Linear Regression-1

Prof. Asim Tewari
IIT Bombay

Characteristic Function

The characteristic function of a random variable is

$$\phi_X(t) \equiv E(e^{itX}) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx$$

$$e^{itX} = \frac{1}{\angle 0} + \frac{itX}{\angle 1} + \frac{(it)^2 X^2}{\angle 2} + \frac{(it)^3 X^3}{\angle 3} + \cdots$$

$$\therefore \quad \phi_X(t) = E\left[\frac{1}{\angle 0} + \frac{itX}{\angle 1} + \frac{(it)^2 X^2}{\angle 2} + \frac{(it)^3 X^3}{\angle 3} + \cdots\right]$$

$$\Rightarrow \quad \phi_X(t) = 1 + it\,E(X) + \frac{(it)^2}{\angle 2} E(X^2) + \frac{(it)^3}{\angle 3} E(X^3) + \cdots$$

Characteristic Function

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

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

i.e. $m_n = E(X^n)$

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

Moment generating Function

The moment generation function of a r.v. is

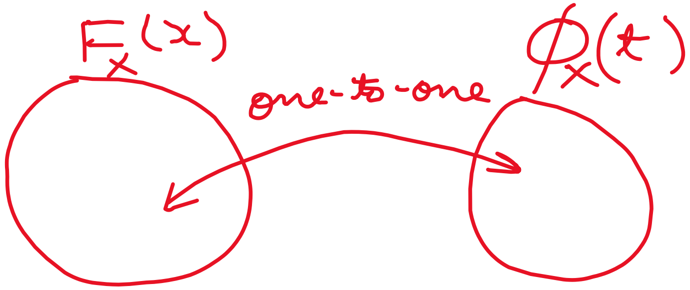
$$M_X(t) = \phi_X(-it) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

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

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

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

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

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

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

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

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

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 $N(0, 1)$

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

Central limit theorem

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

Sample mean $\bar{X}_n \equiv \frac{1}{n} \sum X_i$

$\therefore$ Expected value of sample mean is

$$E(\bar{X}_n) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

Central limit theorem

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

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

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

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

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

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

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

Central limit theorem

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

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

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

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

If now we define

$Y_i = \dfrac{X_i-\mu}{\sigma}$, then

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

Central limit theorem

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

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

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

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

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

$$= E\left(\prod_{k=1}^{n} e^{i\frac{t}{\sqrt{n}} Y_k}\right)$$

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

Central limit theorem

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

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

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

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

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

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

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

Central limit theorem

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

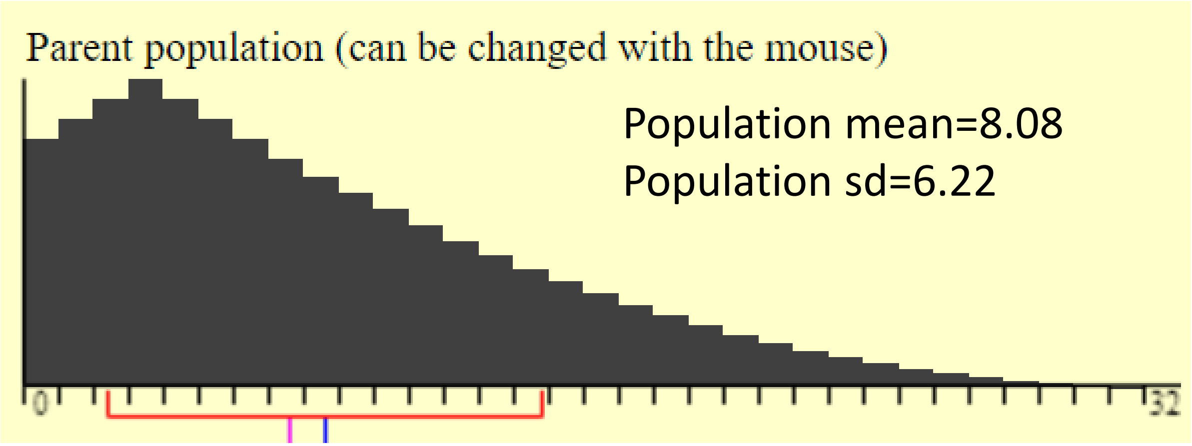
Then mean of Sample mean $E(\overline{X}_n) = \mu$

