Linear Regression-2

Prof. Asim Tewari
IIT Bombay

Linear Regression

$$Y = \underbrace{\beta_0 + \beta_1}\, X + \epsilon$$

0 mean

σ^2 var

Data

X ,	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n

Sample

Limited Data (y_i, x_i) → $\hat{\beta}_0$, $\hat{\beta}_1$

$\hat{\beta}_0$, $\hat{\beta}_1$

Characteristic Function

The characteristic function of a random variable X is

$$\phi_X(t) \equiv E\left(e^{itx}\right) = \int_{-\infty}^{\infty} e^{itx} \underbrace{f_X(x)}_{\text{pdf of } X}\, dx$$

$$e^{itx} = \frac{1}{\angle 0} + \frac{itx}{\angle 1} + \frac{(itx)^2}{\angle 2} + \cdots$$

$$\therefore \phi_X(t) = E\left[\frac{1}{\angle 0} + \frac{itx}{\angle 1} + \frac{(itx)^2}{\angle 2} + \cdots\right]$$

$$= \frac{1}{\angle 0} + \frac{it\,E[x]}{\angle 1} + \frac{(it)^2 E[x^2]}{\angle 2} + \frac{(it)^3 E[x^3]}{\angle 3} + \cdots$$

Characteristic Function

$$\therefore \phi_X(t) = \frac{1}{\lfloor 0} + \frac{itm_1}{\lfloor 1} + \frac{(it)^2 m_2}{\lfloor 2} + \cdots$$

where m_n is the n^{th} moment of the r.v. X

i.e. $m_n \equiv E[X^n]$

$$\phi_X(t)\Big|_{t=0} = 1 \; ; \; \frac{d}{dt}\phi_X(t)\Big|_{t=0} = i m_1 \; ; \; \frac{d^n}{dt^n}\phi_X(t)\Big|_{t=0} = (i)^n m_n$$

Moment generating Function

A moment generating function of a r.v. X is

$$M_X(t) = \phi_X(-it) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx$$

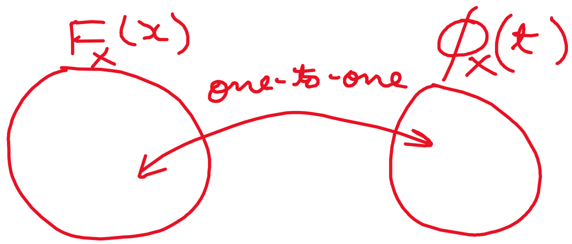
Now

$$\because \quad \left.\frac{d^n}{dt^n}\phi_X(t)\right|_{t=0} = (i)^n m_n \implies \left.\frac{d^n}{dt^n}M_X(t)\right|_{t=0} = m_n = n^{th} \text{ moment of } X$$

Characteristic Function

$$\therefore \phi_X(t) = \frac{1}{0!} + \frac{it}{1!} m_1 + \frac{(it)^2}{2!} m_2 + \frac{(it)^3}{3!} m_3 + \cdots$$

There is a one-to-one correspondance between the cumulative distribution function and the characteristic function.



If the r.v. has a probability density function $f_X(x)$ then

$$f_X(x) = F_X'(x) = \frac{1}{2\pi} \int e^{-itx} \phi_X(t)\, dt$$

Characteristic Function

$$\therefore \phi_X(t) = \frac{1}{\underline{|0}} + \frac{it\, m_1}{\underline{|1}} + \frac{(it)^2}{\underline{|2}} m_2 + \frac{(it)^3}{\underline{|3}} m_3 + \cdots$$

If a r.v. X has $\mu = 0$ and $\sigma^2 = 1$, i.e $X \sim (0, 1)$

$\mu \quad \sigma^2$

then $\phi_X(t) = 1 + 0 - \frac{t^2}{2} + o(t^2)$

For a normal distribution $N(\mu, \sigma^2)$

$$\phi_X(t) = e^{it\mu - \frac{1}{2}\sigma^2 t^2}$$

and for $N(0,1)$; $\boxed{\phi_X(t) = e^{-\frac{t^2}{2}}}$

For $N(\mu, \sigma^2)$

$$f_X = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

For $N(0,1)$

$$f_X = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Central limit theorem

$$X_i \overset{iid}{\sim} (\mu, \sigma^2)$$

Sample mean $\quad \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

$\therefore$ Expected value of the sample mean is

$$E[\overline{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left(\sum_{i=1}^{n} E[X_i]\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} \mu\right) = \frac{1}{n} n\mu = \mu$$

