Lecture 3 - Expectation, inequalities and laws of large numbers

Jan Bouda

FI MU

April 19, 2009

Part I

Functions of a random variable

Functions of a random variable

Given a random variable X and a function $\Phi : \mathbb{R} \rightarrow \mathbb{R}$ we define the transformed random variable $Y = \Phi(X)$ as

- Random variables X and Y are defined on the same sample space, moreover, $\mathbf{Dom}(X) = \mathbf{Dom}(Y)$.
- $\mathbf{Im}(Y) = \{\Phi(x) | x \in \mathbf{Im}(X)\}$.
- The probability distribution of Y is given by

$$p_Y(y) = \sum_{x \in \mathbf{Im}(X); \Phi(x) = y} p_X(x).$$

In fact, we may define it by $\Phi(X) = \Phi \circ X$, where $\circ$ is the usual function composition.

Part II

Expectation

Expectation

- The probability distribution or probability distribution function completely characterize properties of a random variable.
- Often we need description that is less accurate, but much shorter - single number, or a few numbers.
- First such characteristic describing a random variable is the **expectation**, also known as the **mean value**.

Definition

Expectation of a random variable X is defined as

$$E(X) = \sum_i x_i p(x_i)$$

provided the sum is absolutely (!) convergent. In case the sum is convergent, but not absolutely convergent, we say that no finite expectation exists. In case the sum is not convergent the expectation has no meaning.

Median; Mode

- The **median** of a random variable X is any number x such that $P(X < x) \leq 1/2$ and $P(X > x) \geq 1/2$.
- The **mode** of a random variable X is the number x such that

$$p(x) = \max_{x' \in \mathbf{Im}(X)} p(x').$$

Moments

- Let us suppose we have a random variable X and a random variable $Y = \Phi(X)$ for some function Φ. The expected value of Y is
$$E(Y) = \sum_i \Phi(x_i) p_X(x_i).$$
- Especially interesting is the power function $\Phi(X) = X^k$. $E(X^k)$ is known as the kth moment of X. For $k = 1$ we get the expectation of X.
- If X and Y are random variables with matching corresponding moments of all orders, i.e. $\forall k \;\; E(X^k) = E(Y^k)$, then X and Y have the same distributions.
- Usually we center the expected value to 0 – we use moments of $\Phi(X) = X - E(X)$.
- We define the kth central moment of X as
$$\mu_k = E\left([X - E(X)]^k \right).$$

Variance

Definition

The second central moment is known as the **variance** of X and defined as

$$\mu_2 = E\left([X - E(X)]^2\right).$$

Explicitly written,

$$\mu_2 = \sum_i [x_i - E(X)]^2 p(x_i).$$

The variance is usually denoted as σ_X^2 or $Var(X)$.

Definition

The square root of σ_X^2 is known as the **standard deviation** $\sigma_X = \sqrt{\sigma_X^2}$.

If variance is small, then X takes values close to $E(X)$ with high probability. If the variance is large, then the distribution is more 'diffused'.

Expectation revisited

Theorem

Let $X_1, X_2, \dots X_n$ be random variables defined on the same probability space and let $Y = \Phi(X_1, X_2, \dots X_n)$. Then

$$E(Y) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_n} \Phi(x_1, x_2, \dots x_n) p(x_1, x_2, \dots x_n).$$

Theorem (Linearity of expectation)

Let X and Y be random variables. Then

$$E(X + Y) = E(X) + E(Y).$$

Linearity of expectation (proof)

Linearity of expectation.

$$
\begin{aligned}
E(X+Y) &= \sum_i \sum_j (x_i + y_j) p(x_i, y_j) = \\
&= \sum_i x_i \sum_j p(x_i, y_j) + \sum_j y_j \sum_i p(x_i, y_j) = \\
&= \sum_i x_i p_X(x_i) + \sum_j y_j p_Y(y_j) = \\
&= E(X) + E(Y).
\end{aligned}
$$

□

Linearity of expectation

The linearity of expectation can be easily generalized for any linear combination of n random variables, i.e.

Theorem (Linearity of expectation)

Let $X_1, X_2, \ldots X_n$ be random variables and $a_1, a_2, \ldots a_n \in \mathbb{R}$ constants. Then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i).$$

Proof is left as a home exercise :-).

