Introduction to Statistical Inference

Lecture 3: Tricks with Random Variables:

The Law of Large Numbers &

The Central Limit Theorem

Mohammad-Reza A. Dehaqani

dehaqani@ut.a

Large Sample Theory

The most important aspect of probability theory concerns the behavior of sequences of random variables. This part of probability is called large sample theory or limit theory or asymptotic theory. This theory is extremely important for statistical inference.

The basic question is this:
What can we say about the limiting behavior of a sequence of random variables?

$$X_1, X_2, X_3 \dots$$

In the statistical context: What happens as we gather more and more data?

In Calculus, we say that a sequence of real numbers $x_1, x_2, \dots$ converges to a limit x if, for every $\epsilon > 0$, we can find N such that $|x_n - x| < \epsilon$ for all $n > N$.

In Probability, convergence is more subtle.
Going back to calculus, suppose that $x_n = 1/n$. Then trivially, $\lim_{n\to\infty} x_n = 0$.
Consider a probabilistic version of this example: suppose that $X_1, X_2, \dots$ are independent and $X_n \sim \mathcal{N}(0, 1/n)$. Intuitively, X_n is very concentrated around 0 for large n, and we are tempted to say that X_n "converges" to zero. However, $\mathbb{P}(X_n = 0) = 0$ for all n!

Types of Convergence

There are two main types of convergence:
convergence in probability and convergence in distribution

Definition

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

- X_n **converges** to X **in probability**, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\epsilon > 0$
$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$$

- X_n **converges** to X **in distribution**, written $X_n \xrightarrow{\mathcal{D}} X$, if
$$\lim_{n\to\infty} F_n(x) = F(x)$$
for all x for which F is continuous.

Relationships Between the Types of Convergence

Example: Let $X_n \sim \mathcal{N}(0, 1/n)$. Then

- $X_n \xrightarrow{\mathbb{P}} 0$
- $X_n \xrightarrow{\mathcal{D}} 0$

Question: Is there any relationship between $\xrightarrow{\mathbb{P}}$ and $\xrightarrow{\mathcal{D}}$?

Answer: Yes:

$$\boxed{X_n \xrightarrow{\mathbb{P}} X \quad \text{implies that} \quad X_n \xrightarrow{\mathcal{D}} X}$$

Important Remark: The reverse implication does not hold:
convergence in distribution does not imply convergence in probability.

Example: Let $X \sim \mathcal{N}(0, 1)$ and let $X_n = -X$ for all n. Then

- $X_n \xrightarrow{\mathcal{D}} X$
- $X_n \overset{\mathbb{P}}{\nrightarrow} X$

The Law of Large Numbers

The law of large numbers is one of the main achievements in probability. This theorem says that the mean of a large sample is close to the mean of the distribution.

The Law of Large Numbers

Let $X_1, X_2, \ldots$ be an i.i.d. sample and let $\mu = \mathbb{E}[X_1]$ and $\sigma^2 = \mathbb{V}[X_1] < \infty$. Then

$$\boxed{\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathbb{P}} \mu}$$

Useful Interpretation:

The distribution of $\overline{X}_n$ becomes more concentrated around μ as n gets larger.

Let $X_1, X_2, \ldots, X_i \ldots$ be a sequence of independent random variables with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then, for any $\varepsilon > 0$,

$$P(|\overline{X}_n - \mu| > \varepsilon) \to 0 \qquad \text{as } n \to \infty$$

Proof

We first find $E(\overline{X}_n)$ and $\mathrm{Var}(\overline{X}_n)$:

$$E(\overline{X}_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$$

Since the X_i are independent,

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

The desired result now follows immediately from Chebyshev's inequality, which states that

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \qquad \text{as } n \to \infty \qquad \blacksquare$$

Example: Let $X_i \sim \mathrm{Bernoulli}(p)$. The fraction of heads after n tosses is $\overline{X}_n$.

According to the LLN, $\overline{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_i] = p$. It means that, when n is large, the distribution of $\overline{X}_n$ is tightly concentrated around p.

Q: How large should n be so that $\mathbb{P}(|\overline{X}_n - p| < \epsilon) \geq 1 - \alpha$?

Answer: $n \geq \frac{p(1-p)}{\alpha\epsilon^2}$

$$P(|\overline{X}_n - \mu| > \varepsilon) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2}$$

Measurement Error

The Monte Carlo Method

Suppose we want to calculate

$$I(f) = \int_0^1 f(x)dx$$

where the integration cannot be done by elementary means.

