Statistics for Data Science

Wintersemester 2024/25

Vesa Kaarnioja
vesa.kaarnioja@fu-berlin.de

FU Berlin, FB Mathematik und Informatik

Fifth lecture, November 11, 2024

Inequalities and limits

Random sample / i.i.d. random variables

Let $X_1, \ldots, X_n$ be random variables. We call $X_1, \ldots, X_n$ a random sample if the random variables are independent and identically distributed (i.i.d.).

- Independent means that $X_1, \ldots, X_n$ are mutually independent random variables.
- Identically distributed means that $X_1, \ldots, X_n$ all have the same law.

Often, we specify the law (probability distribution) of a random variable X and say that $X_1, \ldots, X_n$ are i.i.d. copies of X.

Example

Let $X \sim \mathcal{N}(0,1)$. Suppose that $X_1, \ldots, X_n$ are i.i.d. copies of X. This means that

- $X_i \sim \mathcal{N}(0,1)$ for all $i = 1, \ldots, n$ ("identically distributed").
- X_i are mutually independent:
  $p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = p_X(x_1) \cdots p_X(x_n)$, where $p_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}$ is the PDF of $X \sim \mathcal{N}(0,1)$ ("independence").

In practice, the terms "random sample" and "i.i.d." are interchangeable.

- We begin by deriving bounds on the probabilities that a random variable X stays away from its mean by a certain distance $t > 0$:

$$\mathbb{P}(|X - \mathbb{E}[X]| > t).$$

- Then we discuss two results, which lie at the heart of statistical inference: the Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). The LLN states that, if $X_1, X_2, \ldots$ are i.i.d. random variables with finite mean, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} \mathbb{E}[X_1]$$

  where the convergence happens in a sense to be specified. The CLT states that, if the i.i.d. random variables $X_1, X_2, \ldots$ have finite variance, then this convergence happens at rate $\mathcal{O}(n^{-1/2})$.
- Together, these two results can be used to obtain *approximate* bounds on the probability that the empirical sum remains away from its mean:

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}[X]| > \frac{t}{\sqrt{n}})$$

  for fixed $t > 0$ and n large.

Inequalities for expected values

Theorem (Cauchy–Schwarz inequality)

Let X and Y be two square-integrable, real-valued random variables.[†] *Then*

$$\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.$$

Proof. If $X = 0$ or $Y = 0$ almost surely, then the claim is trivial. Suppose that $X \neq 0$ and $Y \neq 0$ almost surely. Let $t \in \mathbb{R}$ and note that

$$0 \leq \mathbb{E}[(X + tY)^2] = \mathbb{E}[X^2] + 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2]$$

is a second degree polynomial with respect to t which has at most one real root. Therefore its discriminant must be nonpositive:

$$\begin{aligned} \text{discriminant} \leq 0 \quad &\Leftrightarrow \quad (2\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0 \\ &\Leftrightarrow \quad \mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]. \quad \square \end{aligned}$$

[†] Recall that square-integrability implies that $\mathbb{E}[X^2]$ and $\mathbb{E}[Y^2]$ are well-defined and finite.

Let X be a real-valued random variable. A fundamental problem in statistics is to be able to bound from above the probability $\mathbb{P}(X > t)$ for fixed $t > 0$. Bounds of the following kind are known as "tail bounds".

Theorem (Markov's inequality)

If X is an integrable[†]*, non-negative real-valued random variable and $t > 0$, then*

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Let us consider the case of X being a continuous RV (the discrete case is similar). There holds

$$\mathbb{P}(X > t) = \int_t^\infty f_X(x)\,\mathrm{d}x \overset{(*)}{\leq} \frac{1}{t}\int_t^\infty x f_X(x)\,\mathrm{d}x \leq \frac{1}{t}\int_0^\infty x f_X(x)\,\mathrm{d}x,$$

where $(*)$ follows from $x \geq t \;\Leftrightarrow\; 1 \leq \frac{x}{t}$. Since we assumed that X is non-negative, $f_X(x) = 0$ for $x < 0$, and thus

$$\mathbb{P}(X > t) \leq \frac{1}{t}\int_0^\infty x f_X(x)\,\mathrm{d}x = \frac{1}{t}\int_{-\infty}^\infty x f_X(x)\,\mathrm{d}x = \frac{\mathbb{E}[X]}{t}. \quad \square$$

[†]Recall that this means $\mathbb{E}[|X|] < \infty$.

