Recall :- Estimation

$X_1, X_2, \ldots, X_n$ i.i.d. from a population X

X had pmf / pdf $f(\cdot \mid \theta)$ $\theta \in \rho \subseteq \mathbb{R}^d$

Goal :- Estimate θ

- **Point Estimator :** $g : \mathbb{R}^n \to \mathbb{R}$

(Suitable) $g(X_1, X_2, \ldots, X_n)$ as point estimator for θ.

Unbiased :- $E[g(X_1, X_2, \ldots, X_n)] = \theta$

Consistent :- $Var[g(X_1, \ldots, X_n)] \to 0$ as $n \to \infty$.

Two methods :-

(i) Method of moments.

calculate : $\frac{1}{n} \sum_{i=1}^{n} X_i^k \quad k = 1, \ldots, d$

Equate with true moments : $E[X^k]$, $k = 1, \ldots, d$

Solve the d - equations with d- unknowns to find θ.

(ii) Maximum likelihood Estimate

$$l_n(\theta \mid X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Keep $(X_1, \ldots, X_n)$ fixed and maximize l as a function of θ.

Interval Estimation : Used the Central limit theorem to provide an interval estimate with some confidence for the mean.

Central Limit Theorem

Recall:- [SLLN] $\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i = \mu\right) = 1$

i.e. $\bar{X}_n \longrightarrow \mu$ as $n \to \infty$ w.p. 1

Second order Result: "fluctuations of $\bar{X}_n - \mu$". *

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

$(\sigma \neq 0)$

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

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

*: • $\bar{X}_n \longrightarrow \mu$ w.p. 1 as $n \to \infty$ [SLLN] — 1^{st} order

• $(\bar{X}_n - \mu) := \frac{\sigma}{\sqrt{n}} C_n$ where $C_n \xrightarrow{d} Z$ — 2^{nd} order

Central Limit Theorem

Significance :- $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\sqrt{n}(\frac{S_n}{n} - \mu)}{\sigma} = \frac{S_n - n\mu}{\sqrt{n}\sigma}$

We could rephrase the result as:

* "Sums of i.i.d r.v" ≡ "Normal Random variables"

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

($\sigma \neq 0$)

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

where $S_n = X_1 + X_2 + \dots + X_n$ and $Z \sim$ Normal $(0, 1)$.

*: $\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim Z$ for large $n >>$ $\Leftrightarrow$ $S_n \sim \sqrt{n}\sigma Z + n\mu$ for large n $\Leftrightarrow$ $S_n \sim \text{Normal}(n\mu, n\sigma^2)$ for large n

i.e. $S_n \sim AN(n\mu, n\sigma^2)$

Central Limit Theorem – Special case.

$X_i = \begin{cases} 0 & \text{w.p } 1-p \\ 1 & \text{w.p } p \end{cases} \Rightarrow X_i \overset{d}{\sim} \text{Bernoulli}(p)$ & $S_n = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n,p)$

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

$p \notin \{0,1\}$

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

is close to a Normal distribution.

Berry – Esseen worksheet

Variation

"large n" w.r.t p.

for approximation to Normal

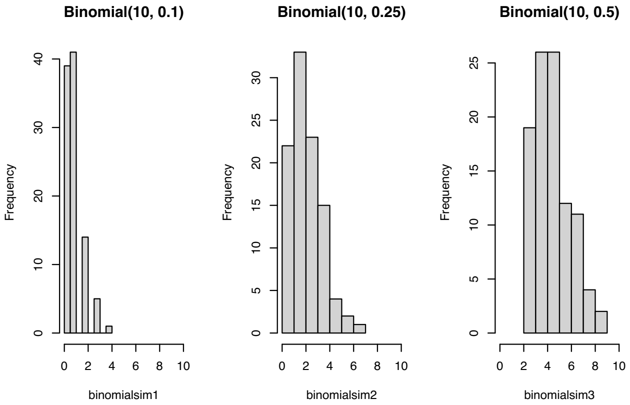
Is $\boxed{S_n \sim AN(np, np(1-p))}$?

