Binormalfordelingen

Definisjon

Noe av hensikten med å innføre begrepet betinget sannsynlighet er at kompliserte modeller ofte kan bygges ut fra enkle betingede modeller. Når man spesifiserer betingelser (hendelser som er gitt) får man ofte et enklere utfallrom å arbeide med. Ved hjelp av setninger for betingede sannsynligheter kan kan vi så bygge mer kompliserte modeller.

Binormalfordelingen, som er den viktigste to-dimensjonale fordelingen, kan også enklest konstrueres ved å ta utgangspunkt i betingede endimensjonale fordelinger.

La U og V være uavhengige standard normalfordelte variable og definer

$$X = U \tag{1}$$

$$Y = aU + bV \tag{2}$$

hvor a og b er konstanter. Siden Y er en lineær funksjon av U og V er både X og Y her normalfordelte. Variansen til Y vil være lik 1 hvis $a^2 + b^2 = 1$. Kovariansen mellom X og Y er lik a, og hvis variansene er 1 så er dette også korrelasjonen mellom X og Y. Ved å velge $a = \rho$ og $b = \sqrt{1 - a^2} = \sqrt{1 - \rho^2}$, dvs.

$$Y = \rho U + \sqrt{1 - \rho^2}V \tag{3}$$

blir X og Y standard normalfordelte variable med korrelasjon ρ.

Simultanfordelingen til (X, Y) kan skrives på formen

$$f_{X,Y}(x, y) = f_{Y|X}(y|x) f_X(x) \tag{4}$$

hvor den betingede fordelinge $f_{Y|X}(y|x)$ er fordelinen til $\rho U + \sqrt{1-\rho^2}V$ når U er gitt (dvs. X er gitt). Siden U og V er uavhengige vil V være standard normalfordelt selv om U er gitt. Når U er gitt lik x er $\rho U + \sqrt{1-\rho^2}V$ (dvs. Y) derfor normalfordelt med forventning ρx og varians $1-\rho^2$. Ved å sette dette inn i (4) finner vi

$$f(x,y) = \frac{1}{\sqrt{2\pi(1-\rho^2)}}\ e^{-\frac{(y-\rho x)^2}{2(1-\rho^2)}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{5}$$

som kan trekkes sammen til

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}e^{-\frac{x^2-2\rho xy+y^2}{2(1-\rho^2)}}. \tag{6}$$

Dette er en standard binormalfordeling med korrelasjon ρ.

Vi kan konstruere den generelle binormale fordelingen på en tilsvarende måte ved å redefinere (X,Y) som

$$X = \sigma_x U + \mu_x \tag{7}$$

$$Y = \sigma_y(\rho U + \sqrt{1-\rho^2}V) + \mu_y. \tag{8}$$

Da har vi at X er $N(\mu_x, \sigma_x^2)$, Y er $N(\mu_y, \sigma_y^2)$ og korrelasjonen mellom X og Y er fremdeles lik ρ. Fordelingen til Y gitt $X = x$ er $N(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1-\rho^2))$. Ved igjen å sette inn i (4) og forenkle uttrykket i eksponenten finner vi

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2}-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right]}. \tag{9}$$

Dette er den generelle binormalfordelingen. Marginalfordelingene til X og Y er $N(\mu_x, \sigma_x^2)$ og $N(\mu_y, \sigma_y^2)$ og vi har sett at den betingede fordelingen til Y gitt $X = x$ er normalfordelingen med forventning og varians

$$\mathrm{E}(Y|X=x) = \mu_y + \frac{\sigma_y}{\sigma_x}\rho(x-\mu_x) \tag{10}$$

$$\mathrm{var}(Y|X=x) = \sigma_y^2(1-\rho^2). \tag{11}$$

Siden uttrykket for $f(x,y)$ er symmetrisk i x og y (vi får samme formel om vi bytter om x og y) har vi også

$$\mathrm{E}(X|Y=y) = \mu_x + \frac{\sigma_x}{\sigma_y}\rho(y-\mu_y) \tag{12}$$

$$\mathrm{var}(X|Y=y) = \sigma_x^2(1-\rho^2). \tag{13}$$

Det kan også vises at dersom (X,Y) er binormalfordelt så er også (Z,W) binormalfordelt når Z og W er lineære uttrykk i X og Y, dvs. $Z = cX + dY$ og $W = eX + fY$.

Sentralgrenseteoremet i to dimensjoner

La nå (X_i, Y_i), $i = 1, 2, \ldots n$ være n uavhengige to-dimensjonale stokastiske variable med samme to-dimensjonale fordeling med forventninger μ_x og μ_y, varianser σ_x^2 og σ_y^2 og korrelasjon ρ. Sentralgrenseteoremet i en dimensjon sier da at fordelingen til $(\bar{X} - \mu_x)\sqrt{n}/\sigma_x$ og $(\bar{Y} - \mu_y)\sqrt{n}/\sigma_y$ vil konvergere mot standard normalfordelingen når n vokser mot uendelig. Ifølge sentralgrenseteoremet i to dimensjoner vil simultanfordelingen til disse to variable konvergere mot standard binormalfordelingen gitt ved likning (6).

