Central limit theorem & Normal distribution

$Z \sim N(0,1)$ Z is said to have normal distribution with mean 0 and variance 1 if

$$\mathbb{P}(Z \leq x) = \int_{-\infty}^{x} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\, dy$$

[improper Riemann integral] → [Numerical Evaluation by normal tables]

- The distribution seems to occur naturally in many biological observations.

  E.g. — Height of individuals
  - # of leafs on a tree

  arise as sum of independent random process

  thus some contributing to low / middle / high

  yielding a frequency plot that is symmetrical

- This is a fact that can be established theoretically.

[De Moivre - Laplace C.L.T.] :- Let $Z_1, Z_2, \ldots Z_n$ be i.i.d Bernoulli(p), $0 < p < 1$ then

$$\hat{p}_n = \frac{\sum_{i=1}^{n} Z_i}{n} \quad \text{- sample proportion}$$

$$\sqrt{n}\left(\frac{\hat{p}_n - p}{\sigma}\right) \xrightarrow{d} N(0,1) \quad \text{— } (\circledast)$$

[i.e. Suppose $W_n = \sqrt{n}\,\frac{(\hat{p}_n - p)}{\sigma}$ & $Z \sim N(0,1)$

$P(W_n \le x) \longrightarrow P(Z \le x)$ as $n \to \infty$)]

for all $x \in \mathbb{R}$

Proof [IDEA] :- Rewrite $(\circledast)$:- if $S_n = \sum_{i=1}^{n} Z_i$

$$\frac{S_n - np}{\sqrt{n}\,\sigma} \xrightarrow{d} N(0,1)$$

Now, $S_n \overset{d}{=} \text{Binomial}(n,p)$

$$\left| P\left(a \le \frac{S_n - np}{\sqrt{n}\,\sigma} < b\right) - \int_a^b \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\,dy \right|$$

$\sigma = p(1-p)$

$$\le \left| \sum_{k = \lceil a\sqrt{npq} + np \rceil}^{\lceil b\sqrt{npq} - 1 + np \rceil} P(S_n = k) - \int_a^b \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\,dy \right|$$

$q = 1 - p$

$$\leq \left| \sum_{k=[a\sqrt{npq}+np]}^{[b\sqrt{npq}-1+np]} \left[P(S_n = k) - \frac{1}{\sqrt{npq}} \phi\left(\frac{k-np}{\sqrt{npq}}\right) \right] \right|$$

$$+ \left| \sum_{k=[a\sqrt{npq}+np]}^{[b\sqrt{npq}-1+np]} \frac{1}{\sqrt{npq}} \phi\left(\frac{k-np}{\sqrt{npq}}\right) - \int_a^b \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy \right|$$

$$= \quad I \quad + \quad II$$

II - Riemann Sum Convergence

I - $\cdot P(S_n = k) = \dfrac{n!}{k!\,(n-k)!} p^k q^{n-k}$

[Using these two]

$\cdot \; n! \sim n^n e^{-n} \sqrt{n} \cdot \sqrt{2\pi}$

$I + II \to 0$ as $n \to \infty$ □

Central Limit Theorem: $X_1, \ldots, X_n$ are i.i.d random variables with $E X_1 = \mu$ & $\mathrm{Var}(X_1) = \sigma^2$

Then

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \xrightarrow{d} N(0,1).$$

Proof :- uses several results ... □
beyond the scope of this class

Central Limit Theorem

Suppose we want to verify the below result via simulations:

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2. Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} Z, \tag{1}$$

where $Z \sim$ Normal $(0, 1)$.

Central Limit Theorem

Suppose we want to verify the below result via simulations:

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2. Then

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{d} Z, \tag{2}$$

where $\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$ and $Z \sim \text{Normal}\ (0, 1)$.

Central Limit Theorem

We could rephrase the result as:

Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ, finite variance σ^2. Then

$$\frac{(S_n - n\mu)}{\sqrt{n}\sigma} \xrightarrow{d} Z, \tag{3}$$

where $S_n = X_1 + X_2 + \ldots + X_n$ and $Z \sim \text{Normal}\ (0, 1)$.

