AUA CS108, Statistics, Fall 2020

Lecture 24

Michael Poghosyan

21 Oct 2020

Contents

Supplements, LLN

Sometimes we are required to calculate limits of the form:

$$\lim_{n \to +\infty} \frac{g(X_1) + g(X_2) + ... + g(X_n)}{n}$$

in the *Probability* or *a.s.* sense, for some nice function g.

Supplements, LLN

Sometimes we are required to calculate limits of the form:

$$\lim_{n\to+\infty} \frac{g(X_1)+g(X_2)+\ldots+g(X_n)}{n}$$

in the *Probability* or *a.s.* sense, for some nice function g.Clearly, under the condition that $\mathbb{E}(g(X_1))$ and $Var(g(X_1))$ are finite, or $\mathbb{E}(|g(X_1)|)<+\infty$, we will have

$$\frac{g(X_1)+g(X_2)+\ldots+g(X_n)}{n} \xrightarrow{\mathbb{P},a.s.} \mathbb{E}(g(X_1)), \qquad n\to+\infty.$$

Supplements, LLN

Sometimes we are required to calculate limits of the form:

$$\lim_{n \to +\infty} \frac{g(X_1) + g(X_2) + ... + g(X_n)}{n}$$

in the *Probability* or *a.s.* sense, for some nice function g.Clearly, under the condition that $\mathbb{E}(g(X_1))$ and $Var(g(X_1))$ are finite, or $\mathbb{E}(|g(X_1)|) < +\infty$, we will have

$$\frac{g(X_1) + g(X_2) + ... + g(X_n)}{n} \xrightarrow{\mathbb{P}, a.s.} \mathbb{E}(g(X_1)), \qquad n \to +\infty.$$

Say, for example,

$$\frac{X_1^2 + X_2^2 + ... + X_n^2}{n} \xrightarrow{\mathbb{P}, a.s.}$$

Supplements, LLN

Sometimes we are required to calculate limits of the form:

$$\lim_{n \to +\infty} \frac{g(X_1) + g(X_2) + ... + g(X_n)}{n}$$

in the *Probability* or *a.s.* sense, for some nice function g.Clearly, under the condition that $\mathbb{E}(g(X_1))$ and $Var(g(X_1))$ are finite, or $\mathbb{E}(|g(X_1)|) < +\infty$, we will have

$$\frac{g(X_1) + g(X_2) + ... + g(X_n)}{n} \xrightarrow{\mathbb{P}, a.s.} \mathbb{E}(g(X_1)), \qquad n \to +\infty.$$

Say, for example,

$$\frac{X_1^2 + X_2^2 + ... + X_n^2}{n} \xrightarrow{\mathbb{P}, a.s.} \mathbb{E}(X_1^2), \qquad n \to +\infty.$$

CLT

The LLN says that the values of $\overline{X}_n$ are concentrated around $\mathbb{E}(X_1)$.

CLT

The LLN says that the values of $\overline{X}_n$ are concentrated around $\mathbb{E}(X_1)$. But it is not giving us an idea about how the values are distributed around that mean value.

CLT

The LLN says that the values of $\overline{X}_n$ are concentrated around $\mathbb{E}(X_1)$. But it is not giving us an idea about how the values are distributed around that mean value. CLT helps us in this regard.

CLT

The LLN says that the values of $\overline{X}_n$ are concentrated around $\mathbb{E}(X_1)$. But it is not giving us an idea about how the values are distributed around that mean value. CLT helps us in this regard.

To give the general idea of the CLT, let us use the following transform: for a r.v. X, let us denote

$$Z = Standardize(X) = \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}} = \frac{X - \mathbb{E}(X)}{SD(X)},$$

the Standardization (normalization, scaling) of X.

CLT

The LLN says that the values of $\overline{X}_n$ are concentrated around $\mathbb{E}(X_1)$. But it is not giving us an idea about how the values are distributed around that mean value. CLT helps us in this regard.

To give the general idea of the CLT, let us use the following transform: for a r.v. X, let us denote

$$Z = Standardize(X) = \frac{X - \mathbb{E}(X)}{\sqrt{Var(X)}} = \frac{X - \mathbb{E}(X)}{SD(X)},$$

the Standardization (normalization, scaling) of X. Clearly,

$$\mathbb{E}(Z) = 0 \qquad \text{and} \qquad Var(Z) = 1.$$

Basic Idea of the CLT

The basic idea of the CLT is the following: if we have a sequence of IID r.v. X_n, and we consider their sum S_n or their average $\overline{X}_n$, then

$$Standardize(S_n) \xrightarrow{D} \mathcal{N}(0,1)$$

and

$$Standardize(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0,1).$$

Basic Idea of the CLT

The basic idea of the CLT is the following: if we have a sequence of IID r.v. X_n, and we consider their sum S_n or their average $\overline{X}_n$, then

$$Standardize(S_n) \xrightarrow{D} \mathcal{N}(0,1)$$

and

$$Standardize(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0,1).$$

Btw, trivially,

$$Standardize(\overline{X}_n) = Standardize(S_n),$$

and these two versions of CLT are the same.

