

Chebyshev inequality and its proof (Lecture 11)

Markov inequality. Let $t > 0$, and X be a **nonnegative random variable whose expected value exists**, that is, for which $X \geq 0$ holds for sure, and $\mathbb{E}(X)$ exists.

Then

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

Chebyshev inequality and its proof (Lecture 11)

Markov inequality. Let $t > 0$, and X be a **nonnegative random variable whose expected value exists**, that is, for which $X \geq 0$ holds for sure, and $\mathbb{E}(X)$ exists. Then

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

Chebyshev inequality. Let $t > 0$, and X be a random **random variable whose standard deviation exists**, that is, for which $D(X)$ exists. Then we have

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{D^2(X)}{t^2}.$$

Proof. Let $Z = (X - \mathbb{E}(X))^2$. This random variable is nonnegative and has finite expectation, hence we can apply the Markov inequality, for the positive number $t^2 > 0$:

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}(X)| \geq t) &= \mathbb{P}((X - \mathbb{E}(X))^2 \geq t^2) = \mathbb{P}(Z \geq t^2) \leq \\ &\stackrel{\text{Markov}}{\leq} \frac{\mathbb{E}(Z)}{t^2} = \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{t^2} = \frac{D^2(X)}{t^2} \end{aligned}$$

based on the definition of variance.

Application of Chebyshev inequality

What is the minimal number of people to ask (supposing that everyone answers honestly), such that, **for every** p , the probability that the proportion of vegetarian people in the sample differs with **at most 1%** from p is **at least 95%**?

n participants, everyone is vegetarian with probability p

X : number of vegetarian people among the participants

Necessary condition:

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0.01\right) \geq 0.95$$

holds for every $0 \leq p \leq 1$ (we do not know p)

Application of Chebyshev inequality

n participants, everyone is vegetarian with probability p

X : number of vegetarian people among the participant

Since X has binomial distribution:

$$\mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n} \cdot np = p; \quad D\left(\frac{X}{n}\right) = \frac{1}{n} \sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}}.$$

Application of Chebyshev inequality

n participants, everyone is vegetarian with probability p

X : number of vegetarian people among the participant

Since X has binomial distribution:

$$\mathbb{E}\left(\frac{X}{n}\right) = \frac{1}{n} \cdot np = p; \quad D\left(\frac{X}{n}\right) = \frac{1}{n} \sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}}.$$

Chebyshev inequality for the random variable X/n :

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \geq 0.01\right) \leq \frac{D^2\left(\frac{X}{n}\right)}{0.01^2} = \frac{p(1-p)}{0.01^2 \cdot n} \leq \frac{1}{4 \cdot 0.01^2 \cdot n},$$

because $p(1-p) \leq 1/4$ always holds (e.g. by the arithmetic-geometric mean).

Application of Chebyshev inequality

n participants, everyone is vegetarian with probability p

X : number of vegetarian people among the participant

Chebyshev inequality implies that

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \geq 0.01\right) \leq \frac{1}{4 \cdot 0.01^2 \cdot n}$$

Necessary condition:

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0.01\right) \geq 0.95,$$

that is, we have

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| > 0.01\right) \leq 0.05.$$

It follows that **the following is sufficient**:

$$\frac{1}{4 \cdot 0.01^2 \cdot n} \leq 0.05 \quad \Leftrightarrow \quad n \geq \frac{1}{4 \cdot 0.01^2 \cdot 0.05} = 50000.$$

Application of Chebyshev inequality

n participants, everyone is vegetarian with probability p

X : number of vegetarian people among the participant

Chebyshev inequality implies that

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \geq 0.01\right) \leq \frac{1}{4 \cdot 0.01^2 \cdot n}$$

Necessary condition:

$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| \leq 0.01\right) \geq 0.95,$$

that is, we have

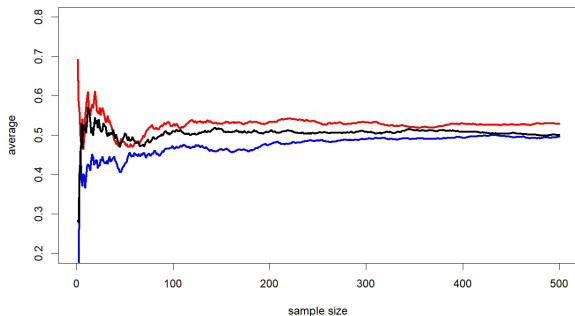
$$\mathbb{P}\left(\left|\frac{X}{n} - p\right| > 0.01\right) \leq 0.05.$$

It follows that **the following is sufficient**:

$$\frac{1}{4 \cdot 0.01^2 \cdot n} \leq 0.05 \quad \Leftrightarrow \quad n \geq \frac{1}{4 \cdot 0.01^2 \cdot 0.05} = 50000.$$

If 0.01 would be replaced by 0.005 (its half), $n \geq 200000$ (four times 50000) would be the lower bound.