Central Limit Theorem

Suppose each X_i was distributed as Bernoulli (p) random variable. Then S_n is a Binomail(n,p) random variable. Let us check for what p does

$$\frac{S_n - np}{\sqrt{np(1-p)}}$$

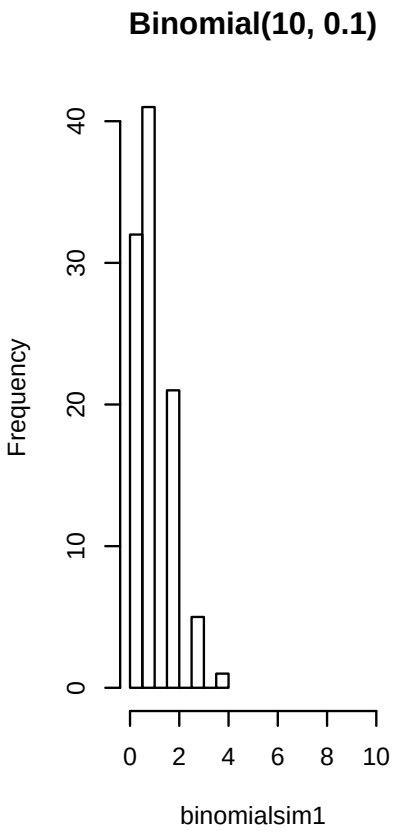
is close to a Normal distribution.

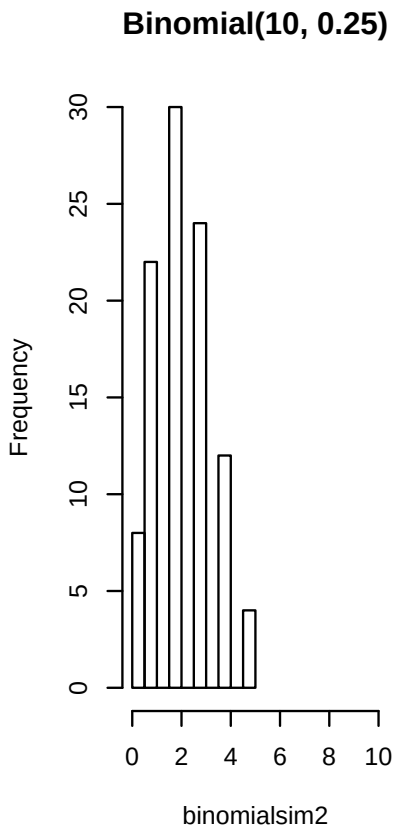
Central Limit Theorem

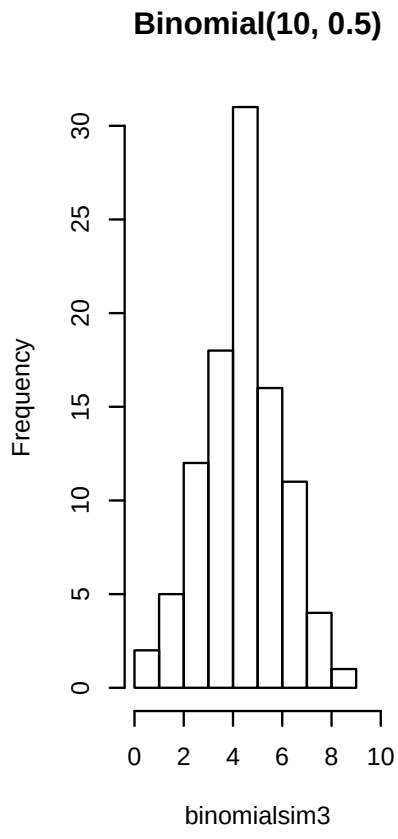
We may simulate Binomial samples either direclty by `rbinom` command or usi ng the `replicate` and `rbinom` command.

```
> binomialsim1 = rbinom(100,10,0.1)
> # generates 100 Binomial (10,0.1) samples
>
> binomialsim2 = replicate(100, rbinom(1,10,0.25))
> # generates 100 Binomial (10,0.25) samples
>
> binomialsim3 = replicate(100, rbinom(1,10,0.5))
> # generates 100 Binomial (10,0.5) samples
>
```

