Introduction to Probability & Statistics
Prof Abhay Gopal Bhatt
Department of Statistics
Indian Statistical Institute Delhi
Week - 9
Lecture - 33
Central Limit Theorem, Distribution of a Linear Combination

sampling distribution of x bar dekh rahe the jab original population normal distributed ho mean μ aur variance σ^2 ke saath, tab x bar bhi normal hota hai mean μ ke saath aur variance chhota ho jata hai σ^2/n; example me jab n=5 aur n=6 tha, tab x bar aur y bar ke variance alag nikle, kyunki sample size alag tha; jaise-jaise n badhta hai, σ^2/n, distribution squeeze mean ke around plots me dikh raha tha ki original normal curve blue hai, phir green wala (n=3) narrower, phir red wala (n=10) aur bhi narrow ho gaya; ye sab isliye tha kyunki population normal assume kiya tha; ab aata hai sabse important aur surprising theorem Central Limit Theorem (CLT): agar x1, x2, ..., xn random sample ho kisi population se jiska mean μ aur variance σ^2 ho (population normal hona aur sample size n sufficiently large ho, tab x bar approximately normal distributed hota hai mean μ aur variance σ^2/n ke saath, aur sample total T approximately normal hota hai mean $n \cdot \mu$ aur variance $n \cdot \sigma^2$ ke saath; yahan population distribution kuch bhi ho Bernoulli, uniform, skewed, bimodal phir bhi sample mean normal ke isi ko humne Bernoulli(0.3) ke plots se n=1 par bilkul non-normal, n=2 par discrete weird shape, n=3 par thoda symmetric, n=5 par aur shape dikhna n=10 par bell curve n=25 par to histogram almost normal curve jaisa aur jaise-jaise n sample mean ka distribution normal shape adopt karta CLT ka magic bina population ko normal bhi sample mean almost normal ho large n kyonki normal distribution to sabhi $-\infty$ se $+\infty$ tak values leta hai, lekin yahan humne jo shuruaat ki thi wo Bernoulli random variable se jisme sirf 0 aur 1 values hoti hain aur unki probabilities thi 0.7 aur 0.3 phir bhi $\bar{X}$ ka distribution dheere-dheere normal curve ki taraf badhne laga; ye jo humne Bernoulli ke liye dekha tha, ye general me bhi sahi hai aur wahi baat Central Limit Theorem (CLT) batati hai; CLT ke teen key conditions hai: (1) mean aur variance population ke exist karne chahiye, (2) observations random sample hone chahiye , aur (3) sample size n sufficiently large hona chahiye “sufficiently large” ek vague term hai, lekin rule of thumb ke hisaab se CLT tab use kiya ja sakta hai jab $n > 30$ ho; jaise-jaise n bada hota hai, approximation behtar hota jata hai; ab example dekhiye: kisi chemical ke batch me impurity quantity random hai jiska population mean 4.0 grams aur std deviation 1.5 grams hai; agar 50 batches independently prepare kiye gaye, to $\bar{X}$ approx normal hota hai mean = 4 aur variance = 2.25/50 ke saath; fir probability compute karte hain ki sample average 3.5 aur 3.8 ke beech me aayega, jiske liye normal standardization karke Z me convert kiya aur tables se P = 0.1645 nikla; dusre example me daily parking challans ek highway stretch par Poisson(50) distributed the iska exact distribution pata hone ke bawajood T = sum Xi ka exact distribution hard hota hai par CLT use karke hum T ko approx normal treat kar sakte hain jiska mean 5000 aur variance 5000 hota hai; phir total challans 4900 aur 5200 ke beech hone ki probability standardization karke Z-values me convert kar ke nikaali jati hai; yani original distribution chahe Bernoulli ho, Poisson ho, ya koi unknown large n ke saath sample mean ya sample total ka