Expectation of independent random variables

Theorem

If X and Y are independent random variables, then

$$E(XY) = E(X)E(Y).$$

Proof.

$$\begin{aligned}
E(XY) =& \sum_i \sum_j x_i y_j p(x_i, y_j) = \\
=& \sum_i \sum_j x_i y_j p_X(x_i) p_Y(y_j) = \\
=& \sum_i x_i p_X(x_i) \sum_j y_j p_Y(y_j) = \\
=& E(X)E(Y).
\end{aligned}$$

Expectation of independent random variables

The expectation of independent random variables can be easily generalized for any n–tuple $X_1, X_2, \dots X_n$ of mutually independent random variables:

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i).$$

If $\Phi_1, \Phi_2, \dots \Phi_n$ are functions, then

$$E\left[\prod_{i=1}^{n} \Phi_i(X_i)\right] = \prod_{i=1}^{n} E[\Phi_i(X_i)].$$

Variance revisited

Theorem

Let σ_X^2 be the variance of the random variable X. Then

$$\sigma_X^2 = E(X^2) - [E(X)]^2.$$

Proof.

$$\begin{aligned}\sigma_X^2 =& E\left([X - E(X)]^2\right) = E\left(X^2 - 2XE(X) + [E(X)]^2\right) = \\ =& E(X^2) - E[2XE(X)] + [E(X)]^2 = \\ =& E(X^2) - 2E(X)E(X) + [E(X)]^2.\end{aligned}$$

□

Covariance

Definition

The quantity

$$E\left([X - E(X)][Y - E(Y)]\right) = \sum_{i,j} p_{x_i,y_j} \left[x_i - E(X)\right] \left[y_j - E(Y)\right]$$

is called the **covariance** of X and Y and denoted $Cov(X, Y)$.

Theorem

Let X and Y be independent random variables. Then the covariance of X and Y $Cov(X, Y) = 0$.

Covariance

Proof.

$$\begin{aligned}
Cov(X,Y) =& E\big([X-E(X)][Y-E(Y)]\big) = \\
=& E[XY - YE(X) - XE(Y) + E(X)E(Y)] = \\
=& E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) = \\
=& E(X)E(Y) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) = 0
\end{aligned}$$

$\square$

- Covariance measures linear (!) dependence between two random variables.
- E.g. when $X = aY$, $a \neq 0$, using $E(X) = aE(Y)$ we have

$$Cov(X,Y) = aVar(Y) = \frac{1}{a}Var(X).$$

Covariance

In general it holds that

$$0 \leq Cov^2(X, Y) \leq Var(X)Var(Y).$$

Definition

We define the **correlation coefficient** $\rho(X, Y)$ as the normalized covariance, i.e.

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

It holds that $-1 \leq \rho(X, Y) \leq 1$.

Covariance

It may happen that X is completely dependent on Y and yet the covariance is 0, e.g. for $X = Y^2$ and a suitably chosen Y.

Variance

Theorem

If X and Y are independent random variables, then

$$Var(X+Y) = Var(X) + Var(Y).$$

Proof.

$$
\begin{aligned}
&Var(X+Y) = E\big([(X+Y) - E(X+Y)]^2\big) = \\
=&E\big([(X+Y) - E(X) - E(Y)]^2\big) = E\big([(X-E(X)) + (Y - E(Y))]^2\big) = \\
=&E\big([X-E(X)]^2 + [Y-E(Y)]^2 + 2[X-E(X)][Y-E(Y)]\big) = \\
=&E\big([X-E(X)]^2\big) + E\big([Y-E(Y)]^2\big) + 2E\big([X-E(X)][Y-E(Y)]\big) = \\
=&Var(X) + Var(Y) + 2E\big([X-E(X)][Y-E(Y)]\big) = \\
=&Var(X) + Var(Y) + 2Cov(X,Y) = Var(X) + Var(Y).
\end{aligned}
$$

□

Variance

- If X and Y are not independent, we obtain (see proof on the previous transparency)

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

- The additivity of variance can be generalized to a set $X_1, X_2, \ldots X_n$ of mutually independent variables and constants $a_1, a_2, \ldots a_n \in \mathbb{R}$ as

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i).$$

Part III

Conditional Distribution and Expectation

Conditional probability

Using the derivation of conditional probability of two events we can derive conditional probability of (a pair of) random variables.

Definition

The **conditional probability distribution** of random variable Y given random variable X (their joint distribution is $p_{X,Y}(x,y)$) is

$$\begin{aligned} p_{Y|X}(y|x) =& P(Y=y|X=x) = \frac{P(Y=y, X=x)}{P(X=x)} = \\ =& \frac{p_{X,Y}(x,y)}{p_X(x)} \end{aligned} \tag{1}$$

provided $p_X(x) \neq 0$.

