February 14, 2018

CS 361: Probability & Statistics

Random variables

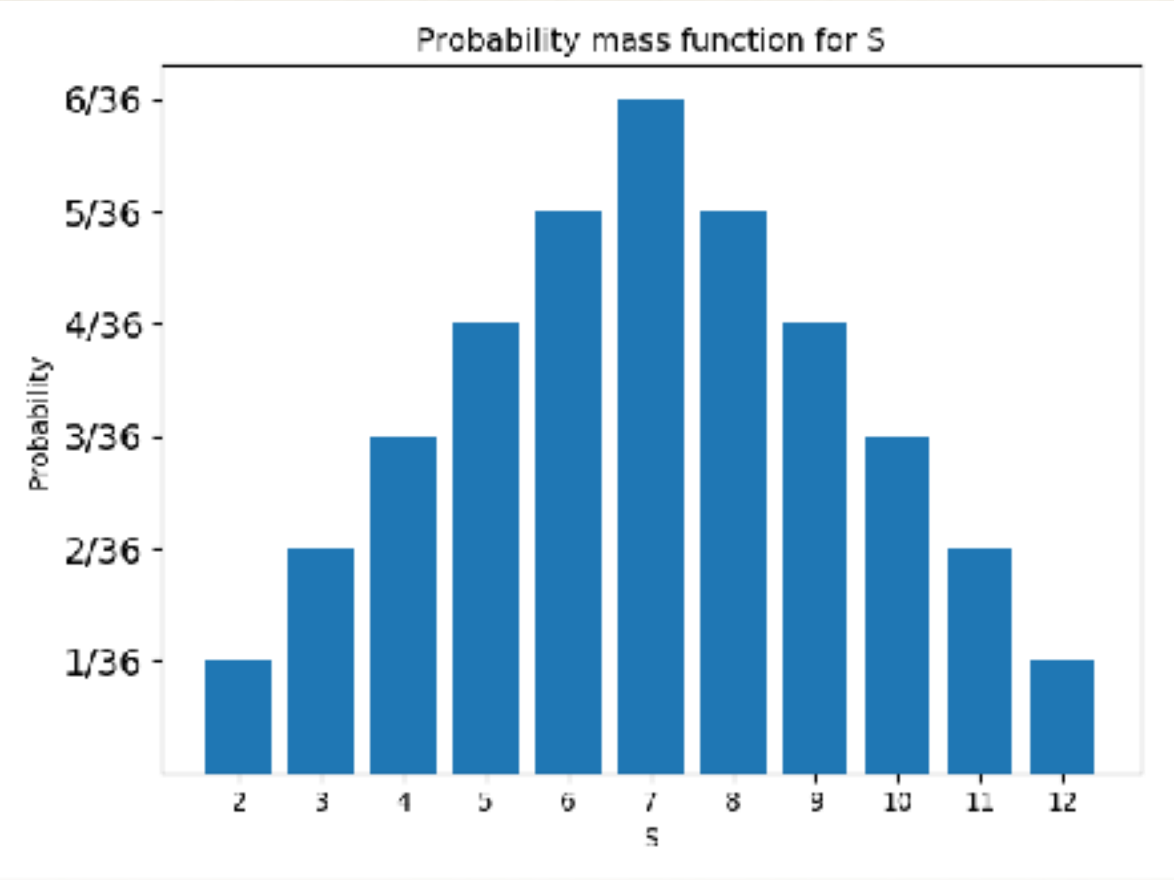
Sum and difference

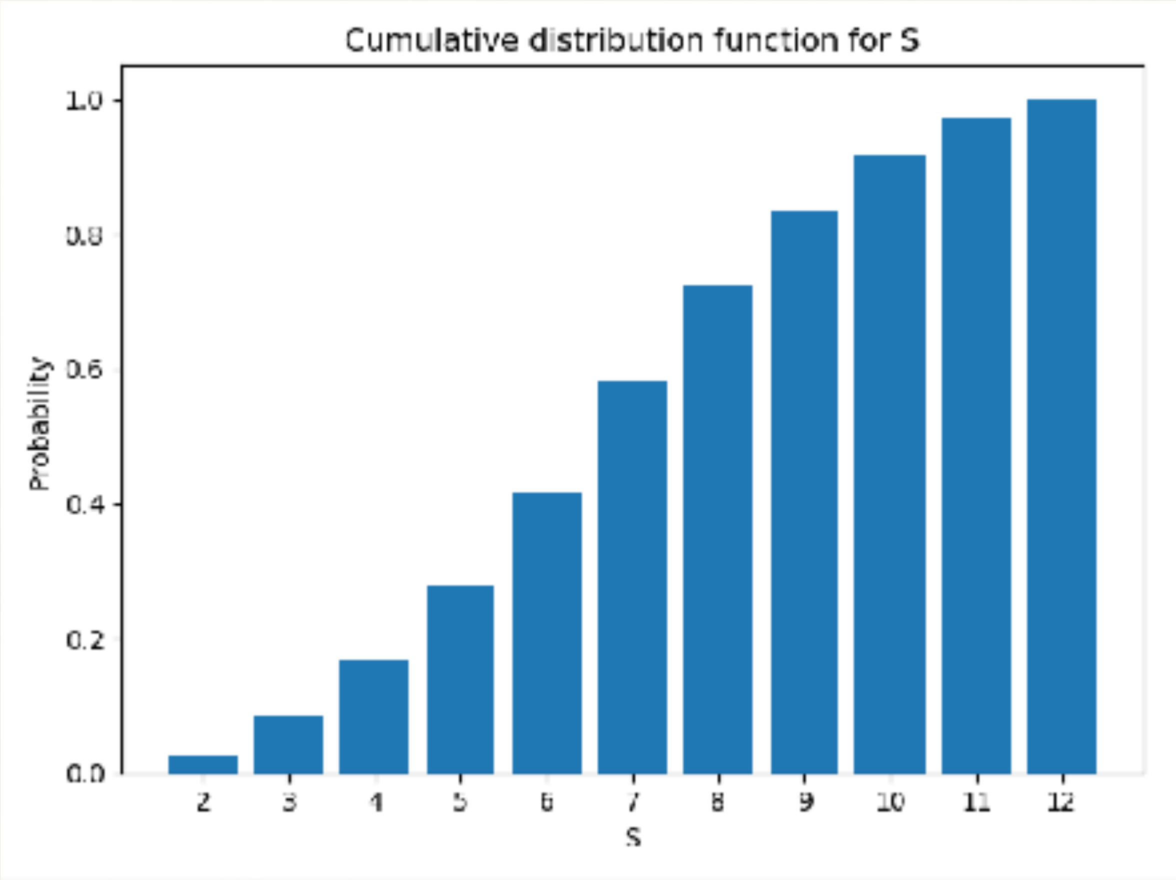
Throw two dice. The number of spots on the first die is a random variable X, the number on the second is a random variable Y. Let S be the random variable given by S = X + Y and D = X - Y

What is the probability distribution of S?

S	P(s)
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

Sum and difference





Sum and difference

Throw two dice. The number of spots on the first die is a random variable X, the number on the second is a random variable Y. Let S be the random variable given by $S = X + Y$ and $D = X - Y$

What is the probability distribution of D?

D	P(d)
-5	1/36
-4	2/36
-3	3/36
-2	4/36
-1	5/36
0	6/36
1	5/36
2	4/36
3	3/36
4	2/36