The Monte Carlo method works as follows:

1. Generate independent uniform random variables on [0,1], $X_1, \ldots, X_n \sim U[0,1]$
2. Compute $Y_1 = f(X_1), \ldots, Y_n = f(X_n)$. Then $Y_1, \ldots, Y_n$ are i.i.d.
3. By the law of large numbers $\overline{Y}_n$ should be close to $\mathbb{E}[Y_1]$. Therefore:

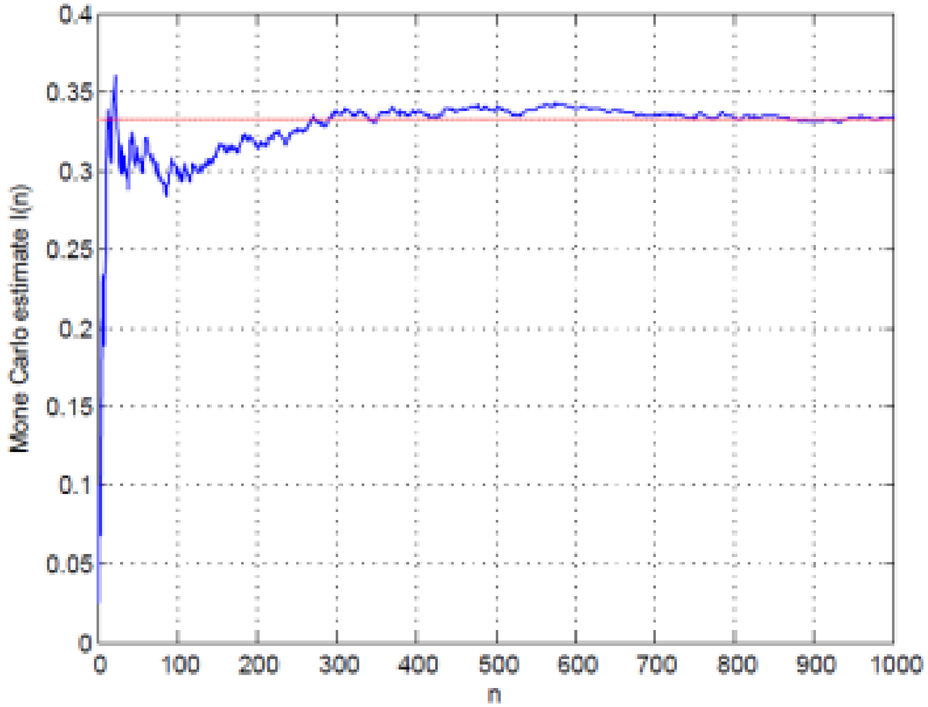
$$\boxed{\frac{1}{n}\sum_{i=1}^{n} f(X_i) = \overline{Y}_n \approx \mathbb{E}[Y_1] = \mathbb{E}[f(X_1)] = \int_0^1 f(x)dx}$$

Monte Carlo method: Example

Suppose we want to compute the following integral:

$$I = \int_0^1 x^2 dx$$

- From calculus: $I = 1/3$
- Using Monte Carlo method: $I(n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, where $X_i \sim U[0,1]$



Let $X_1, X_2, \ldots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \ldots,$ and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every point at which F is continuous. ■

THEOREM A *Continuity Theorem*

Let F_n be a sequence of cumulative distribution functions with the corresponding moment-generating function M_n. Let F be a cumulative distribution function with the moment-generating function M. If $M_n(t) \to M(t)$ for all t in an open interval containing zero, then $F_n(x) \to F(x)$ at all continuity points of F. ■

Moment-generating functions

Definition

The moment-generating function (MGF) of a random variable $X \sim f(x)$ is

$$M(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x) dx$$

(if the expectation is defined)

Important Properties of MGFs:

- If $\exists\ \varepsilon > 0$ such that $M(t)$ exists for all $t \in (-\varepsilon, \varepsilon)$, then $M(t)$ uniquely determines the probability distribution, $M(t) \rightsquigarrow f(x)$.
- If $M(t)$ exists in an open interval containing zero, then

$$M^{(r)}(0) = \mathbb{E}[X^r] \qquad \text{(hence the name)}$$

To find moments $\mathbb{E}[X^r]$, we must do integration.
Knowing the MGF allows to replace integration by differentiation.

