LAW OF LARGE NUMBERS, MONTE CARLO METHODS, AND EMPIRICAL DISTRIBUTIONS

NIUSHAN GAO
DEPARTMENT OF MATHEMATICS
RYERSON UNIVERSITY

This note briefly reviews laws of large numbers and two of their applications to statistics: Monte Carlo methods and Empirical distributions.

1. Law of Large Numbers

1.1. **Weak law of large numbers.** The term of weak law refers to convergence in probability in the context of laws of large numbers. We decompose the proof of weak law into several short lemmas. The first one provides a typical way to yield convergence in probability.

Lemma 1.1. *For a sequence (X_n) of rvs in L^2, if $V[X_n] \longrightarrow 0$, then*

$$X_n - \mathbb{E}[X_n] \xrightarrow{pr} 0.$$

Proof. For any $\varepsilon > 0$, by Chebyshev's inequality,

$$\mathbb{P}\Big(\Big|X_n - \mathbb{E}[X_n]\Big| > \varepsilon\Big) \leq \frac{\mathbb{E}\Big[\big(X_n - \mathbb{E}[X_n]\big)^2\Big]}{\varepsilon^2} = \frac{V[X_n]}{\varepsilon^2} \longrightarrow 0.$$

□

The second one encourages us to do truncations. Two sequences of rvs, (X_n) and (Y_n), are said to be ***equivalent*** if $\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq Y_n) < \infty$.

Lemma 1.2. (1) *If (X_n) and (Y_n) are equivalent, then $\frac{1}{n}\sum_{k=1}^{n} X_k$ converges a.s. (resp., in probability) if and only if $\frac{1}{n}\sum_{k=1}^{n} Y_k$ converges a.s. (resp., in probability). The limits also coincide in the case of convergence.*

(2) *Let (X_n) a sequence of identically distributed, integrable random variables. Let $Y_n = X_n \mathbb{1}_{\{|X_n| \leq n\}}$ for each $n \in \mathbb{N}$. Then (X_n) and (Y_n) are equivalent.*

Proof. (1). Assume that (X_n) and (Y_n) are equivalent. By Borel-Cantelli Lemma,

$$\mathbb{P}\Big(\limsup_n \{X_n \neq Y_n\}\Big) = 0.$$

Take any $\omega \in \big(\limsup_n \{X_n \neq Y_n\}\big)^c = \liminf_n \{X_n = Y_n\}$. There exists n_0, depending on ω, such that for any $n \geq n_0$, $\omega \in \{X_n = Y_n\}$, i.e., $X_n(\omega) = Y_n(\omega)$, implying that

$$\lim_n \frac{1}{n}\sum_{k=1}^{n}(X_k(\omega) - Y_k(\omega)) = 0.$$

Date: April 23, 2019.

These two observations together give that $\frac{1}{n}\sum_{k=1}^{n}(X_k - Y_k)$ converges to 0 a.s. and thus in probability. The assertions in (1) now follow immediately.

(2) holds because

$$\sum_{n=1}^{\infty}\mathbb{P}(X_n \neq Y_n) = \sum_{n=1}^{\infty}\mathbb{P}(|X_n| > n) = \sum_{n=1}^{\infty}\mathbb{P}(|X_1| > n) \leq \mathbb{E}[|X_1|] < \infty.$$

□

The nice properties of truncated rvs are contained in the next lemma.

Lemma 1.3. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of identically distributed integrable rvs. Let $Y_n = X_n\mathbb{1}_{\{|X_n|\leq n\}}$ for each $n \in \mathbb{N}$. Then*