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

Now by central limit theorem we get that

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

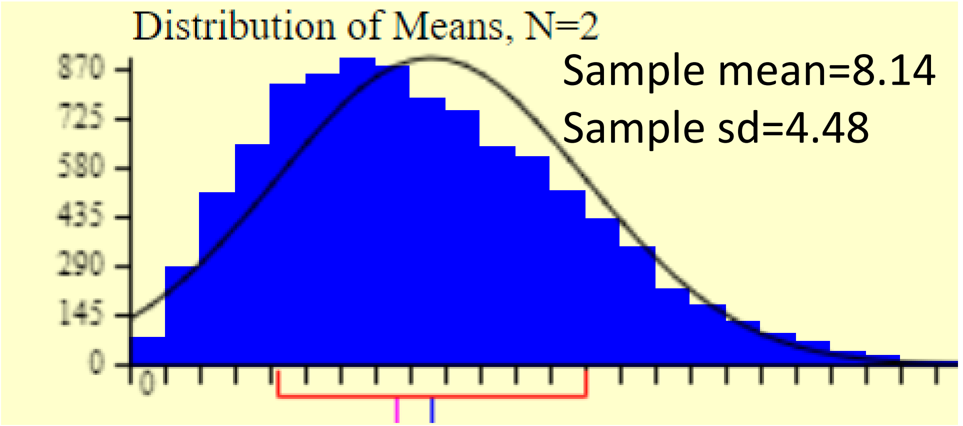
Central limit theorem

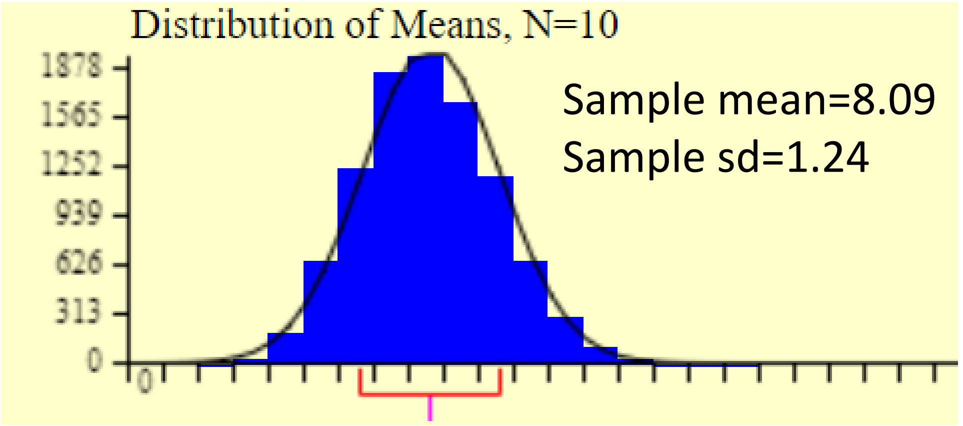


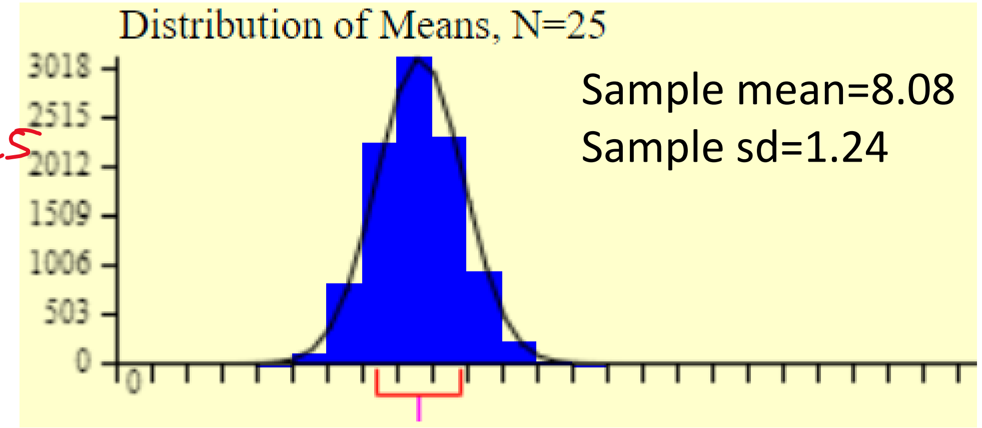
Sample size = 2

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

Application of Central limit theorem

From a population of $X_i \overset{iid}{\sim} (\mu, \sigma^2)$ we draw a sample of size n

