Lecture 8:

The Central Limit Theorem (CLT for short)

A truely amazing mathematical fact.

CLT: If $X_1, X_2, \ldots, X_n$ are independent R.V.s, each with the same distribution, then $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ is approximatly a Normal random variable for large n (something like $n > 30$), No matter what PDF or PMF each X_i has.

Basically the average of a bunch of independent copies of R.V.s starts behaving like a Normal R.V.

Example:

Let $X_1, X_2, \ldots, X_{100}$ be independent R.V.s s.t. the PMF of X_i is

$$P_{X_i}(x) = \begin{cases} \binom{22}{x}\left(\frac{1}{3}\right)^x\left(\frac{2}{3}\right)^{22-x} & \text{for } x = 0, 1, \ldots, 22 \\ 0 & \text{otherwise} \end{cases}$$

$\uparrow$

Kind nasty PMF, i.e. I wouldn't want to calculate $P(X > 7)$ by hand.

Yet the CLT says $\bar{X}$ is approx Normal so I can use the Gaussian PDF to approximate $P(\bar{X} > 7)$ easily.

First Note

$$E(X_i) = \frac{22}{3}$$

$$\text{var}(X_i) = 22\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)$$

You don't need to rember these formulas from stat 13.

$$\therefore E(\bar{X}) = E\left(\frac{1}{100}(X_1 + \cdots + X_{100})\right)$$

$$\overset{MF1}{=} \frac{1}{100} E(X_1 + \cdots + X_{100})$$

$$\overset{MF1}{=} \frac{1}{100}\left(\underset{\frac{22}{3}}{\underset{\uparrow}{E(X_1)}} + \cdots + \underset{\frac{22}{3}}{\underset{\uparrow}{X_{100}}}\right)$$

$$= \frac{1}{\cancel{100}}\left(\cancel{100} \cdot \frac{22}{3}\right)$$

$$= \frac{22}{3}$$

$$\text{var}(\bar{X}) = \frac{1}{100^2}\left(100 \cdot 22\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)$$

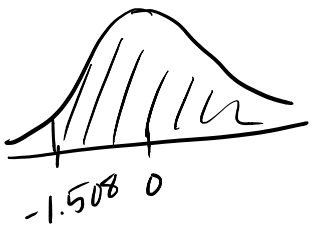
$$= \frac{22}{100}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)$$

$$\therefore \bar{X} \underset{\text{by the CRT}}{\underset{\uparrow}{\approx}} N\left(\frac{22}{3}, \frac{22}{100}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\right)$$

$\therefore\ P(\bar{X} > 7)$

$$= P\Bigg(\underbrace{\frac{\bar{X} - 22/3}{\sqrt{\left(\frac{22}{100}\right)\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}}}_{\approx Z \sim N(0,1)} > \underbrace{\frac{7 - \frac{22}{3}}{\sqrt{\frac{22}{100}\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}}}_{-1.508} \Bigg)$$

$$\approx P(Z > -1.508)$$

$=$ 

$= 1 - (\text{table @ } 1.508)$

$= 0.9342$

$\therefore\ P(\bar{X} > 7) \approx 0.9342$

The exact value is

$$P(\bar{X} > 7) = 0.9316$$

$X_1 + \cdots + X_{100} \sim Bin\left(100 \cdot 22, \frac{1}{3}\right)$

$\therefore\ P(\bar{X} > 7) = P(X_1 + \cdots + X_{100} > 700)$

In the previous problem we found $E(\bar{X})$ & $var(\bar{X})$ by hand.

If you don't want to do this every time you can use the following more detailed version of the CLT

CLT: Let $X_1, \dots, X_n$ be independent R.V.s, all with the same PMF or PDF. Suppose each X_i satisfies:

$$E(X_i) = \mu$$
$$var(X_i) = \sigma^2$$

Then for large n

$$\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

Note: MF1 & MF2 already gives you $E(\bar{X}) = \mu$ & $var(\bar{X}) = \frac{\sigma^2}{n}$.

why: From before

$$E(X_1 + \cdots + X_n) = n\mu$$
$$var(X_1 + \cdots + X_n) = n\sigma^2$$

$$\therefore\ E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \overset{MF1}{=} \frac{1}{n} E(X_1 + \cdots + X_n)$$
$$= \frac{1}{\cancel{n}} \cancel{n}\mu = \mu$$

$$var(\bar{X}) = var\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \overset{MF2}{=} \left(\frac{1}{n}\right)^2 var(X_1 + \cdots + X_n) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Example:

Suppose the ave height of Cali residents is 170.18 cm with a s.d. of 6.35.
Let $X_1, \dots, X_{57}$ be the height (in cm) of 57 randomly selected (w/Rep) Cali Residents.