Central limit theorem

Variance of the sample mean $\mathrm{Var}(\bar{X}_n)$

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

$$\Rightarrow \mathrm{Var}(\bar{X}_n) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Mean of sample mean $E(\bar{X}_n) = \mu$

and Variance of sample mean $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$

Central limit theorem

Now we define $Z_n \equiv \dfrac{n\bar{X}_n - n\mu}{\sigma\sqrt{n}}$

$$\Rightarrow Z_n = \frac{n \frac{1}{n}\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

$$= \frac{\sum_{i=1}^{n} (X_i - \mu)}{\sigma\sqrt{n}}$$

If now we define

$$Y_i = \frac{X_i - \mu}{\sigma} \quad \text{then}$$

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

Central limit theorem

$\therefore \; Y_i = \frac{X_i - \mu}{\sigma}$; $E(Y_i) = 0$ and

$Var(Y_i) = Var\left(\frac{X_i}{\sigma}\right) = \frac{1}{\sigma^2}\sigma^2 = 1$

$\therefore \; \phi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$

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

$\therefore \; \phi_{Z_n}(t) = E\left[e^{it\left(\frac{Y_1 + Y_2 + \cdots Y_n}{\sqrt{n}}\right)}\right]$

$= E\left[\prod_{k=1}^{n} e^{i\frac{t Y_k}{\sqrt{n}}}\right] = \prod_{k=1}^{n} \underbrace{E\left[e^{i\frac{t Y_k}{\sqrt{n}}}\right]}_{\phi_Y(t/\sqrt{n})}$

$= \left[\phi_Y(t/\sqrt{n})\right]^n$

Central limit theorem

$X_i \overset{iid}{\sim} (\mu, \sigma^2)$; $Y_i = \dfrac{X_i - \mu}{\sigma}$

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

$$\therefore \phi_{Z_n}(t) = \left[\phi_Y(t/\sqrt{n})\right]^n = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n$$

As we increase the sample size n, we get the limit

$$\lim_{n\to\infty} \phi_{Z_n}(t) = \lim_{n\to\infty}\left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n$$

$$= e^{-t^2/2}$$

} This is same as the characteristic function for $N(0,1)$

Hence, $\lim_{n\to\infty} Z_n = N(0,1)$

Central limit theorem

$X_i \overset{iid}{\sim} (\mu, \sigma^2)$; $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

$x_i = \frac{X_i - \mu}{\sigma}$

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

$$\lim_{n \to \infty} Z_n \sim N(0, 1) \Rightarrow \lim_{n \to \infty} \sqrt{n} \frac{(\bar{X}_n - \mu)}{\sigma} \sim N(0, 1)$$

$$\Rightarrow \lim_{n \to \infty} \sqrt{n} (\bar{X}_n - \mu) \sim N(0, \sigma^2)$$

$$\Rightarrow \lim_{n \to \infty} (\bar{X}_n - \mu) \sim N\left(0, \frac{\sigma^2}{n}\right)$$

$$\Rightarrow \lim_{n \to \infty} \bar{X}_n \sim \mu + N\left(0, \frac{\sigma^2}{n}\right) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Central limit theorem

For $X_i \overset{iid}{\sim} (\mu, \sigma^2)$ if we define sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

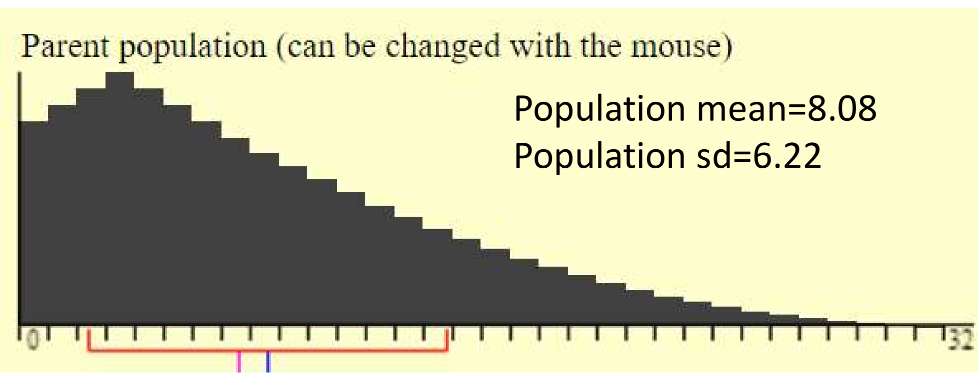
then mean of sample mean $(\bar{X}_n)$ is $E[\bar{X}_n] = \mu$

and variance of sample mean $(\bar{X}_n)$ is $Var(\bar{X}_n) = \frac{\sigma^2}{n}$