5	1/36

Sum and difference, joint distribution

Throw two dice. The number of spots on the first die is a random variable X, the number on the second is a random variable Y. Let S be the random variable given by S = X + Y and D = X - Y

What does is the joint distribution of S and D?

$$\frac{1}{36} \times \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

TABLE 5.1: *A table of the joint probability distribution of S (vertical axis; scale $2, \ldots, 12$) and D (horizontal axis; scale $-5, \ldots, 5$) from example 5.4*

Sum and difference, independence

Throw two dice. The number of spots on the first die is a random variable X, the number on the second is a random variable Y. Let S be the random variable given by $S = X + Y$ and $D = X - Y$

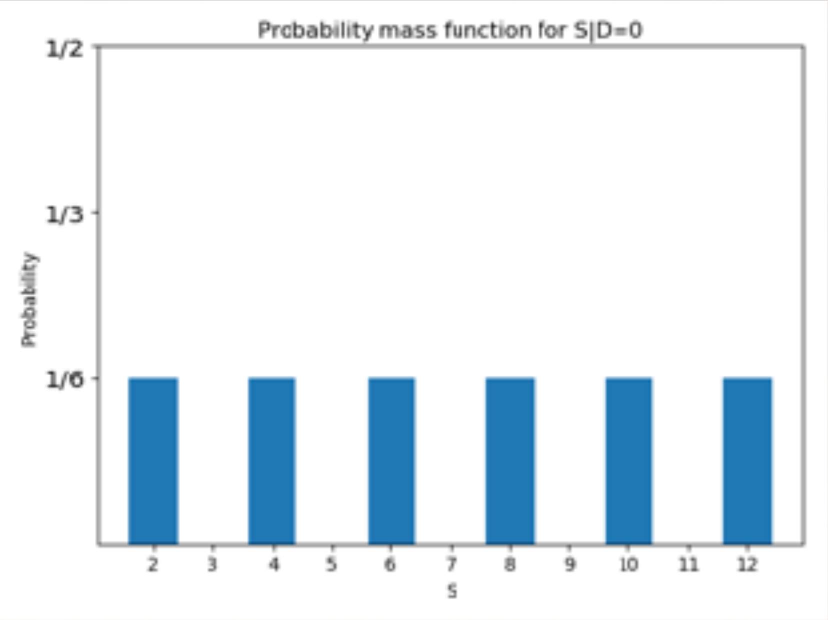
Are X and Y independent?

