is obtained by letting $c = k\sigma$, where k is positive, which yields

$$P(|X - \mu| \geq k\sigma) \leq \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

- The probability that a random variable that is more than k standard deviations away from its mean is at most $\frac{1}{k^2}$.

Comparisons

- The Chebyshev inequality tends to be more powerful than the Markov inequality
  - the bounds are more accurate, because it also uses information on the variance of X
- The mean and the variance of a random variable are only a rough summary of its properties
  - we cannot expect the bounds to be close approximations of the exact probabilities.

Example 2. uninformative case

- Let X be uniformly distributed in $[0,4]$.
- Let us use the Chebyshev inequality

  $$\mathrm{P}(|X-\mu| \geq c) \leq \sigma^2/c^2$$

  to bound the probability that $|X-2| \geq 1$.
- We have $\sigma^2 = \frac{(b-a)^2}{12} = \frac{16}{12} = \frac{4}{3}$, and thus

  $$P(|X-2| \geq 1) \leq \frac{4}{3}$$

  which is uninformative.

Example 2. uninformative case

- let X be exponentially distributed with parameter $\lambda = 1$, so that $\mathbf{E}[X] = \mathrm{var}(X) = 1$.
- For $c > 1$, Chebyshev inequality yields

$$\begin{aligned} \mathrm{P}(X \geq c) &= \mathrm{P}(X - 1 \geq c - 1) \\ &\leq \mathrm{P}(|X - 1| \geq c - 1) \\ &\leq \frac{1}{(c-1)^2} \end{aligned}$$

This is again conservative compared to the exact answer $P(X \geq c) = e^{-c}$

Example 3. Upper Bounds

- When X is known to take values in a range $[a, b]$, we claim that $\sigma^2 \leq (b-a)^2/4$.

- If σ^2 is unknown, we may use the bound $(b-a)^2/4$ in the Chebyshev inequality,

$$\mathrm{P}(|X-\mu| \geq c) \leq \frac{(b-a)^2}{4c^2}, \qquad \text{for all } c > 0$$

Example 3. Upper Bounds (Proof)

- For the claim $\sigma^2 \leq (b-a)^2/4$, note that for any constant γ we have

$$\mathbf{E}[(X-\gamma)^2] = \mathbf{E}[X^2] - 2\mathbf{E}[X]\gamma + \gamma^2$$

- This quadratic is minimized when $\gamma = \mathbf{E}[X]$.
- It follows that, for all $\lambda \in [a, b]$,

$$\sigma^2 = \mathbf{E}[(X - \mathbf{E}[X])^2] \leq \mathbf{E}[(X-\gamma)^2]$$

Example 3. Upper Bounds (Proof)

- By letting $\gamma = (a+b)/2$, we obtain

$$\sigma^2 \leq \mathbf{E}\left[\left(X - \frac{a+b}{2}\right)^2\right]$$

$$= \mathbf{E}[(X-a)(X-b)] + \frac{(b-a)^2}{4} \leq \frac{(b-a)^2}{4}$$

  - The equality above is verified by calculation,
  - The last inequality follows from the fact

$$(x-a)(x-b) \leq 0$$

$\forall x \in [a,b]$.

Example 3. Upper Bounds (Proof)

- The bound may be quite conservative, but in the absence of further information about X, it cannot be improved.

- Indeed, it is satisfied with equality when X takes the two extreme values a and b with equal probability $1/2$.

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

Weak law of large numbers

- The weak law of large numbers asserts that the sample mean of a large number of i.i.d. random variables is close to the true mean.
  - i.i.d.: independent identically distributed.
  - This holds with high probability.
- Consider a sequence X_1, X_2, ... of i.i.d. random variables with mean μ and variance σ^2,
- The sample mean is $M_n = \frac{X_1 + \cdots + X_n}{n}$, while the true mean is μ.

- We have

$$\mathbf{E}[M_n] = \frac{\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]}{n} = \frac{n\mu}{n} = \mu.$$

- and

$$\begin{aligned} \mathrm{var}(M_n) &= \frac{\mathrm{var}(X_1 + \cdots + X_n)}{n^2} \\ &= \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)}{n^2} \quad \text{(independence)} \\ &= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

- We apply the Chebyshev inequality and obtain

$$\mathrm{P}(|M_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}, \text{ for any } \epsilon > 0$$

- For any fixed $\epsilon > 0$, the right-hand side of this inequality goes to zero as n increases.
- This is the weak law of large numbers.