Moment-generating functions

Important Properties of MGFs: (continuation)

- If X has the MGF $M_X(t)$ and $Y = a + bX$, then

$$M_Y(t) = e^{at} M_X(bt)$$

- If X and Y are independent, then

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

- If X and Y have a joint distribution, then their joint MGF is defined as

$$M_{X,Y}(s,t) = \mathbb{E}[e^{sX+tY}]$$

X and Y are independent if and only if

$$M_{X,Y}(s,t) = M_X(s) M_Y(t)$$

Let $\lambda_1, \lambda_2, \ldots$ be an increasing sequence with $\lambda_n \to \infty$, and let $\{X_n\}$ be a sequence of Poisson random variables with the corresponding parameters.

standardizing the random variables-

$$Z_n = \frac{X_n - E(X_n)}{\sqrt{\mathrm{Var}(X_n)}}$$

$$= \frac{X_n - \lambda_n}{\sqrt{\lambda_n}}$$

$$M_{X_n}(t) = e^{\lambda_n(e^t - 1)}$$

$$e^x = \textstyle\sum_{k=0}^{\infty} \frac{x^k}{k!},$$

$$M_{Z_n}(t) = e^{-t\sqrt{\lambda_n}} M_{X_n}\left(\frac{t}{\sqrt{\lambda_n}}\right)$$

$$= e^{-t\sqrt{\lambda_n}} e^{\lambda_n(e^{t/\sqrt{\lambda_n}} - 1)}$$

$$\Longrightarrow$$

$$\lim_{n\to\infty} \log M_{Z_n}(t) = \frac{t^2}{2}$$

$$\lim_{n\to\infty} M_{Z_n}(t) = e^{t^2/2}$$

The Central Limit Theorem

Suppose that $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2. The **central limit theorem** (CLT) says that $\overline{X}_n$ has a distribution which is approximately Normal. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance.

The Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

Useful Interpretation:

- Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution.

The Central Limit Theorem

$$\boxed{Z_n \equiv \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

There are several forms of notation to denote the fact that the distribution of Z_n is converging to a Normal. They all mean the same thing:

$$Z_n \;\dot{\sim}\; \mathcal{N}(0,1)$$

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

$$\overline{X}_n - \mu \;\dot{\sim}\; \mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$$

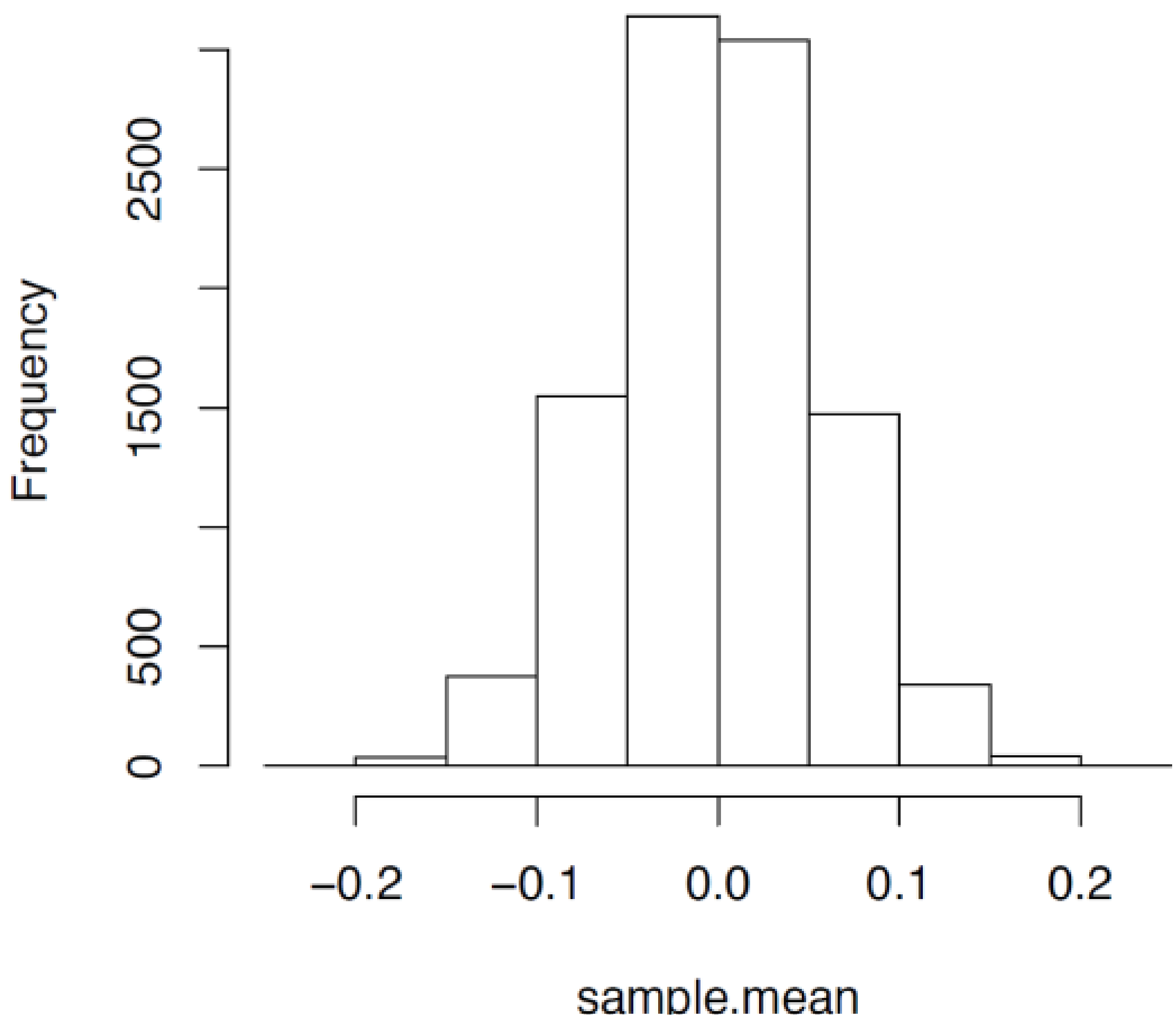
$$\sqrt{n}(\overline{X}_n - \mu) \;\dot{\sim}\; \mathcal{N}(0, \sigma^2)$$