Conditional distribution function

Definition

The **conditional probability distribution function** of random variable Y given random variable X (their joint distribution is $p_{X,Y}(x,y)$) is

$$F_{Y|X}(y|x) = P(Y \leq y | X = x) = \frac{P(Y \leq y \text{ and } X = x)}{P(X = x)} \qquad (2)$$

for all values $P(X = x) > 0$. Alternatively we can derive it from the conditional probability distribution as

$$F_{Y|X}(y|x) = \frac{\sum_{t \leq y} p(x,t)}{p_X(x)} = \sum_{t \leq y} p_{Y|X}(t|x).$$

Conditional expectation

We may consider $Y|(X = x)$ to be a new random variable that is given by the conditional probability distribution $p_{Y|X}$. Therefore, we can define its mean and moments.

Definition

The **conditional expectation** of Y given $X = x$ is defined

$$E(Y|X = x) = \sum_y yP(Y = y|X = x) = \sum_y yp_{Y|X}(y|x). \tag{3}$$

Analogously can be defined conditional expectation of a transformed random variable $\Phi(Y)$, namely the conditional kth moment of Y: $E(Y^k|X = x)$. Of special interest will be the conditional variance

$$Var(Y|X = x) = E(Y^2|X = x) - [E(Y|X = x)]^2.$$

Conditional expectation

We can derive the expectation of Y from the conditional expectations. The following equation is known as the **theorem of total expectation**:

$$E(Y) = \sum_x E(Y|X = x)p_X(x). \tag{4}$$

Analogously, the **theorem of total moments** is

$$E(Y^k) = \sum_x E(Y^k|X = x)p_X(x). \tag{5}$$

Example: Random sums

Let $N, X_1, X_2, \ldots$ be mutually independent random variables. Let us suppose that $X_1, X_2, \ldots$ have identical probability distribution $p_X(x)$, mean $E(X)$, and variance $Var(X)$. We also know the values $E(N)$ and $Var(N)$. Let us consider the random variable defined as a sum

$$T = X_1 + X_2 + \cdots + X_N.$$

In what follows we would like to calculate $E(T)$ and $Var(T)$. For a fixed value $N = n$ we can easily derive the conditional expectation of T by

$$E(T|N = n) = \sum_{i=1}^{n} E(X_i) = nE(X). \tag{6}$$

Using the theorem of total expectation we get

$$E(T) = \sum_{n} nE(X)p_N(n) = E(X)\sum_{n} np_N(n) = E(X)E(N). \tag{7}$$

Example: Random sums

It remains to derive the variance of T. Let us first compute $E(T^2)$. We obtain

$$E(T^2|N=n) = Var(T|N=n) + [E(T|N=n)]^2 \tag{8}$$

and

$$Var(T|N=n) = \sum_{i=1}^{n} Var(X_i) = nVar(X) \tag{9}$$

since $(T|N=n) = X_1 + X_2 + \cdots + X_n$ and $X_1, \ldots, X_n$ are mutually independent.

We substitute (6) and (9) into (8) to get

$$E(T^2|N=n) = nVar(X) + n^2 E(X)^2. \tag{10}$$

Example: Random sums

Using the theorem of total moments we get

$$
\begin{aligned}
E(T^2) &= \sum_n \left(nVar(X) + n^2[E(X)]^2\right) p_N(n) \\
&= \left(Var(X)\sum_n np_N(n)\right) + \left([E(X)]^2 \sum_n p_N(n)n^2\right) \\
&= Var(X)E(N) + E(N^2)[E(X)]^2.
\end{aligned}
\tag{11}
$$

Finally, we obtain

$$
\begin{aligned}
Var(T) =& E(T^2) - [E(T)]^2 = \\
=& Var(X)E(N) + E(N^2)[E(X)]^2 - [E(X)]^2[E(N)]^2 = \\
=& Var(X)E(N) + [E(X)]^2 Var(N).
\end{aligned}
\tag{12}
$$

Part IV

Inequalities

Markov inequality

It is important to derive as much information as possible even from a partial description of random variable. The mean value already gives more information than one might expect, as captured by Markov inequality.

Theorem (Markov inequality)

Let X be a nonnegative random variable with finite mean value $E(X)$. Then for all $t > 0$ it holds that

$$P(X \geq t) \leq \frac{E(X)}{t}$$

Markov inequality

Proof.

