Random Variables and Expectation

Random Variables

- Variable describing the outcome of an experiment with random outcomes
- An example is https://www.geogebra.org/m/n5IHi3Ij

- We use large capital letters X to denote the variable and small capital letters x to denote its value in one experiment
- Random variable is in contrast to Deterministic Variable which has a fixed value. An example of deterministic variable is the velocity of a particular car at a particular instance

Discrete and Continuous Random Variables

- Random variables whose set of possible values can be written either as a finite sequence $x_1, \ldots, x_n$ or an infinite sequence $x_1, \ldots$ is said to be discrete (e.g. the set of positive integers)

  Example: Throwing a dice, the discrete random variable is $X = \{1, 2, 3, 4, 5, 6\}$

  Example: If each trial has a probability p of success, the number of trials until success occurs is a discrete random variable $X = \{1, 2, \ldots\}$

- Random variables that take on a continuum of possible values is said to be continuous (e.g. the set of real numbers)

  Example: The amount of time in hours that a computer functions before breaking down is a continuous random variable

Probability Mass Function

- A random variable whose set of possible values is a sequence is said to be discrete. For a discrete random variable X, the probability mass function $p(a)$ of X is

$$p(a) = P\{X = a\}$$

- $p(a)$ is positive for at most a countable number of values of a. If $x_1, x_2, \ldots$ have positive probability,

$$p(x_i) > 0 \quad i = 1,2, \ldots$$

$$p(x) = 0 \quad \text{all other values of } x$$

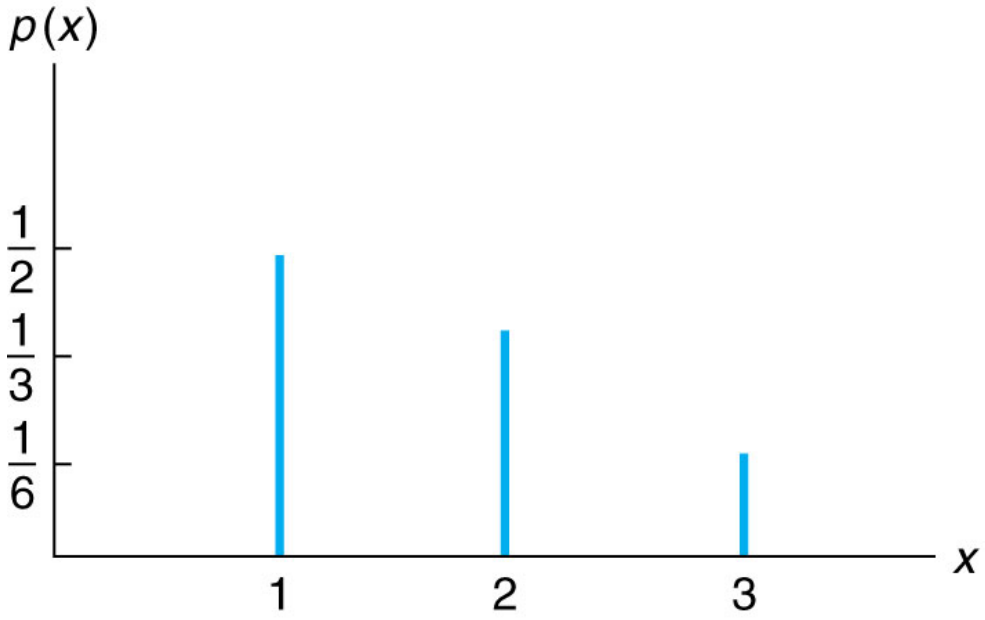
- The sum of the probabilities equals to 1

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

Example

$X = \{1, 2, 3\}$

$P(1) = \frac{1}{2}$ $\quad$ $\mathrm{P}(2) = \frac{1}{3}$ $\quad$ $\mathrm{P}(3) = 1 - \mathrm{P}(1) - \mathrm{P}(2) = \frac{1}{6}$



Probability Density Function

- A random variable whose set of possible values is continuous is said to be continuous. For a continuous random variable, the probability density function (PDF) $f(x)$ is a non-negative function, defined for all real $x \in (-\infty, \infty)$, having the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)\, dx$$