Convergence of the average



Average of a sample with uniform distribution on the interval $[0, 1]$, as a function of the sample size until $n = 500$, for three different samples

Types of convergence

Sequences of random variables might converge with respect to **different definitions**.

The sequence of random variables Z_1, Z_2, \dots , **converges in probability** to random variable Z if for every $\varepsilon > 0$ the following holds:

$$\mathbb{P}(|Z_n - Z| > \varepsilon) \rightarrow 0$$

as $n \rightarrow \infty$.

Types of convergence

Sequences of random variables might converge with respect to **different definitions**.

The sequence of random variables Z_1, Z_2, \dots , **converges in probability** to random variable Z if for every $\varepsilon > 0$ the following holds:

$$\mathbb{P}(|Z_n - Z| > \varepsilon) \rightarrow 0$$

as $n \rightarrow \infty$.

The sequence of random variables Z_1, Z_2, \dots , converges **with probability 1** to random variable Z if

$$\mathbb{P}(\omega \in \Omega : Z_n(\omega) \rightarrow Z(\omega) \text{ as } n \rightarrow \infty) = 1.$$

Types of convergence

Sequences of random variables might converge with respect to **different definitions**.

The sequence of random variables Z_1, Z_2, \dots , **converges in probability** to random variable Z if for every $\varepsilon > 0$ the following holds:

$$\mathbb{P}(|Z_n - Z| > \varepsilon) \rightarrow 0$$

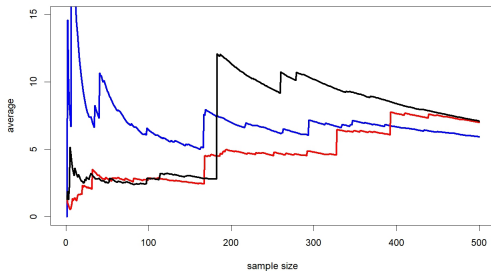
as $n \rightarrow \infty$.

The sequence of random variables Z_1, Z_2, \dots , converges **with probability 1** to random variable Z if

$$\mathbb{P}(\omega \in \Omega : Z_n(\omega) \rightarrow Z(\omega) \text{ as } n \rightarrow \infty) = 1.$$

The sequence converges with probability 1 \Rightarrow it converges in probability, but the other direction is not true.

Behavior of the average



The average as the function of sample size in a case when the expected value does not exist

Weak law of large numbers: proof

Let X_1, \dots, X_n be independent identically distributed random variables with finite variance. Let $m = \mathbb{E}(X_1)$ and $\sigma = D(X_1)$.

As we have seen earlier:

$$\mathbb{E}(\bar{X}) = m; \quad D^2(\bar{X}) = \frac{\sigma^2}{n}.$$

Weak law of large numbers: proof

Let X_1, \dots, X_n be independent identically distributed random variables with finite variance. Let $m = \mathbb{E}(X_1)$ and $\sigma = D(X_1)$.

As we have seen earlier:

$$\mathbb{E}(\bar{X}) = m; \quad D^2(\bar{X}) = \frac{\sigma^2}{n}.$$

Chebyshev inequality implies that for every $\varepsilon > 0$ we have

$$\mathbb{P}(|\bar{X} - m| > \varepsilon) \leq \frac{D^2(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2 n} \rightarrow 0 \quad (n \rightarrow \infty).$$

Hence $\bar{X} \rightarrow m = \mathbb{E}(X_1)$ in probability.

Laws of large numbers

Theorem (Weak law of large numbers)

Let X_1, X_2, \dots be independent and identically distributed random variables. Suppose that $D(X_1) < \infty$. Then for every $\varepsilon > 0$ we have

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

that is, $\bar{X}_n \rightarrow \mathbb{E}(X_1)$ in probability.

Laws of large numbers

Theorem (Weak law of large numbers)

Let X_1, X_2, \dots be independent and identically distributed random variables. Suppose that $D(X_1) < \infty$. Then for every $\varepsilon > 0$ we have

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X_1)| > \varepsilon) \rightarrow 0 \quad (n \rightarrow \infty),$$

that is, $\bar{X}_n \rightarrow \mathbb{E}(X_1)$ in probability.

Theorem (Strong law of large numbers)

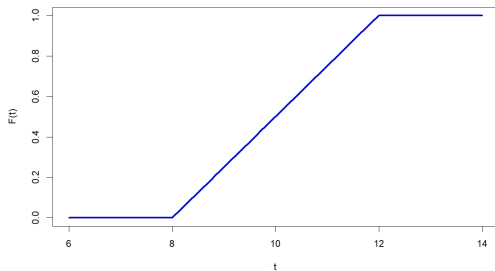
Let X_1, X_2, \dots be independent and identically distributed random variables. Suppose that $m = \mathbb{E}(X_1) < \infty$. Then we have

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} \rightarrow \mathbb{E}(X_1) = m$$

holds with probability 1 as $n \rightarrow \infty$.