The Weak Law of Large Numbers

- *Theorem*. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with mean μ, then $\forall \epsilon > 0$, we have

$$\mathrm{P}(|M_n - \mu| \geq \epsilon) \to 0, \text{ as } n \to \infty$$

  - Recall:

$$M_n = \frac{X_1 + \cdots + X_n}{n}.$$

Intuitively

- The weak law of large numbers states that for large n, the bulk of the distribution of M_n is concentrated near μ.
- If we consider a positive length interval $[\mu - \epsilon, \mu + \epsilon]$ around μ, then there is high probability that M_n will fall in that interval.
- As $n \to \infty$, this probability converges to 1.
  - If ϵ is very small, we need to wait longer (i.e. need a larger value of n) to assert that M_n is highly likely to fall in that interval.

Example 4. Probabilities and Frequencies

- Consider an event A defined in the context of some probabilistic experiment.
- Let $p = \mathrm{P}(A)$ be the probability of this event.
- Consider n independent repetitions of the experiment, and let M_n be the fraction of time that event A occurs;
- In this context, M_n is often called the empirical frequency of A.

Example 4. Probabilities and Frequencies

- Note that

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

where X_i is 1 whenever A occurs, and 0 otherwise. In particular, $\mathbf{E}[X_i] = p$.

- The weak law shows that when n is large, the empirical frequency is most likely to be within ϵ of p.

Example 4. Probabilities and Frequencies

- Loosely speaking, this allows us to conclude that empirical frequencies are faithful estimates of p.

- Alternatively, this is a step towards interpreting the probability p as the frequency of occurrence of A.

Example 5. Polling

- Let p be the fraction of voters who support a particular candidate for office.
- We interview n "randomly selected" voters and record M_n, the fraction of them that support the candidate.
- We view M_n as our estimate of p and would like to investigate its properties.

Example 5. Polling

- We interpret "randomly selected" to mean that the n voters are chosen independently and uniformly from the given population.
- Thus, the reply of each person interviewed can be viewed as an independent Bernoulli random variable X, with success probability p and variance $\sigma^2 = p(1-p)$.

Example 5. Polling

- The Chebyshev inequality yields

$$P(|M_n - p| \geq \epsilon) \leq \frac{1}{4n\epsilon^2}$$

- E.g. if $\epsilon = 0.1$ and $n = 100$, we obtain

$$P(|M_{100} - p| \geq 0.1) \leq \frac{1}{4 * 100 * 0.1^2} = 0.25$$

  - With a sample size of $n = 100$, the probability that our estimate is incorrect by more than 0.1 is no larger than 0.25.

Example 5. Polling (2)

- Suppose now that we impose some tight specifications on our poll.
- We would like to have high confidence (probability at least 95%) that our estimate will be very accurate (within 0.01 of p).

- *Question*: *How many voters should be sampled?*

Example 5. Polling (2)

- The only guarantee that we have at this point is the inequality

$$\mathrm{P}(|M_n - p| \geq 0.01) \leq \frac{1}{4 * n * 0.01^2}$$

- To satisfy the above specifications:

$$\frac{1}{4 * n * 0.01^2} \leq 1 - 0.95 = 0.05$$

which yields $n \geq 50{,}000$

Example 5. Polling (2)

- This n satisfies our specifications, but turns out to be fairly conservative.
  - because it is based on the rather loose Chebyshev inequality.
- We'll give a finer bound later.

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

CONVERGENCE IN PROBABILITY

- We can interpret the weak law of large numbers as stating that

" M_n converges to μ."

- However, since $M_1, M_2, \ldots$ is a sequence of random variables, not numbers, the meaning of convergence has to be made precise.

Convergence of a Deterministic Sequence

- Let $a_1, a_2, \ldots$ be a sequence of real numbers, and let a be another real number. We say that a_n converges to a, or

$$\lim_{n\to\infty} a_n = a,$$

  if for every $\epsilon > 0$ there exists some n_0 such that

$$|a_n - a| \leq \epsilon, \qquad \text{for all } n \geq n_0$$

- Intuitively, for any accuracy level ϵ, a_n must be within ϵ of a, when n is large enough.

