Introduction to Probability & Statistics

Prof Abhay Gopal Bhatt
Department of Statistics
Indian Statistical Institute Delhi

Week - 10
Lecture - 35

Unbiased Estimators

ye example humne pichhle class ke end me dekha tha jisme ek parameter μ tha “true average time spent everyday by students in physical exercise,” aur uske liye humne teen estimators liye the: sample mean ($\bar{x}$), sample median ($\tilde{x}$), aur $\hat{\mu}_3 = (\min x_i + \max x_i)/2$; ye teeno sensible estimators lagen, par kaunsa choose kare aur kyun? pehle note karte hain ki kisi bhi random sample ke liye agar population ka mean μ hai, to expected value of sample mean hamesha hota hai μ matlab $E[\bar{x}] = \mu$, isliye sample mean ek strongly preferable estimator hai; isi type ke estimator ko unbiased estimator kehte hain, matlab agar kisi parameter θ ka estimator $\hat{\theta}$ ho, aur expected value of $\hat{\theta} = \theta$ ho, to $\hat{\theta}$ unbiased estimator hai; agar expected $\hat{\theta} - \theta \neq 0$ ho to usko bias kehte hain agar ye difference negative ho to negative bias, positive ho to positive bias; extreme ideal case me agar $\hat{\theta}$ always $= \theta$ ho to wo perfect estimator hota, par real life me $\hat{\theta}$ ek random variable hai jo sample par depend karta hai kisi sample ke liye estimate $\hat{\theta}$ θ se zyada ho sakta hai, kisi aur sample ke liye kam; important ye hai ki on average, expected value equal to θ ho; isi ko hum unbiasedness ke roop me prefer karte hain; ek known proposition humne pehle bhi dekha sample mean $\bar{x}$ is an unbiased estimator of population mean μ; ye isliye important hai kyunki sample mean kabhi μ se upar kabhi μ se neeche ho sakta hai, par long-run average me $\bar{x}$ population mean ke paas hi rahta hai isi liye $\bar{x}$ ek desirable unbiased estimator mana jata hai. parameter θ kabhi usse kam hota hai, kabhi zyada, lekin average me agar dekha jaye to wo theek θ se match karta hai isliye unbiased estimator prefer kiya jata hai; ab hum is concept ko examples ke context me dekhte hain: Bernoulli distribution me agar $x_1, x_2, \ldots, x_n$ iid Bernoulli-p ho, to $T = \sum x_i$ hota hai total number of heads in n tosses, jiska distribution hai binomial (n, p), aur $E[T] = np$, isliye $E[T/n] = p$, matlab sample proportion (i.e., $\bar{x}$) unbiased estimator hai p ka; doosra example: uniform distribution on $(0, \theta)$ me sample $x_1,\ldots,x_n$ iid uniform$(0,\theta)$; yahan maximum observation $x_{(n)} = \max\{x_1,\ldots,x_n\}$ ek sensible estimator lagta hai θ ka, kyunki sab $x \leq \theta$ hote hain; lekin mathematically expected value of $x_{(n)} = n/(n+1) \cdot \theta$ jo hamesha θ se chhota hota hai, isliye $x_{(n)}$ negatively biased estimator hai; bias $= E[x_{(n)}] - \theta = (n/(n+1)\theta - \theta) = -(\theta / (n+1)) < 0$; isko correct karne ke liye hum unbiased estimator define kar sakte hain: $\hat{\theta}_1 = ((n+1)/n)\, x_{(n)}$, tab $E[\hat{\theta}_1] = \theta$ bilkul exact; matlab humne $x_{(n)}$ ka bias quantify kiya aur appropriate constant se multiply karke ek unbiased estimator banaya; ye approach general principle ko highlight karta hai: sensible statistic choose karo, uska expectation analyze karo, aur agar usme bias ho to mathematically adjust karke unbiased version construct karo ye hi reason hai ki statistical estimation me “unbiasedness” ek important desirable property mani jati hai.aur hamne unbiased estimator dhoond liya θ ke liye $n+1/n \cdot x_{(n)}$ ek unbiased estimator hai θ ka to jab bhi humein unbiased estimator mil jata hai, hum usko prefer karte hain; isliye main formally principle likh raha hoon: “jo estimator unbiased ho, wo preferable hota hai”; agar ek hi parameter ke liye multiple estimators available ho, to hum unme se woh choose karenge jo unbiased ho yani unbiasedness estimator ki achchi property hai; humein pata hai ki agar parameter population mean μ ho, to sample mean $\bar{x}$ ek unbiased estimator hai, kyunki $E[\bar{x}] = \mu$; ab variance ke liye: x1, x2, …, xn ek random sample hai population se jiska mean μ aur variance σ^2 hai; $\bar{x}$ to μ ka unbiased estimator ho hi gaya, lekin ab humein estimator chahiye σ^2 ke liye; sample variance

