Central limit Theorem and Confidence Intervals

Sample from population

$X_1, X_2, \ldots, X_n$ i.i.d. X

[with replacement ; without replacement]

- Suppose $\mu = E[X]$, $\sigma^2 = \text{var}[X]$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \equiv \ldots \text{estimate} \ldots = \mu$$

[Justification] required

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \equiv \ldots \text{estimate} \ldots = \sigma^2$$

[Justification] required

- - Summary : $\bar{X}$, σ_x^2, median, min, max of the distribution
  - Plots : Histogram, box plots, q-q plots.

- Empirical distribution :- $S = \{X_1, X_2, \ldots, X_n\}$
  - include repeat observations

$$f_n(t) = \frac{\#\{i : X_i = t\}}{n}$$

- Discrete p.m.f. on S ≡ inference based on this is called descriptive statistics

Ex: Suppose Y (r.v.) has p.m.f $f_n(\cdot)$

$\mathbb{P}(Y=t) = f_n(t)$

$E[Y] = \bar{X}$, $\text{Var}[Y] = ?$

$Z \sim \text{Normal}(0,1)$ if

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

Normal Tables to evaluate numerically

Z has p.d.f

$$f_Z(z) = \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \quad z \in \mathbb{R}$$

Central limit Theorem

- Distribution occurs naturally.
- arises as sum of independent processes
  - # of leafs in a tree
  - height of individuals.

[Suitable interpretation]

Normal Distribution: PDF

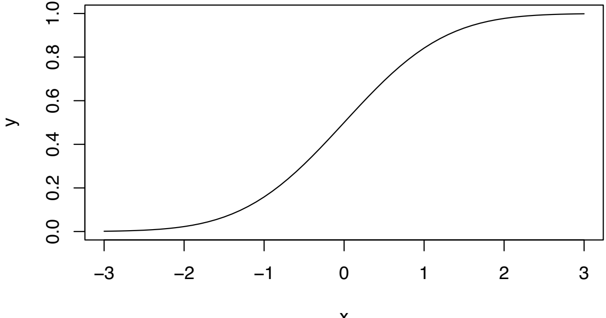
You can calculate the values of the normal density function using the the `dnorm` command.

```
> dnorm(0)
[1] 0.3989423
> dnorm(1)
[1] 0.2419707
> dnorm(0, mean=4, sd=3)
[1] 0.05467002
```

Normal Distribution: CDF

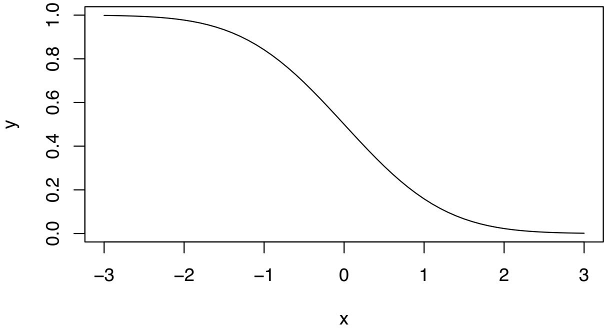
You can calculate the values of the cummulative distribution function of the normal using the the `pnorm` command.

```
> pnorm(0)
> pnorm(1)
> x = seq(-3,3, by=0.1); y = pnorm(x) ;plot(x,y, type="l")
```



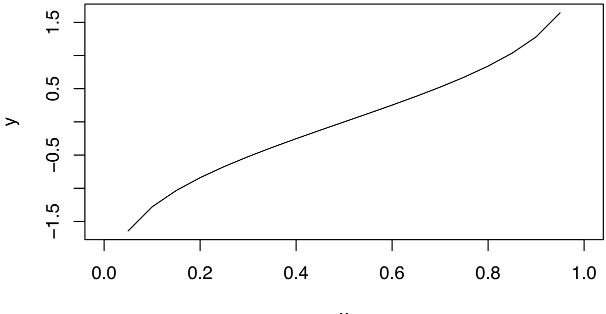
Normal Distribution: Tail Probabilities

```
> pnorm(0, lower.tail=FALSE)
> pnorm(1, lower.tail=FALSE)
> x = seq(-3,3, by=0.1); y = pnorm(x, lower.tail=FALSE)
> plot(x,y, type="l")
```