Convergence in Probability

- Let $Y_1, Y_2, \ldots$ be a sequence of random variables, and let a be a real number.

- We say that the sequence Y_n converges to a in probability, if for every $\epsilon > 0$, we have

$$\lim_{n \to \infty} \mathrm{P}(|Y_n - a| \geq \epsilon) = 0.$$

- Given this definition, the weak law of large numbers simply states that the sample mean converges in probability to the true mean μ.

- More generally, the Chebyshev inequality implies the following:
- If all Y_n have the same mean μ and $\mathrm{var}(Y_n)$ converges to 0, then Y_n converges to μ in probability.

- Suppose that Y_1, Y_2 ... have a PMF or a PDF and converge in probability to a.

- Then "almost all" of the PMF or PDF of Y_n is concentrated within ϵ of a for large values of n.

- It is also instructive to rephrase the above definition as follows.
- For every $\epsilon > 0$ and for every $\delta > 0$, there exists some n_0 such that

$$\mathrm{P}(|Y_n - a| \geq \epsilon) \leq \delta, \qquad \text{for all } n \geq n_0.$$

- Last slide: $\mathrm{P}(|Y_n - a| \geq \epsilon) \leq \delta$.
- Let's refer to ϵ as the accuracy level, and δ as the confidence level.
- The definition takes the following intuitive form.
- For any given level of accuracy and confidence, Y_n will be equal to a, within these levels of accuracy and confidence, provided that n is large enough.

Example 6.

- Consider a sequence of independent random variables X_n that are uniformly distributed in the interval $[0,1]$.
- Let $Y_n = \min\{X_1, \ldots X_n\}$.
- The sequence of values of Y_n cannot increase as n increases.
  - Minimum over more numbers is smaller.
- Y_n occasionally decreases
  - whenever a value of X_n that is smaller than the preceding values is obtained.

- Thus, we expect that Y_n converges to zero.
- Indeed, for $\epsilon > 0$, we have using the independence of the X_n,

$$\begin{aligned} P(|Y_n - 0| \geq \epsilon) &= P(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) \\ &= P(X_1 \geq \epsilon) \cdots P(X_n \geq \epsilon) \\ &= (1-\epsilon)^n \end{aligned}$$

- In particular,

$$\lim_{n\to\infty} P(|Y_n - 0| \geq \epsilon) = \lim_{n\to\infty} (1-\epsilon)^n = 0$$

- Since this is true for every $\epsilon > 0$, Y_n converges to zero (in probability).

Example 7.

- Let Y be an exponentially distributed random variable with parameter $\lambda = 1$.
- For any positive integer n, let $Y_n = Y/n$.
- Note: These random variables are dependent.

- We wish to investigate whether the sequence Y_n converges to zero.

- For $\epsilon > 0$, we have

$$\begin{aligned} \mathrm{P}(|Y_n - 0| \geq \epsilon) &= \mathrm{P}(Y_n \geq \epsilon) \\ &= \mathrm{P}(Y \geq n\epsilon) = e^{-n\epsilon} \end{aligned}$$

- In particular,

$$\lim_{n\to\infty} \mathrm{P}(|Y_n - 0| \geq \epsilon) = \lim_{n\to\infty} e^{-n\epsilon} = 0.$$

- Since this is the case for every $\epsilon > 0, Y_n$ converges to zero, in probability.

Example 8.

- One might believe that if a sequence Y_n converges to a number a, then $\mathbf{E}[Y_n]$ must also converge to a.
- The following example shows that this need not be the case.
- This illustrates some of the limitations of the notion of convergence in probability.

Example 8.

- Consider a sequence of discrete random variables Y_n with the following distribution:

$$\mathrm{P}(Y_n = y) = \begin{cases} 1 - \dfrac{1}{n} & \text{for } y = 0, \\ \dfrac{1}{n} & \text{for } y = n^2, \\ 0 & \text{otherwise.} \end{cases}$$

Example 8.

- For every $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathrm{P}(|Y_n| \geq \epsilon) = \lim_{n\to\infty} \frac{1}{n} = 0$$

- Thus Y_n converges to zero in probability.
- On the other hand, $\mathbf{E}[Y_n] = \frac{n^2}{n} = n$, which goes to infinity as n increases.

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

Sample mean

- According to the weak law of large numbers,
- the distribution of the sample mean

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

  is increasingly concentrated in the near vicinity of the true mean μ.
- In particular, its variance tends to zero.