$$\sum_{n=1}^{\infty}\frac{\mathbb{V}[Y_n]}{n^2} < \infty.$$

Consequently,

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{k=1}^{n}V[Y_k] = 0.$$

Proof. Let F be the CDF of X_n's. Then

$$\begin{aligned}
\sum_{n=1}^{\infty}\frac{V[Y_n]}{n^2} \leq& \sum_{n=1}^{\infty}\frac{\mathbb{E}[Y_n^2]}{n^2} = \sum_{n=1}^{\infty}\frac{1}{n^2}\int_{\{|x|\leq n\}}x^2\mathrm{d}F(x) = \sum_{n=1}^{\infty}\frac{1}{n^2}\sum_{k=1}^{n}\int_{\{k-1<|x|\leq k\}}x^2\mathrm{d}F(x)\\
=&\sum_{k=1}^{\infty}\sum_{n=k}^{\infty}\frac{1}{n^2}\int_{\{k-1<|x|\leq k\}}x^2\mathrm{d}F(x) \leq \sum_{k=1}^{\infty}\frac{2}{k}\int_{\{k-1<|x|\leq k\}}x^2\mathrm{d}F(x)\\
\leq&\sum_{k=1}^{\infty}\frac{2}{k}\int_{\{k-1<|x|\leq k\}}k|x|\mathrm{d}F(x) = 2\sum_{k=1}^{\infty}\int_{\{k-1<|x|\leq k\}}|x|\mathrm{d}F(x)\\
=&2\mathbb{E}[|X|] < \infty.
\end{aligned}$$

The second assertion follows from Kronecker's Lemma below on convergence of numbers, whose proof can be found in a mathematical analysis textbook and we omit. □

Lemma (Kronecker)**.** *Let $(x_n)_{n\in\mathbb{N}}$ be a sequence of real numbers, $(a_n)_{n\in\mathbb{N}}$ be a sequence of positive real numbers increasing to ∞. If $\sum_{n=1}^{\infty}\frac{x_n}{a_n}$ converges to a real number, then*

$$\frac{1}{a_n}\sum_{k=1}^{n}x_k \longrightarrow 0.$$

We are now ready to present the proof of the weak law of large numbers.

Theorem 1.4 (Weak LLN)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of pairwise independent, identically distributed, integrable rvs. Then*

$$\frac{1}{n}\sum_{k=1}^{n}X_k \xrightarrow{pr} \mu,$$

where μ is the mean of X_i's.

Proof. Let $Y_n = X_n \mathbb{1}_{\{|X_n| \le n\}}$ for each $n \in \mathbb{N}$. By Lemma 1.2, it suffices to prove that

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \xrightarrow{pr} \mu.$$

Moreover,

$$\mathbb{E}[Y_n] = \mathbb{E}[X_n \mathbb{1}_{\{|X_n| \le n\}}] = \int_{|x| \le n} x \mathrm{d}F(x) = \mathbb{E}[X_1 \mathbb{1}_{\{|X_1| \le n\}}] \longrightarrow \mathbb{E}[X_1] = \mu$$

by the Dominated Convergence Theorem, where F is the CDF of X_n's. Thus

$$\frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[Y_n] \longrightarrow \mu.$$

Therefore, it suffices to prove that

$$T_n := \frac{\sum_{k=1}^{n}(Y_k - \mathbb{E}[Y_k])}{n} \xrightarrow{pr} 0.$$

Note that Y_n's are also pairwise independent and thus are uncorrelated. Hence,

$$\mathbb{V}[T_n] = \frac{1}{n^2}\mathbb{V}\Big[\sum_{k=1}^{n}(Y_k - \mathbb{E}[Y_k])\Big] = \frac{1}{n^2}\sum_{k=1}^{n}\mathbb{V}[Y_k] \longrightarrow 0,$$

by Lemma 1.3. Thus by Lemma 1.1, $T_n = T_n - \mathbb{E}[T_n] \xrightarrow{pr} 0$. □

1.2. **Strong law of large numbers.** The term of strong law refers to a.s. convergence. Again, we split the proof into several lemmas.

Lemma 1.5. *Let (X_n) be a sequence of independent mean-zero rvs in L^2. Put $S_n = \sum_{k=1}^{n} X_k$ for each $n \in \mathbb{N}$. Then for any $\varepsilon > 0$ and $n \in \mathbb{N}$,*

$$\mathbb{P}\Big(\max_{1 \le k \le n} |S_k| > \varepsilon\Big) \le \frac{\mathbb{V}[S_n]}{\varepsilon^2}.$$