Basic Idea of the CLT

The basic idea of the CLT is the following: if we have a sequence of IID r.v. X_n, and we consider their sum S_n or their average $\overline{X}_n$, then

$$Standardize(S_n) \xrightarrow{D} \mathcal{N}(0,1)$$

and

$$Standardize(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0,1).$$

Btw, trivially,

$$Standardize(\overline{X}_n) = Standardize(S_n),$$

and these two versions of CLT are the same.

So for large n, the Distribution of the $Standardize(S_n)$ or $Standardize(\overline{X}_n)$ is approximately Standard Normal.

Basic Idea of the CLT

The basic idea of the CLT is the following: if we have a sequence of IID r.v. X_n, and we consider their sum S_n or their average $\overline{X}_n$, then

$$Standardize(S_n) \xrightarrow{D} \mathcal{N}(0,1)$$

and

$$Standardize(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0,1).$$

Btw, trivially,

$$Standardize(\overline{X}_n) = Standardize(S_n),$$

and these two versions of CLT are the same.

So for large n, the Distribution of the $Standardize(S_n)$ or $Standardize(\overline{X}_n)$ is approximately Standard Normal. And this **independent of the initial Distribution of** X_k!

Basic Idea of the CLT

The basic idea of the CLT is the following: if we have a sequence of IID r.v. X_n, and we consider their sum S_n or their average $\overline{X}_n$, then

$$Standardize(S_n) \xrightarrow{D} \mathcal{N}(0,1)$$

and

$$Standardize(\overline{X}_n) \xrightarrow{D} \mathcal{N}(0,1).$$

Btw, trivially,

$$Standardize(\overline{X}_n) = Standardize(S_n),$$

and these two versions of CLT are the same.

So for large n, the Distribution of the $Standardize(S_n)$ or $Standardize(\overline{X}_n)$ is approximately Standard Normal. And this **independent of the initial Distribution of X_k**!

Easy and beautiful, isn't it?

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) =$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

Now we use

$$\mathbb{E}(S_n) =$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

Now we use

$$\mathbb{E}(S_n) = n \cdot \mu, \qquad Var(S_n) =$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

Now we use

$$\mathbb{E}(S_n) = n \cdot \mu, \qquad Var(S_n) = n \cdot \sigma^2.$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

Now we use

$$\mathbb{E}(S_n) = n \cdot \mu, \qquad Var(S_n) = n \cdot \sigma^2.$$

Then,

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma}.$$

CLT, in the Sums form

Assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

We consider sums

$$S_n = X_1 + X_2 + ... + X_n.$$

We Standardize S_n:

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}}.$$

Now we use

$$\mathbb{E}(S_n) = n \cdot \mu, \qquad Var(S_n) = n \cdot \sigma^2.$$

Then,

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma}.$$

The CLT states:

$$\frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma} \xrightarrow{D} \mathcal{N}(0,1).$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$.

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) =$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

We use

$$\mathbb{E}(\overline{X}_n) =$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

We use

$$\mathbb{E}(\overline{X}_n) = \mu, \qquad Var(\overline{X}_n) =$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

We use

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

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

We use

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

Then,

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

CLT, in the Averages form

Again, assume X_n be a sequence of IID r.v. with finite expectation $\mu = \mathbb{E}(X_i)$ and variance $\sigma^2 = Var(X_i)$. We consider the Averages

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}.$$

We Standardize $\overline{X}_n$:

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}}.$$

We use

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

Then,

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

The CLT states:

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

Two forms of CLT

Of course, these two forms of the CLT are the same: we have

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma}$$

and

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Two forms of CLT

Of course, these two forms of the CLT are the same: we have

$$Standardize(S_n) = \frac{S_n - \mathbb{E}(S_n)}{\sqrt{Var(S_n)}} = \frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma}$$

and

$$Standardize(\overline{X}_n) = \frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\sqrt{Var(\overline{X}_n)}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}.$$

Now,

$$\frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma} = \frac{n \cdot (\frac{S_n}{n} - \mu)}{\sqrt{n} \cdot \sigma} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}},$$