Siden vilkårlige lineærkombinasjoner av komponentene også er binormalfordelt blir da også $(\bar{X}, \bar{Y})$ og $(\sum X_i, \sum Y_i)$ tilnærmet binormalfordelt når n er stor.

Eksempel a - Multiplikativ vekst

Naturlig vekst, for eksempel individers lengde og vekt, eller vekst i størrelsen til en populasjon, kan ofte uttrykkes med en såkalt multiplikativ modell. La størrelsen ved tid (alder) $t = 0$ være S_0 og anta modellen

$$S_{t+1} = S_t X_t, \tag{14}$$

hvor faktorene X_t som størrelsen multipliseres med i løpet av en tidsenhet (f. eks et år) antas å være uavhengige med samme fordeling. Ved å løse dette rekursivt finner vi da at

$$S_{t+1} = S_0 X_1 X_2 \dots X_t. \tag{15}$$

Ved å ta logaritmen på begge sider finner vi

$$\ln S_{t+1} = \ln S_0 + \sum_{i=1}^{t} \ln X_i. \tag{16}$$

Det følger da av sentralgrenseteoremet i en dimensjon at $\ln S_{t+1}$ er tilnærmet $N(\ln S_0 + \mu_1 t, \sigma_1^2 t)$, hvor $\mu_1 = \mathrm{E}(\ln X_i)$ og $\sigma_1^2 = \mathrm{var}[\ln(X_i)]$.

La nå W_t være en annen størrelse som vokser og oppfyller tilsvarende modell

$$\ln W_{t+1} = \ln W_0 + \sum_{i=1}^{t} \ln(Y_i)$$

hvor faktorene Y_i her er uavhengige med samme fordeling. Her har vi tilsvarende at $\ln W_{t+1}$ er tilnærmet $N(\ln W_0 + \mu_2 t, \sigma_2^2 t)$, hvor $\mu_2 = \mathrm{E}(\ln Y_i)$ og $\sigma_2^2 = \mathrm{var}[\ln(Y_i)]$.

Dersom de to størrelsene vokser i det samme miljøet er det rimelig å tro at X_i og Y_i vil være avhengige. Da vil $(\ln S_t, \ln W_t)$ være tilnærmet binormalfordelt med korrelasjon $\rho = \mathrm{corr}(\ln X_i, \ln Y_i)$.

Legg merke til at S_t og W_t nå vil være tilnærmet lognormalfordelte. Simultanfordelingen til (S_t, Y_t), som vi ikke skal utlede her, kalles da en bilognormalfordeling.

Eksempel b - Høydefordeling for kvinner og menn

Anta at høyden til kvinner og menn er hhv. $N(170, 7^2)$ og $N(180, 7^2)$ og at andelen kvinner i en stor populasjon er 0.5. Vi betrakter et tilfeldig utvalg av n personer fra populasjonen. Til hver person kan vi knytte en todimensjonal variabel (X, Y), der X er høyden og Y en indikatorvariabel for kjønn definert slik at $Y = 1$ for en kvinne og $Y = 0$ for en mann. Siden populasjonen er stor kan utvalget betraktes som n uavhengige todimensjonale observasjoner $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Andelen av kvinner i utvalget kan nå uttrykkes som middelverdien $\bar{Y}$, mens $\bar{X}$ er middelhøyden i utvalget.

Ifølge sentralgrenseteoremet er da $(\bar{X}, \bar{Y})$ tilnærmet binormalfordelt når n er stor. La oss først finne parametrene i denne fordelingen. Vi vet fra før at korrelasjonen mellom middelverdiene generelt er lik korrelasjonen mellom X_i og Y_i.

Forventningen til $(X|Y)$ kan uttrykkes som

$$\mathrm{E}(X|Y) = 170Y + 180(1 - Y). \tag{17}$$

(Sett inn $Y = 0$ og $Y = 1$ for å kontrollere at dette er riktig). Setningen om dobbeltforventning gir da

$$\mathrm{E}X = 170\mathrm{E}Y + 180(1 - \mathrm{E}Y) = 175 \tag{18}$$

siden $\mathrm{E}Y = 0.5$. Vi kan finne $\mathrm{E}X^2$ på samme måten

$$\mathrm{E}(X^2|Y) = (170^2 + 7^2)Y + (180^2 + 7^2)(1 - Y) \tag{19}$$

som gir ubetinget $\mathrm{E}X^2 = 30699$ og $\mathrm{var}(X) = \mathrm{E}X^2 - (\mathrm{E}X)^2 = 74$. Vi ser at den ubetingede variansen til X er mye større enn den betingede variansen som er antatt lik 7^2, fordi den ubetingede også tar hensyn til variasjonen mellom menn og kvinner. Vi har nå at at $\bar{X}$ er tilnærmet $N(175, 74/n)$. Videre har vi at $\sum Y_i$ er $\mathrm{bin}(n/2, n/4)$ slik at $\bar{Y}$ er tilnærmet $N(\frac{1}{2}, \frac{1}{4n})$.