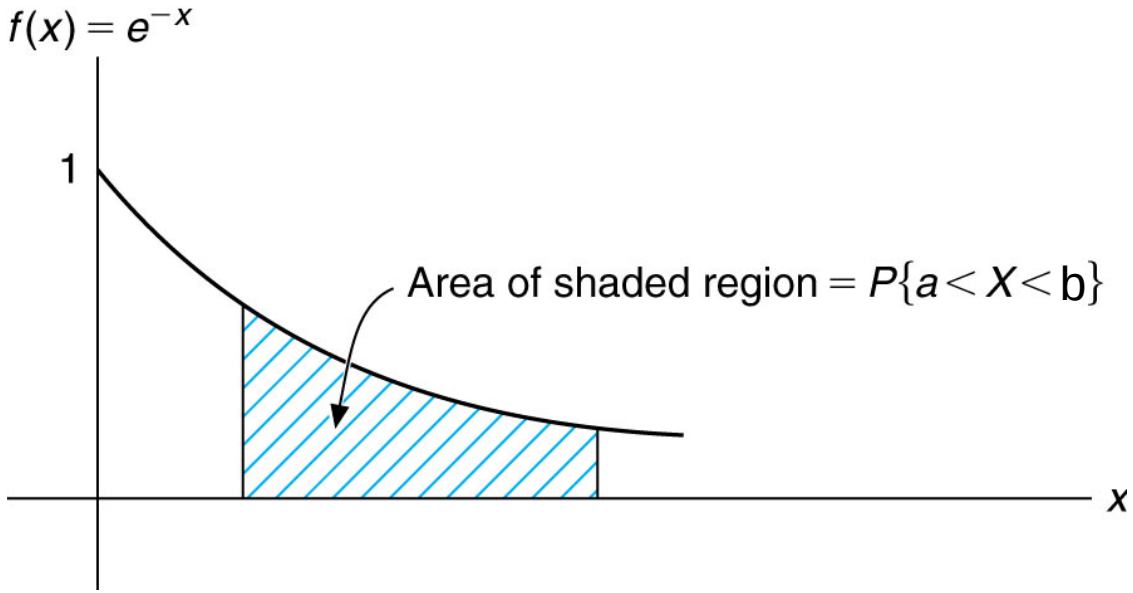
- The "area under the function" is equal to 1, i.e.,

$$1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x) dx$$

- Probability Mass Function can be considered as a special case of Probability Distribution Function of which $f(x)$ is a collection of impulse functions

Example: An exponential decaying continuous random variable

$$f(x) = \begin{cases} e^{-x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$



Note that the "area under the curve" $\int_{-\infty}^{\infty} f(x)dx = 1$

Cumulative distribution function

- The cumulative distribution function (or simply distribution function) of the random variable X is defined for any real number x by,

$$F(x) = P\{X \leq x\}$$

- Denoted by X$\sim F$

- Relationship between the cumulative distribution and the probability distribution function is

$$F(a) = P\{X \in (-\infty, a)\} = \int_{-\infty}^{a} f(x)dx$$

Jointly Distributed Random Variables

- To specify the relationship between two random variables X and Y, we define the joint cumulative probability distribution function of X and Y by

$$F(x, y) = P\{X \leq x, Y \leq y\}$$

- The distribution function of X – call it F_X - can be obtained from $F(x, y)$

$$F_X(x) = P\{X \leq x\} = P\{X \leq x, Y < \infty\} = F(x, \infty)$$

- Similarity,

$$F_Y(y) = F(\infty, y)$$

- Similar expressions for cumulative probability distribution and joint cumulative probability distribution exists in the special case of probability mass function

Independent Random Variables

- The random variables X and Y are said to be independent if for any two sets of real numbers A and B

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}$$

- In other words, X and Y are independent if, for all A and B, the events $E_A = \{X \in A\}$ and $E_B = \{Y \in B\}$ are independent

Expectation

- If X is a discrete random variable taking on the possible values $x_1, x_2, ...$, then the expectation, expected value, or mean of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_i x_i P\{X = x_i\} = \sum_x xp(x)$$

  where $p(x_i) = P\{X = x_i\}$ is the probability mass function
- The expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.
- Using the frequency interpretation of probability, the expected value is the average value if an infinite sequence of independent replications of an experiment is performed
- If X is a continuous random variable,

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx$$

Example

Find $E[X]$ where X is the outcome when we roll a fair die.

$$P(1) = P(2) = \cdots = P(6) = \frac{1}{6}$$

$$E[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = 3.5$$

Properties of the Expected Value

Let $g(x)$ be a (deterministic) real-valued function of x

For the discrete case, since $E[X] = \sum_x xp(x)$

$$E[g(X)] = \sum_x g(x)p(x)$$

For the continuous case, since $E[X] = \int_{-\infty}^{\infty} xf(x)dx$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

