

Transformasjoner av stokastiske variabler

Notasjon – merkelapper på fordelingene

Sannsynlighetstettheten og den kumulative fordelingen til en stokastisk variabel X betegnes hhv. f_X og F_X . Indeksen er altså en merkelapp som forteller hvilken stokastisk variabel som har den aktuelle fordelingen.

Funksjoner av stokastiske variabler

Hvis X er høyden til en vilkårlig valgt norsk mann målt i centimeter vil fordelingen til X være sentrert rundt verdien 180 (cm). Hvis vi innfører Y som høyden målt i meter vil verdiene gruppere seg rundt 1.80 (meter). Den matematiske sammenhengen mellom X og Y er da

$$Y = X/100. \tag{1}$$

Den variable Y er en funksjon av X , og den er da også en stokastisk variabel, men den vil nødvendigvis ha en helt annen fordeling enn X . Til ethvert enkeltutfall av eksperimentet (en mann trukket fra populasjonen) svarer det en verdi for X , men samtidig også en verdi for Y .

Et annet eksempel er å la X betegne oppnådd konsentrasjon av H^+ -ioner etter et gjennomført kjemisk eksperiment. Hvis Y er PH-verdien gjelder da sammenhengen

$$Y = -\log X. \tag{2}$$

De variable X og Y vil ha helt forskjellige verdier, og de vil også ha helt forskjellige fordelinger.

De samme betraktningene vil også være gyldige for en vilkårlig funksjonssammenheng mellom X og Y

$$X = g(Y). \quad (3)$$

Vi sier da at Y er en transformasjon av X definert ved funksjonen $g(\cdot)$. I sannsynlighetsregning og statistikk vil det ofte være nødvendig å studere slike transformasjoner av stokastiske variabler. Vi skal i det følgende se på hvordan vi kan finne fordelingen til $Y = g(X)$ når vi kjenner fordelingen til X .

Generell behandling av transformasjoner

Å finne fordelingen til Y når fordelingen til X er kjent gjøres enklest ved først å finne den kumulative fordelingen til Y som er

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y). \quad (4)$$

Videre kan ulikheten $g(X) \leq y$ alltid løses m.h.p. X . I eksemplet med høydene målt i centimeter og meter (likning 1) finner vi

$$\{g(X) \leq y\} = \{X/100 \leq y\} = \{X \leq 100y\} \quad (5)$$

slik at den kumulative fordelingen til Y da blir $F_Y(y) = P(X \leq 100y) = F_X(100y)$. Den kumulative fordelingen til Y er her uttrykt ved den (kjente) kumulative fordelingen til X . Vi kan nå finne sannsynlighetstettheten til Y ved å derivere m.h.p. y på begge sider. Vi må da bruke kjerneregelen som gir

$$f_Y(y) = F'_Y(y) = \frac{d}{dy} F_X(100y) = f_X(100y) \frac{d}{dy} (100y) = 100 f_X(100y). \quad (6)$$

Uttrykket $f_X(100y)$ betyr her at vi skal erstatte den variable x i formelen $f_X(x)$ med $100y$.

I eksemplet med det kjemiske eksperimentet (likning 2) finner vi tilsvarende

$$\{g(X) \leq y\} = \{-\log X \leq y\} = \{X \geq 10^{-y}\} \quad (7)$$

slik at $F_Y(y) = P(X \geq 10^{-y}) = 1 - P(X < 10^{-y}) = 1 - F_X(10^{-y})$. Sannsynlighetstettheten til Y blir da

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}[1 - F_X(10^{-y})] = f_X(10^{-y})10^{-y} \ln(10). \quad (8)$$

Vi har her brukt at den deriverte av funksjonen 10^{-y} er $-10^{-y} \ln(10)$.

De to transformasjonene ovenfor (likning 1 og 2) er definert ved strengt monotone funksjoner $g(\cdot)$. Løsningen på ulikheten $g(X) \leq y$ blir da et intervall på X -aksen. Generelt kan løsningen bli et sammensatt område på tall-linjen. For transformasjonen $Y = g(X) = X^2$ finner vi f. eks.

$$\{g(X) \leq y\} = \{X^2 \leq y\} = \{-\sqrt{y} \leq X \leq \sqrt{y}\}. \quad (9)$$

Den kumulative fordelingen til Y blir da

$$F_Y(y) = P(g(X) \leq y) = F_X(\sqrt{y}) - F_X(-\sqrt{y}) \quad (10)$$

som gir sannsynlighetstettheten

$$f_Y(y) = \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})]. \quad (11)$$