Central Limit Theorem

We may simulate Binomial samples either directly by `rbinom` command or using the `replicate` and `rbinom` command.

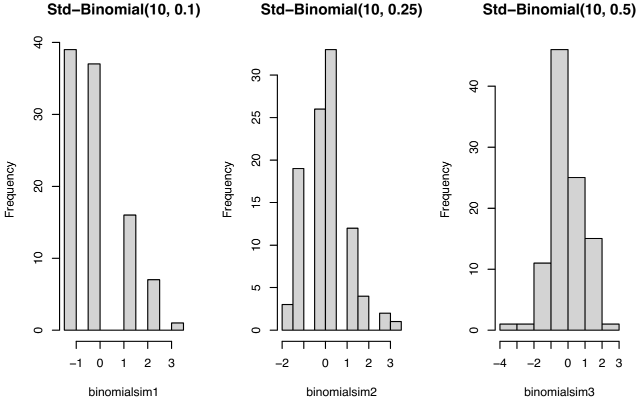
```
> binomialsim1 = rbinom(100,10,0.1)
> # generates 100 Binomial (10,0.1) samples
>
> binomialsim2 = replicate(100, rbinom(1,10,0.25))
> # generates 100 Binomial (10,0.25) samples
>
> binomialsim3 = replicate(100, rbinom(1,10,0.5))
> # generates 100 Binomial (10,0.5) samples
>
```

Histogram of all three simulations



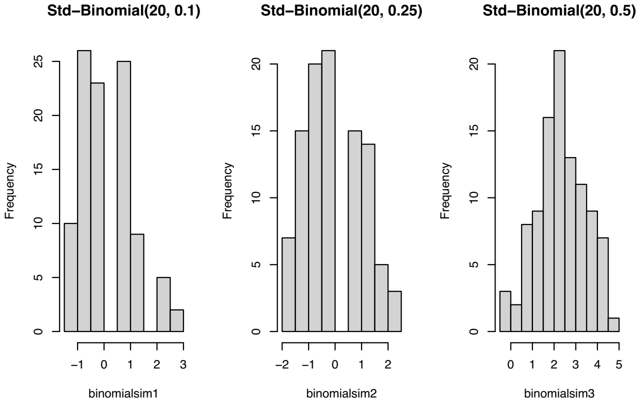
From the above it seems that at $n = 10$ the symmetry is achieved when $p = 0.5$ and not at $p = 0.1$ and $p = 0.25$

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



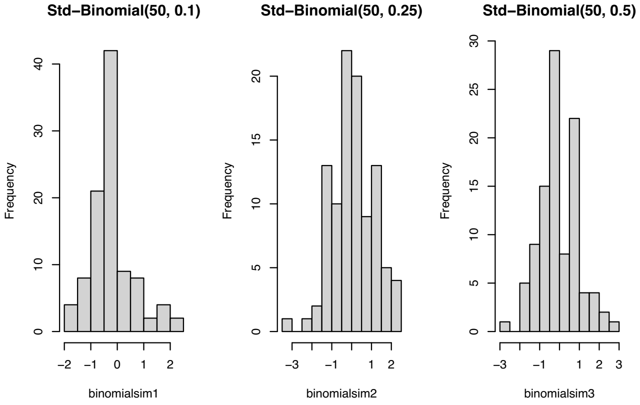
Perhaps $n = 10$ is not large enough to see the Central Limit Theorem occuring.

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

Ber(0) or Ber(1) : only failure or only success : NOT symmetric

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.

Confidence Intervals — Recall

$Z \sim \text{Normal}(0,1)$ $\quad \mathbb{P}(|Z| \leq 1.96) \simeq 0.95$

Using the Central Limit Theorem for large n we have

$$P\left(\left| \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right| \leq 1.96\right) \approx 0.95$$

which is the same as saying

$$P\left(\mu \in \left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X}\right)\right) \approx 0.95$$

The interval $\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X}\right)$ is called the 95% confidence interval for μ.