If X is a square-integrable random variable, then we can bound the probability that X is at least distance $t > 0$ away from the average.

Theorem (Chebyshev's inequality)

Let X be a square-integrable random variable. For all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Proof. By Markov's inequality,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 > t^2) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{t^2} = \frac{\mathrm{Var}(X)}{t^2},$$

where we applied Markov's inequality $\mathbb{P}(Y > t') \leq \frac{\mathbb{E}[Y]}{t'}$ to the non-negative random variable $Y = |X - \mathbb{E}[X]|^2$ and $t' = t^2$. □

Let $\sigma = \sqrt{\mathrm{Var}(X)}$. It is sometimes useful to rewrite the Chebyshev inequality in the form (set $t = k\sigma$)

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

If $k = 2$, then $1 - \frac{1}{k^2} = 75\%$.

If $k = 3$, then $1 - \frac{1}{k^3} \approx 88.9\%$.

In practice, expected value and variance must be estimated. Chebyshev's inequality can be used to evaluate the rareness of a single observation.

The Chebyshev inequality can be useful in situations where we only know the mean and variance of X. On the other hand, it is quite a rough bound. If we know the distribution of X, the probability $\mathbb{P}(|X - \mathbb{E}[X]| > t)$ can be computed more precisely, typically leading to much better bounds.

Example

Let X be a random variable with mean $\mathbb{E}[X] = 0$ and variance $\mathrm{Var}(X) = 1$. Suppose that we wish to estimate $\mathbb{P}(|X| > 2)$.

If the mean and variance is all we know about the random variable, then Chebyshev's inequality gives a very rough bound:

$$\mathbb{P}(|X| > 2) \leq \frac{1}{2^2} = \frac{1}{4} = 0.25.$$

If X is a Gaussian random variable, i.e., in this case we would have $X \sim \mathcal{N}(0, 1)$, then we know precisely

$$\mathbb{P}(|X| > 2) = 2\Phi(-2) = 0.04550\ldots$$

If X is a Gaussian random variable, then the probabilities $\mathbb{P}(|X| > t)$, $t > 0$, can be computed numerically using the CDF. Unfortunately, the CDF does not have a closed form expression. Sometimes the following bound is useful.

Theorem (Mill's inequality)

Let $X \sim \mathcal{N}(0,1)$. Then for all $t > 0$,

$$\mathbb{P}(|X| > t) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-\frac{1}{2}t^2)}{t}.$$

Proof. Let $X \sim \mathcal{N}(0,1)$ and $t > 0$. Then

$$\begin{aligned}\mathbb{P}(|X| > t) &= \int_{-\infty}^{-t} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s + \int_{t}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s \\ &= 2 \int_{t}^{\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s.\end{aligned}$$

It is enough to bound this last integral.

Arguing similarly as in the proof of Markov's inequality,

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s \leq \int_t^\infty \frac{s}{t} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s = \frac{1}{t\sqrt{2\pi}} \int_t^\infty s \, \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s.$$

For this integral, we have

$$\int_t^\infty s \, \mathrm{e}^{-\frac{s^2}{2}} \, \mathrm{d}s = -\left[\mathrm{e}^{-\frac{s^2}{2}}\right]\Big|_{s=t}^{s=\infty} = \mathrm{e}^{-\frac{t^2}{2}},$$

which yields the assertion. □

The previous result can be generalized to Gaussian random variables with arbitrary variance via the whitening transform.