Histogram of all three simulations

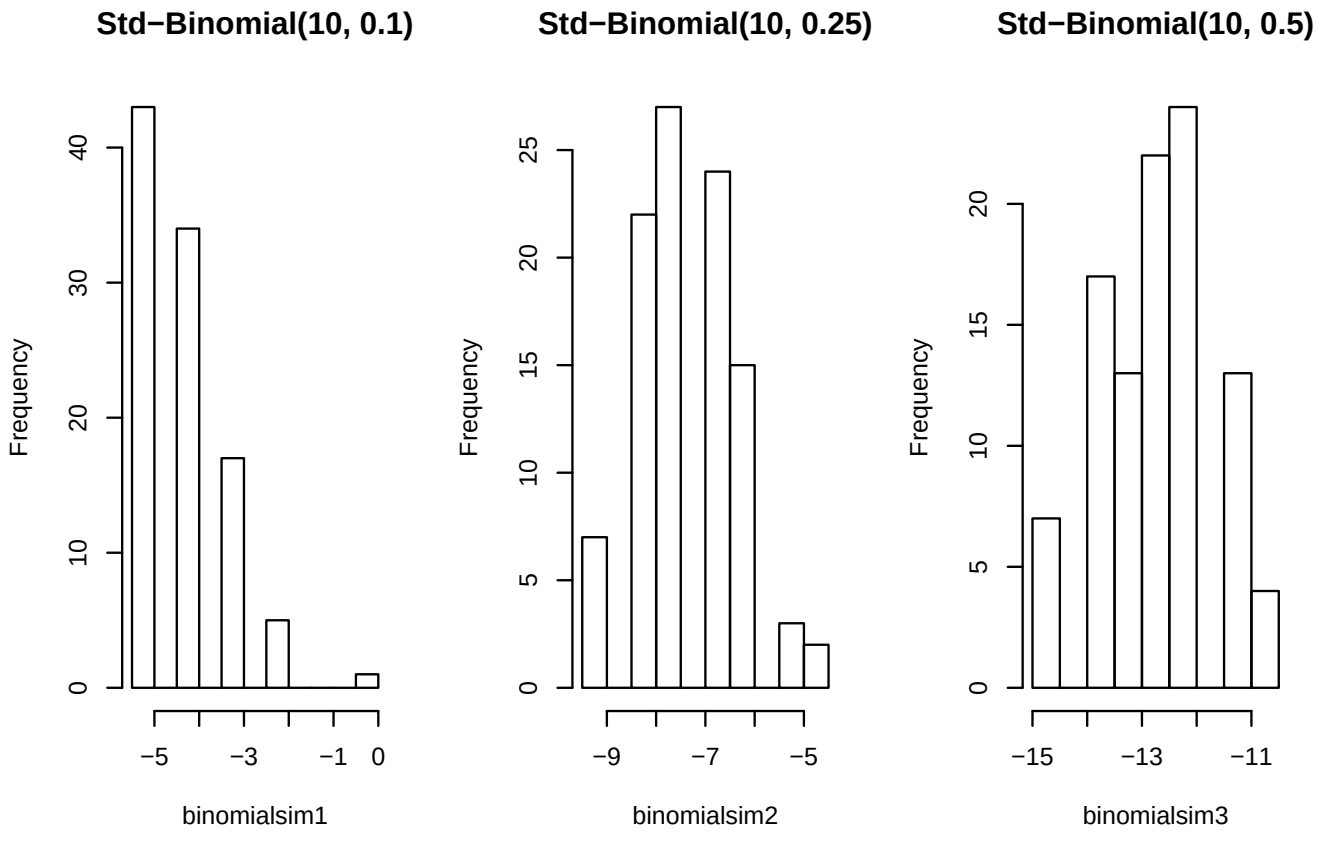






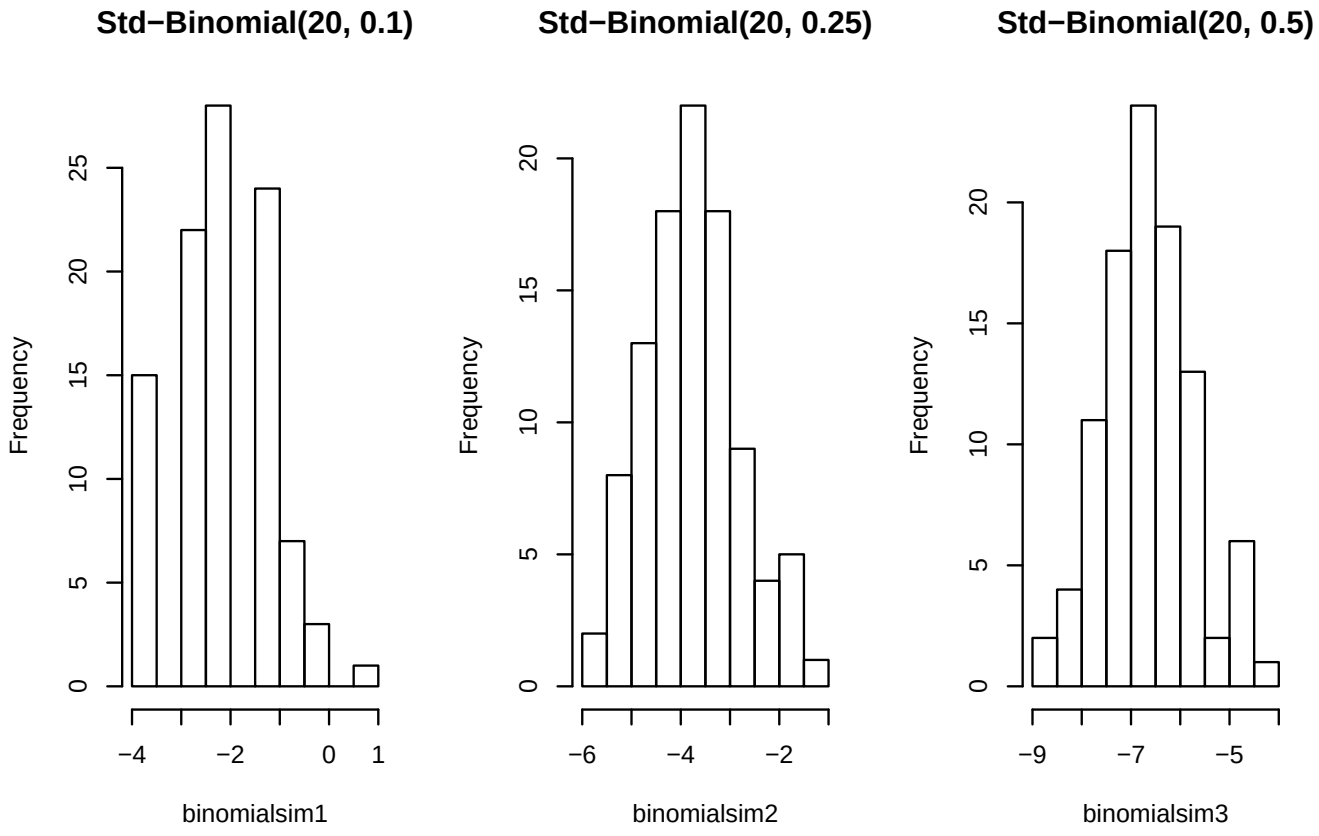
From the above it seems that at $n = 10$ the symmetry is achieved when $p = 0.5$ and not at $p = 0.1$ and $p = 0.25$