How about S and D?

Sum and difference, conditional

Throw two dice. The number of spots on the first die is a random variable X, the number on the second is a random variable Y. Let S be the random variable given by S = X + Y and D = X - Y

What is P(S | D=0)?



What is P(D | S=11)

D	P(d)
-1	1/2
1	1/2

Indicator random variables

Any arbitrary function that looks at an outcome and spits out a number is a random variable

One useful random variable that we can define is the indicator random variable for an event A

This is a random variable that looks at an outcome and spits out a 1 if the outcome is in event A and 0 otherwise

Example

Flip a fair coin 3 times and consider the indicator random variable X for the event A “2 heads have come up”

X outputs a 0 for the outcome TTH, X outputs a 1 for the outcome HHT

Expected value

Expected value: motivation

Think about the game we mentioned earlier. Flip a coin, heads with probability p, tails with (1-p). When it's heads you pay me q, tails I pay you r.

Should we bother to play this game?

Recall our relative frequency interpretation of probabilities

$$Npq - N(1-p)r$$

Approx payout for N games

$$pq - (1-p)r$$

Average payout per game

Expected value

Definition: 5.6 *Expected value*

Given a discrete random variable X which takes values in the set $\mathcal{D}$ and which has probability distribution P, we define the expected value

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} x P(X = x).$$

This is sometimes written which is $\mathbb{E}_P[X]$, to clarify which distribution one has in mind

The expected value of a random variable X is a weighted average of the values that X can take on

Example

- Flip a fair coin, $P(T)=P(H)=1/2$, if it comes up heads you pay me 1, if it comes up tails, I pay you 1. Let X be the random variable corresponding to my income
- $$\begin{aligned} E[X] &= 1/2(1) + 1/2(-1) \\ &= 0 \end{aligned}$$

Example

- Flip a fair coin, heads you pay me 2, tails I pay you 1
- $E[X] = (1/2)(2) - (1/2)(1)$
  $= 1/2$
- My "expected income" from a game is 1/2, which we note is not an amount I could actually earn in one run of the game
- Playing this game is good for me and bad for you

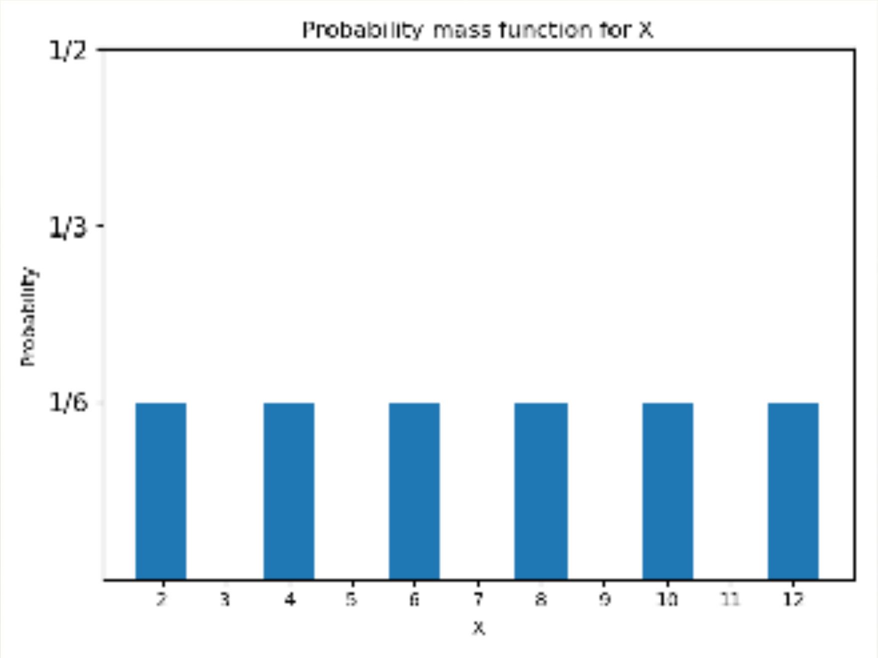
Expectation

- If f is a function of a random variable, then f(X) is a random variable as well. And we can write the expectation of f(X) as

$$E[f] = \sum_{x \in D} f(x) P(X = x)$$

Example

The probability mass function for a random variable X is depicted below, what is E[X^2]?