Sample sum and normalized mean

- On the other hand, the variance of the sum
$$S_n = X_1 + \cdots + X_n = nM_n$$
increases to infinity.
- And the distribution of S_n cannot be said to converge to anything meaningful.
- An intermediate view is obtained by considering the deviation $S_n - n\mu$ of S_n, and scaling it by a factor proportional to $1/\sqrt{n}$.

Formally

- Let $X_1, \cdots, X_n$ be a sequence of independent identically distributed random variable with mean μ and variance σ^2

- Define

$$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

Mean and variance

- An easy calculation yields

$$\mathbf{E}[Z_n] = \frac{\mathbf{E}[X_1 + \cdots + X_n] - n\mu}{\sigma\sqrt{n}} = 0$$

- For variance, we have

$$\mathrm{var}(Z_n) = \frac{\mathrm{var}(X_1 + \cdots + X_n)}{(\sigma\sqrt{n})^2} = \frac{n\sigma^2}{n\sigma^2} = 1$$

The Central Limit Theorem

- *Theorem* (*The Central Limit Theorem*) The CDF of $Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$ converges to standard normal CDF

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$$

in the sense that

$$\lim_{n \to \infty} \mathrm{P}(Z_n \leq z) = \Phi(z)$$

Generality

- The central limit theorem is surprisingly general.
- Besides independence, and the implicit assumption that the mean and variance are finite, it places no other requirement on the distribution of the X_i,
  - which could be discrete, continuous, or mixed.

Importance - conceptual

- The theorem is of tremendous importance for several reasons.
  - Both conceptual and practical.
- Conceptual: It indicates that *the sum of a large number of independent random variables is approximately normal.*
- As such, it applies to many situations in which a random effect is the sum of a large number of small but independent random factors.

- Noise in many natural or engineered systems has this property.
- In a wide array of contexts, it has been found empirically that the statistics of noise are well-described by normal distributions.
- The central limit theorem provides a convincing explanation for this phenomenon.

Importance - practical

- Practical: It *eliminates the need for detailed probabilistic models*.
- Rather, it allows the calculation by referring to the normal CDF.
- Furthermore, these calculations only require the knowledge of means and variances.

Approximations Based on CLT

- The central limit theorem allow us to calculate probabilities related to Z_n as if Z_n is normal.
- Since normality is preserved under linear transformations, this is equivalent to treating S_n as a normal variable with mean $n\mu$ and variance $n\sigma^2$

Approximations Based on CLT

- Let $S_n = X_1 + \cdots + X_n$, where the X_i are i.i.d. random variables with mean μ, variance σ^2.
- If n is large, the probability $\mathrm{P}(S_n \leq c)$ can be approximated by treating S_n as if it were normal, according to the following procedure.

I. Calculate the mean $n\mu$ and variance $n\sigma^2$.

II. Calculate $z = (c - n\mu)/\sigma\sqrt{n}$.

III. Use the approximation

$$\mathrm{P}(S_n \leq c) \approx \Phi(z).$$

Approximations example: Plane

- We load on a plane 100 packages whose weights are independent and uniform between 5 and 50.
- *Question: What is the probability that the total weight exceeds 3000?*
- Let S_{100} be the sum of weights. We first find the mean and variance of the weight of a single package

$$\mu = \frac{5+50}{2} = 27.5, \sigma^2 = \frac{(50-5)^2}{12} = 168.75$$

Approximations example: Plane

- We then calculate

$$z = \frac{3000 - n\mu}{\sigma\sqrt{n}} = 1.92$$

- Then

$$\mathrm{P}(S_n \leq 3000) \approx \Phi(1.92) \approx 0.9726$$

  - by checking the standard normal table
- Hence the desired probability (that the total weight exceeds 3000) is about 0.0274.

Approximations example: Machine

- A machine process parts, one at a time, in a time independently and uniformly distributed in [1,5].
- We will approximate the probability the machine processes at least 100 parts in 320 time units.
- Let X_i be the processing time of the i-th part, and let

$$S_n = X_1 + \cdots + X_n$$

be the total processing time of first 100 parts.