Theorem

Let $\mu \in \mathbb{R}$, $\sigma > 0$, and let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for all $t > 0$,

$$\mathbb{P}(|X - \mathbb{E}[X]| > t) \leq \sqrt{\frac{2\sigma^2}{\pi}} \frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t}.$$

Proof. By the whitening transform, the random variable $Y = \frac{1}{\sigma}(X - \mu) \sim \mathcal{N}(0, 1)$, so

$$\mathbb{P}(|X - \mu| > t) = \mathbb{P}(|Y| > \sigma^{-1} t),$$

and the result follows from Mill's inequality with t replaced by $\sigma^{-1} t$. □

Limit theorems

We will state two fundamental limit theorems for sums of i.i.d. random variables. To do so, we will first need to define what we mean by convergence of a sequence of random variables.

Definition (Convergence in probability)

Let X be a real-valued random variable and let $(X_n)_{n\geq 0}$ be a sequence of real-valued random variables. We say that X_n converges to X in probability, and write $X_n \xrightarrow{P} X$, if for any $\varepsilon > 0$, there holds

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

In other words, X_n converges to X in probability if the probability that X_n is separated from X by any (even very small) non-zero distance vanishes as n grows.

Another, *weaker* form of convergence involves the CDFs F_{X_n} and F_X of the RVs X_n and X, respectively.

Definition

Let X be a real-valued random variable and let $(X_n)_{n\geq 0}$ be a sequence of real-valued random variables. We say that X_n converges to X in distribution (or in law), and write $X_n \xrightarrow{d} X$, if for any $x \in \mathbb{R}$ where F_X is continuous, there holds

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

- If X is a continuous random variable, then F_X is everywhere continuous, and the above condition simply means that F_{X_n} converges pointwise to F_X.
- If X is discrete, then F_X will be discontinuous at every point x such that $\mathbb{P}(X = x) > 0$. The above definition says that, when checking whether X_n converges in distribution to X, we do not need to look at these points of discontinuities.
- That X_n converges in distribution to X means that $\mathbb{P}(X_n \leq x) \xrightarrow{n\to\infty} \mathbb{P}(X \leq x)$ for all points x where F_X does not jump. It is only a statement about the probability distributions of X_n and X. In particular, it does not say at all that X_n is close to X when n is large.

Proposition

1. *If X_n and X are square-integrable and*

$$\mathbb{E}[|X_n - X|^2] \xrightarrow{n\to\infty} 0, \tag{1}$$

   then $X_n \xrightarrow{P} X$. The converse is false in general.
2. *If X_n converges to X in probability, then X_n also converges to X in law. The converse is false in general.*
3. *If X is constant, i.e., there exists $a \in \mathbb{R}$ such that $X = a$ almost surely, then*

$$X_n \xrightarrow{P} X \quad \Leftrightarrow \quad X_n \xrightarrow{d} X.$$

The convergence (1) is called "convergence in quadratic mean", and written $X_n \xrightarrow{q.m.} X$. By the above proposition, convergence in quadratic mean is strictly stronger than convergence in probability, and convergence in probability is strictly stronger than convergence in distribution.

Proof. We only prove the first claim. Assume that $X_n \overset{q.m.}{\to} X$. Then for all $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X|^2 > \varepsilon^2) \leq \frac{\mathbb{E}[|X_n - X|^2]}{\varepsilon^2},$$

where the last inequality is a consequence of Markov's inequality. By assumption, $\mathbb{E}[|X_n - X|^2] \xrightarrow{n\to\infty} 0$. Hence, by the above inequality, we get $\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow{n\to\infty} 0$. This proves that $X_n \overset{P}{\to} X$. That the converse implication is false in general can be shown by counterexample (left as an exercise). □

The Law of Large Numbers (LLN)

Before stating the LLN, we need a technical, but intuitive, lemma.

Lemma

Let X and Y be real-valued random variables which are equal in law. Then, for any real-valued map such that $f(X)$ is integrable, we have $\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$.

Proof. Let us prove the claim for discrete RVs (the continuous case is similar just by replacing PMFs with PDFs and sums by integrals). Let X and Y be discrete. Then $p_X = p_Y$, so for all integrable functions f, by the law of the unconscious statistician, there holds

$$\mathbb{E}[f(X)] = \sum_{x \in E} p_X(x) f(x) = \sum_{y \in E} p_Y(y) f(y) = \mathbb{E}[f(Y)]. \quad \square$$

The above result implies also that if X and Y are equal in law, then

$$\mathbb{E}[X] = \mathbb{E}[Y], \quad \mathbb{E}[X^2] = \mathbb{E}[Y^2], \quad \mathrm{Var}(X) = \mathrm{Var}(Y),$$

provided that these quantities are well-defined.