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \;\dot{\sim}\; \mathcal{N}(0,1)$$

Sample mean of IID uniform

```
nreps = 10000
sample.mean = numeric(nreps)
n = 100
for (i in 1:nreps) {
    X = runif(n, min=-1, max=1)
    sample.mean[i] = mean(X)
}
hist(sample.mean)
```

Histogram of sample.mean



How good is this approximation? Here's a comparison of CDF values, for sample size $n = 10$:*

Normal	Exact
0.01	0.009
0.25	0.253
0.50	0.500
0.75	0.747
0.99	0.991

It's already very close! In general, accuracy depends on

- Sample size n,
- Skewness of the distribution of X_i, and
- Heaviness of tails of the distribution of X_i

The Central Limit Theorem: Remarks

- The CLT asserts that the CDF of $\overline{X}_n$, suitably normalized to have mean 0 and variance 1, converges to the CDF of $\mathcal{N}(0,1)$.

  Q: Is the corresponding result valid at the level of PDFs and PMFs?

  Broadly speaking the answer is yes, but some condition of smoothness is necessary (generally, $F_n(x) \to F(x)$ does not imply $F_n'(x) \to F'(x)$).

- The CLT does not say anything about the rate of convergence.
- The CLT tells us that in the long run we know what the distribution must be.
  - Even better: it is always the same distribution.
    - Still better: it is one which is remarkably easy to deal with, and for which we have a huge amount of theory.

Historic Remark:

- For the special case of Bernoulli variables with $p = 1/2$, CLT was proved by **de Moivre** around **1733**.
- General values of p were treated later by **Laplace**.
- The first rigorous proof of CLT was discovered by **Lyapunov** around **1901**.

The Central Limit Theorem: Example

- Suppose that the number of errors per computer program has a Poisson distribution with mean $\lambda = 5$. $f(k|\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$
- We get $n = 125$ programs; n is sample size
- Let $X_1, \ldots, X_n$ be the number of errors in the programs, $X_i \sim \mathrm{Poisson}(\lambda)$.
- Estimate probability $\mathbb{P}(\overline{X}_n \leq \lambda + \epsilon)$, where $\epsilon = 0.5$.

Answer:

$$\mathbb{P}(\overline{X}_n \leq \lambda + \epsilon) \approx \Phi\left(\epsilon\sqrt{\frac{n}{\lambda}}\right) = \Phi(2.5) \approx 0.994$$

The central limit theorem tells us that

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \,\dot{\sim}\, \mathcal{N}(0,1)$$

However, in applications, we rarely know σ. We can estimate σ^2 from $X_1, \ldots, X_n$ by sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

Question: If we replace σ with S_n is the central limit theorem still true?

Answer: Yes!