Proof. For each $n \in \mathbb{N}$, set $\mathscr{F}_n = \sigma(X_k : 1 \le k \le n)$. Then

$$\mathbb{E}[S_{n+1}|\mathscr{F}_n] = \mathbb{E}[S_n + X_{n+1}|\mathscr{F}_n] = S_n + \mathbb{E}[X_{n+1}|\mathscr{F}_n] = S_n,$$

where we use the fact that since X_{n+1} is independent from $\mathscr{F}_n$, $\mathbb{E}[X_{n+1}|\mathscr{F}_n] = \mathbb{E}[X_{n+1}] = 0$. It follows that $\{(S_n); (\mathscr{F}_n)\}$ is a martingale, and thus $\{(S_n^2); (\mathscr{F}_n)\}$ is a positive submartingale. Thus by Doob's maximal inequality,

$$\mathbb{P}\Big(\max_{1 \le k \le n} |S_k| > \varepsilon\Big) = \mathbb{P}\Big(\max_{1 \le k \le n} S_k^2 > \varepsilon^2\Big) \le \frac{\mathbb{E}[S_n^2]}{\varepsilon^2} = \frac{\mathbb{V}[S_n]}{\varepsilon^2}.$$

□

Lemma 1.6. *Let (X_n) be a sequence of independent mean-zero rvs in L^2. Suppose that $\sum_{n=1}^{\infty} \mathbb{V}[X_n] < \infty$. Then $\sum_{n=1}^{\infty} X_n$ converges a.s.*

Proof. For any $m \in \mathbb{N}$, take $n_m \in \mathbb{N}$ such that

$$\sum_{k=n_m+1}^{\infty} \mathbb{V}[X_k] < \frac{1}{m^4}.$$

For any $n' > n_m$, by Lemma 1.5, we have

$$\mathbb{P}\Big(\max_{n_m+1\le k\le n'} |S_k - S_{n_m}| > \frac{1}{m}\Big) \le \frac{\mathbb{V}\big[S_{n'} - S_{n_m}\big]}{\frac{1}{m^2}} = m^2 \sum_{n_m+1\le k\le n'} \mathbb{V}[X_k] < \frac{1}{m^2}.$$

Putting

$$A_m = \{\max_{k\ge n_m+1} |S_k - S_{n_m}| > \frac{1}{m}\}$$

and letting $n' \to \infty$ above, we have

$$\mathbb{P}(A_m) \le \frac{1}{2^m},$$

so that $\sum_{m=1}^{\infty} \mathbb{P}(A_m) < \infty$ and thus by Borel-Catenlli Lemma,

$$\mathbb{P}(\limsup_m A_m) = 0.$$

Now take any $\omega \notin \limsup_m A_m$, there exists some $m \in \mathbb{N}$ such that $\omega \notin A_m$, which is equivalent to that $|S_k(\omega) - S_{n_m}(\omega)| \le \frac{1}{m}$ for any $k \ge n_m$. Therefore,

$$|S_k(\omega) - S_l(\omega)| \le \frac{2}{m}, \quad \text{for any } k, l \ge n_m.$$

This proves that the partial sums of the series $\sum_{n=1}^{\infty} S_n(\omega)$ are Cauchy, and hence the series is convergent. This completes the proof. □

Theorem 1.7 (Strong LLN)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed, integrable rvs. Then*

$$\frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{a.s.} \mu,$$

where μ is the mean of X_i's.

Proof. Let $Y_n = X_n \mathbb{1}_{\{|X_n|\le n\}}$ for each $n \in \mathbb{N}$. As in the weak case, it suffices to prove that

$$T_n := \frac{\sum_{k=1}^{n}(Y_k - \mathbb{E}[Y_k])}{n} \xrightarrow{a.s.} 0.$$

By Lemmas 1.3 and 1.6, $\sum_{n=1}^{\infty} \frac{Y_n - \mathbb{E}[Y_n]}{n}$ converges a.s. By Kronecker's Lemma, $T_n \xrightarrow{a.s.} 0$. □