Approximations example: Machine

- Then we need to calculate $\mathrm{P}(S_{100} \leq 320)$.
- Note that $\mu = \mathbf{E}[X_i] = 3, \sigma^2 = \mathrm{var}(X_i) = 4/3$,

$$z = \frac{320 - n\mu}{\sigma\sqrt{n}} = 1.73$$

- Hence

$$\mathrm{P}(S_{100} \leq 320) \approx \Phi(1.73) \approx 0.9582$$

Approximations example: Polling

- We poll n voters and record the fraction M_n of those polled who are in favor of a particular candidate.
- If p is fraction of the entire voter population that supports this candidate, then

$$M_n = \frac{X_1 + \cdots + X_n}{n},$$

where X_i is a Bernoulli random variable with parameter p.

Approximations example: Polling

- M_n has mean p and variance $p(1-p)/n$.
- We're interested in the probability $\mathrm{P}(|M_n - p| \geq \epsilon)$.
  - The probability that the polling error is larger than some desired accuracy ϵ.
- Because of the symmetry of the normal PDF around the mean, we have

$$\mathrm{P}(|M_n - p| \geq \epsilon) \approx 2\mathrm{P}(M_n - p \geq \epsilon)$$

Approximations example: Polling

- The variance $p(1-p)/n$ of $M_n - p$ depends on p and is therefore unknown.
- Note that the probability of a large deviation from the mean *increases with the variance*.
- Thus, we can obtain an upper bound on $\mathrm{P}(M_n - p \geq \epsilon)$ by assuming that $M_n - p$ has the largest possible variance $1/4n$, which corresponds to $p = 1/2$.

Approximations example: Polling

- We evaluate the normalized value $z = \frac{\epsilon}{1/(2\sqrt{n})}$
- And use the normal approximation

$$\mathrm{P}(M_n - p \geq \epsilon) \leq 1 - \Phi(2\epsilon\sqrt{n})$$

- For example: $n = 100$ and $\epsilon = 0.1$. We observe, for any p

$$\mathrm{P}(M_{100} - p \geq \epsilon) \leq 2 - 2\Phi(2\epsilon\sqrt{n}) = 0.046$$

- This is much smaller (more accurate) than the estimate of 0.25 that was obtained using the Chebyshev inequality.

Approximations example: Polling

- We consider a reverse problem.
- How large a sample size n is needed if we wish our estimate M_n to be within 0.01 of p with probability at least 0.95?
- Similar to previous calculations, we have

$$2 - 2\Phi(2\epsilon\sqrt{n}) \leq 0.05$$

  which leads to $n \geq 9604$.
- This is significantly better than the 50,000 that we found using Chebyshev's inequality.

Approximations - Example

- The normal approximation is increasingly accurate as n tends to infinity.
- But in practice we are generally faced with specific and finite values of n.
- It would be useful to know how large n should be before the approximation can be trusted.

- But there are no simple and general guidelines. Much depends on whether the distribution of the X_i is close to normal and, in particular, whether it is symmetric.
- For example, if the X_i are uniform, then S_8 is already very close to normal.
- But if the X_i are, say, exponential, a significantly larger n will be needed before S_n is close to a normal one.

Approximation to the Binomial

- Consider a binomial random variable S_n with parameters n and p.
- It can be viewed as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$, with common parameter p:

$$S_n = X_1 + \cdots + X_n$$

- Recall that for each X_i

$$\mu = p, \qquad \sigma = \sqrt{p(1-p)}$$

Approximation to the Binomial

- We will now use the CLT approximation for $\mathrm{P}(\{k \leq S_n \leq l\})$,
  - where k and l are given integers.
- We express the event of interest in terms of a standardized random variable, using the equivalence