Now by Central limit theorem, we get that

$$\lim_{n \to \infty} \bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

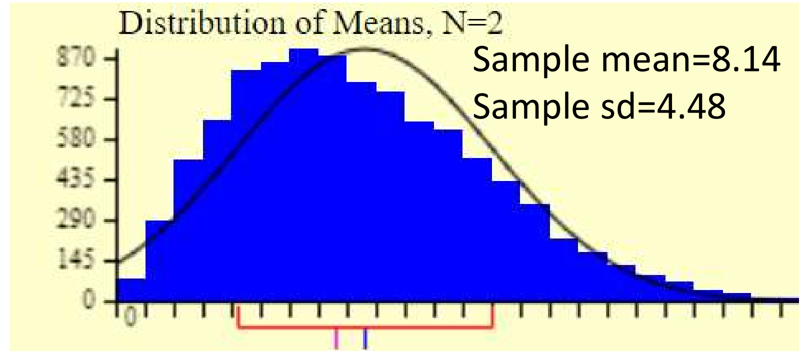
Central limit theorem

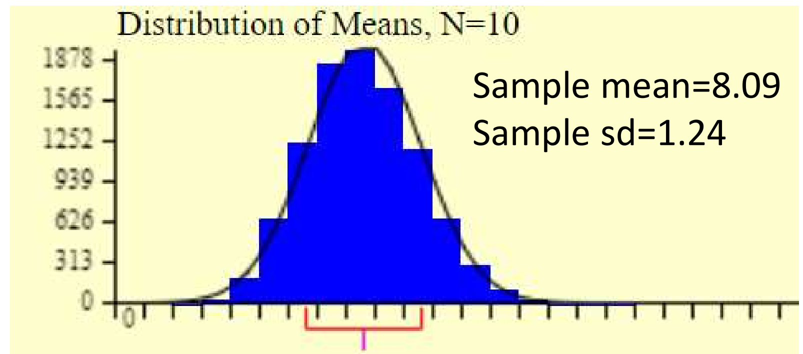


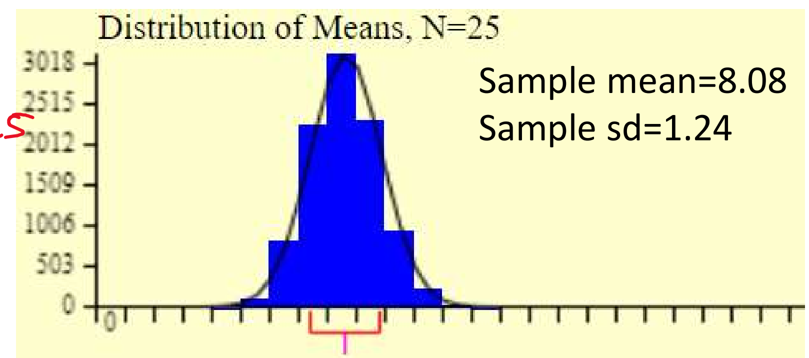
Sample size = 2

Sample size = 10

Sample size = 25



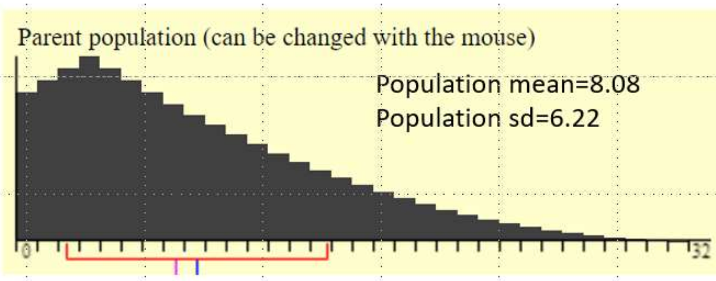




As $n \uparrow$ the distribution of sample mean approaches a normal distribution.

