Central Limit Theorem

Overview

Central Limit Theorem (CLT)

“Central" result in Statistics

Mild conditions

Normalized sum of random variables is roughly Normal

Discrete, continuous , mixed distributions

Explains why Bell Curve so popular

Allows for simple probability estimation

Seeing is believing

Mixed Distributions

Central Limit Theorem

iid - independent identically distributed

μ - mean

σ - standard deviation

Let $X_1, X_2, X_3, \ldots$ be iid with finite μ and σ.

As $n \to \infty$, the distribution of $\frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$ approaches $\mathcal{N}(0,1)$.

$-n\mu$

$/\sigma$

Normalization

Without loss of generality, assume $\mu=0$, $\sigma=1$

$X_1, X_2, X_3, \ldots$ iid with $\mu=0$ and $\sigma=1$. As $n \to \infty$, the distribution of $\frac{X_1+X_2+\ldots+X_n}{\sqrt{n}}$ approaches $\mathcal{N}(0,1)$.

Greatly generalizes the Weak Law of Large Numbers

WLLN  CLT

X_1, X_2, X_3,… iid with μ=0 and σ=1

WLLN

As n→∞, $\frac{X_1+X_2+\ldots+X_n}{n} \to 0$

$E=0$

$V(X_1+\ldots+X_n)=n$

$V\left(\frac{X_1+\ldots+X_n}{n}\right)=\frac{n}{n^2}=\frac{1}{n}\to 0$

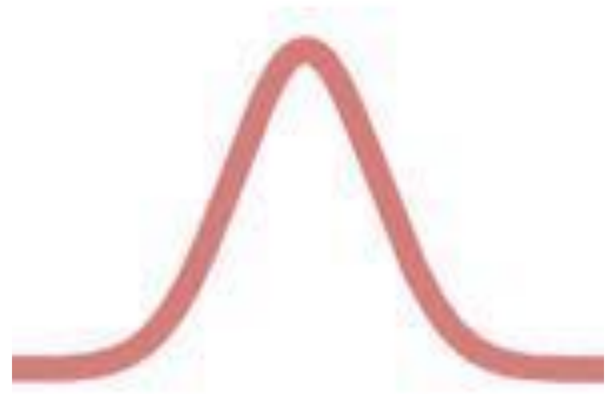
CLT

As n→∞, the distribution of $\frac{X_1+X_2+\ldots+X_n}{\sqrt{n}} \to \mathcal{N}(0,1)$ distribution