Theorem

Assume the same conditions as the CLT. Then,

$$\boxed{\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{D}} Z \sim \mathcal{N}(0,1)}$$

Summary

Theorem (LLN)

Suppose $X_1, \dots, X_n$ *are IID, with* $\mathbb{E}[X_1] = \mu$ *and* $\mathrm{Var}[X_1] < \infty$. *Let* $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. *Then, for any fixed* $\varepsilon > 0$, *as* $n \to \infty$,

$$\mathbb{P}[|\bar{X}_n - \mu| > \varepsilon] \to 0.$$

A sequence of random variables $\{T_n\}_{n=1}^{\infty}$ **converges in probability** to a constant $c \in \mathbb{R}$ if, for any fixed $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}[|T_n - c| > \varepsilon] \to 0.$$

So the LLN says $\bar{X}_n \to \mu$ in probability.

Theorem (CLT)

Suppose $X_1, \ldots, X_n$ are IID, with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2 < \infty$. Let $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. Then, for any fixed $x \in \mathbb{R}$, as $n \to \infty$,

$$\mathbb{P}\left[\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \leq x\right] \to \Phi(x),$$

where Φ is the CDF of the $\mathcal{N}(0,1)$ distribution.

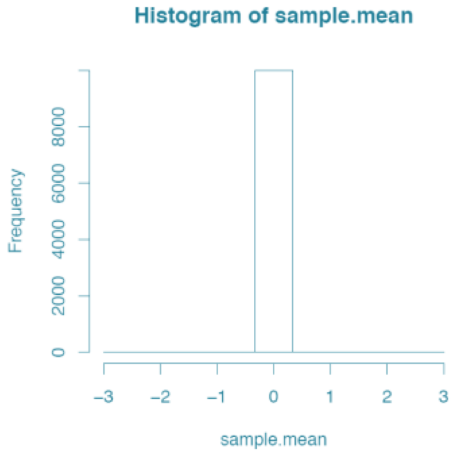
$\{T_n\}_{n=1}^{\infty}$ **converges in distribution** to a probability distribution with CDF F if, for every $x \in \mathbb{R}$ where F is continuous, as $n \to \infty$,

$$\mathbb{P}[T_n \leq x] \to F(x).$$

We sometimes write $T_n \to Z$ in distribution, where Z is a random variable having this distribution F. So the CLT says $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to Z$ in distribution where $Z \sim \mathcal{N}(0,1)$.

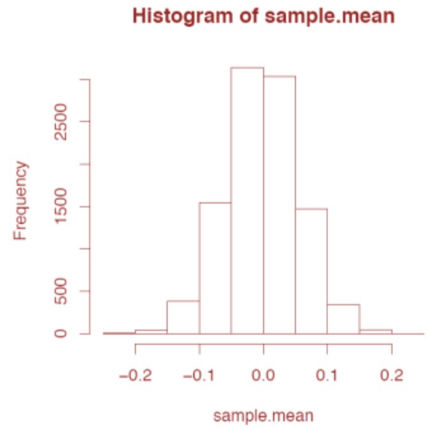
The Difference is in Scaling

$X_1, \ldots, X_{100} \sim \text{Uniform}(-1, 1)$. $\bar{X}_{100}$ across 10000 simulations:



This illustrates the LLN, that is, $\bar{X}_n \to 0$ in probability.

Here's the exact same histogram, on a different scale:



This illustrates the CLT, that is, $\sqrt{3n}\bar{X}_n \to \mathcal{N}(0, 1)$ in distribution. (Here $\text{Var}[X_1] = \frac{1}{3}$.)

```
nreps = 10000
sample.mean = numeric(nreps)
n = 100
for (i in 1:nreps) {
    X = runif(n, min=-1, max=1)
    sample.mean[i] = mean(X)
}
hist(sample.mean)
```

Multivariate Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d. random vectors with mean μ and covariance matrix Σ:

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \\ \vdots \\ X_{ki} \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mathbb{E}[X_{1i}] \\ \mathbb{E}[X_{2i}] \\ \vdots \\ \mathbb{E}[X_{ki}] \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \mathbb{V}[X_{1i}] & \mathrm{Cov}(X_{1i}, X_{2i}) & \ldots & \mathrm{Cov}(X_{1i}, X_{ki}) \\ \mathrm{Cov}(X_{2i}, X_{1i}) & \mathbb{V}[X_{2i}] & \ldots & \mathrm{Cov}(X_{2i}, X_{ki}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_{ki}, X_{1i}) & \ldots & \mathrm{Cov}(X_{ki}, X_{k-1i}) & \mathbb{V}[X_{ki}] \end{pmatrix}$$

Let $\overline{X}_n = (\overline{X}_{1n}, \ldots, \overline{X}_{kn})^T$. Then

$$\boxed{\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)}$$

Reference

The slides contents come from USC mathematical statistics and Stanford course + John A. Rice's book