Ref: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

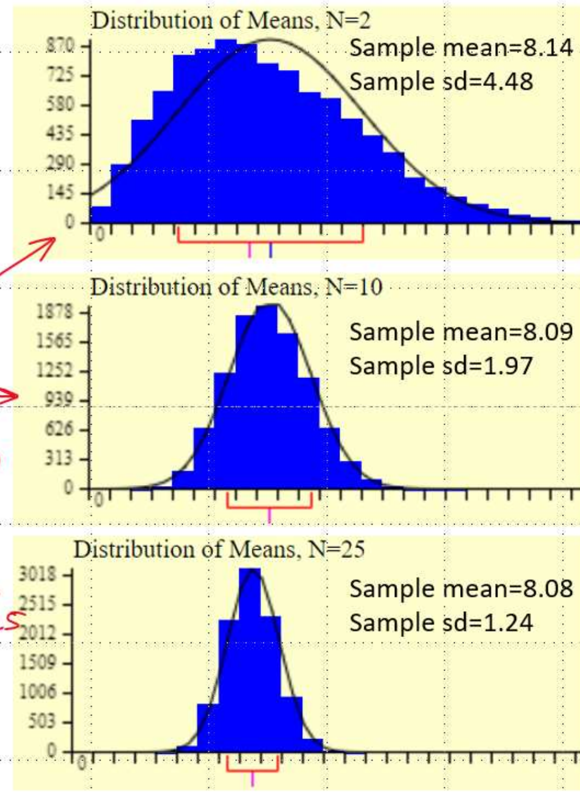
Central limit theorem



Sample size = 2

Sample size = 10

Sample size = 25



$E(\bar{X}_2) = 8.14 \simeq \mu = 8.08$

$\sqrt{Var(\bar{X}_2)} = 4.48 \simeq \frac{\sigma}{\sqrt{n}} = 4.40$

$E(\bar{X}_{10}) = 8.09 \simeq \mu = 8.08$

$\sqrt{Var(\bar{X}_{10})} = 1.97 \simeq \frac{\sigma}{\sqrt{n}} = 1.97$

$E(\bar{X}_{25}) = 8.08 = \mu = 8.08$

$\sqrt{Var(\bar{X}_{25})} = 1.24 = \frac{\sigma}{\sqrt{n}} = 1.24$

As $n \uparrow$ the distribution of sample mean approaches a normal distribution.

Application of Central limit theorem

Population $\xrightarrow{n}$ Sample

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$X_i \overset{iid}{\sim} (\mu, \sigma^2)$$

Find population mean from a sample of size n.

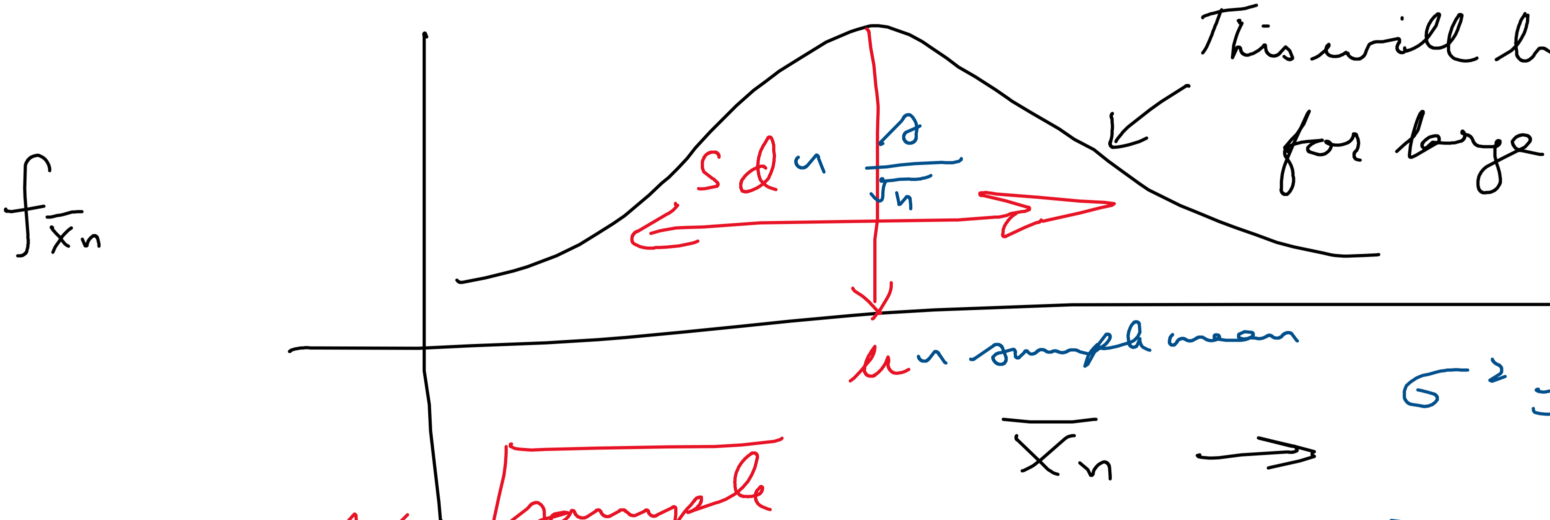
$$\text{Population mean } \mu = E(\text{sample mean}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