Let us define the random variable Y_t (for fixed t) as

$$Y_t = \begin{cases} 0 & \text{if } X < t \\ t & X \geq t. \end{cases}$$

Then Y_t is a discrete random variable with probability distribution $p_{Y_t}(0) = P(X < t)$, $p_{Y_t}(t) = P(X \geq t)$. We have

$$E(Y_t) = tP(X \geq t).$$

The observation $X \geq Y_t$ gives

$$E(X) \geq E(Y_t) = tP(X \geq t),$$

what is the Markov inequality. $\square$

Chebyshev inequality

In case we know both mean value and variance of a random variable, we can use much more accurate estimation

Theorem (Chebyshev inequality)

Let X be a random variable with finite variance. Then

$$P\big[|X - E(X)| \geq t\big] \leq \frac{Var(X)}{t^2},\ t > 0$$

or, alternatively, substituting $X' = X - E(X)$

$$P(|X'| \geq t) \leq \frac{E(X'^2)}{t^2},\ t > 0.$$

We can see that this theorem is in agreement with our interpretation of variance. If σ is small, then there is large probability of getting outcome close to $E(X)$, if σ is large, then there is large probability of getting outcomes farther from the mean.

Chebyshev inequality

Proof.

We apply the Markov inequality to the nonnegative variable $[X - E(X)]^2$ and we replace t by t^2 to get

$$P\left[(X - E(X))^2 \geq t^2\right] \leq \frac{E\left([X - E(X)]^2\right)}{t^2} = \frac{\sigma^2}{t^2}.$$

We obtain the Chebyshev inequality using the fact that the events $[(X - E(X))^2 \geq t^2] = [|X - E(X)| \geq t]$ are the same. $\square$

Kolmogorov inequality

Theorem (Kolmogorov inequality)

Let $X_1, X_2, \dots X_n$ be independent random variables. We put

$$S_k = X_1 + \cdots + X_k, \tag{13}$$

$$m_k = E(S_k) = E(X_1) + \cdots + E(X_k), \tag{14}$$

$$s_k^2 = Var(S_k) = Var(X_1) + \cdots + Var(X_k). \tag{15}$$

For every $t > 0$ it holds that

$$P\Big[|S_1 - m_1| < ts_n \wedge |S_2 - m_2| < ts_n \wedge \dots |S_n - m_n| < ts_n\Big] \geq 1 - t^{-2}. \tag{16}$$

Kolmogorov inequality

Comparing to Chebyshev inequality we see that the Kolmogorov inequality is considerably stronger since Chebyshev inequality implies only

$$\forall i = 1 \dots n \quad P(|S_i - m_i| < ts_i) \geq 1 - t^{-2}.$$

We used rewriting the Chebyshev inequality as

$$P\left[|X - E(X)| < t'\right] \geq 1 - \frac{Var(X)}{t'^2}, \; t > 0$$

and the substitution $X = S_i$, $t' = s_i t$.

Kolmogorov inequality

Proof.

We want to estimate the probability p that at least one of the terms in Eq. (16) does not hold (this is complementary event to Eq. (16)) and to verify the statement of the theorem, i.e. to test whether $p \leq t^{-2}$.

Let us define n random variables Y_k, $k = 1, \ldots, n$ as

$$Y_k = \begin{cases} 1 & \text{if } |S_k - m_k| \geq ts_n \text{ and } \forall v = 1, 2 \ldots k-1,\ |S_v - m_v| < ts_n, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

In words, Y_k equals 1 at those sample points in which the kth inequality of (16) is the first to be violated. For any sample point at most one of Y_v is one and the sum $Y_1 + Y_2 + \cdots + Y_n$ is either 0 or 1. Moreover, it is 1 if and only if at least one of the inequalities (16) is violated. □

Kolmogorov inequality

Proof.

Therefore,

$$p = P(Y_1 + Y_2 \cdots Y_n = 1). \tag{18}$$

We know that $\sum_k Y_k \leq 1$ and multiplying both sides of this inequality by $(S_n - m_n)^2$ gives

$$\sum_k Y_k (S_n - m_n)^2 \leq (S_n - m_n)^2. \tag{19}$$

Through taking expectation of each side over S_n we get

$$\sum_{k=1}^{n} E[Y_k (S_n - m_n)^2] \leq s_n^2. \tag{20}$$

□

Kolmogorov inequality

Proof.

We introduce the substitution

$$U_k = (S_n - m_n) - (S_k - m_k) = \sum_{v=k+1}^{n} [X_v - E(X_v)]$$

to evaluate respective summands of the left-hand side of Eq. (20) as

$$\begin{aligned} E[Y_k(S_n - m_n)^2] =& E[Y_k(U_k + S_k - m_k)^2] = \\ =& E[Y_k(S_k - m_k)^2] + 2E[Y_k U_k(S_k - m_k)] + E(Y_k U_k^2). \end{aligned} \quad (21)$$

Now, since U_k depends only on $X_{k+1}, \ldots, X_n$ and Y_k and S_k depend only on $X_1, \ldots, X_k$, we have that U_k is independent of $Y_k(S_k - m_k)$. Therefore, $E[Y_k U_k(S_k - m_k)] = E[Y_k(S_k - m_k)]E(U_k) = 0$ since $E(U_k) = 0$. □

Kolmogorov inequality

Proof.