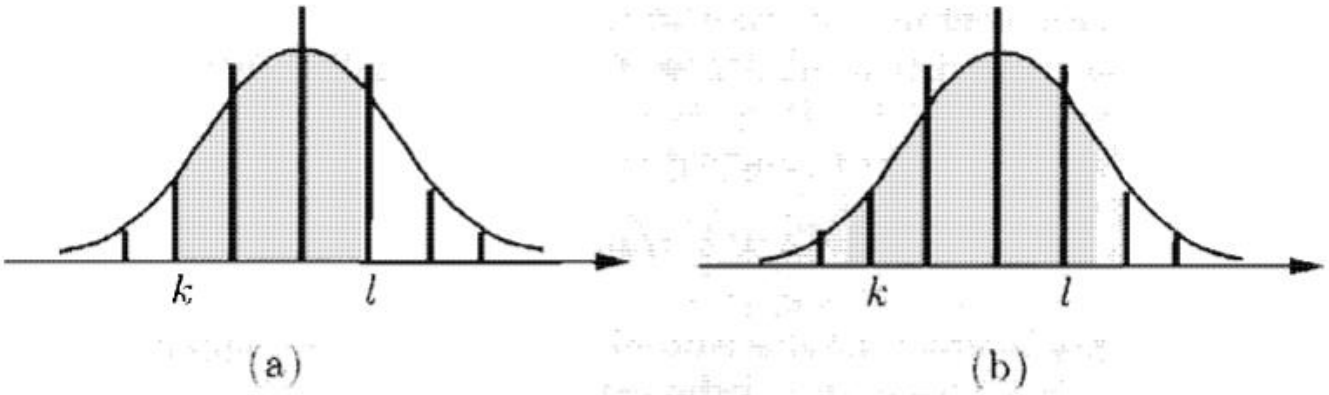
$$k \leq S_n \leq l \Leftrightarrow \frac{k - n\mu}{\sigma\sqrt{n}} \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{l - n\mu}{\sigma\sqrt{n}}$$

- By CLT, $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ approximates standard normal,

$$\begin{aligned} &\mathrm{P}(k \le S_n \le l) \\ &= \mathrm{P}\left(\frac{k-n\mu}{\sigma\sqrt{n}} \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le \frac{l-n\mu}{\sigma\sqrt{n}}\right) \\ &\approx \Phi\left(\frac{l-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{k-n\mu}{\sigma\sqrt{n}}\right) \end{aligned}$$

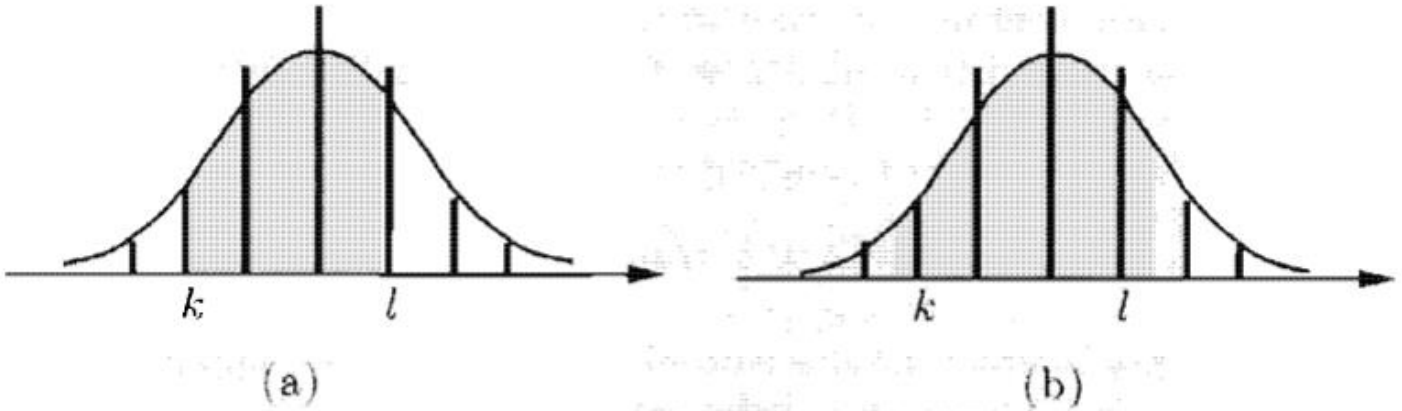
Approximation to the Binomial

- A first approximation (a) of a binomial probability $\mathrm{P}(k \leq S_n \leq l)$ is obtained by integrating the area under the normal PDF from k to l.
- An issue happens when $k = l$: $\mathrm{P}(S_n = k)$ will be approximated as 0.



Approximation to the Binomial

- A possible remedy (b) is to integrate normal PDF between $k - 1/2$ and $l + 1/2$, to approximate $\mathrm{P}(k \leq S_n \leq l)$.
- If so, $\mathrm{P}(S_n = k)$ is no longer 0.