When $g(x)$ is a linear function of x

$g(x) = ax + b$

Proposition: $E[aX + b] = aE[X] + b$

Proof: In the discrete case,

$$E[aX + b] = \sum_x (ax + b)p(x)$$
$$= a\sum_x xp(x) + b\sum_x p(x)$$
$$= aE[X] + b$$

The proof is similar for the continuous case

Expected Value of Sums of Random Variables (linearity of expectation)

Proposition: $E[X+Y] = E[X] + E[Y]$

Proof:

$$E[X+Y] = \sum_x \sum_y (x+y)p(x,y) = \sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y)$$

$$= \sum_x x \sum_y p(x,y) + \sum_y y \sum_x p(x,y)$$

$$= \sum_x xp(x) + \sum_y yp(y) = E[X] + E[Y]$$

The proof of the continuous case is similar

$$E[X+Y+Z] = E[(X+Y)+Z]$$
$$= E[X+Y] + E[Z] \quad \text{(by the previous proof)}$$
$$= E[X] + E[Y] + \mathrm{E}[Z]$$

In general,

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

Example

Suppose there are 20 different types of coupons and suppose that each time one obtains a coupon it is equally likely to be any one of the types. Compute the expected number of different types that are contained in a set for 10 coupons.

Let X denote the number of different types in the set of 10 coupons. We compute $E[X]$ by using the representation

$$X = X_1 + \cdots + X_{20}$$

where

$$X_i = \begin{cases} 1 & if\ at\ least\ one\ type\ i\ coupon\ is\ in\ the\ set\ of\ 10 \\ 0 & otherwise \end{cases}$$

$$E[X_i] = 1P\{X_i = 1\} + 0P\{X_i = 0\} = P\{X_i = 1\}$$
$$= P\{at\ least\ one\ type\ i\ coupon\ is\ in\ the\ set\ of\ 10\}$$
$$= 1 - P\{no\ type\ i\ coupons\ are\ in\ the\ set\ of\ 10\}$$
$$= 1 - \left(\frac{19}{20}\right)^{10}$$

$$E[X] = E[X_1] + \cdots + E[X_{20}] = 20\left[1 - \left(\frac{19}{20}\right)^{10}\right] = 8.025$$

Expected Value as the Best Least Square Predictor

Suppose the value of a random variable X is to be predicted. If we predict that X will equal c, then the square of the "error" involved will be $(X - c)^2$

$$E[(X-c)^2] = E[(X-\mu+\mu-c)^2]$$
$$= E[(X-\mu)^2 + 2(\mu-c)(X-\mu) + (\mu-c)^2]$$
$$= E[(X-\mu)^2] + 2(\mu-c)E[X-\mu] + (\mu-c)^2$$

Since $E[X-\mu] = \mathrm{E}[X] - \mu = 0$

$$= E[(X-\mu)^2] + (\mu - c^2)$$
$$\geq E[(X-\mu)^2]$$

Hence the best predictor of a random variable, in terms of minimizing its mean square error, is just its mean

Variance

- The expected value $E[X]$ is also denoted by the mean μ
- If X is a random variable with mean μ, then the variance of X, denoted by $Var(X)$, is

$$Var(X) = E[(X - \mu)^2]$$

- An alternative formula for $Var(X)$ is

$$\begin{aligned} Var(X) &= E[(X - \mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - (E[X])^2 \end{aligned}$$

Property

$$Var(aX + b) = a^2 Var(X)$$

Proof

$$\begin{aligned} Var(aX + b) &= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - a\mu - b)^2] \\ &= E[(aX - a\mu)^2] \\ &= E[a^2(X - \mu)^2] \\ &= a^2 E[(X - \mu)^2] \\ &= a^2 Var(X) \end{aligned}$$

Standard Deviation

The quantity $\sqrt{Var(X)}$ is called the standard deviation of X

It has the same unit as the mean

Example

Compute $Var(X)$ and standard deviation of X when X represents the outcome when we roll a fair die

$$P\{X = i\} = \frac{1}{6} \qquad i = 1, \ldots, 6$$

$$E[X^2] = \sum_{i=1}^{6} i^2 P\{X = i\} = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + \cdots + 6^2\left(\frac{1}{6}\right) = \frac{91}{6}$$

Since $E[X] = \frac{7}{2}$ from previous calculations

$$Var(X) = E[X^2] - (E[X])^2 = \frac{35}{12} \approx 2.917$$

Standard deviation $= \sqrt{Var(X)} \approx 1.7$

Covariance and Variance of Sums of Random Variables

The covariance of two random variables X and Y, written $Cov(X,Y)$, is defined by

$$Cov(X,Y) = E\left[(X-\mu_x)(Y-\mu_y)\right]$$

μ_x and μ_y are the means of X and Y

$$\begin{aligned} Cov(X,Y) &= E\left[XY - \mu_x Y - \mu_y X + \mu_x\mu_y\right] \\ &= E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x\mu_y \\ &= E[XY] - \mu_x\mu_y - \mu_y\mu_x + \mu_x\mu_y \\ &= E[XY] - \mathrm{E}[X]E[Y] \end{aligned}$$