$$S^2 = (\frac{1}{n-1}) \sum_{i=1}^{n} (X_i - \bar{x}^2)$$

$$= \left(\frac{1}{n-1}\right) [\sum_{i=1}^{n} X_i^{\,2} - \; n\bar{X}^2]$$

$$E[S^2] = \left(\frac{1}{n-1}\right) [\sum_{i=1}^{n} X_i^{\,2} - \; n\bar{X}^2]$$

$$= \left(\frac{1}{n-1}\right) [n(\sigma^2 + \mu^2) - \; n(\frac{\sigma^2}{n} + \mu^2)$$

$$= \sigma^2$$

ye ek statistic hai, aur agar iski expectation nikaalein to wo exactly σ^2 aati hai — matlab S^2 bhi unbiased estimator hai σ^2 ka; lekin agar hum denominator ko n rakhen (i.e., $S'^2 = 1/n \sum(x_i - \bar{x})^2$), to ye biased estimator hota hai iska expectation $(n-1)/n \cdot \sigma^2 < \sigma^2$ hota hai yani negative bias; ab uniform(0,θ) wale example par lautkar: humne maximum statistic x(n) se θ ke liye estimator banaya $\hat{\theta} = x(n)$, jo biased nikal gaya; isliye humne x(n) ka expected value $n/(n+1)\theta$ use karke corrected estimator $n+1/n\ x(n)$ banaya, jo unbiased ho gaya; doosra unbiased estimator bhi mila: $\hat{\theta} = 2\bar{x}$ kyunki uniform distribution me population mean $\theta/2$ hota hai, to $\bar{x}$ ka expectation $\theta/2$ aur isliye $E[2\bar{x}] = \theta$; ab ek hi parameter θ ke liye do unbiased estimators mil gaye to kaunsa best? dono ka mean to θ hi hai, lekin hum inka variance compare karte hain jo estimator ka variance kam ho, uski values θ ke aas-paas tightly clustered hoti hain, yani wo estimator zyada reliable hota hai; isliye jab multiple unbiased estimators ho, tab selection ke liye criterion hota hai minimum variance, jo humein batata hai kaunsa estimator practical estimation me zyada accurate hoga. to pehle ham variance of $\hat{\theta}_2$ dekhte hain kyunki $\hat{\theta}_2 = 2\cdot\bar{x}$ aur $\bar{x}$ ka variance humein pata hai, so $\mathrm{Var}(\hat{\theta}_2) = \mathrm{Var}(2\cdot\bar{x}) = 2^2\cdot\mathrm{Var}(\bar{x}) = 4\cdot(\sigma^2/n)$; ab uniform(0,θ) distribution ke liye population variance hota hai $(\theta-0)^2/12 = \theta^2/12$, to $\mathrm{Var}(\hat{\theta}_2) = 4\cdot(\theta^2/12)/n = \theta^2/(3n)$; ab $\mathrm{Var}(\hat{\theta}_1)$ nikaalne ke liye jahan $\hat{\theta}_1 = (n+1)/n \cdot x(n)$ hume x(n) ke distribution ka use karna padta hai, jo is course me detail me nahi kar rahe par result jo milta hai wo yeh: $Var(\hat{\theta}_1) = \theta^2 / [n(n+2)]$; ab compare karte hain dono variances: hum check karte hain $\theta^2/[n(n+2)] < \theta^2/(3n)$, θ^2 cancel ho jata hai, to condition banti hai: $1/[n(n+2)] < 1/(3n) \Rightarrow n(n+2) > 3n \Rightarrow n > 1$, jo hamesha true hai kyunki sample size $n > 1$ hi hota hai; therefore $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$ for all $n>1$, isi wajah se $\hat{\theta}_1$ is preferred over $\hat{\theta}_2$; ab doosra principle milta hai: among all unbiased estimators of θ, choose the one which has minimum variance; is type ke estimator ko kehte hain minimum variance unbiased estimator, ya short form me MVUE; ek remark hamesha MVUE exist ho, yeh nahi hai, kabhi–kabhi kisi parameter ke liye MVUE hota hi nahi; lekin ek important theorem hai: agar $x_1, x_2, \ldots, x_n$ iid normal with mean μ and variance σ^2 ho, to sample mean $\bar{x}$ is not only unbiased for μ but is actually the MVUE for μ, yaani sabhi unbiased estimators me uska variance sabse kam hota hai isliye normal population ke case me μ ke estimation ke liye best estimator hota hai $\bar{x}$; isi powerful theorem ke saath hum is topic ko yahin stop karte hain thank you.