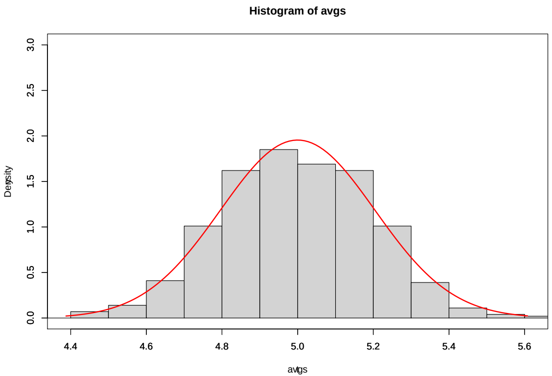
so

$$Standardize(S_n) = Standardize(\overline{X}_n).$$

Hence, the above two versions of CLT are the same, just one is in terms of S_n, the other one is in terms of $\overline{X}_n$.

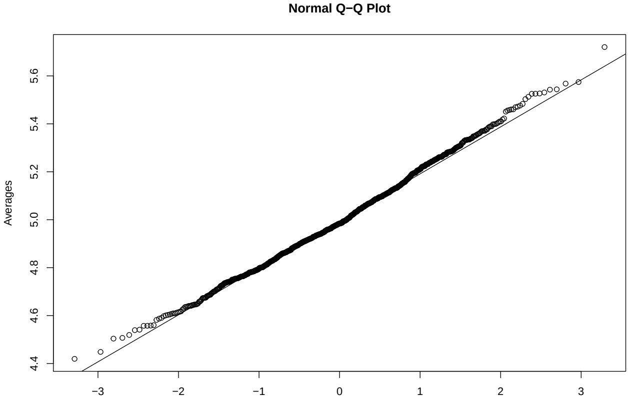
CLT, Visually

```r
n <- 600 # Sample Size
m <- 1000 # no of Samples
rate <- 0.2
x <- rexp(n*m, rate = rate)
theo.mean <- 1/rate #theoretical mean
theo.sd <- 1/rate #theoretical SD
m <- matrix(x, ncol = m); d <- data.frame(m)
avgs <- sapply(d, mean)
a = theo.mean-3*theo.sd/sqrt(n); b = theo.mean+3*theo.sd/sqrt(n)
hist(avgs, freq = F, xlim = c(a, b), ylim=c(0,3))
par(new = T)
t <- seq(a,b, 0.01)
y <- dnorm(t, mean = theo.mean, sd = theo.sd/sqrt(n))
plot(t,y, type = "l", col="red", lwd = 2, xlim = c(a,b), ylim=c(0,3))
```



CLT, Visually, v2

```
n <- 600 # Sample Size
m <- 1000 # no of Samples
rate <- 0.2
x <- rexp(n*m, rate = rate)
m <- matrix(x, ncol = m); d <- data.frame(m)
avgs <- sapply(d, mean)
qqnorm(avgs, ylab = "Averages"); qqline(avgs)
```



CLT, Roughly

In a non-rigorous way, we can write, for large n (here $\approx$ means approximately distributed as):

$$\frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma} \approx \mathcal{N}(0,1) \qquad \textit{and} \qquad \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \approx \mathcal{N}(0,1).$$

CLT, Roughly

In a non-rigorous way, we can write, for large n (here $\approx$ means approximately distributed as):

$$\frac{S_n - n \cdot \mu}{\sqrt{n} \cdot \sigma} \approx \mathcal{N}(0,1) \qquad \textit{and} \qquad \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx \mathcal{N}(0,1).$$

or

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2) \qquad \textit{and} \qquad \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

CLT, Roughly -2

Let us summarize:

CLT, Roughly -2

Let us summarize:

- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **are Normally Distributed**, i.e., $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then

CLT, Roughly -2

Let us summarize:

- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **are Normally Distributed**, i.e., $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \qquad \textit{and} \qquad \overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right);$$

so we know the **exact Distributions** of S_n and $\overline{X}_n$.

CLT, Roughly -2

Let us summarize:

- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **are Normally Distributed**, i.e., $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then

  $$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \qquad and \qquad \overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right);$$

  so we know the **exact Distributions** of S_n and $\overline{X}_n$.
- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **from any Distribution** (but the same Distribution), then

CLT, Roughly -2

Let us summarize:

- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **are Normally Distributed**, i.e., $X_k \sim \mathcal{N}(\mu, \sigma^2)$, then

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right);$$

  so we know the **exact Distributions** of S_n and $\overline{X}_n$.

- If X_k-s are independent, have the Mean $\mathbb{E}(X_k) = \mu$ and $Var(X_K) = \sigma^2$, and **from any Distribution** (but the same Distribution), then

$$S_n \approx \mathcal{N}(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right);$$

  and we know the **asymptotic Distributions** (approximate Distributions for large n) of S_n and $\overline{X}_n$.