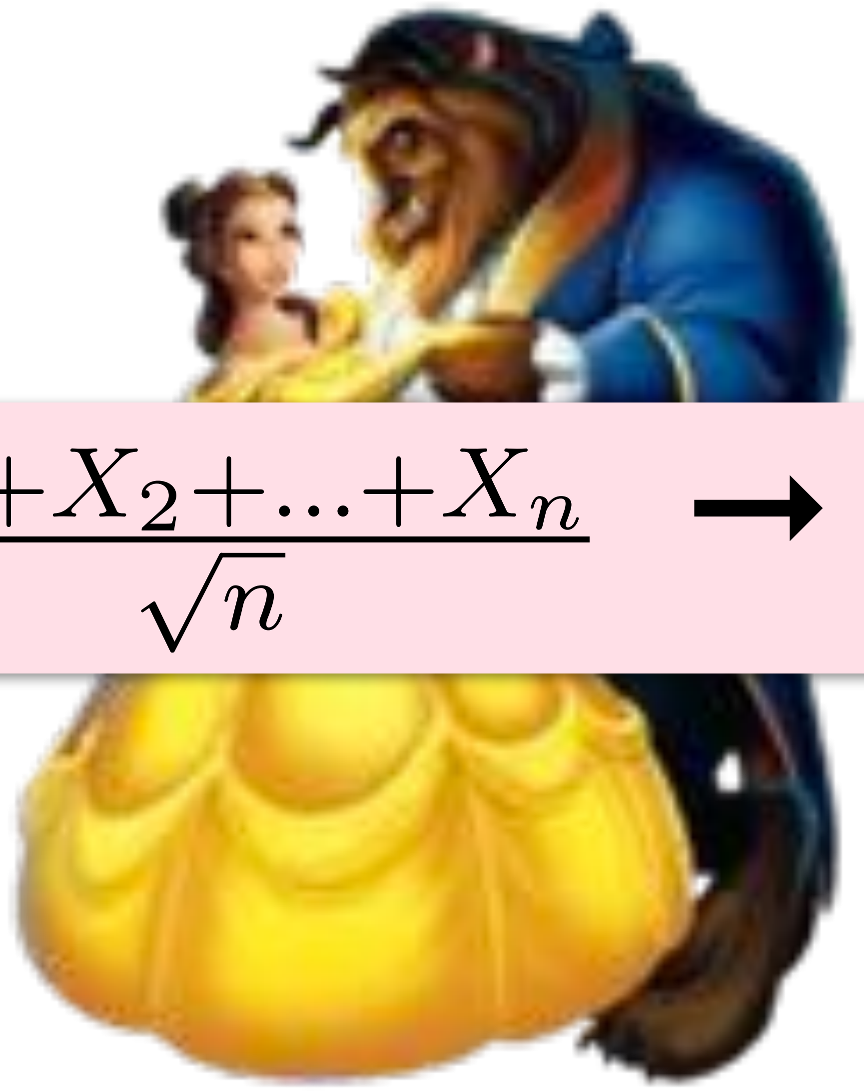
$E=0$

$V\left(\frac{X_1+X_2+\ldots+X_n}{\sqrt{n}}\right)=\frac{n}{n}=1$

More importantly



CLT



As n→∞, the distribution of $\frac{X_1+X_2+...+X_n}{\sqrt{n}}$ → $\mathcal{N}(0,1)$ distribution

$$V\left(\frac{X_1+X_2+...+X_n}{\sqrt{n}}\right)=\frac{n}{n}=1$$

Normalize by

$\gg \sqrt{n}$ V → 0

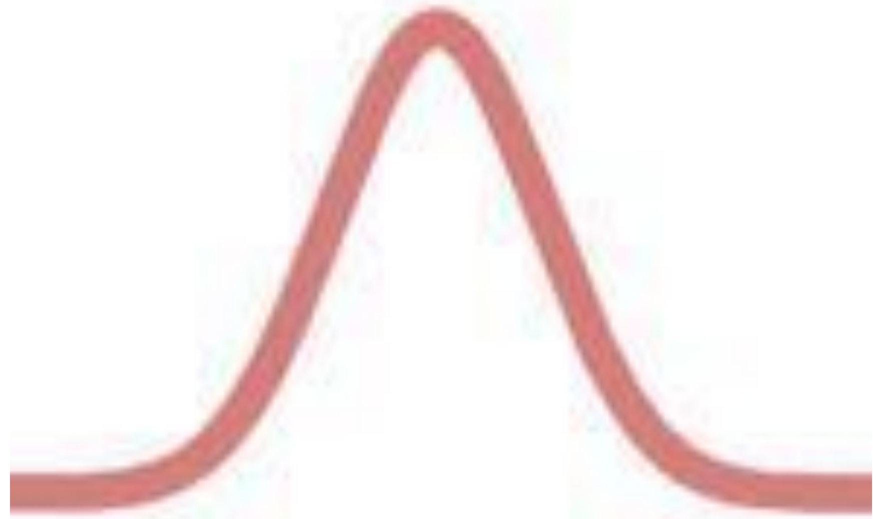
$\ll \sqrt{n}$ V → ∞

$\sqrt{n}$ V = 1

For any X_i

“Normal”

Converge to $\mathcal{N}(0,1)$



CLT

Does the extra resolution help?

ADD

Show example where following not useful

Markov (≤)

Chebyshev (≤ sigma)

WLLN (weak - in the wrong way)

Store Income

At a store

Customer's spending

$\mu = 80$

$\sigma = 40$

P(a customer spends ≤ 72)

10% below average

???

Care more about total revenue

100 Customers

P(average spending ≤ 72)

Can answer!

Thanks to CLT!