Properties of Covariance

From the definition,

$Cov(X, Y) = Cov(Y, X)$

$Cov(X, X) = Var(X)$

$Cov(aX, Y) = aCov(X, Y)$ for any constant a

$$Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$$

Proof:

$$Cov(X_1 + X_2, Y)$$
$$= E[(X_1 + X_2)Y] - E[X_1 + X_2]E[Y]$$
$$= E[X_1 Y] + E[X_2 Y] - (E[X_1] + E[X_2])E[Y]$$
$$= E[X_1 Y] - E[X_1]E[Y] + E[X_2 Y] - E[X_2]E[Y]$$
$$= Cov(X_1, Y) + Cov(X_2, Y)$$

It can be generalized to

$$Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_j) \qquad (1)$$

$$Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y) + \text{Cov}(\text{Y},\text{X})$$

Proof

As $Cov(X,X) = Var(X)$, $Var(X+Y) = Cov(X+Y, X+Y)$

From equation (1) above,

$$\begin{aligned} Var(X+Y) &= Cov(X,X) + Cov(Y,Y) + Cov(X,Y) + Cov(Y,X) \\ &= Var(X) + Var(Y) + Cov(X,Y) + \text{Cov}(\text{Y},\text{X}) \end{aligned}$$

It can be generalized to

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} Cov\left(X_i, X_j\right)$$

Covariance for independent variables

If X and Y are independent random variables,

$$Cov(X, Y) = 0$$

Proof

$Cov(X, Y) = E[XY] - E[X]E[Y]$

In the discrete case

$$E[XY] = \sum_j \sum_i x_i y_j P\{X = x_i, Y = y_j\}$$

$$= \sum_j \sum_i x_i y_j P\{X = x_i\} P\{Y = y_j\}$$

$$= \sum_j y_j P\{Y = y_j\} \sum_i x_i P\{X = x_i\} = E[Y]E[X]$$

A similar argument holds for the continuous case

For independent X and Y

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

Proof

$$Var(X+Y) = Var(X) + Var(Y) + Cov(X,Y) + \mathrm{Cov(Y,X)}$$

For independent X and Y, $Cov(X,Y) = Cov(Y,X) = 0$

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

In general, for independent X_i

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

The covariance of two random variables is important as an indicator of the relationship between them. Consider the situation where X and Y are indicator variables for whether or not the events A and B occur.

$$X = \begin{cases} 1 & if\ A\ occurs \\ 0 & otherwise \end{cases} \qquad Y = \begin{cases} 1 & if\ B\ occurs \\ 0 & otherwise \end{cases}$$

$$XY = \begin{cases} 1 & if\ X = 1, Y = 1 \\ 0 & otherwise \end{cases}$$

$$\begin{aligned} Cov(X,Y) &= E[XY] - E[X]E[Y] \\ &= P\{X = 1, Y = 1\} - P\{X = 1\}P\{Y = 1\} \end{aligned}$$

$$\begin{aligned} Cov(X,Y) > 0 &\Leftrightarrow P\{X = 1, Y = 1\} > P\{X = 1\}P\{Y = 1\} \\ &\Leftrightarrow \frac{P\{X = 1, Y = 1\}}{P\{X = 1\}} > P\{Y = 1\} \\ &\Leftrightarrow P\{Y = 1 | X = 1\} > P\{Y = 1\} \end{aligned}$$

- $Cov(X, Y) > 0$ implies that the outcome $X = 1$ makes it more likely that $Y = 1$.
- $Cov(X, Y) < 0$ implies that the outcome $X = 1$ makes it less likely that $Y = 1$.
- By symmetry, the argument stills holds if X and Y are exchanged
- In general, a positive value of $Cov(X, Y)$ is an indication that Y tends to increase as X does, whereas a negative value indicates the opposite

Correlation

The strength of the relationship between X and Y is indicated by the correlation between X and Y , a dimensionless quantity obtained by dividing the covariance by the product of the standard deviations of X and Y