Strengt monotone transformasjoner

Fordelingen til Y kan uttrykkes ved en enkel generell formel for alle strengt monotone transformasjoner. Anta først at $g(X)$ er en strengt voksende funksjon. Løsningen på likningen $g(X) = y$ m.h.p. X blir da en funksjon av y (den inverse funksjonen av $g(\cdot)$) som vi skriver

$$X = h(y). \quad (12)$$

I eksempelet med høyder målt i centimeter og meter (likning 1) blir da $h(y) = 100y$, mens kjemi-eksempelet (likning 2) gir $h(y) = 10^y$. Siden funksjonen er strengt voksende finner vi da

$$\{g(X) \leq y\} = \{X \leq h(y)\} \quad (13)$$

slik at kumulativ fordeling til Y blir

$$F_Y(y) = P(g(X) \leq y) = P(X \leq h(y)) = F_X(h(y)). \quad (14)$$

Sannsynlighetstettheten blir da

$$f_Y(y) = F'_Y(y) = \frac{d}{dy} F_X(h(y)) = f_X(h(y))h'(y). \quad (15)$$

Tilsvarende regning for en strengt avtagende funksjon $g(\cdot)$ gir

$$\{g(X) < y\} = \{X > h(y)\} \quad (16)$$

og kumulativ fordeling

$$F_Y(y) = P(g(X) \leq y) = P(X \geq h(y)) = 1 - P(X < h(y)) = 1 - F_X(h(y)). \quad (17)$$

Ved å derivere begge sider av likning 17 m.h.p. y finner vi nå sannsynlighetstettheten

$$f_Y(y) = \frac{d}{dy} [1 - F_X(h(y))] = -f_X(h(y))h'(y). \quad (18)$$

For en strengt avtagende funksjon $g(\cdot)$ er også den inverse funksjonen $h(y)$ strengt avtagende slik at den deriverte $h'(y)$ er negativ. Vi kan da skrive $-h'(y)$ som $|h'(y)|$. For voksende $h(y)$ kan vi tilsvarende skrive $h'(y)$ som

$|h'(y)|$. Ved å velge å uttrykke resultatene ved $|h'(y)|$ ser vi da at formelene for $F_Y(y)$ blir like for strengt voksende og strengt avtagende transformasjoner. **For strengt monotone (voksende eller avtagende) transformasjoner gjelder da altså**

$$\mathbf{f}_Y(\mathbf{y}) = \mathbf{f}_X(\mathbf{h}(\mathbf{y}))|\mathbf{h}'(\mathbf{y})|. \quad (19)$$

Noen eksempler på transformasjoner

(a) *Normalfordeling.*

La nå X være standard normalfordelt, dvs.

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad (20)$$

Vi vil finne fordelingen til den variable $Y = X^2$. Den variable X kan ta verdier på hele tall-linjen, mens Y åpenbart ikke kan ta negative verdier. Ved å sette inn i den generelle formelen vi fant for denne transformasjonen (likning (11)) finner vi for $y \geq 0$

$$f_Y(y) = \frac{1}{2\sqrt{2\pi y}} [e^{-(\sqrt{y})^2/2} + e^{-(\sqrt{y})^2/2}] = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad (21)$$

mens sannsynligheten er null for negative y (fordi Y ikke kan ta negative verdier).

(b) *Lineære transformasjoner.*

Spesielt når vi analyserer normalfordelte variable blir det ofte nødvendig å se på lineære transformasjoner $g(X) = aX + b$. Hvis a ikke er lik null er dette en strengt monoton transformasjon med invers funksjon $h(y) = (y - b)/a$. Fordelingen til Y finnes da ved å sette inn i likning 19

$$f_Y(y) = f_X((y - b)/a)|1/a|. \quad (22)$$

Spesielt får vi f.eks. at hvis X er standard normalfordelt, dvs. $f_X(x) = (2\pi)^{-1/2}e^{-x^2/2}$, blir (for $a > 0$) $f_Y(y) = (2\pi)^{-1/2}a^{-1}e^{-(y-b)^2/(2a^2)}$.

(c) *Generering av eksponensielt fordelte variabler med datamaskin.*

De fleste datamaskiner har innebygde rutiner ("pseudo random number generators") som gir en variabel en tilfeldig verdi mellom 0 og 1, eller mer presist, den variable X blir tilordnet en verdi som er rektangulært fordelt på intervallet $[0,1]$:

$$f_X(x) = 1 \text{ for } 0 \leq x \leq 1, \quad (23)$$

og lik null når x ikke ligger i intervallet $[0,1]$.

Vi lar så maskinen beregne $Y = -\frac{1}{\lambda} \ln X$, der λ er et gitt positivt tall. Hvilken fordeling får da den variable Y ?

Dette er en monoton transformasjon og den inverse funksjonen er $h(y) = e^{-\lambda y}$. Vi finner igjen fordelingen til Y ved å sette inn i likning (19), og finner

$$f_Y(y) = f_X(e^{-\lambda y}) | -\lambda e^{-\lambda y} | = \lambda e^{-\lambda y}.$$

(c) *Lognormalfordelingen.*

La X være $N(\mu, \sigma^2)$ og definer $Y = e^X$. Vi ser at Y bare tar positive verdier. Siden dette er en monoton transformasjon kan vi bruke formelen (19) med $h(y) = \ln(y)$ som gir

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}y} e^{-\frac{(\ln(y)-\mu)^2}{2\sigma^2}}. \quad (24)$$

Dette kalles lognormalfordelingen med parametre μ og σ^2 . Tolkningen av parametrene er da at $E(\ln Y) = \mu$ og $\text{var}(\ln Y) = \sigma^2$. Det kan vises at

$$EY = e^{\mu+\sigma^2/2} \quad (25)$$

og at

$$\text{var}(Y) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1). \quad (26)$$