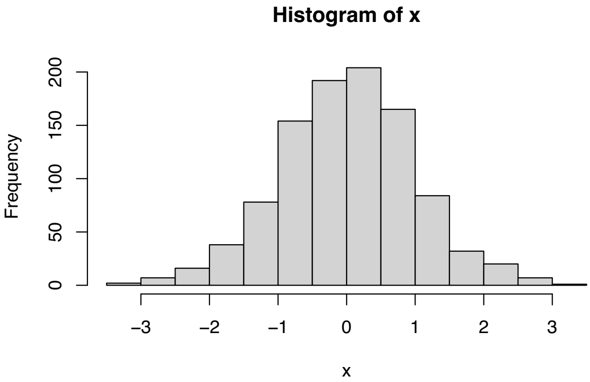
Normal Distribution: quantiles

```
> qnorm(0.68); qnorm(0.95);qnorm(0.997)
[1] 0.4676988
[1] 1.644854
[1] 2.747781
> x = seq(0,1, by=0.05); y = qnorm(x);plot(x,y, type="l")
```



Normal Distribution: samples

```
> x=rnorm(1000)
> hist(x)
```



- Symmetric about mean = 0

Normal Distribution: Key Probabilities 68-95-99

```
> pnorm(1) - pnorm(-1) # within one standard deviation
[1] 0.6826895
> pnorm(2) - pnorm(-2) # within two standard deviation
[1] 0.9544997
> pnorm(3) - pnorm(-3) # within three standard deviation
[1] 0.9973002
```

Central Limit Theorem - Recall from Earlier Worksheet

$S_n = \sum_{i=1}^{n} X_i$ with $X_i \sim$ Bernoulli(p) independent for $i \geq 1$

Q:- How good is the Normal approximation?

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

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

is close to a Normal distribution.

Noted in Worksheet:

$$\left| \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \leq x \right) - \mathbb{P}(Z \leq x) \right| \leq \frac{C_1}{\sqrt{n}} \longrightarrow 0 \text{ as } n \to \infty$$