Approximate the prob that the combined height is less than 9600 cm.

combined height $= Y = X_1 + \cdots + X_{57}$

$\frac{1}{57} Y = \bar{X}$ is approx $N\left(170.18, \frac{6.35^2}{57}\right)$

want $\Pr(Y < 9600 \text{ cm})$

$= \Pr\left(\frac{1}{57} Y < \frac{9600}{57}\right)$

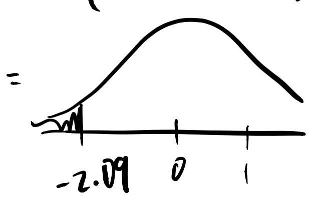
$= \bar{X}$ is approx $N\left(170.18, \frac{6.35^2}{57}\right)$

$= \Pr\left(\bar{X} < \frac{9600}{57}\right)$

$\approx \Pr\left(Z < \frac{\frac{9600}{57} - 170.18}{\left(6.35/\sqrt{57}\right)}\right)$ $= -2.09$

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

$= \Pr(Z < -2.09)$

$=$ 

$=$ (table @ $a = 2.09$)

$= 0.0183$

Lets do a hardish problem.

e.g. $X_1, \dots, X_{92}$ are iid

s.t.

$E(X_i) = -\frac{1}{2}$

$\text{var}(X_i) = 10.$

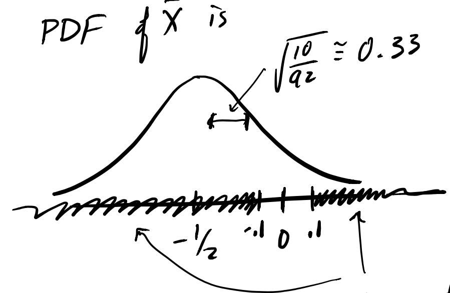
means they are independent and have the same PMF or PDF

Approximate $P(|\bar{X}| > 0.1)$.

Answer: By the CLT

$\bar{X} \approx N\left(-\frac{1}{2}, \frac{10}{92}\right)$

$\therefore$ PDF of $\bar{X}$ is



this is the event that $|\bar{X}| > 0.1$

$\therefore P(|\bar{X}| > 0.1)$

$= P(\bar{X} < -0.1) + P(\bar{X} > 0.1)$

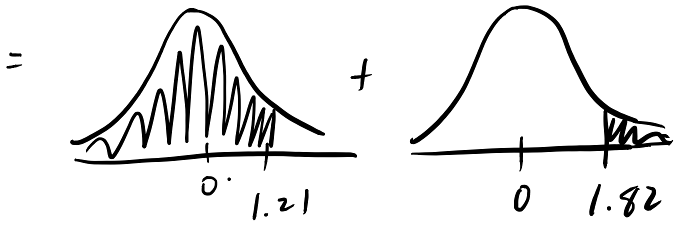
$= P\left(\frac{\bar{X} - (-\frac{1}{2})}{\sqrt{\frac{10}{92}}} < \frac{-0.1 - (-\frac{1}{2})}{\sqrt{\frac{10}{92}}}\right)$

$+ P\left(\frac{\bar{X} - (-\frac{1}{2})}{\sqrt{\frac{10}{92}}} > \frac{0.1 - (-\frac{1}{2})}{\sqrt{\frac{10}{92}}}\right)$

by the CLT

$\cong P(Z < 1.21)$

$+ P(Z > 1.82)$



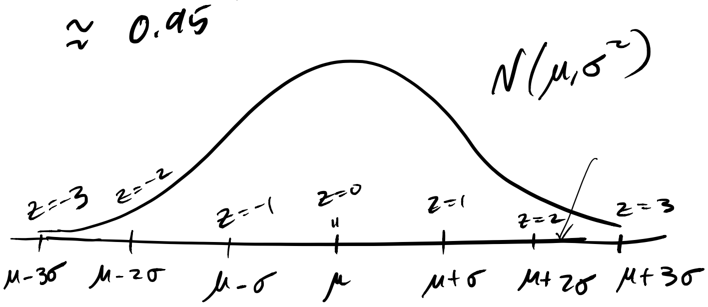
$= 1 - (\text{table} @ 1.21) + (\text{table} @ 1.82)$

$= 0.921.$

Empirical Rule

The CLT explains why so many R.V.s observed in nature look approx Normal.