2. Applications

2.1. **Monte Carlo Simulations.** The SLLN provides a numerical method for computing the expectation $\mathbb{E}[X]$ of a rv via simulations. Recall that a distribution function $F : \mathbb{R} \to \mathbb{R}$ is an increasing, right-continuous function such that $F(-\infty) = 0$ and $F(\infty) = 1$. Recall also that there is a bijection between distribution functions and probability measures on $\mathbb{R}$ via:

$$\mu\big((a, b]\big) = F(b) - F(a).$$

We will call a distribution function F of interest a ***(univariate) population***. A ***random sample*** drawn from the population F is a sequence (X_n) of independent rvs all having F as their CDF. A ***sample*** drawn from the population F is a sequence (x_n) of real numbers, which is a ***realization*** of a random sample (X_n), namely, there exists ω such that

$$x_n = X_n(\omega) \quad \text{for each } n.$$

Suppose that the ***population mean*** $m := \int_{\mathbb{R}} x \, \mathrm{d}F(x)$ is finite. Let (X_n) be a random sample drawn from F. Then $\mathbb{E}[|X_n|] = \int_{\mathbb{R}} |x| \, \mathrm{d}F(x) < \infty$, so that the SLLN is applicable to the sequence (X_n). Thus, for a sample (x_n) drawn from F, the ***sample means*** converge to the population mean[1]:

$$\frac{1}{n} \sum_{k=1}^{n} x_k \longrightarrow \mathbb{E}[X_1] = \int_{\mathbb{R}} x \, \mathrm{d}F(x) = m, \quad \text{as } n \to \infty.$$

In reality, the sample drawn from the population is of course a finite sequence, say, $x_1, x_2, \ldots, x_n$, where n is called the ***sample size***. When the size n is large enough, we have the following approximation:

$$\frac{1}{n} \sum_{k=1}^{n} x_k \approx m.$$

This algorithm of computing a population parameter using random sampling is typically referred to as ***Monte Carlo methods***. For example, once we have a way to generate from F a sample, also called ***random numbers*** in the context of Monte Carlo methods, we can evaluate the population mean m by $\frac{1}{n} \sum_{k=1}^{n} x_k$ as above.

Most computational software contain random number generators for classical distributions, such as uniform distribution and normal distributions. For example, $x = \mathrm{rand}(n, 1)$ returns n random numbers from the uniform distribution on $(0, 1)$:

```
>> x=rand(10,1)

x =

    0.1622
    0.7943
    0.3112
    0.5285
```

[1]Not an accurate assertion, since for any random sample, the convergence may fail on a set of probability 0. But for all practical purposes, probability-zero events are regarded as never happening.

```
    0.1656
    0.6020
    0.2630
    0.6541
    0.6892
    0.7482
```

When n is large, we can see that the sample mean $\frac{1}{n}\sum_{k=1}^{n} x_k$ is indeed close to the population mean $\mu = \int_{(0,1)} x\mathrm{d}x = \frac{1}{2}$. We simulate 5 samples of size one million and calculate the respective sample means; all of these five sample means are close to $\mu = 0.5$:

```
>> y=zeros(5,1);
for k=1:5
    x=rand(1000000,1);
    y(k)=mean(x);
end
y

y =

    0.5002
    0.4999
    0.5000
    0.5001
    0.4998
```

Of course, we can use the Monte Carlo methods to compute population parameters other than the mean. Suppose that the population F has a finite second moment, i.e., $\int_{\mathbb{R}} x^2\mathrm{d}F(x) < \infty$. Let (X_n) be a random sample drawn from F. Then $\mathbb{E}[X_n^2] = \int_{\mathbb{R}} x^2\mathrm{d}F(x) < \infty$, which further implies that $\mathbb{E}[|X_n|] < \infty$ by the Cauchy-Schwartz inequality. Thus the the SLLN applies to both (X_n) and (X_n^2), namely,

$$\frac{1}{n}\sum_{k=1}^{n} X_n \xrightarrow{a.s.} \mathbb{E}[X_1] = \int_{\mathbb{R}} x\mathrm{d}F(x), \qquad \frac{1}{n}\sum_{k=1}^{n} X_n^2 \xrightarrow{a.s.} \mathbb{E}[X_1^2] = \int_{\mathbb{R}} x^2\mathrm{d}F(x).$$

It follows that for a sample (x_n),

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2 - \Big(\frac{1}{n}\sum_{k=1}^{n} x_k\Big)^2 \longrightarrow \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = V[X_1],$$

where the far right term is clearly equal to

$$V[F] :== \int_{\mathbb{R}} x^2\mathrm{d}F(x) - \Big(\int_{\mathbb{R}} x\mathrm{d}F(x)\Big)^2,$$

called the ***population variance***. Thus $V[F]$ is evaluated by

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2 - \Big(\frac{1}{n}\sum_{k=1}^{n} x_k\Big)^2 = \frac{\sum_{k=1}^{n}(x_k - \overline{x})^2}{n},$$

for some large enough n; here $\overline{x} := \frac{1}{n}\sum_{k=1}^{n} x_k$ is the sample mean.