Approximation to the Binomial

- Plugging $\mu = p, \sigma = \sqrt{p(1-p)}$, we get the following *De Moivre-Laplace Approximation to the Binomial.*
- If S_n is a binomial random variable with parameters n and p, n is large, and k, l are nonnegative integers, then

$$\mathrm{P}(k \le S_n \le l) \approx \Phi\left(\frac{l + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

Approximation to the Binomial

- When p is close to 1/2, in which case the PMF of the X_i is symmetric, the above formula yields a very good approximation for n as low as 40 or 50.
- When p is near 1 or near 0, the quality of the approximation drops, and a larger value of n is needed to maintain the same accuracy.

Approximation to the Binomial

- For example, let S_n be a binomial random variable with $n = 36$ and $p = 0.5$, the exact calculation

$$\mathrm{P}(S_n \leq 21) = \sum_{k=0}^{21} \binom{36}{k} (0.5)^{36} = 0.8785.$$

- Using CLT approximation,

$$\mathrm{P}(S_n < 21) \approx \Phi\left(\frac{21.5 - np}{\sqrt{np(1-p)}}\right) = 0.879.$$

Content

- Markov and Chebyshev Inequalities
- The Weak Law of Large Numbers
- Convergence in Probability
- The Central Limit Theorem
- The Strong Law of Large Numbers

The Strong Law of Large Numbers

- The strong law of large numbers is similar to the weak law in that it also deals with the convergence of the sample mean to the true mean.
- It is different, however, because it refers to another type of convergence.
- The following is a general statement of the strong law of large numbers.

The Strong Law of Large Numbers

- Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables with mean μ.
- Then, the sequence of sample means $M_n = (X_1 + \cdots + X_n)/n$ converges to μ, with probability 1, in the sense that

$$\mathrm{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1$$

The Strong Law of Large Numbers

- In order to interpret the strong law of large numbers, we need to go back to our original description of probabilistic models in terms of sample spaces.
- The contemplated experiment is infinitely long and generates a sequence of values, one value for each one of the random variables in the sequence $X_1, X_2, \ldots$

The Strong Law of Large Numbers

- Thus, it is best to think of the sample space as *a set of infinite sequences* $(x_1, x_2, \dots)$ *of real numbers*: any such sequence is a possible outcome.
- Let us now consider the set A consisting of those sequences $(x_1, x_2, \dots)$ whose long-term average is μ, i.e.,

$$(x_1, x_2, \dots) \in A \quad \Leftrightarrow \lim_{n\to\infty} \frac{x_1 + \cdots + x_n}{n} = \mu$$

The Strong Law of Large Numbers

- The strong law of large numbers states that all of the probability is concentrated on this subset A of the sample space.

- Equivalently, the collection of outcomes that do not belong to A (infinite sequences whose long-term average is not μ) has probability zero.

The Strong Law of Large Numbers

- The difference between the weak and the strong law is subtle and deserves close scrutiny.
- The weak law states that the probability $\mathrm{P}(|M_n - \mu| \geq \epsilon)$ of a significant deviation of M_n from μ goes to zero as $n \to \infty$.
- Still, for any finite n, this probability can be positive and M_n may once in a while deviates significantly from μ.

- The weak law provides no conclusive information on the number of such deviations.
- But the strong law does.
- According to the strong law, with probability 1, M_n converges to μ.
- This implies that for any given $\epsilon > 0$, the probability that the difference $|M_n - \mu|$ will exceed ϵ an infinite number of times is equal to zero.

Probabilities and Frequencies

- Consider an event A defined in terms of some probabilistic experiment.
  - Denote $\mu = P(A)$.
- Consider a sequence of independent repetitions of the same experiment.
- Let M_n be the fraction of the first n repetitions in which A occurs.

Probabilities and Frequencies

- The strong law of large numbers asserts that M_n converges to μ, with probability 1.
  - $\mathrm{P}\left(\lim_{n\to\infty} M_n = \mu\right) = 1$
- In contrast, the weak law of large numbers asserts that M_n converges to $\mathrm{P}(A)$ in probability.
  - $\forall \epsilon > 0,\ \mathrm{P}(|M_n - \mu| \geq \epsilon) \to 0,$ as $n \to \infty$.

Probabilities and Frequencies

- We have often talked intuitively about the probability of an event A as the frequency with which it occurs in an infinitely long sequence of independent trials.
- The strong law backs this intuition and establishes that the long-term frequency of occurrence of A is indeed equal to $\mathrm{P}(A)$, with essential certainty.