Also explains The empirical Rule:
"The chance a bell shaped R.V. X falls within $E(X) \pm 2 \cdot sd(X)$ is ≈ 0.95"



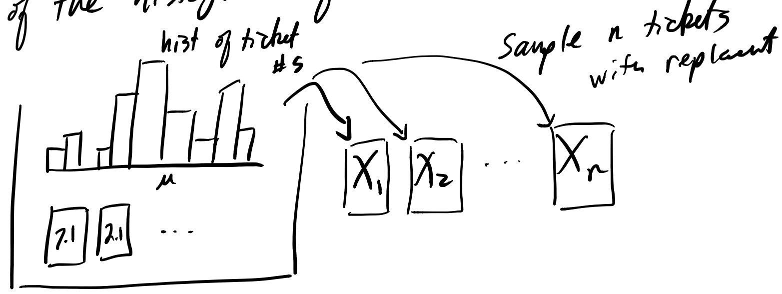
Prob X lands here ≈ 0.68

Prob X lands here ≈ 0.95

Prob X lands here ≈ 0.997

Box model picture of the CLT

In a box model $E(X)$ & $sd(X)$ just represent the ave & standard deviation of the histogram of ticket numbers

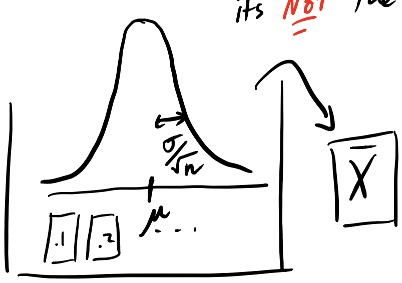


Let μ = Ave ticket #
σ = s.d of ticket #s

$\therefore E(X_i) = \mu$, $var(X_i) = \sigma^2$ & X_i's are indep R.V.s.

What about $\bar{X}$?

It is a R.V. $\underbrace{\text{so what Box models } \bar{X}?}_{\text{its Not the box that gives } X_i}$



So $\bar{X}$ behaves like a single draw from a different box ... one where the histogram of ticket numbers is a perfect Normal bell shape.

e.g. each student grabs 100 tickets from a box full of tickets Records the ave & puts them back. what is the histogram of the students Ave look like.

Z-scores of $\bar{X}$

Suppose $X_1, \dots, X_{250}$ are draws with Replacent from a box of numbers s.t.

$$E(X_i) = 5$$

$$sd(X_i) = 15.$$

Would you think it is unusual if I told you $\bar{X}$ was observed to be 0.9?

Since $E(\bar{X}) = 5 \qquad (\mu)$

$$sd(\bar{X}) = \frac{15}{\sqrt{250}} \qquad \left(\sqrt{\frac{\sigma^2}{n}}\right)$$

The Z-score for $\bar{X}$ is

$$Z = \frac{\bar{X} - 5}{\left(\frac{15}{\sqrt{250}}\right)}$$

The value $\bar{X} = 0.9$ corresponds to

$$Z = \underbrace{\frac{0.9 - 5}{\left(\frac{15}{\sqrt{250}}\right)} = -4.321}_{\text{extremely unlikely Z-score.}}$$

Lets finish with one more CLT example:

Suppose I take a handful of 100 6-sided, roll them all & Let Y be the ave roll.

Approximate $P(Y > 3)$.

Since $Y = \frac{X_1 + \dots + X_{100}}{100}$ where X_i represents the amount showing on the i^{th} die (all indep) the CLT says

$$Y \approx N\left(\mu, \frac{\sigma^2}{100}\right)$$

where μ, σ^2 are the expected value and var of an individual Roll

one die

x	$P_X(x)$	$xP_X(x)$	$x^2P_X(x)$
1	1/6	1/6	1/6
2	⋮	2/6	$2^2/6$
⋮	⋮	⋮	⋮
6	1/6	⋮	$6^2/6$
		3.5	$15.16\bar{6}$

$\therefore \mu = 3.5$

$$\sigma^2 = 15.16\bar{6} - (3.5)^2$$

$$= 2.91\bar{6}\bar{6}$$

$\therefore Y \approx N(3.5, 0.0291\bar{6})$

$$\therefore P(Y > 3) \approx P\Big(\underset{N(0,1)}{Z} > \underbrace{\frac{3 - 3.5}{\sqrt{0.0291\bar{6}}}}_{-2.92}\Big)$$

$$= 0.9982$$