We can use tricks to generate random numbers to compute more sophisticated probabilistic terms. Say, let's compute the expectation of

$$X = \frac{2^U}{e^{\sqrt{|Z|}}},$$

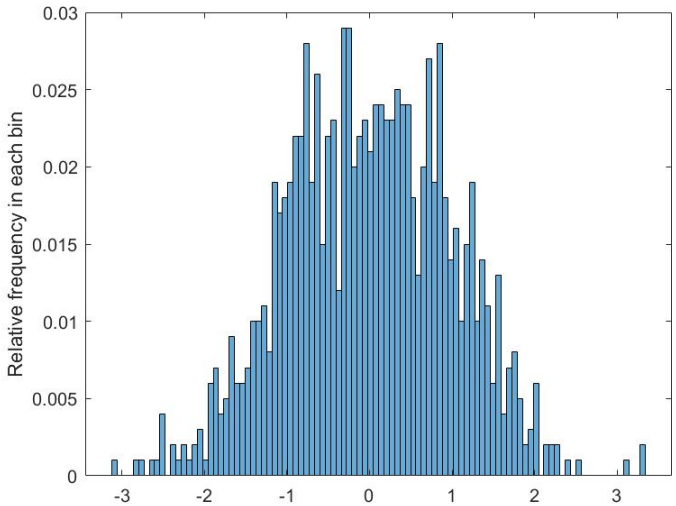
where U and Z are independent, U is uniform on $(0, 1)$, and Z is standard normal. We simulate a sample of size one million for the uniform distribution and the standard normal, respectively, aggregate them to produce random numbers for F, where F is the CDF of X, and then take the new sample mean:

```
>> u=rand(1000000,1);
z=randn(1000000,1);
x=2.^u./exp(sqrt(abs(z)));
mean(x)

ans =

    0.6736
```

One can run these codes a few times and will see that the answer is stable around 0.673.

2.2. **Empirical Distributions.** In Statistics, one often draws a histogram of data, which shows "distribution" of the data, to infer the distribution of the population. For example, the following codes in Matlab simulate 1000 random numbers from the standard normal distribution and produces the histogram of these numbers with 50 bins, Figure 1.

```
>> x=randn(1000,1);
>> histogram(x,50)
ylabel('Relative frequency in each bin')
```



One sees that the histogram does demonstrate a shape like the graph of the density of the standard normal distribution. We now study why this happens.

In general, suppose that we collect n ***observations***, i.e., a sample of size n, from the population, which we denote by $x_1, x_2, \dots, x_n$. In the histogram, one first cuts the x-axis into several bins. Then the histogram captures the relative frequencies of observations that belong to each bin $(a_j, b_j]$:

$$\frac{\#\{k : a_j < x_k \leq b_j\}}{n}.$$

We consider the following function $F_n : \mathbb{R} \to \mathbb{R}$ defined by

$$F_n(x) = \frac{\#\{k : x_k \leq x\}}{n}, \quad x \in \mathbb{R}.$$

Clearly, $F_n(x)$ is the relative frequency of the observations $x_1, \dots, x_n$ that belong to the interval $(-\infty, x]$. In this notation, the relative frequency in a bin $(a_j, b_j]$ can be expressed by

$$F_n(b_j) - F_n(a_j).$$

Thus, the histogram is produced by the values of F_n at the end points of the bins.