$$E[X^2] = 2^2(1/6) + 4^2(1/6) + 6^2(1/6) + 8^2(1/6) + 10^2(1/6) + 12^2(1/6)$$

E[X^2] = 66.67

Linearity of expectation

- For random variables X and Y, and constant *k* we have

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

$$E[kX] = kE[X]$$

Mean and variance

- The expected value of a random variable is also called the **mean**
- The **variance** of a random variable is defined as

$$\mathrm{var}[X] = E[(X - E[X])^2]$$

Properties of variance

Useful Facts: 5.3 *Variance*

1. For any constant k, $\mathsf{var}[k] = 0$
2. $\mathsf{var}[X] \geq 0$
3. $\mathsf{var}[kX] = k^2 \mathsf{var}[X]$
4. if X and Y are independent, then $\mathsf{var}[X+Y] = \mathsf{var}[X] + \mathsf{var}[Y]$

You will prove some of these

Standard deviation

Definition: 5.13 *Standard deviation*

The **standard deviation** of a random variable X is defined as

$$\mathsf{std}\left(\{X\}\right) = \sqrt{\mathsf{var}[X]}$$

Covariance

- The **covariance** of X and Y is defined as

$$\mathrm{cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$$

- Note that cov(X,X) = var[X]

Another expression for variance

By definition we have

$$var[X] = E[(X - E[X])^2]$$

Which we rewrite to reduce confusion

$$var[X] = E[(X - \mu_X)^2]$$

Expanding

$$var[X] = E[X^2 - 2X\mu_X + \mu_x^2]$$

Using linearity of expectation

$$var[X] = E[X^2] - 2\mu_X E[X] + \mu_x^2$$

Or

$$var[X] = E[X^2] - 2E[X]E[X] + E[X]^2$$

Giving our final expression

$$var[X] = E[X^2] - (E[X])^2$$

Another expression for covariance

By definition, we have

$$\mathrm{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Rewriting

$$\mathrm{cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$
$$\mathrm{cov}(X,Y) = E[XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y]$$

Using linearity of expectation

$$\mathrm{cov}(X,Y) = E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X\mu_Y$$

Which is

$$\mathrm{cov}(X,Y) = E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y]$$

Giving our final result

$$\boxed{\mathrm{cov}(X,Y) = E[XY] - E[X]E[Y]}$$

A couple of results

Useful Facts: 5.6 *Variance and Covariance*

1. if X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.
2. if X and Y are independent, then $\mathsf{cov}(X, Y) = 0$.

Proofs are straightforward, in the book

Example: coin flip

- We flip a biased coin that has P(H)=p and P(T) = 1-p and let X be a random variable with value 1 for heads and 0 for tails
- What is E[X]?

$$1p + 0(1-p) = p$$

- What is Var[X]?

$$\mathrm{var}[X] = E[X^2] - E[X]^2$$

$$E[X^2] = 1^2 p + 0^2(1-p)$$

$$E[X]^2 = p^2$$

$$\mathrm{var}[X] = p - p^2$$

Overloaded terms?

- You may have noticed we are using some terms like mean, variance, standard deviation, that we have used before
- Earlier we developed concepts for datasets and now we have concepts for random variables
- They aren't very different concepts
- If we suppose that each data item in a dataset has probability 1/N and consider the random variable given by reporting the value of the data item. These concepts will be equal

$$\mathbb{E}[x] = \sum_i x_i p(x_i) = \frac{1}{N} \sum_i x_i = \mathsf{mean}\left(\{x\}\right).$$

Towards the weak law of large numbers

Markov's inequality

This theorem says that for any random variable X and any value a, we have

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

A random variable is unlikely to have an absolute value much larger than the mean of its absolute value

If, for instance, we took a = 10 E[| X |], we'd get

$$P(\{|X| \geq 10E[|X|]\}) \leq .10$$

Indicator functions

- To make Markov's inequality easy to prove we will recall the useful notion of an indicator function

Definition: 5.16 *Indicator functions*

An indicator function for an event is a function that takes the value zero for values of X where the event does not occur, and one where the event occurs. For the event $\mathcal{E}$, we write

$$\mathbb{I}_{[\mathcal{E}]}(X)$$

for the relevant indicator function.