distribution normal ke bahut paas aa jata hai isi liye CLT statistics me itna powerful aur fundamental result hai. kyonki normal distribution to sabhi $-\infty$ se $+\infty$ tak values leta hai, lekin yahan humne jo shuruat ki thi wo Bernoulli random variable se jisme sirf 0 aur 1 values hote the aur unki probabilities 0.7 aur 0.3 thi phir bhi $\bar{X}$ ka distribution dheere-dheere normal curve ki tarah dikhne laga; ye jo humne Bernoulli ke liye dekha, ye general case me bhi sahi hota hai aur wahi baat Central Limit Theorem (CLT) batati hai; CLT ke 3 main conditions hai: (1) population ka mean aur variance exist kare, (2) sample IID (independent and identically distributed) ho, (3) sample size n sufficiently large ho "sufficiently large" term vague hai par rule-of-thumb ke hisaab se CLT tab reliable hota hai jab $n > 30$ ho; jaise-jaise n badhta hai, approximation aur accurate hota jata hai; ab example me dekhte hain: chemical batches me impurity random variable hai jiska population mean 4.0 grams aur std deviation 1.5 grams hai; agar 50 batches independently prepare hue, to $\bar{X}$ approx normal hota hai mean = 4 aur variance = 2.25/50 ke saath; isse probability nikali jaati hai ki sample average 3.5 aur 3.8 ke beech rahe — normal standardization karke Z me convert karke tables se value 0.1645 milti hai; dusre example me parking challans Poisson(50) distributed the isme exact T distribution complex hota, lekin CLT ke through T approx normal treat hota hai jiska mean 5000 aur variance 5000; phir T ke 4900 aur 5200 ke beech hone ki probability find karne ke liye standardization kiya, jahan right-side value $\Phi(2.83)=0.9977$ aur left-side value $\Phi(-1.41)=0.0793$, jiska final answer = 0.9977 – 0.0793 = 0.9184; yani CLT hume yeh powerful ability deta hai ki chahe original population Bernoulli ho, Poisson ho, ya unknown jab n bada hota hai to sample mean ya sample total ka distribution normal ke bahut paas aa jata hai, sahi mean aur sahi variance ke saath isi wajah se CLT statistics ka perhaps sabse important theorem maana jata hai. to shayad alag color se likh loon, x1, x2, x3 iska matlab hai: x1 hai number of cars jo randomly chosen day par road 1 se expressway me enter karti hain, x2 road 2 se aur x3 road 3 se; ye teeno random variables hain; agar hume expected value chahiye total number of cars ki, to expected value of (x1 + x2 + x3) = E[x1] + E[x2] + E[x3] = 800 + 1000 + 600 = 2400; ab variance dekhte hain agar assume kare ki x1, x2, x3 independent hain, tab Var(x1 + x2 + x3) = Var(x1) + Var(x2) + Var(x3) = $16^2 + 25^2 + 18^2$ = 1205, aur standard deviation = $\sqrt{1205} \approx 34.7131$; lekin agar ye independent nahi hain aur additional information di jaye ki Cov(x1, x2)=80, Cov(x1, x3)=90, Cov(x2, x3)=100, tab expected value same rahega 2400, par variance change ho jayega: Var(t)= Var(x1)+Var(x2)+Var(x3) + 2Cov(x1,x2)+2Cov(x1,x3)+2Cov(x2,x3) = $16^2 + 25^2 + 18^2$ + 2(80+90+100) = 1745; yaani expected value ke liye independence matter nahi karta, lekin variance ke liye independence important hai, kyunki covariance terms add hote hain; aur jaise shuruaat me kaha tha, linear combination ek general concept hai: agar humare paas random variables x1, x2, …, xn ho, to linear combination = a1x1 + a2x2 + … + anxn; sample mean $\bar{x}$ bhi ek linear combination hai jahan ai = 1/n; sample total T bhi ek linear combination hai jahan ai = 1; yani $\bar{x}$ aur T general linear combinations ke special cases hain, aur hamara theorem kisi bhi general linear combination ke liye apply hota hai — isi tarah se hum expected values aur variances compute kar sakte hain independent ho ya na ho, covariance values pata honi chahiye.