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

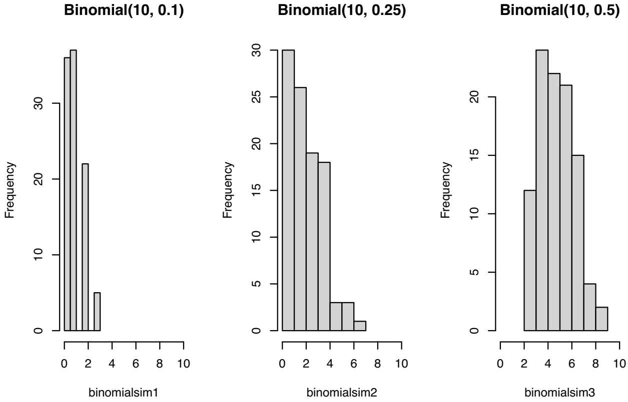
Central Limit Theorem

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
>
```

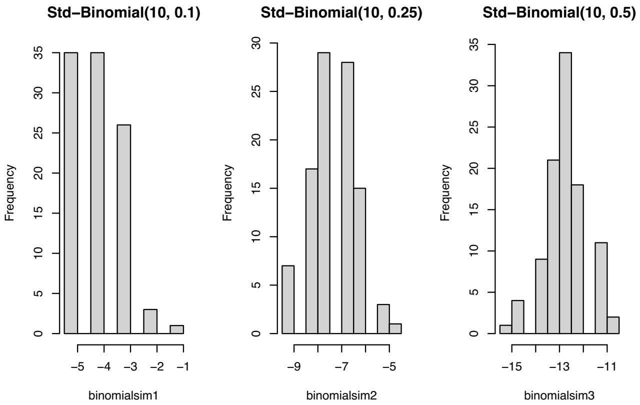
Histogram of all three simulations

$n = 100$



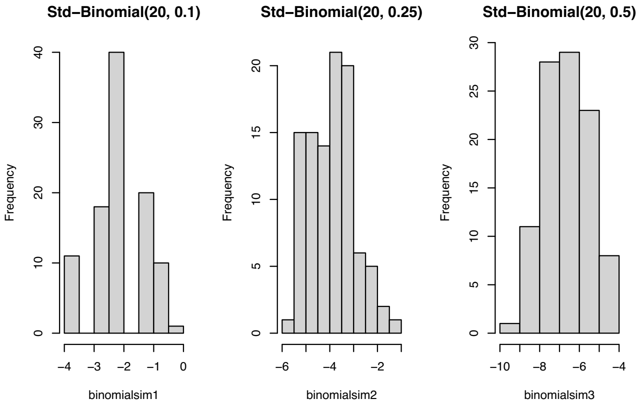
From the above it seems that at $n = 10$ the symmetry is achieved when $p = 0.5$ and not at $p = 0.1$ and $p = 0.25$

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



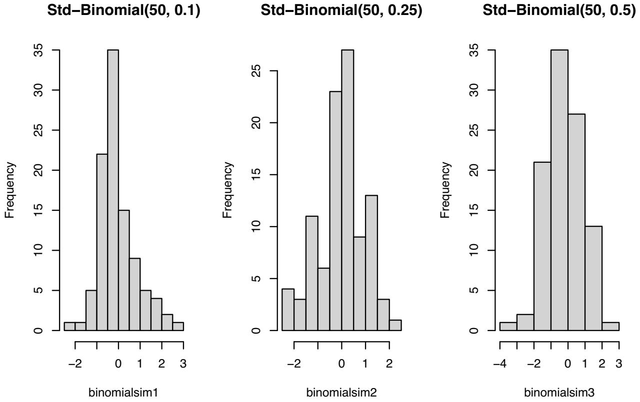
Perhaps $n = 10$ is not large enough to see the Central Limit Theorem occuring.

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.

Central Limit Theorem – "True in general for sums"

Sample: $X_1, X_2, \ldots, X_n$ i.i.d X $E[X] = \mu$; $Var[X] = \sigma^2$

We could rephrase the result as:

Fundamental Result

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

i.e. $\left| \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n}\sigma} \le x \right) - \mathbb{P}(Z \le x) \right| \longrightarrow 0$ as $n \to \infty$

"occurs naturally as sums" $\Longleftrightarrow$ "$S_n \sim N(n\mu, n\sigma^2)$"

Central Limit Theorem

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \frac{\sqrt{n}\left(\frac{S_n}{n} - \mu\right)}{\sigma} = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

Re phrase :-

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

ie. $\left| \mathbb{P}\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \le x\right) - \mathbb{P}(Z \le x) \right| \longrightarrow 0$ as $n \to \infty$

Confidence Interval

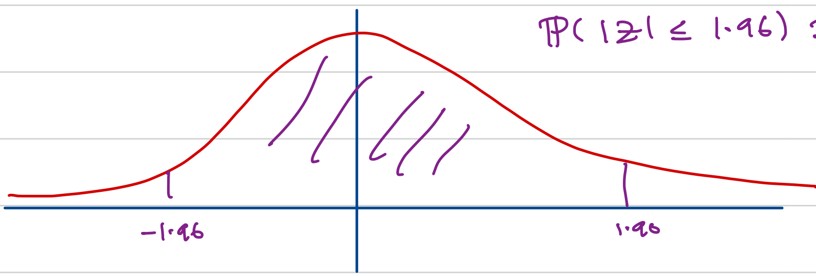
$X_1, \ldots, X_n$ i.i.d μ - mean

σ^2 - variance

$$\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \underset{\text{Central limit Theorem}}{=} \underset{(0,1)}{\text{Normal}} - Z$$

$$\mathbb{P}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le x\right) \text{ "="} \ \mathbb{P}(Z \le x)$$

$\mathbb{P}(|Z| \le 1.96) \simeq 0.95$



$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

$$\left|\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right| \underset{\text{Central limit Theorem}}{=} \underset{(0,1)}{\text{Normal}} - |Z|$$

$$\left|\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right| \le 1.96 \iff -1.96 \le \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \le 1.96$$

$$\iff (-1.96)\sigma \le \sqrt{n}(\bar{X}-\mu) \le (1.96)\sigma$$

$$\iff \bar{X} - \frac{1.96\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{1.96\sigma}{\sqrt{n}}$$

Confidence Intervals

Sample $x_1, \ldots, x_n$ from population. Assume i.i.d. & $E[X] = \mu$, σ – known

Compute: $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$

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.

↳ dependent on sample. and is valid if σ is known.

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.

Confidence Intervals

The below is code for finding the confidence interval for a data x.

```
> 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)
+ }
```

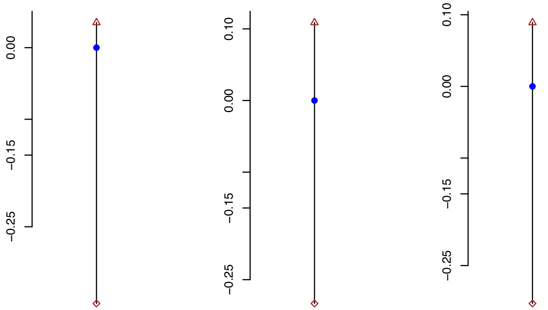
Three Confidence Intervals for Normal(0,1)