General CLT Application

$$X_1, \ldots, X_n \perp\!\!\!\perp, \sim \text{ any distribution with mean } \mu, \text{ and stdv } \sigma$$

$$\overline{X^n} \stackrel{\text{def}}{=} \frac{X_1 + \ldots + X_n}{n} \qquad P(\overline{X^n} \le \alpha) = ?$$

$$Z_n \stackrel{\text{def}}{=} \frac{(X_1 + \ldots + X_n) - n\mu}{\sigma\sqrt{n}}$$

Typically ≥ 30

Roughly

CLT

For sufficiently large n

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

$$Z_n = \frac{n\overline{X^n} - n\mu}{\sigma\sqrt{n}} = \frac{n(\overline{X^n} - \mu)}{\sigma\sqrt{n}} = \frac{\overline{X^n} - \mu}{\sigma/\sqrt{n}}$$

$$P(\overline{X^n} \le \alpha) = P\left(Z_n \le \frac{\alpha - \mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{\alpha - \mu}{\sigma/\sqrt{n}}\right)$$

General formula for average of random variables

CDF

$X \sim N(0,1)$

$$\Phi(x) \triangleq F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy$$

No known formula

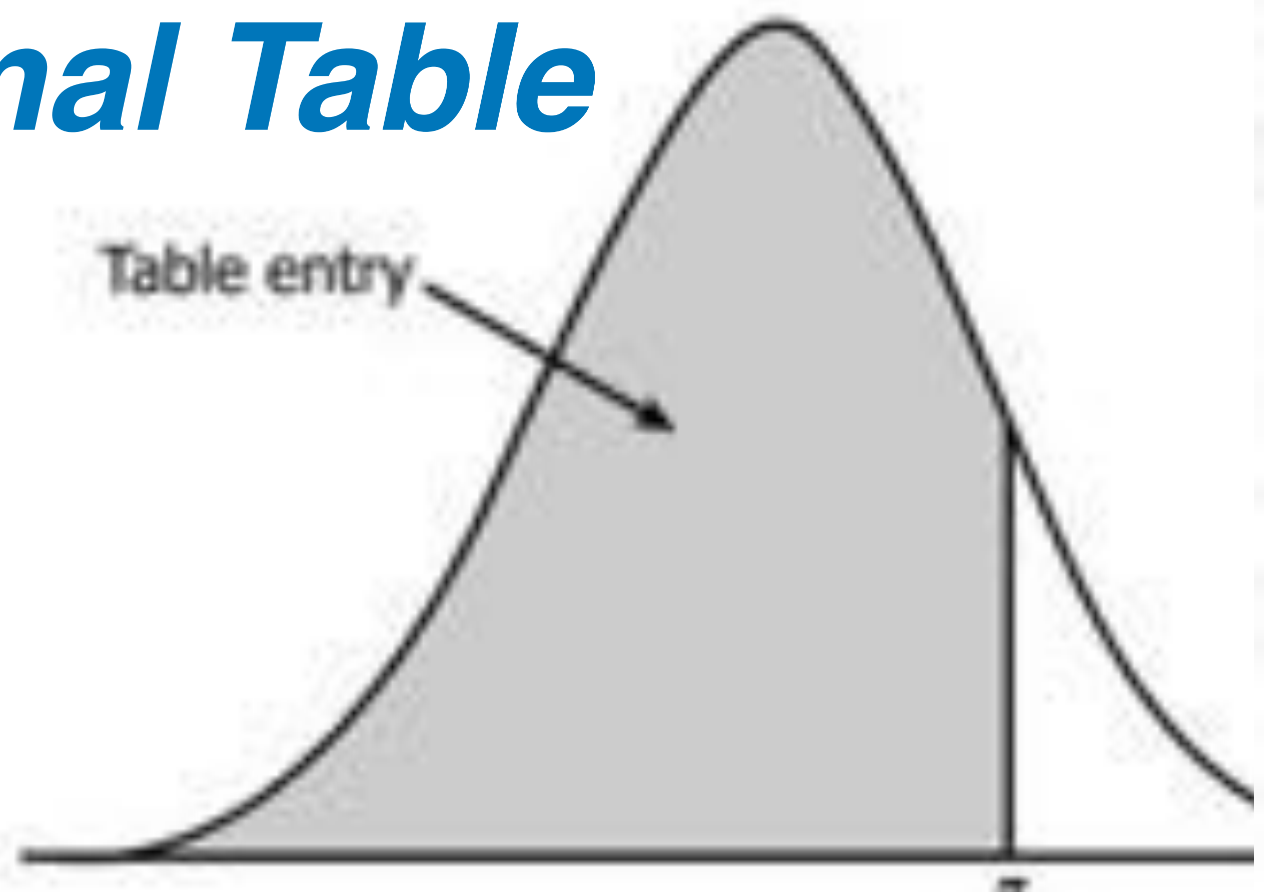
Instead use table or computer

Table for each μ, σ?

1 suffices!

Standard Normal Table

Z Table



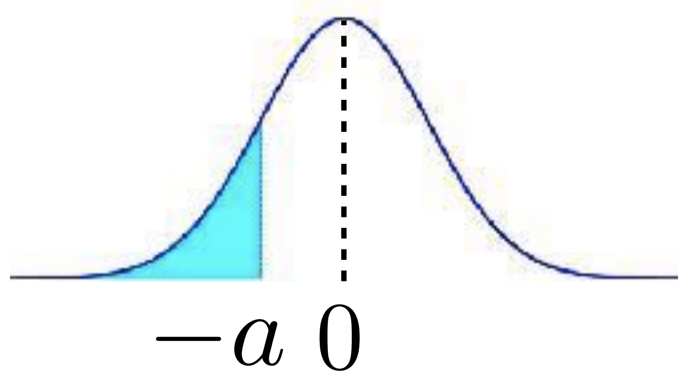
Store

Customer's spending X_i $\perp\!\!\!\perp$ $\mu = 80$ $\sigma = 40$ $n = 100$ Customers

P(average spending ≤ 72) 10% below average

$$P(\overline{X^n} \le \alpha) \approx \Phi\left(\frac{\alpha-\mu}{\sigma/\sqrt{n}}\right)$$

$$P(\overline{X^{100}} \le 72) \approx \Phi\left(\frac{72-80}{40/\sqrt{100}}\right) = \Phi(-\tfrac{8}{4}) = \Phi(-2)$$



$$P(X \le -a) = \Phi(-a) = 1 - \Phi(a)$$