Let $(X_n)_{n\geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X. By this we mean that $(X_n)_{n\geq 1}$ is a sequence of i.i.d. real-valued random variables having the same law as X. For all $n \geq 1$, let $\overline{X}_n$ denote the sample mean of $X_1, \ldots, X_n$:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

If X_i are integrable, then by linearity of the expected value, there holds

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mathbb{E}[X].$$

Heuristically, we expect $\overline{X}_n$ to converge to $\mathbb{E}[X]$ when $n \to \infty$. This is made precise by the following theorem.

Theorem (Weak Law of Large Numbers)

Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X. If X_i are integrable, then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X].$$

Proof. For simplicity, we provide a proof in the special case where the X_n are also square-integrable. Then

$$\mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] = \mathbb{E}[(\overline{X}_n - \mathbb{E}[\overline{X}_n])^2] = \mathrm{Var}(\overline{X}_n).$$

Now

$$\mathrm{Var}(\overline{X}_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \frac{1}{n^2}\mathrm{Var}(X_i),$$

where the second equality holds since the X_i are independent. Now, by the technical lemma we proved prior to this result, $\mathrm{Var}(X_i) = \mathrm{Var}(X)$ for all i, so we get

$$\mathrm{Var}(\overline{X}_n) = \frac{\mathrm{Var}(X)}{n} \xrightarrow{n\to\infty} 0.$$

Hence $\mathbb{E}[(\overline{X}_n - \mathbb{E}[X])^2] \xrightarrow{n\to\infty} 0$, therefore $\overline{X}_n$ converges to $\mathbb{E}[X]$ in quadratic mean, and hence also in probability. □

A stronger statement holds with the same assumptions.

Theorem (Strong Law of Large Numbers)

Let $(X_i)_{i\geq 1}$ be a sequence of i.i.d. copies of a real-valued random variable X. If X_i are integrable, then

$$\mathbb{P}\big(\{\omega \in \Omega \mid \overline{X}_n(\omega) \xrightarrow{n\to\infty} \mathbb{E}[X]\}\big) = 1.$$

That is, $\overline{X}_n \to \mathbb{E}[X]$ almost surely.

Remark. The significance of the LLN is that it provides a concrete way of approximating the value of $\mathbb{E}[X]$ by sampling values of X a large number of times and taking the sample average.

Example

Let $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$ be independent for some $p \in (0,1)$. Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_1] = p.$$

In other words, if we keep throwing a coin with parameter p a large number of times, the rate of success will converge in probability to p. If the coin is fair, i.e., $p = 1/2$, then the rate of success approaches $1/2$ for n large.

Example

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be independent for some $\mu \in \mathbb{R}$ and $\sigma > 0$. Then

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_1] = \mu.$$

The LLN implies that $\overline{X}_n = \mathbb{E}[X_1] + \varepsilon_n$, where ε_n is some remainder satisfying $\varepsilon_n \xrightarrow{P} 0$. The obvious question to consider is to ask, how fast does ε_n converge to 0?

The Central Limit Theorem

Let $X_1, X_2, \ldots$ be a sequence i.i.d. real-valued random variables. We assume that the X_i are square-integrable and denote by μ and σ^2 their mean and variance, respectively. Thus, for all i,

$$\mathbb{E}[X_i] = \mu, \quad \mathrm{Var}(X_i) = \sigma^2.$$

As we saw in the previous section,

$$\mathbb{E}[\overline{X}_n] = \mu, \quad \mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}.$$

We can perform an affine transformation on $\overline{X}_n$ in order to set its expectation and variance to 0 and 1, respectively. This can be achieved as follows:

1. We center it by subtracting its mean $\mathbb{E}[\overline{X}_n]$,
2. We normalize it by dividing it by its standard deviation $\sqrt{\mathrm{Var}(\overline{X}_n)}$.