Confidence Intervals

- $\mathbb{P}\left(\mu \in \left(\frac{-1.96\sigma}{\sqrt{n}} + \bar{X}, \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right)\right) \simeq 0.95$

95% confidence interval for μ is $\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X}\right)$

Meaning: for n large if we did m (large) repeated trials and computed the above interval for each trial then true mean would belong to approximately 95% of m intervals calculated.

Confidence Intervals

with
$Var(x) = 1$

The below is code for finding the confidence interval for a data x. — $(x_1, \ldots x_n)$

```
> cifn = function(x, alpha=0.95){
+ z = qnorm( (1-alpha)/2, lower.tail=FALSE)
+ sdx = sqrt(1/length(x))
+ c(mean(x) - z*sdx, mean(x) + z*sdx)
+ }
```

$z = 1.96$

$sdx = 1/\sqrt{n}$

$\left(\bar{x} - \frac{1.96}{\sqrt{n}}, \quad \bar{x} + \frac{1.96}{\sqrt{n}}\right)$

Three Confidence Intervals for Normal(0,1)

100 samples from Normal (0,1)

```
> x1 = rnorm(100,0,1);y = cifn(x1)
```

$X^1_1, X^1_2, \ldots, X^1_{100}$

```
> y
[1] -0.1624472  0.2295456
```

100 samples from Normal (0,1)

```
> x2 = rnorm(100,0,1);z = cifn(x2)
```

$X^2_1, X^2_2, \ldots, X^2_{100}$

```
> z
[1] -0.2167657  0.1752271
```

100 samples from Normal (0,1)

```
> x3 = rnorm(100,0,1);w = cifn(x3)
```

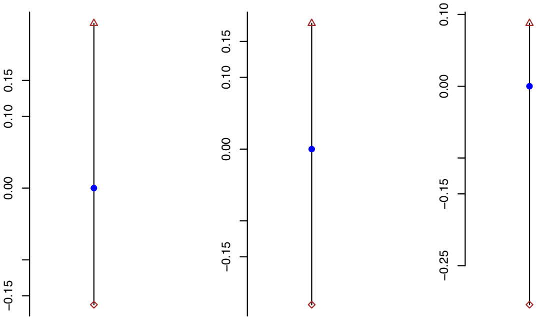
$X^3_1, X^3_2, \ldots, X^3_{100}$

```
> w
[1] -0.30436422  0.08762858
```

Does 0 belong to all the three confidence intervals ?

Confidence Intervals Plots

The below is a plot of the three confidence intervals computed in the previous slide.



In this simulation 0 belongs to all three intervals.

Confidence Intervals : 10 Trials

We generate 10 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

10 trials

```
> normaldata = replicate(10, rnorm(100,0,1),
+ simplify=FALSE)
> cidata = sapply(normaldata, cifn)
```

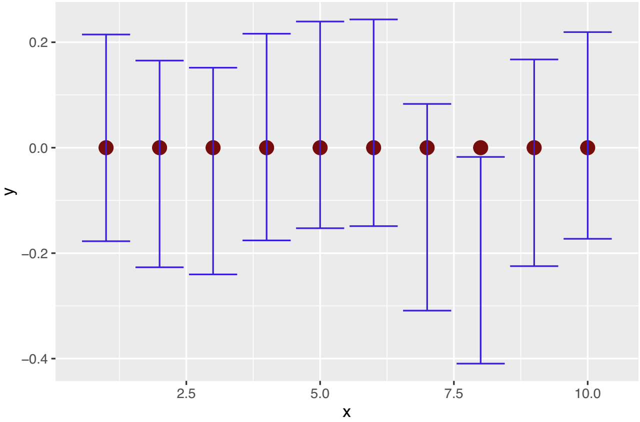
apply cifn to each trial

It is easy to check how many of them contain 0.