$$= 1 - \Phi(2) \approx 1 - 0.9772 = 0.0228 \approx 2.3\%$$

Change to insurance outlays so ≤ a

Do we need to artificially introduce multiple instances?

Not necessarily, some phenomena are MI by themselves

In communication, noise is gaussian

What about normalization by sqrt n - why would we do that

Central Limit Theorem

How the CLT is Made

Convergence

$X_1, X_2, X_3, \ldots$ iid $\mu=0$ $\sigma=1$ $Z_n \stackrel{\text{def}}{=} \frac{X_1+X_2+\ldots+X_n}{\sqrt{n}}$

As n→∞, the distribution of Z_n approaches $\mathcal{N}(0,1)$.

Cumulative distribution function (CDF)

$Z \sim \mathcal{N}(0,1)$

$$F_{Z_n}(x) = P(Z_n \leq x) = P\left(\frac{X_1+\ldots+X_n}{\sqrt{n}} \leq x\right)$$

$$F_Z(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt \stackrel{\text{def}}{=} \Phi(x)$$

$$F_{Z_n}(x) \xrightarrow[n\to\infty]{} F_Z(x) \quad \forall x$$

CDF of standard normal distribution

Convergence “in distribution”

Moment Generating Functions

$$M_X(t) \stackrel{\text{def}}{=} E[e^{tX}]$$

Scaling and independent addition

$$M_{aX}(t) = E(e^{t\cdot aX}) = E(e^{at\cdot X}) = M_X(at)$$

$$M_{X+Y}(t) = E[e^{t(X+Y)}] = E[e^{tX}\cdot e^{tY}] = E[e^{tX}]\cdot E[e^{tY}] = M_X(t)\cdot M_Y(t)$$

Derivatives of M (at 0) are moment of X

$$M_X^{(n)}(0) = E[X^n]$$

$$E[e^{0X}]$$

$$M'_X(0) = E[X] \qquad M''_X(0) = E[X^2] \qquad M_X(0) = E[X^0] = 1$$

Standard Normal $\quad Z \sim \mathcal{N}(0,1) \qquad M_Z(t) = e^{\frac{t^2}{2}}$

Continuity

CDF - cumulative distribution fun.