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

$Corr(X,Y)$ has a value between -1 and +1

Moment Generating Functions

The moment generating function $\phi(t)$ of the random variable X is defined for all values t by

$$\phi(t) = E[e^{tX}] = \begin{cases} \sum_x e^{tx} p(x) & if\ X\ is\ discrete \\ \int_{-\infty}^{\infty} e^{tx} f(x)dx & if\ X\ is\ continuous \end{cases}$$

We call $\phi(t)$ the moment generating function because all of the moments of X can be obtained by successively differentiating $\phi(t)$. For example,

$$\phi'(t) = \frac{d}{dt} E[e^{tX}] = E\left[\frac{d}{dt}(e^{tX})\right] = E[Xe^{tX}]$$

Hence $\phi'(0) = E[X]$

Similarly, compute $\phi''(t)$

$$\phi''(t) = \frac{d}{dt}\phi'(t) = \frac{d}{dt}E[Xe^{tX}] = E\left[\frac{d}{dt}(Xe^{tX})\right] = E[X^2e^{tX}]$$

and so

$$\phi''(0) = E[X^2]$$

In general, the nth derivative of $\phi(t)$ evaluated at $t = 0$ equals $E[X^n]$. That is

$$\phi^n(0) = E[X^n] \qquad n \geq 1$$

An important property of moment generating function is that the moment generating function of the sum of independent random variables is just the product of the individual moment generating functions. To see this, suppose that X and Y are independent and have moment generating functions $\phi_X(t)$ and $\phi_Y(t)$ respectively. Then $\phi_{X+Y}(t)$, the moment generating function of $X + Y$, is

$$\phi_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E[e^{tX}e^{tY}]$$
$$= E[e^{tX}]E[e^{tY}] \quad \text{since } X \text{ and } Y \text{ are independent}$$
$$= \phi_X(t)\phi_Y(t)$$

Another important result is that the moment generating function uniquely determines the distribution. That is, there exists a one-to-one correspondence between the moment generating function and the distribution function of a random variable

Markov's Inequality

If X is a random variable that takes only nonnegative values, then for any value $a > 0$

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

Proof

We give a proof for the case where X is continuous with density f

$$\begin{aligned} E[X] &= \int_0^\infty xf(x)dx \\ &= \int_0^a xf(x)dx + \int_a^\infty xf(x)dx \\ &\geq \int_a^\infty xf(x)dx \\ &\geq \int_a^\infty af(x)dx \\ &= a\int_a^\infty f(x)dx = aP\{X \geq a\} \end{aligned}$$

Chebyshev's Inequality

If X is a random variable with mean μ and variance σ^2, then for any value $k > 0$

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

Proof

Since $(X - \mu)^2$ is a nonnegative random variable, we can apply Markov's inequality(with $a = k^2$) to obtain

$$P\{(X - \mu)^2 \geq k^2\} \leq \frac{E[(X - \mu)^2]}{k^2}$$

However, since $(X - \mu)^2 \geq k^2$ if and only if $|X - \mu| \geq k$. The equation is equivalent to

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

The importance of Markov's and Chebyshev's inequalities is that they enable us to derive bounds on probabilities when only the mean, or both the mean and the variance, of the probability distribution are known. Of course, if the actual distribution were known, then the desired probabilities could be exactly computed and we would not need to resort to bounds.

Example

Suppose that it is known that the number of items produced by a factory during a week is a random variable with mean 50.

a) What can be said about the probability that this week's production will exceed 75?

b) If the variance of a week's production is 25, what can be said about the probability that this week's production will be between 40 and 60?

a) By Markov's inequality

$$P\{X > 75\} \leq \frac{E[X]}{75} = \frac{50}{75} = \frac{2}{3}$$

b) By Chebyshev's inequality

$$P\{|X - 50| \geq 10\} \leq \frac{\sigma^2}{10^2} = \frac{1}{4}$$

Hence

$$P\{|X - 50| < 10\} \geq 1 - \frac{1}{4} = \frac{3}{4}$$

The Weak Law of Large Numbers

Let $X_1, X_2, \ldots,$ be a sequence of independent and identically distributed random variables, each having mean $E[X_i] = \mu$. Then for any $\epsilon > 0$,

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \to 0 \qquad as\ n \to \infty$$

Proof (for finite σ only)

$$E\left[\frac{X_1 + \cdots + X_n}{n}\right] = \mu$$

$$Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{\sigma^2}{n}$$

It follows from Chebyshev's inequality that

$$P\left\{\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| > \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2}$$

Interpretation

$\frac{X_1+\cdots+X_n}{n}$ can be interpreted as the sample mean of n samples of the same random variable X

The weak law informs us that no matter how small is ϵ, for large n, the sample mean will with probability close to 1 be identical to the actual mean

References

- Text book, Ch. 4