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

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

Now by central limit theorem

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

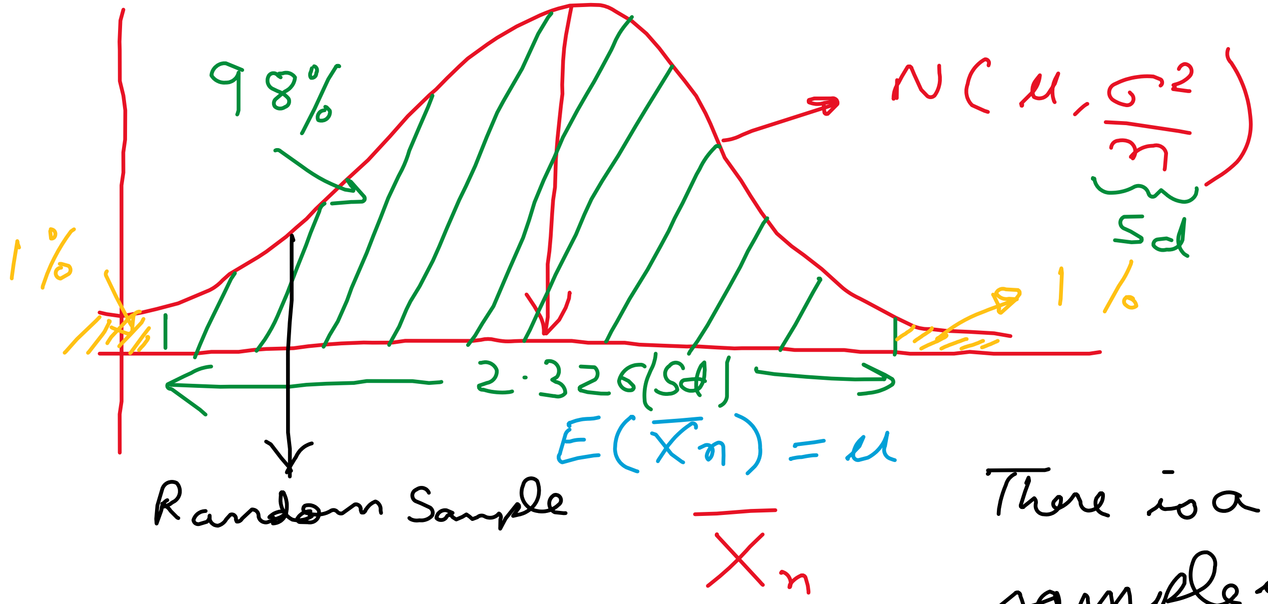
An unbiased estimate of σ^2 is s^2 i.e. sample variance

An unbiased estimate of this is $\bar{X}_n$ i.e. sample mean

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



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

There is a 98% probability that a random sample is with in $\pm$ 2.326 (sd) of μ

$\boxed{\therefore \text{ we can say with 98\% confidence that } \mu = \bar{X}_n \pm 2.326 \text{ sd}}$

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

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

Therefore we can say with C confidence

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

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

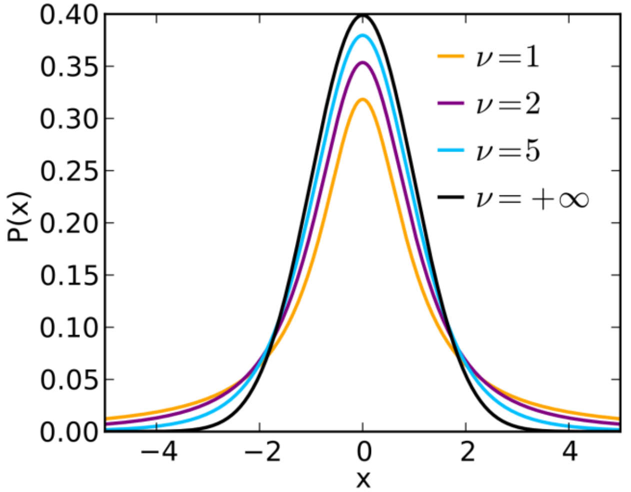
If $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ then we can define

$Z_n \equiv \dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ and by CLT as $n \to \infty$ $Z_n \to N(0,1)$

Similarly, it can be shown that if $t \equiv \dfrac{\bar{X}_n - \mu}{S/\sqrt{n}}$

$\boxed{\text{P.S: As } n \to \infty \\ \text{t-dist} \to N(0,1)}$

then the r.v. t follows t-distribution with $\nu = n-1$ degrees of freedom



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

Confidence Interval

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

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

Sample variance S is an unbiased estimate of σ

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