MGF - moment generating fun.

Random Variables	CDF	MGF
$Z_1, Z_2, Z_3, \ldots$	$F_{Z_n}(t)$	$M_{Z_n}(t)$
Z	$F_Z(t)$	$M_Z(t)$

If $M_{Z_n}(t) \to M_Z(t) \quad \forall t$

then $F_{Z_n}(t) \to F_Z(t)$ wherever $F_Z(t)$ is continuous

Plan

$Z \sim \mathcal{N}(0,1)$ $\qquad$ $M_Z(t) = e^{\frac{t^2}{2}}$

Show $M_{Z_n}(t) \underset{n\to\infty}{\longrightarrow} e^{\frac{t^2}{2}}$ Imply $F_{Z_n}(t) \underset{n\to\infty}{\longrightarrow} \Phi(t)$

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}}$$

$$M_{\frac{X_i}{\sqrt{n}}}(t) = M_{X_i}\left(\frac{t}{\sqrt{n}}\right) = M\left(\frac{t}{\sqrt{n}}\right)$$

$$M_{Z_n}(t) = \prod_{i=1}^{n} M_{\frac{X_i}{\sqrt{n}}}(t) = \prod_{i=1}^{n} M\left(\frac{t}{\sqrt{n}}\right) = \left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n$$

Show $\left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n \underset{n\to\infty}{\longrightarrow} e^{\frac{t^2}{2}}$

L'Hôpital's Rule

Johann Bernoulli

Seek $\lim\limits_{x\to a} \frac{f(x)}{g(x)}$ a - anything, finite or $\pm\infty$

$$\lim f(x) = 6 \quad \lim g(x) = 3 \longrightarrow \quad \lim \frac{f(x)}{g(x)} = \frac{6}{3} = 2$$

$$\lim f(x) = \lim g(x) = 0 \quad \text{or} \quad \lim f(x),\ \lim g(x) = \pm\infty \longrightarrow \quad \lim \frac{f(x)}{g(x)} \text{ unclear}$$

$$\text{If } \lim \frac{f'(x)}{g'(x)} \text{ exists} \qquad \text{then } \lim \frac{f(x)}{g(x)} = \lim \frac{f'(x)}{g'(x)}$$

Hospital → Hôpital → Ô

$$\lim_{n\to\infty}\left[M\left(\frac{t}{\sqrt{n}}\right)\right]^n = e^{\frac{t^2}{2}} \quad \leftrightarrow \quad \lim_{n\to\infty} n\cdot\ln M\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2}$$

$$\lim_{n\to\infty} n\cdot\ln M\left(\frac{t}{\sqrt{n}}\right) = \lim_{u\to 0}\frac{\ln M(tu)}{u^2}$$