```
> x1 = rnorm(100,0,1);y = cifn(x1)
> y
[1] -0.35705304  0.03493976
> x2 = rnorm(100,0,1);z = cifn(x2)
> z
[1] -0.2832489  0.1087439
> x3 = rnorm(100,0,1);w = cifn(x3)
> w
[1] -0.30294682  0.08904598
```

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.



Confidence Intervals : 10 Trials

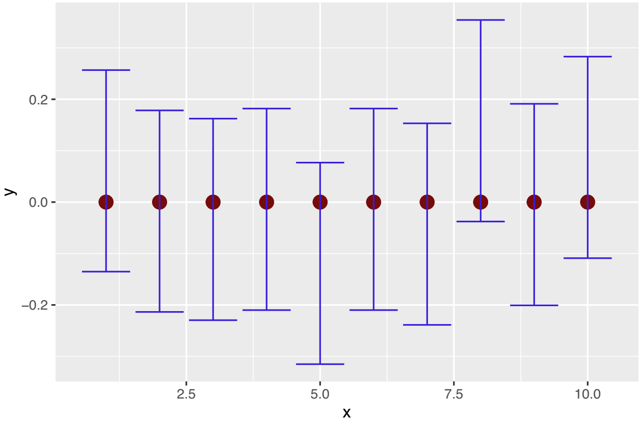
We generate 10 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

```
> normaldata = replicate(10, 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
TRUE
  10
```

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.

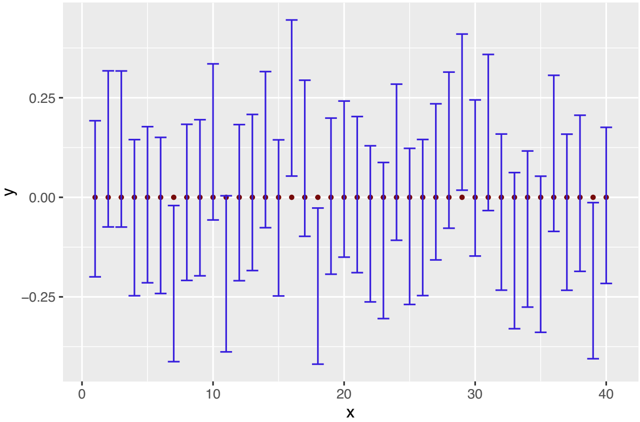
```
> normaldata = replicate(40, rnorm(100,0,1),
+ simplify=FALSE)
> cidata = sapply(normaldata, cifn)
```

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

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

```
TRUEIN
FALSE  TRUE
    5    35
```

Confidence Intervals: 40 trials Plot



Confidence Intervals : 100 Trials

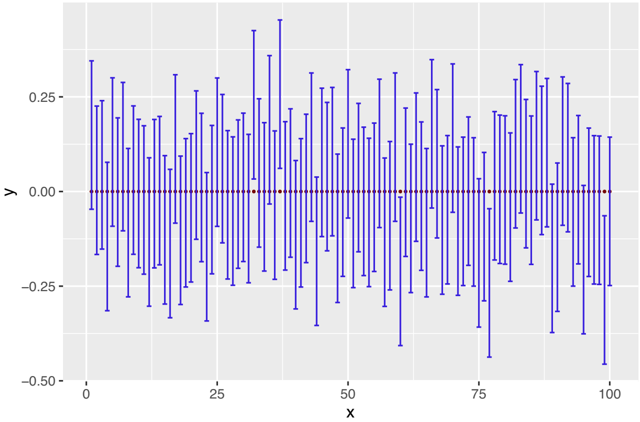
We generate 100 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

```
> 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
    5    95
```

Confidence Intervals : 100 Trials



Confidence Intervals : 1000 Trials

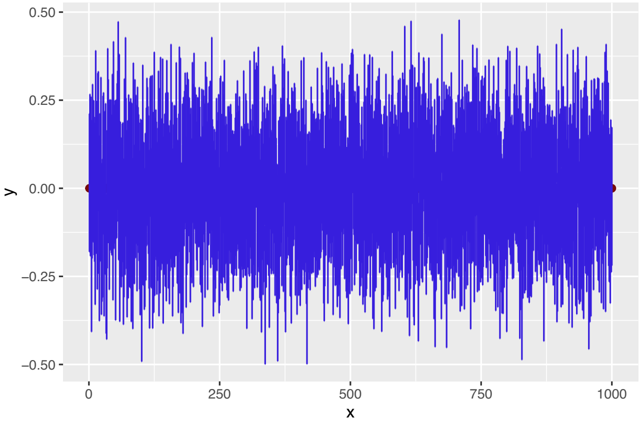
We generate 1000 trials of 100 samples from Normal(0,1) and compute the confidence intervals using the function defined earlier.

```
> normaldata = replicate(1000, 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
   54   946
```

Confidence Intervals : 1000 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.