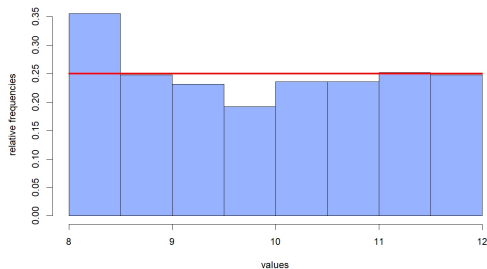
In the second version, a stronger statement follows from a weaker condition.

Uniform distribution



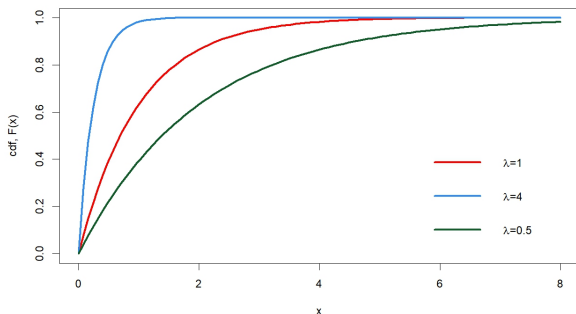
Cumulative distribution function of the uniform distribution on the interval $[8, 12]$

Uniform distribution



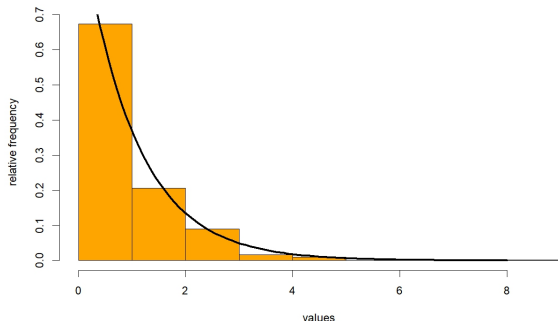
Histogram of a sample of size 500 with uniform distribution on the interval [8, 12]

Exponential distribution



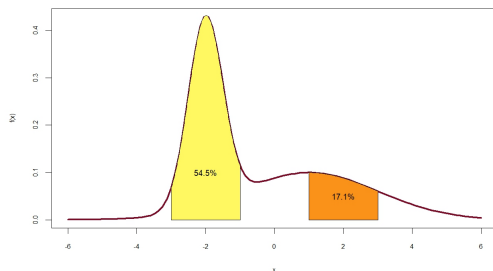
Cumulative distribution function of exponential distribution with different parameters: ($\lambda = \frac{1}{2}, 1$, and 4)

Exponential distribution



Histogram of a sample of size 500 from exponential distribution with parameter $\lambda = 1$

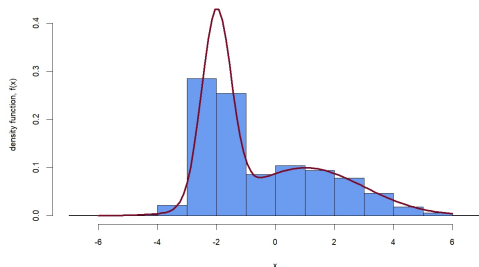
Density function



If random variable X has density function f (shown on the figure): $\mathbb{P}(-3 \leq X \leq -1) = \int_{-3}^{-1} f(x)dx = 54.5\%$;

$\mathbb{P}(1 \leq X \leq 3) = \int_1^3 f(x)dx = 17.1\%$.

Density function



A density function and the histogram of a sample of size 1000 from the same distribution;

higher values of the density function \rightarrow higher frequency;

sample: independent random variables such that all have density function f

Density function: definition

Random variable $X : \Omega \rightarrow \mathbb{R}$ has **density function** $f : \mathbb{R} \rightarrow \mathbb{R}$ if

$$\mathbb{P}(X \leq t) = \int_{-\infty}^t f(x) dx$$

holds for every real number $t \in \mathbb{R}$.

Density function: definition

Random variable $X : \Omega \rightarrow \mathbb{R}$ has **density function** $f : \mathbb{R} \rightarrow \mathbb{R}$ if

$$\mathbb{P}(X \leq t) = \int_{-\infty}^t f(x) dx$$

holds for every real number $t \in \mathbb{R}$.

There exists random variables that do not have density function, for example, the discrete ones. If X **has density function**, then we say that X is **absolutely continuous**.

If random variable X has density function f , then for all real numbers $a < b$ we have

$$\mathbb{P}(a < X < b) = \mathbb{P}(a \leq X \leq b) = \int_a^b f(x) dx.$$