Thus from Eq. (21) we get

$$E[Y_k(S_n - m_n)^2] \geq E[Y_k(S_k - m_k)^2] \tag{22}$$

observing that $E(Y_k U_k^2) \geq 0$.
We know that $Y_k \neq 0$ only if $|S_k - m_k| \geq ts_n$, so that $Y_k(S_k - m_k)^2 \geq t^2 s_n^2 Y_k$ and therefore

$$E[Y_k(S_k - m_k)^2] \geq t^2 s_n^2 E(Y_k). \tag{23}$$

□

Kolmogorov inequality

Proof.

Combining (20), (22) and (23) we get

$$\begin{aligned} s_n^2 &\geq \sum_{k=1}^{n} E[Y_k(S_n - m_n)^2] \geq \sum_{k=1}^{n} E[Y_k(S_k - m_k)^2] \geq \\ &\geq \sum_{k=1}^{n} t^2 s_n^2 E(Y_k) = t^2 s_n^2 E(Y_1 + \cdots + Y_n) \end{aligned} \tag{24}$$

and from (18) using $P(Y_1 + \cdots + Y_n = 1) = E(Y_1 + \cdots + Y_n)$ we finally obtain

$$pt^2 \leq 1. \tag{25}$$

□

Part V

Laws of Large Numbers

(Weak) Law of Large Numbers

Theorem ((Weak) Law of Large Numbers)

Let $X_1, X_2, \ldots$ be a sequence of mutually independent random variables with a common probability distribution. If the expectation $\mu = E(X_k)$ exists, then for every $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right) = 0.$$

In words, the probability that the average S_n/n differs from the expectation by less then arbitrarily small ϵ goes to 0.

Proof.

WLOG we can assume that $\mu = E(X_k) = 0$, otherwise we simply replace X_k by $X_k - \mu$. This induces only change of notation. □

(Weak) Law of Large Numbers

Proof.

In the special case $Var(X_k)$ exists, the law of large numbers is a direct consequence of the Chebyshev inequality; we substitute $X = X_1 + \cdots + X_n = S_n$ to get

$$P(|S_n - \mu| \geq t) \leq \frac{Var(X_k)n}{t^2}. \tag{26}$$

We substitute $t = \epsilon n$ and observe that with $n \to \infty$ the right-hand side tends to 0 to get the result. However, in case $Var(X_k)$ exists, we can apply the more accurate central limit theorem. The proof without the assumption that $Var(X_k)$ exists follows. □

(Weak) Law of Large Numbers

Proof.

Let δ be a positive constant to be determined later. For each k we define a pair of random variables ($k = 1 \dots n$)

$$U_k = X_k, V_k = 0 \qquad \text{if } |X_k| \leq \delta n \tag{27}$$

$$U_k = 0, V_k = X_k \qquad \text{if } |X_k| > \delta n \tag{28}$$

By this definition

$$X_k = U_k + V_k. \tag{29}$$

□

(Weak) Law of Large Numbers

Proof.

To prove the theorem it suffices to show that both

$$\lim_{n\to\infty} P(|U_1 + \cdots + U_n| > \frac{1}{2}\epsilon n) = 0 \qquad (30)$$

and

$$\lim_{n\to\infty} P(|V_1 + \cdots + V_n| > \frac{1}{2}\epsilon n) = 0 \qquad (31)$$

hold, because $|X_1 + \cdots + X_n| \leq |U_1 + \cdots + U_n| + |V_1 + \cdots + V_n|$.
Let us denote all possible values of X_k by $x_1, x_2, \ldots$ and the corresponding probabilities $p(x_i)$. We put

$$a = E(|X_k|) = \sum_i |x_i| p(x_i). \qquad (32)$$

□

(Weak) Law of Large Numbers

Proof.

The variable U_1 is bounded by δn and X_1 and therefore

$$U_1^2 \leq X_1 \delta n.$$

Taking expectation on both sides gives

$$E(U_1^2) \leq a\delta n. \tag{33}$$

Variables $U_1, \ldots U_n$ are mutually independent and have the same probability distribution. Therefore,

$$\begin{aligned} E[(U_1 + \cdots + U_n)^2] - [E(U_1 + \cdots + U_n)]^2 = Var(U_1 + \cdots + U_n) = \\ = nVar(U_1) \leq nE(U_1^2) \leq a\delta n^2. \end{aligned} \tag{34}$$

□

(Weak) Law of Large Numbers

Proof.