In other words, we set

$$Y_n = \frac{1}{\sqrt{\mathrm{Var}(\overline{X}_n)}}(\overline{X}_n - \mathbb{E}[\overline{X}_n]) = \sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - \mu).$$

With this procedure, we obtain a random variable Y_n which is centered and normalized, i.e., which satisfies

$$\mathbb{E}[Y_n] = 0, \quad \mathrm{Var}(Y_n) = 1.$$

The following theorem shows that, for n large, the distribution of Y_n is actually close to $\mathcal{N}(0,1)$.

Theorem (Central Limit Theorem)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. real-valued, square-integrable random variables with mean μ and variance σ^2. Then

$$\sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1).$$

Remark. The CLT implies that, for all $a \in \mathbb{R}$,

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - \mu) \leq a\right) \xrightarrow{n\to\infty} \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}x^2}\,\mathrm{d}x.$$

Remark. One may loosely formulate the CLT as saying that

$$\sqrt{\frac{n}{\sigma^2}}(\overline{X}_n - \mu) \overset{d}{\approx} \mathcal{N}(0,1)$$

for n large. In other words,

$$\overline{X}_n \overset{d}{\approx} \mu + \sqrt{\frac{\sigma^2}{n}} Z,$$

where $Z \sim \mathcal{N}(0,1)$. Thus, by the coloring transform,

$$\overline{X}_n \overset{d}{\approx} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Example

Let $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$ be independent for some $p \in (0,1)$. We know from the LLN that

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X] = p.$$

Since $\mathrm{Var}(X) = p(1-p)$, the CLT further implies that

$$\sqrt{\frac{n}{p(1-p)}}(\overline{X}_n - p) \xrightarrow{d} \mathcal{N}(0,1),$$

or, loosely speaking,

$$\overline{X}_n \overset{d}{\approx} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{for } n \text{ large.}$$

Example (Continued)

By approximating $\overline{X}_n$ using the Gaussian distribution, we can make inferences about the spread of $\overline{X}_n$. For example, if $p = \frac{1}{2}$ and $n = 10^4$, we can use the Gaussian approximation to derive a confidence interval $\mathcal{I}$ such that $\mathbb{P}(\overline{X}_n \in \mathcal{I}) \approx 0.95$. Since n is large, we can use the Gaussian approximation

$$\overline{X}_n \approx \mu + \sigma_n Z, \quad Z \sim \mathcal{N}(0,1),$$

where $\mu = p$ and $\sigma_n = \sqrt{\frac{p(1-p)}{n}}$. We wish to find $a > 0$ such that

$$\int_{\mu-a}^{\mu+a} \frac{1}{\sqrt{2\pi\sigma_n^2}} \mathrm{e}^{-\frac{1}{2\sigma_n^2}(x-\mu)^2} \, \mathrm{d}x = 0.95.$$

Using the change of variables $z = \frac{x-\mu}{\sigma_n}$, where $\mathrm{d}x = \sigma_n \, \mathrm{d}z$, we obtain

$$\int_{-a/\sigma_n}^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \, \mathrm{d}z = 0.95 \quad \Leftrightarrow \quad 2\int_{0}^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \, \mathrm{d}z = 0.95$$

$$\Leftrightarrow \quad \int_{-\infty}^{a/\sigma_n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \, \mathrm{d}z = \frac{1}{2} + \frac{0.95}{2}.$$

Example (Continued)

Using the CDF $\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}z^2} \, \mathrm{d}z$, we obtain

$$\Phi\left(\frac{a}{\sigma_n}\right) = \frac{1}{2} + \frac{0.95}{2} \quad \Leftrightarrow \quad a = \sigma_n \, \Phi^{-1}\left(\frac{1}{2} + \frac{0.95}{2}\right).$$

Plugging in the values $\mu = p = \frac{1}{2}$ and $\sigma_n = \sqrt{\frac{p(1-p)}{n}} = \frac{1}{200}$ yields the interval

$$\mathcal{I} = (\mu - a, \mu + a) = (0.4902, 0.5098).$$