$$\text{Population variance } \sigma^2 = E(\text{sample variance}) = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(\overline{X}_n - X_i)^2\right)$$

$$\text{CLT } \lim_{n\to\infty} \overline{X}_n = N\left(\mu, \frac{\sigma^2}{n}\right)$$

Application of Central limit theorem

$X_i \overset{iid}{\sim} (\mu, \sigma^2)$ | $\bar{X}_n$



This will be $N(\mu, \frac{\sigma^2}{n})$ for large n (CLT)

$\sigma^2 = E(s^2)$

$s^2 = \frac{1}{n-1} \Sigma(\bar{X}_n - x_i)^2$

$\sigma/\sqrt{n} \sim \frac{s}{\sqrt{n}} \leftarrow \sqrt{\text{sample variance}}$

Application of Central limit theorem

$X_i \overset{iid}{\sim} (\mu, \sigma^2)$ | $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$

CLT $\lim_{n \to \infty} \bar{X}_n = N(\mu, \frac{\sigma^2}{n})$

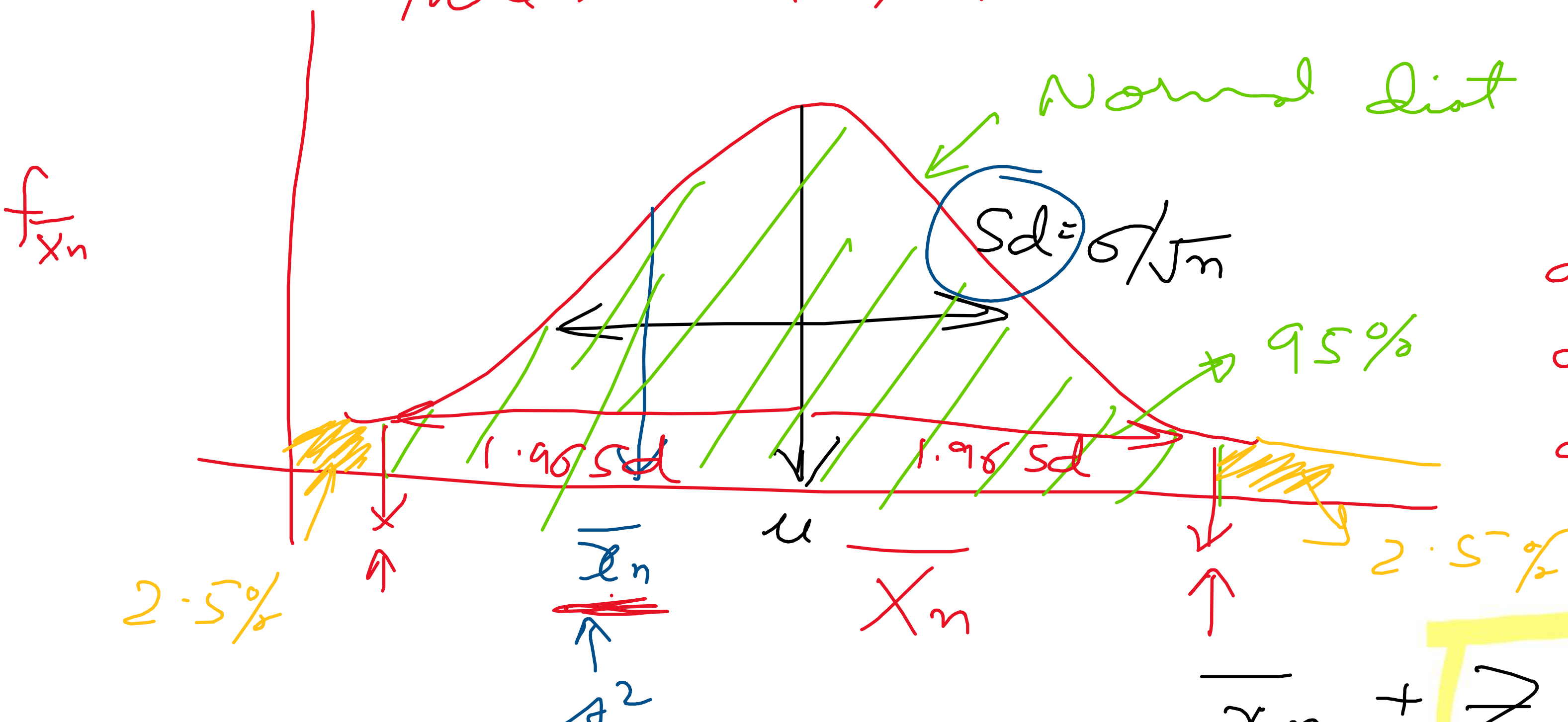
$\mu = \bar{x}_n \pm z^x \frac{s}{\sqrt{n}}$