- We have following fairly immediately, since I is 0 everywhere but on $\mathcal{E}$ $\mathbb{E}[\mathbb{I}_{[\mathcal{E}]}] = P(\mathcal{E})$

Proving Markov

Proving

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

First note that for a>0 we have

$$a\mathbb{I}_{[\{|X|\geq a\}]}(X) \leq |X|$$

Since I(X) will be 0 for any X < a and 1 for any value >= a

Taking expectations, we get

$$E[a\mathbb{I}_{[\{|X|\geq a\}]}(X)] \leq E[|X|]$$

We can pull out the a due to the linearity of expectation

$$E[\mathbb{I}_{[\{|X|\geq a\}]}(X)] \leq \frac{E[|X|]}{a}$$

And finally using

$$\mathbb{E}\left[\mathbb{I}_{[\mathcal{E}]}\right] = P(\mathcal{E})$$

Our left hand side is

$$P(|X| \geq a)$$

Which gives the desired result

Chebyshev's inequality

For any random variable X and value a

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\mathrm{var}[X]}{a^2}$$

Or if we let σ be the standard deviation of X and substitute $a = k\sigma$ we have

$$P(\{|X - E[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

The probability that X is greater than k standard deviations from the mean is small. Look familiar?

Proving Chebyshev

Proving

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\mathrm{var}[X]}{a^2}$$

Write U for the random variable $(X - E[X])^2$

Markov's inequality tells us then that for any w, we will have

$$P(\{|U| \geq w\}) \leq \frac{E[|U|]}{w}$$

Using $w = a^2$ we have

$$P(\{|U| \geq w\}) = P(\{|X - E[X]| \geq a\}$$

Also from the definition of U, w, and variance we have

$$\frac{E[|U|]}{w} = \frac{\mathrm{var}[X]}{a^2}$$

So substituting we get, as desired

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\mathrm{var}[X]}{a^2}$$

Sampling

So far we have been working on a number of problems where we suppose ahead of time that we know the distributions of the outcomes and hence the random variables in our experiments.

We eventually want to get to a place where we don't make that assumption and we guess what the distribution is after observing some runs of an experiment

In this context and others we will refer to our observations of experiments as **trials** or **samples** from the underlying distribution

Sampling: IID

If we have a set of data items x_i meeting the following:

a) they are independent

b) they were generated by the same process

c) the histogram of a very large set of the items looks increasingly like the probability distribution P(X) of some random variable X

We call this set of data items **independent, identically distributed samples of P(X)** or IID for short

Expectation of iid samples

Assume we have a set of N iid samples of a probability distribution P(X)

$$X_N = \frac{\sum_{i=1}^{N} x_i}{N}$$

X_N is a random variable, and if we take the expectation of both sides we get

$$E[X_N] = E\left[\frac{\sum_{i=1}^{N} x_i}{N}\right]$$

Using linearity, we get

$$E[X_N] = \frac{1}{N}\sum_{i=1}^{N} E[x_i]$$

But since x_i is a sample drawn from X

$$E[X_N] = \frac{1}{N}\sum_{i=1}^{N} E[X]$$

Simplifying we get

$$E[X_N] = E[X]$$

Variance of iid samples

Assume that X has a finite variance given by σ^2 . Let's find the variance of X_N

$$\mathrm{var}[X_N] = \mathrm{var}\left[\frac{\sum_{i=1}^{N} x_i}{N}\right]$$

Recall the property of variance $\mathrm{var}[kX] = k^2\mathrm{var}[X]$ to get

$$\mathrm{var}[X_N] = \frac{1}{N^2}\mathrm{var}\left[\sum_{i=1}^{N} x_i\right]$$

Since the x_i are drawn independently from X we can use the fact that the variance of a sum of independent variables can be broken up into a sum of variances

$$\mathrm{var}[X_N] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{var}[x_i]$$

Substituting

$$\mathrm{var}[X_N] = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2$$

And simplifying, we get

$$\mathrm{var}[X_N] = \frac{\sigma^2}{N}$$

Weak law of large numbers

With $X_N = \frac{\sum_{i=1}^{N} x_i}{N}$

The **weak law of large numbers** states that, if $P(X)$ has finite variance, then for any positive number ϵ

$$\lim_{N \to \infty} P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) = 0.$$

Equivalently, we have

$$\lim_{N \to \infty} P(\{|X_N - \mathbb{E}[X]| < \epsilon\}) = 1.$$

Each form means that, for a large enough set of IID samples, the average of the samples (i.e. X_N) will, with high probability, be very close to the expectation $\mathbb{E}[X]$.