Det gjenstår å finne korrelasjonen mellom X og Y. Vi finner da først kovariansen mellom X og Y ved igjen å bruke setningen om dobbeltforventning.

Vi ser at

$$\mathrm{E}(XY|Y) = 170Y \tag{20}$$

(kontroller for $Y = 0$ og $Y = 1$). Ubetinget finner vi dermed $\mathrm{E}(XY) = \mathrm{EE}(XY|Y) = \mathrm{E}(170Y) = 85$. Kovariansen blir da $\mathrm{cov}(X, Y) = 85 - 175\frac{1}{2} = -2.5$ og korrelasjonen $\rho = -2.5/\sqrt{74\frac{1}{4}} = -0.5812$. Dette er da også korrelasjonen mellom $\bar{X}$ og $\bar{Y}$ og vi har dermed beregnet alle parametrene i binormaltilnærmelsen.

Anta nå at vi har et utvalg av $n = 100$ personer og har observert $\bar{X} = 173.2$. Dette skulle tyde på at det er flere kvinner i utvalget enn menn (fordi observasjonen er mindre enn den ubetingede forventingen 175). Hendelsen {flere kvinner enn menn} kan uttrykkes som $\{\bar{Y} > 0.5\}$. Vi har vist at $\bar{Y}$ gitt $\bar{X} = 173.2$ er normalfordelt med forventning og varians gitt ved likningene (10) og (11). Innsatt finner vi at den betingede fordelingen er $N(0.5608, 0.04069^2)$. Herav finner vi at

$$P(\bar{Y} > 0.5|\bar{X} = 173.2) = 1 - \Phi(\frac{0.5 - 0.5608}{0.04069}) = \Phi(1.494) = 0.933. \tag{21}$$

Eksempel c - Arvbarhet

La P betegne en fenotype, dvs. en kvantitativ størrelse som kan måles på et individ. I kvantitativ genetikk kan denne uttrykkes som $P = G + E$, der G er avlsverdien ("breeding value") som er genetisk bestemt, og E er en miljøkomponent som er bestemt av andre ting enn individets gener. Det er vanlig å skalere disse størrelsene (og eventuelt korrigere for forskjeller mellom kjønnene) slik at forventningsverdien er null. Det er ofte realistisk å anta at E er uavhengig av G, som gir

$$\mathrm{var}(P) = \sigma_G^2 + \sigma_E^2 \tag{22}$$

hvor $\sigma_G^2 = \mathrm{var}(G)$ og $\sigma_E^2 = \mathrm{var}(E)$. Videre er ofte G og E normalfordelte. Ifølge teorien ovenfor blir da (P, G) binormalfordelt. Kovariansen mellom P og G er $\mathrm{cov}(G + E, G) = \sigma_G^2$ og korrelasjonen blir dermed

$$\rho = \mathrm{corr}(P, G) = \frac{\mathrm{cov}(P, G)}{\sqrt{\mathrm{var}(P)\mathrm{var}(G)}} = \frac{\sigma_G^2}{\sqrt{\sigma_G^2 \sigma_P^2}} = \frac{\sigma_G}{\sigma_P}. \tag{23}$$

Ifølge likning (12) har vi da

$$\mathrm{E}(G|P) = \frac{\sigma_G}{\sigma_P}\rho P = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_E^2} P = h^2 P, \tag{24}$$

hvor $h^2 = \sigma_G^2/(\sigma_G^2 + \sigma_E^2)$ kalles arvbarheten til den kvantitative karakteren vi studerer.

Hvis det nå virker en seleksjon på populasjon slik at individene som overlever har fenotypisk middel S vil da den midlere avlsverdien til disse være $R = h^2 S$.

Ifølge den kvantitative genetiske teorien til Fisher vil avkommets avlsverdi i en populasjon (under visse forutsetninger som vi ikke skal diskutere her) være $G_{\mathrm{avkom}} = \frac{1}{2}(G_{\mathrm{far}} + G_{\mathrm{mor}}) + \varepsilon$ hvor $\mathrm{var}(\varepsilon) = \frac{1}{2}\sigma_G^2$. Vi ser at dette sikrer at $\mathrm{var}(G_{\mathrm{avkom}}) = \sigma_G^2$ slik at variansen ikke endres fra generasjon til generasjon.

Fra denne modellen finner vi for eksempel

$$\mathrm{cov}(P_{\mathrm{avkom}}, P_{\mathrm{mor}}) = \frac{1}{2}\sigma_G^2 \tag{25}$$

slik at korrelasjonen mellom P_{avkom} og P_{mor} blir $\frac{1}{2}h^2$. Siden P_{avkom} og P_{mor} har samme varians og forventningsverdiene er null finner vi da fra (12) og (13) at

$$\mathrm{E}(P_{\mathrm{avkom}}|P_{\mathrm{mor}}) = \frac{1}{2}h^2 P_{\mathrm{mor}} \tag{26}$$

og

$$\mathrm{var}(P_{\mathrm{avkom}}|P_{\mathrm{mor}}) = (1 - \frac{1}{4}h^4)\sigma_P^2. \tag{27}$$