C	z^x
95	1.96
98	2.326
99	2.576

$\sim N(\bar{x}_n, \frac{s^2}{n})$

Application of Central limit theorem

There is a 95% prob. that μ is within ± 1.96 sd of $\overline{x}_n$.



C	Z^*
95%	1.96
98%	2.326
99%	2.576

$$\mu = \overline{x}_n \pm \boxed{Z^*}\, sd \Big\}\ \text{with } C\% \text{ prob.}$$

$$Sd = \frac{\sigma}{\sqrt{n}} \sim \frac{s}{\sqrt{n}}$$

Application of Central limit theorem

$X_i \overset{iid}{\sim} (\mu, \sigma^2)$ If we take a sample of size n then $\bar{X}_n = \sum_{i=1}^{n} \frac{x_i}{n}$

Then for large n, CLT $\Rightarrow$ $\bar{X}_n \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$

C	z^*
99%	2.576
98%	2.326
95%	1.96

Therefore we can say with C confidence

that $\mu = \bar{X}_n \pm z^*(c)\, \underbrace{(Sd)}_{= \frac{\sigma}{\sqrt{n}} = \frac{E(S)}{\sqrt{n}}}$

Sample variance

$$\therefore \boxed{\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \pm z^*(c)\, \frac{\frac{1}{n-1}\sum(\bar{x}_n - x_i)^2}{\sqrt{n}}}$$

t-distribution

If $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

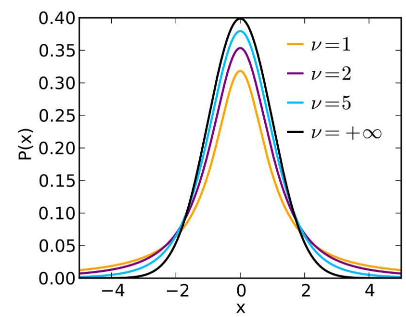
$$\bar{Z}_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

CLT as $n \to \infty$ $Z_n \to N(0,1)$

$t = \dfrac{\bar{X}_n - \mu}{S/\sqrt{n}}$, then the r.v. 't' follows a distribution

known as t- distribution with

$\nu = n-1$ degrees of freedom



pdf of t-distribution $= \dfrac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\dfrac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

Confidence Interval

CLT based : If $X_i \overset{iid}{\sim} (\mu, \sigma^2)$

Sample variance S is an unbiased estimate of σ

then $\mu = \overline{X} \pm z^*(C) \frac{\sigma}{\sqrt{n}}$

C	z^*
99%	2.676
98%	2.326
95%	1.96

t-distribution: If $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$

then $\mu = \overline{X} \pm t^*(C) \frac{S}{\sqrt{n}}$

For 95% C

n	n-1	t^*
6	5	2.571
11	10	2.228
31	30	2.042
∞	∞	1.960

If n >30 use CLT, else use t-distribution