$M(0) = E(X_i^0) = 1$

$M^{(i)}(0) = E(X^i)$

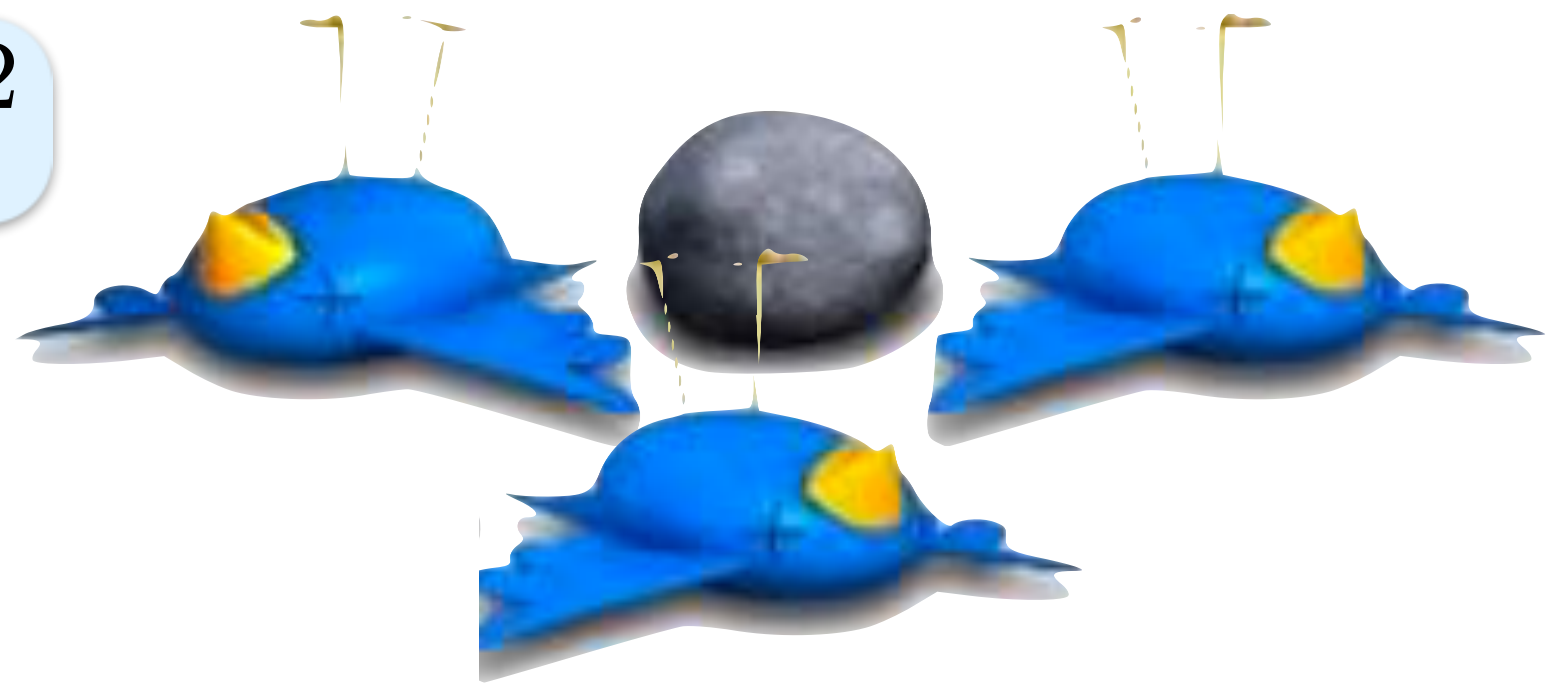
$\infty\cdot 0 = ?$

$u = \frac{1}{\sqrt{n}}$

$\frac{0}{0} \rightarrow$ Ô

n

$1/\sqrt{n} \rightarrow u^2$



Ô

$$= \lim_{u\to 0}\frac{M'(tu)\cdot t}{M(tu)\cdot 2u}$$

t - constant

$$= \frac{t}{2}\lim_{u\to 0}\frac{M'(tu)}{u}$$

$M'(0) = E(X_i^1) = \mu = 0$

Ô

$$= \frac{t}{2}\lim_{u\to 0}\frac{M''(tu)\cdot t}{1}$$

$M''(0) = E(X_i^2) = \sigma^2 = 1$

$$= \frac{t^2}{2}$$

Assembling

$X_i \quad \text{iid}, \quad \mu = 0, \quad \sigma^2 = 1$

$Z_n = \frac{X_1 + \ldots + X_n}{\sqrt{n}}$

$$\lim_{n \to \infty} M_{Z_n}(t) = \lim_{n \to \infty} \left[M\left(\frac{t}{\sqrt{n}} \right) \right]^n = e^{\frac{t^2}{2}} = M_Z(t)$$

$Z \sim \mathcal{N}(0, 1)$

$F_{Z_n} \to F_Z = \Phi$

Z_n approach Standard Normal distribution

General iid X_i just define $Z_n = \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$

Generalizations

Showed for iid

Any original distribution, need just finite μ, σ, M

Discrete

Continuous

Mixed

More generally

Mild conditions

Independent

Weakly dependent

Multi-dimensional

Normal distribution very prevalent

How the CLT is Made