On the other hand, $\lim_{n\to\infty} E(U_1) = E(X_1) = 0$ and for sufficiently large n we have

$$[E(U_1 + \cdots U_n)]^2 = n^2[E(U_1)]^2 \leq n^2 a\delta \tag{35}$$

and for sufficiently large n we get from Eq. (34) that

$$E[(U_1 + \cdots + U_n)^2] \leq 2a\delta n^2. \tag{36}$$

Using the Chebyshev inequality we get the result (30) observing that

$$P(|U_1 + \cdots + U_n| > 1/2\epsilon n) \leq \frac{8a\delta}{\epsilon^2}. \tag{37}$$

By choosing sufficiently small δ we can make the right-hand side arbitrarily small to get (30). $\square$

(Weak) Law of Large Numbers

Proof.

In case of (31) note that

$$P(V_1 + V_2 + \cdots V_n \neq 0) \leq \sum_{i=1}^{n} P(V_i \neq 0) = nP(V_1 \neq 0). \tag{38}$$

For arbitrary $\delta > 0$ we have

$$P(V_1 \neq 0) = P(|X_1| > \delta n) = \sum_{|x_i| > \delta n} p(x_i) \leq \frac{1}{\delta n} \sum_{|x_i| > \delta n} |x_i| p(x_i). \tag{39}$$

The last sum tends to 0 as $n \rightarrow \infty$ and therefore also the left side tends to 0. This statement is even stronger than (31) and it completes the proof. □

Central Limit Theorem

In case the variance exists, instead of the law of large numbers we can apply more exact central limit theorem.

Theorem (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be a sequence of mutually independent identically distributed random variables with a finite mean $E(X_i) = \mu$ and a finite variance $Var(X_i) = \sigma^2$. Let $S_n = X_1 + \cdots + X_n$. Then for every fixed β it holds that

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \beta\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta} e^{-\frac{1}{2}y} dy \tag{40}$$

In case the mean does not exist, neither the central limit theorem nor the law of large number applies. Nevertheless, we still may be interested in a number of such cases, see e.g. (Feller, p. 246).

Law of Large Numbers, Central Limit Theorem and Variables with Different Distributions

- Let us make a small comment on the generality of the law of large numbers and the central limit theorem, namely we will relax the requirement that the random variables $X_1, X_2, \ldots$ are identically distributed.
- Let us denote in the following transparencies

$$s_n^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2. \tag{41}$$

- The law of large numbers holds if and only if X_k are uniformly bounded, i.e. there exists a constant A such that $\forall k \ |X_k| < A$.
- A sufficient condition (but not necessary) is that

$$\lim_{n\to\infty} \frac{s_n}{n} = 0. \tag{42}$$

  In this case our proof applies.

Law of Large Numbers, Central Limit Theorem and Variables with Different Distributions

Theorem (Lindeberg theorem)

The central limit theorem holds for any sequence of random variables $X_1, X_2 \ldots$ with finite means $\mu_1, \mu_2, \ldots$ and variance if and only if for every $\epsilon > 0$ the random variables U_k defined by

$$U_k = \begin{cases} X_k - \mu_k & \text{if } |X_k - \mu_k| \leq \epsilon S_n \\ 0 & \text{if } |X_k - \mu_k| > \epsilon S_n \end{cases} \tag{43}$$

satisfy

$$\lim_{s_n \to \infty} \frac{1}{s_n^2} \sum_{k=1}^{n} E(U_k^2) = 1. \tag{44}$$

This theorem e.g. implies that every sequence of uniformly bounded random variables obeys the central limit theorem.

Strong Law of Large Numbers

The (weak) law of large number implies that large values $|S_n - m_n|/n$ occur infrequently. In many practical situation we require the stronger statement that $|S_n - m_n|/n$ remains small for all sufficiently large n.

Definition (Strong Law of Large Numbers)

We say that the sequence $X_1, X_2, \ldots$ obeys the strong law of large numbers if to every pair $\epsilon > 0,\ \delta > 0$ there exists an $n \in \mathbb{N}$ such that

$$P\left(\forall r : \frac{|S_n - m_n|}{n} < \epsilon \wedge \frac{|S_{n+1} - m_{n+1}|}{n+1} < \epsilon \wedge \ldots \frac{|S_{n+r} - m_{n+r}|}{n+r} < \epsilon\right) \geq 1 - \delta. \tag{45}$$

It remains to determine the conditions when the strong law of large numbers holds.