One can easily see that F_n is an increasing, right continuous function such that $F_n(-\infty) = 0$ and $F_n(\infty) = 1$. That is, F_n is also a distribution function. It is called the ***empirical distribution***, because it is the distribution of the empirical evidence $x_1, \dots, x_n$. The assertion that histograms can be used to approximate the population distribution is mathematically equivalent to that whenever n is large enough, $F_n(x)$ is close to $F(x)$ at every $x \in \mathbb{R}$, or

$$\sup_{x \in \mathbb{R}} \big| F_n(x) - F(x) \big| \quad \text{is small, whenever } n \text{ is large.}$$

Theorem 2.1 (Central Statistical Theorem)**.** *Let (X_n) be a random sample drawn from the population F. For each $n \in \mathbb{N}$ and $x \in \mathbb{R}$, put*

$$F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{\{X_k \leq x\}}.^2$$

Then

$$\mathbb{P}\Big(\lim_n \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = 0 \Big) = 1.$$

That is, out a set of probability 0, $\big(F_n(x)\big)$ converges to $F(x)$, uniformly in x.

Clearly, at any realization (x_n) of (X_n), the two ways of defining $F_n(x)$ coincide.

Proof. We only provide the proof of the following weaker version. At every $x \in \mathbb{R}$, outside a set of probability, we have $F_n(x) \longrightarrow F(x)$. This is easy! Fix $x \in \mathbb{R}$. Since X_n's are iid, the random variables $\mathbb{1}_{\{X_n \leq x\}}$'s are iid too. In fact, their common distribution is as follows:

$$\mathbb{P}\Big(\mathbb{1}_{\{X_n \leq x\}} = 1 \Big) = \mathbb{P}(X_n \leq x) = F(x),$$

and

$$\mathbb{P}\Big(\mathbb{1}_{\{X_n \leq x\}} = 0 \Big) = \mathbb{P}(X_n > x) = 1 - F(x).$$

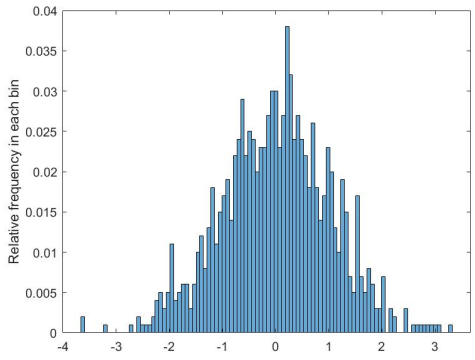
[2]Note that $F_n(x)$ is in fact a rv depending on the random sample.

Thus by the SLLN,

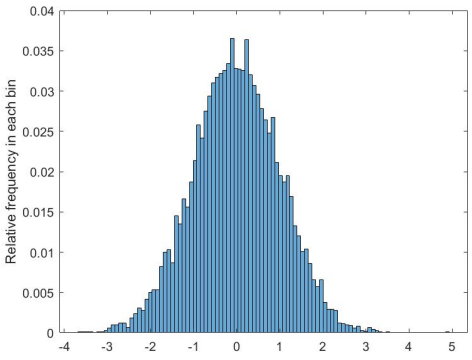
$$F_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k \le x\}} \xrightarrow{a.s.} \mathbb{E}[\mathbb{1}_{\{X_1 \le x\}}] = \mathbb{P}(X_1 \le x) = F(x).$$

□

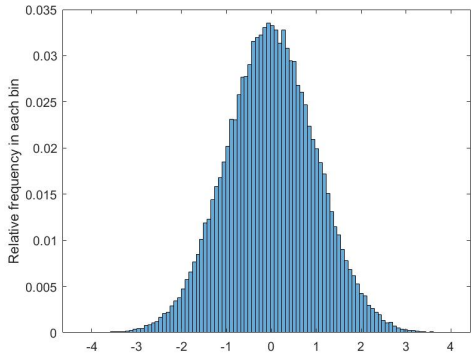
The following are histograms with 100 bins of four simulated samples from the standard normal distribution of sizes $10^3, 10^4, 10^5, 10^6$, respectively.



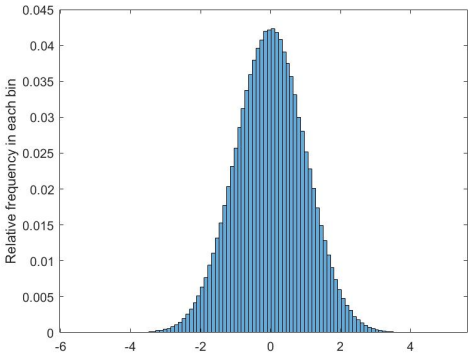
(A) $n = 10^3$



(B) $n = 10^4$



(C) $n = 10^5$



(D) $n = 10^6$