Convergence with Probability 1

- The convergence concept behind the strong law is different than the notion employed in the weak law.
- We provide here a definition and some discussion of this new convergence concept.

Convergence with Probability 1

- Let $Y_1, Y_2, \cdots$ be a sequence of random variables (not necessarily independent).
- Let c be a real number.
- We say that , Y_n converges to c with probability 1 (or almost surely) if

$$\mathrm{P}\left(\lim_{n\to\infty} Y_n = c\right) = 1$$

Convergence with Probability 1

- Similar to our earlier discussion, a proper interpretation of this type of convergence involves a sample space consisting of infinite sequences.
- All of the probability is concentrated on those sequences that converge to c.
- This does not mean that other sequences are impossible, only that their total probability is zero.

Convergence with Probability 1

- Let $X_1, X_2, \ldots$ be a sequence of independent random variables that are uniformly distributed in $[0,1]$. Let

$$Y_n = \min(X_1, X_2, \cdots, X_n).$$

- We will show that Y_n converges to 0 with probability 1.

Convergence with Probability 1

- We first observe Y_n is nonincreasing, i.e., $Y_n \geq Y_{n+1}$.
- Since $Y_n \geq 0$ is bounded, it has a limit Y.
- For any small $\epsilon > 0$, $Y > \epsilon$ iff $X_i > \epsilon, \forall i$.
- Therefore,

$$\mathrm{P}(Y > \epsilon) = \mathrm{P}(X_i > \epsilon, \forall i) = (1-\epsilon)^n$$

- Letting $n \to \infty$, we have $\mathrm{P}(Y > \epsilon) = 0$
- Hence, $\mathrm{P}(Y = 0) = 1$.

Convergence with Probability 1

- Convergence with probability 1 implies convergence in probability, but the converse is not necessarily true.

- Our last example illustrates the difference between convergence in probability and convergence with probability 1.

Convergence with Probability 1

- Consider a *discrete-time arrival process*.
- The set of time is partitioned into consecutive intervals of the form

$$I_k = \{2^k, 2^k + 1, \cdots 2^{k+1} - 1\}.$$

- Note the length of I_k increases as k increases.
- During each interval I_k, there is exactly one arrival, and all times within an interval are equally likely.
  - The arrivals are assumed to be independent.

Convergence with Probability 1

- Define $Y_n = 1$ if there is an arrival at time n, and $Y_n = 0$ if there is no arrival.
- We have $\mathrm{P}(Y_n \neq 0) = 1/2^k$, if $n \in I_k$.
- Therefore,

$$\lim_{n\to\infty} \mathrm{P}(Y_n \neq 0) = \lim_{k\to\infty} \frac{1}{2^k} = 0.$$

- We conclude that Y_n converges to 0 in probability.

Convergence with Probability 1

- However, the total number of arrivals is infinite.
- Therefore, Y_n is unity for infinitely many values of n.
- So the event $\lim_{n\to\infty} Y_n = 0$ has 0 probability.
- It doesn't converge with probability 1.

Convergence with Probability 1

- Intuitively, the following is happening.
- At any given time, there is only a small, and diminishing with n, probability of a substantial deviation from 0,
  - which implies convergence in probability.
- On the other hand, given enough time, a substantial deviation from 0 is certain to occur.
  - For this reason, we do not have convergence with probability 1.

Summary

- *Weak law of large numbers*
  - $\forall \epsilon > 0,\ \mathrm{P}(|M_n - \mu| \geq \epsilon) \to 0,$ as $n \to \infty$.
  - Indicates that the sample mean M_n is very likely to be close to the true mean μ, as the sample size increases.
  - Based on the Chebyshev inequality.

Summary

- *Central limit theorem*
  - $\lim_{n\to\infty} \mathrm{P}(Z_n \leq z) = \Phi(z)$, where
    - $Z_n = \frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$
    - Φ is the CDF of the standard normal.
  - Asserts that the sum of a large number of independent random variables is approximately normal.
  - Can be used for approximation.

Summary

- *Strong law of large numbers*.
  - $\mathrm{P}\left(\lim_{n\to\infty} \frac{X_1+\cdots+X_n}{n} = \mu\right) = 1$
  - Makes a more emphatic connection of probabilities and relative frequencies,
  - Is often an important tool in theoretical studies.