Standardised Histograms: Binomial n=10 and p=0.1, 0.25, 0.5



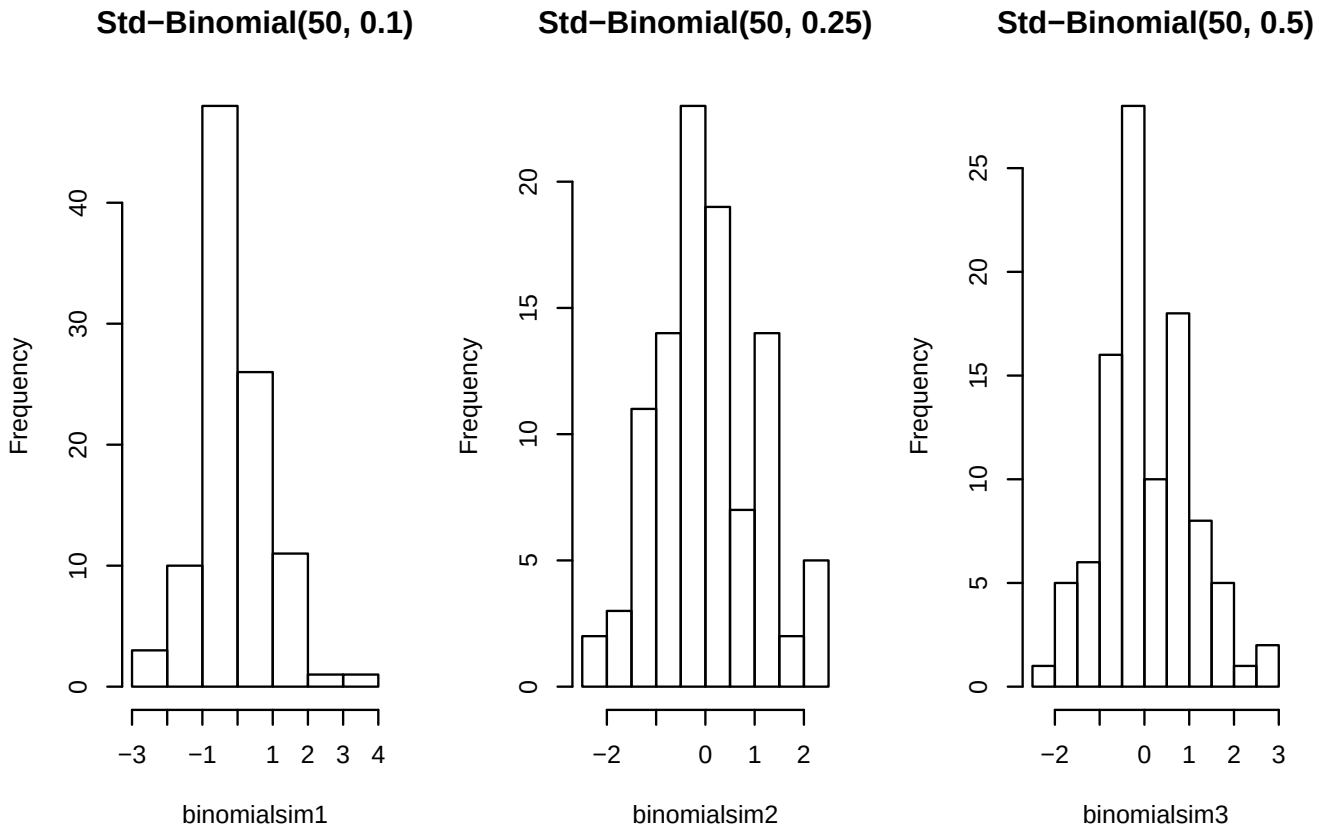
Perhaps $n = 10$ is not large enough to see the Central Limit Theorem occuring.

Standardised Histograms: Binomial n=20 and p=0.1, 0.25, 0.5



$n = 20$ is better.

Standardised Histograms: Binomial n=50 and p=0.1, 0.25, 0.5



$n = 50$ we get closer to Normal distribution

Role of n versus p

Binomial Random variable is close to Normal when the distribution is symmetric. That is when p is close to 0.5. Otherwise the general rule that we can apply is that when

$$np \geq 5 \text{ and } n(1-p) \geq 5.$$

then Binomial(n,p) is close to Normal distribution.