Strong Law of Large Numbers

Theorem (Kolmogorov criterion)

Let $X_1, X_2, \ldots$ be a sequence of random variables with corresponding variances $\sigma_1^2, \sigma_2^2, \ldots$. Then the convergence of the series

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} \tag{46}$$

is a sufficient condition for the strong law of large numbers to apply.

Strong Law of Large Numbers

Proof.

Let A_v be the event that for at least one n such that $2^{v-1} < n \leq 2^v$ the inequality

$$\frac{|S_n - m_n|}{n} < \epsilon \tag{47}$$

does not hold. It suffices to prove that for all sufficiently large v and all r it holds that

$$P(A_v) + P(A_{v+1}) + \cdots + P(A_{v+r}) < \delta, \tag{48}$$

i.e. the series $\sum P(A_v)$ converges. The event A_v implies that for some n in range $2^{v-1} < n \leq 2^v$

$$|S_n - m_n| \geq \epsilon n \geq \epsilon 2^{v-1} \tag{49}$$

$\square$

Strong Law of Large Numbers

Proof.

Using $s_n^2 \leq s_{2^v}^2$ and Kolmogorov inequality with $t = \epsilon 2^{v-1}/s_{2^v}$ we get

$$P(A_v) \leq P(|S_n - m_n| \geq \epsilon 2^{v-1}) \leq 4\epsilon^{-2} s_{2^v}^2 2^{-2v}. \tag{50}$$

Hence (observe that $\sum_{k=1}^{2^v} \sigma_k^2 = s_{2^v}^2$)

$$\begin{aligned}\sum_{v=1}^{\infty} P(A_v) \leq \sum_{v=1}^{\infty} 4\epsilon^{-2} s_{2^v}^2 2^{-2v} \leq 4\epsilon^{-2} \sum_{v=1}^{\infty} 2^{-2v} \sum_{k=1}^{2^v} \sigma_k^2 \\ = 4\epsilon^{-2} \sum_{k=1}^{\infty} \sigma_k^2 \sum_{2^v \geq k} 2^{-2v} \leq 8\epsilon^{-2} \sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2}.\end{aligned} \tag{51}$$

□

Strong Law of Large Numbers

Before formulating our final criterion for the strong law of large numbers, we have to introduce two auxiliary lemmas we will use in the proof of the main theorem.

Lemma (First Borel-Cantelli lemma)

Let $\{A_k\}$ be a sequence of events defined on the same sample space and $a_k = P(A_k)$. If $\sum_k a_k$ converges, then to every $\epsilon > 0$ it is possible to find an (sufficiently large) integer n such that for all integers r the probability

$$P(A_{n+1} \cap A_{n+2} \cap \cdots \cap A_{n+r}) \leq \epsilon. \tag{52}$$

Strong Law of Large Numbers

Proof.

First it is important to determine n so that $a_{n+1} + a_{n+2} + \cdots \leq \epsilon$. This is possible since $\sum_k a_k$ converges. The lemma follows since

$$P(A_{r+1} \cap A_{r+2} \cap \cdots \cap A_{r+n}) \leq a_{r+1} + a_{r+2} + \cdots a_{r+n} \leq \epsilon. \tag{53}$$

□

Strong Law of Large Numbers

Lemma

For any $v \in \mathbb{N}^+$ it holds that

$$v \sum_{k=v}^{\infty} \frac{1}{k^2} \leq 2. \tag{54}$$

Strong Law of Large Numbers

Proof.

Let us consider the series $\sum_{k=1}^{\infty} \frac{1}{k(k+1)}$. Using $\frac{1}{k(k+1)} = \frac{1}{k} - \frac{1}{k+1}$ we get that

$$s_n = \sum_{k=1}^{n} \frac{1}{k(k+1)} = 1 - \frac{1}{n+1}.$$

We calculate the limit to get

$$\sum_{k=2}^{\infty} \frac{1}{k(k-1)} = \sum_{k=1}^{\infty} \frac{1}{k(k+1)} = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \left(1 - \frac{1}{n+1}\right) = 1.$$

□

Strong Law of Large Numbers

Proof.

We proceed to obtain

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = 1 + \sum_{k=2}^{\infty} \frac{1}{k^2} \le 1 + \sum_{k=2}^{\infty} \frac{1}{k(k-1)} = 2,$$

and, analogously,

$$2\sum_{k=2}^{\infty} \frac{1}{k^2} \le 2\sum_{k=2}^{\infty} \frac{1}{k(k-1)} = 2.$$

$\square$

Strong Law of Large Numbers

Proof.