$[a,b] \ni 0$
$\Rightarrow a*b<0$

```
> TRUEIN = cidata[1,]*cidata[2,]<0
> table(TRUEIN)
TRUEIN
FALSE  TRUE
    1     9
```

Confidence Intervals : 10 Trials



Confidence Intervals: 40 Trials

We generate 10 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

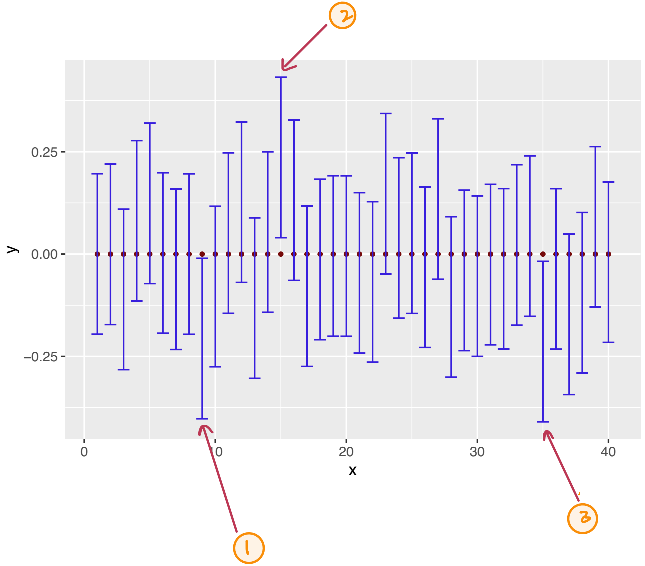
```
> normaldata = replicate(40, rnorm(100,0,1),
+ simplify=FALSE)
> cidata = sapply(normaldata, cifn)
```

It is easy to check how many of them contain 0.

```
> TRUEIN = cidata[1,]*cidata[2,]<0
> table(TRUEIN)
```

```
TRUEIN
FALSE  TRUE
    3    37
```

Confidence Intervals: 40 trials Plot



Confidence Intervals : 100 Trials

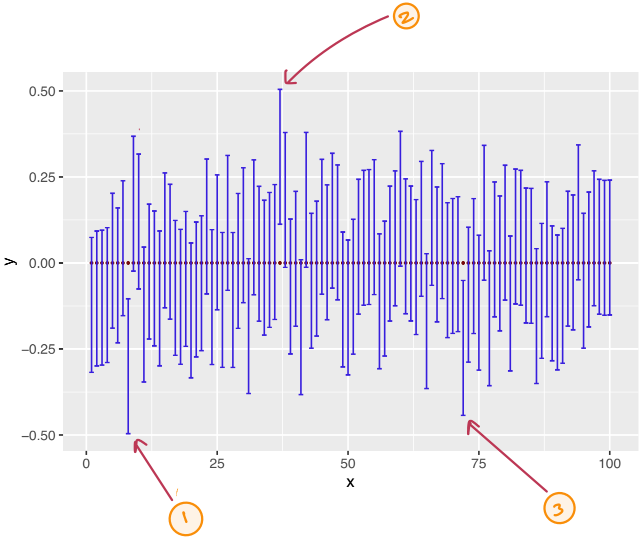
We generate 100 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

```
> normaldata = replicate(100, rnorm(100,0,1),
+ simplify=FALSE)
> cidata = sapply(normaldata, cifn)
```

It is easy to check how many of them contain 0.

```
> TRUEIN = cidata[1,]*cidata[2,]<0
> table(TRUEIN)
TRUEIN
FALSE  TRUE
    3    97
```

Confidence Intervals : 100 Trials



Confidence Intervals

95% confidence interval for μ is $\left(-\frac{1.96\sigma}{\sqrt{n}} + \bar{X}, \frac{1.96\sigma}{\sqrt{n}} + \bar{X}\right)$

Meaning: for n large if we did m (large) repeated trials and computed the above interval for each trial then true mean would belong to approximately 95% of m intervals calculated.

Thus numerically the above meaning seems to hold for a Normal population.