Finally, we get the result for $v > 2$

$$v\sum_{k=v}^{\infty}\frac{1}{k^2} \leq v\sum_{k=v}^{\infty}\frac{1}{k(k-1)} = v\left(\left(\sum_{k=2}^{\infty}\frac{1}{k(k-1)}\right) - \left(\sum_{k=2}^{v-1}\frac{1}{k(k-1)}\right)\right) =$$

$$= v\left(1 - \left(\sum_{k=1}^{v-2}\frac{1}{k(k+1)}\right)\right) = \frac{v}{v-1} \leq 2. \tag{55}$$

□

Strong Law of Large Numbers

Theorem

Let $X_1, X_2, \dots$ be a sequence of mutually independent random variables with common probability distribution $p(x_i)$ and the mean $\mu = E(X_i)$ exists. Then the strong law of large numbers applies to this sequence.

Proof.

Let us introduce two new sequences of random variables defined by

$$U_k = X_k, V_k = 0 \qquad \text{if } |X_k| < k \tag{56}$$

$$U_k = 0, V_k = X_k \qquad \text{if } |X_k| \geq k. \tag{57}$$

U_k are mutually independent and we will show that they satisfy the Kolmogorov criterion. □

Strong Law of Large Numbers

Proof.

For $\sigma_k^2 = Var(U_k)$ we get

$$\sigma_k^2 \leq E(U_k^2) = \sum_{|x_i|<k} x_i^2 p(x_i). \tag{58}$$

Let us put for abbreviation

$$a_v = \sum_{v-1\leq|x_i|<v} |x_i| p(x_i). \tag{59}$$

□

Strong Law of Large Numbers

Proof.

Then the serie $\sum_v a_v$ converges since $E(X_k)$ exists. Moreover, from (58)

$$\sigma_k^2 \leq a_1 + 2a_2 + 3a_3 + \cdots + ka_k \tag{60}$$

observing that

$$\sum_{v-1 \leq |x_i| < v} |x_i|^2 p(x_i) \leq v \sum_{v-1 \leq |x_i| < v} |x_i| p(x_i) = va_v.$$

Strong Law of Large Numbers

Proof.

Thus

$$\sum_{k=1}^{\infty} \frac{\sigma_k^2}{k^2} \leq \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{v=1}^{k} v a_v = \sum_{v=1}^{\infty} v a_v \sum_{k=v}^{\infty} \frac{1}{k^2} \overset{(*)}{<} 2 \sum_{v=1}^{\infty} a_v < \infty. \tag{61}$$

To see $(*)$ recall that from the previous lemma for any $v = 1, 2, \ldots$ we have $2 \geq v \sum_{k=v}^{\infty} \frac{1}{k^2}$. Therefore, the Kolmogorov criterion holds for $\{U_k\}$. □

Strong Law of Large Numbers

Proof.

Now,

$$E(U_k) = \mu_k = \sum_{|x_i|<k} x_i p(x_i), \tag{62}$$

$\lim_{k\to\infty} = \mu$ and hence $\lim_{n\to\infty}(\mu_1 + \mu_2 + \cdots + \mu_n)/n = \mu$.
From the strong law of large numbers for $\{U_k\}$ we obtain that with probability $1 - \delta$ or better

$$\forall n > N : \left| n^{-1} \sum_{k=1}^{n} U_k - \mu \right| < \epsilon \tag{63}$$

provided N is sufficiently large. It remains to prove that the same statement holds when we replace U_k by X_k. It suffices to show that we can chose N sufficiently large so that with probability close to unity the event $U_k = X_k$ occurs for all $k > N$. □

Strong Law of Large Numbers

Proof.

This holds if with probability arbitrarily close to one only finitely many variables V_k are different from zero, because in this case there exists k_0 such that $\forall k > k_0 \; V_k = 0$ with probability 1. By the first Borel-Cantelli lemma this is the case when the series $\sum P(V_k \neq 0)$ converges. It remains to verify the convergence. We have

$$P(V_n \neq 0) = \sum_{|x_i| \geq n} p(x_i) \leq \frac{a_{n+1}}{n} + \frac{a_{n+2}}{n+1} + \cdots \tag{64}$$

and hence

$$\sum_{n=1}^{\infty} P(V_n \neq 0) \leq \sum_{n=1}^{\infty} \sum_{v=n}^{\infty} \frac{a_{v+1}}{v} = \sum_{v=1}^{\infty} \frac{a_{v+1}}{v} \sum_{n=1}^{v} 1 = \sum_{v=1}^{\infty} a_{v+1} < \infty, \tag{65}$$

since $E(X)$ exists. This completes our proof. $\square$