Valg af fordeling for stokastiske variable

Indhold

1 Diskrete fordelinger

Karakteristikum: Den stokastiske variabel, X har kun *diskrete udfald*.

De mulige udfald ordnes (oftest) som heltallige udfald, dvs, at X typisk kan antage værdier som heltallene $0, 1, 2, \ldots$.

Det mest almindelige diskrete udfald er, hvor man *tæller* et vist antal af bestemte hændelser eller udfald.

2 De mest benyttede diskrete fordelinger

2.1 Poissonfordelingen: Pois(λ)

Biologi
X = Optalt antal insekter af en bestemt art på en bestemt plante. X kan antage værdierne $0, 1, 2, \ldots$.

Biologi
X = Optalt antal dyr af en bestemt art på et bestemt areal (f.eks. grønne løvfrøer på én kvadratkm. mose).

Biokemi
X = Optalt antal svampe-vækstpunkter i et felt på en petriskål med en vandprøve.

Trafik
X = Optalt antal cyklister, som passerer et bestemt sted i et bestemt tidsrum.

Telefoni
X = Optalt antal SMS-beskeder, som en bestemt mobilsendemast skal håndtere i et bestemt tidsrum, f.eks. i ét minut.

Sundhed/sygdom
X = Antal af en bestemt mikroorganisme (f.eks. en bestemt virus) i én ml. blod fra en (potentielt) smittet patient.

Fysik
X = Antal γ-partikler i et vist tidsrum fra en radioaktiv kilde (en bestemt mængde, f.eks. 1 mg).

Approximation/grænse
Hvis $X \in \text{Binomial}(n,p) = Bin(n,p)$ for n meget stor og p meget lille: $Bin(n,p) \Rightarrow Pois(np)$, se nedenfor og bogen side 129(126).

I alle eksemplerne er X heltallig og i princippet ubegrænset, omend sandsynligheden for meget store X-værdier kan være hastigt aftagende.

$$X \in Pois(\lambda) \Longleftrightarrow P_r\{X = x\} = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \;\; ; \;\; x = 0, 1, 2, \ldots$$

$$P_r\{X \leq x\} = \sum_{i=0}^{x} \frac{\lambda^i \cdot e^{-\lambda}}{i!} \;\; ; \;\; x = 0, 1, 2, \ldots$$

$$E\{X\} = \lambda \;\text{ og }\; V\{X\} = \lambda$$

Approximation for store λ-værdier: $Pois(\lambda) \Rightarrow N(\lambda, \lambda) \Rightarrow$

$$P_r\{X \leq x\} \simeq P_r\{N(\lambda,\lambda) \leq x + \frac{1}{2}\} = \Phi\left(\frac{x + 1/2 - \lambda}{\sqrt{\lambda}}\right)$$

$$P_r\{x_1 \leq X \leq x_2\} \simeq \Phi\left(\frac{x_2 + 1/2 - \lambda}{\sqrt{\lambda}}\right) - \Phi\left(\frac{x_1 - 1/2 - \lambda}{\sqrt{\lambda}}\right)$$

Hvis $X_1 \in Pois(\lambda_1)$ og $X_2 \in Pois(\lambda_2) \Rightarrow X_1 + X_2 \in Pois(\lambda_1 + \lambda_2)$. Kan generaliseres til flere Poissonfordelte variable. Parameteren λ kaldes ofte Poisson*intensiteten*.

2.2 Binomialfordelingen: Bin(n,p)

Biologi
Undersøg n individer af en bestemt art insekter (bananfluer f.eks.) og optæl X = antal (blandt de n) af en bestemt fænotype (med et bestemt fremtoningspræg, øjenfarve f.eks.). X kan antage værdierne $0, 1, 2, \ldots, n$ dvs. heltallig, men *begrænset* til max. n.

Sundhed/Sygdom
Undersøg n personer og optæl X = antal personer (blandt de n), som har et bestemt antistof (for en vis sygdom f.eks.) i blodet. X kan igen antage værdierne $0, 1, 2, \ldots, n$ dvs. heltallig, men *begrænset* til max. n, det samlede antal undersøgte individer.

Biokemi
Foretag udstrygning af en vandprøver fra et vandværk på n petriskåle. Optæl X = antallet af petriskåle (blandt de n), som udviser vækst. p angiver sandsynligheden for vækst på én skål.

Trafik
Stop n cyklister, som passerer et bestemt sted på en cykelsti, og registrér X = antal heraf, som kører uden cykellygte. p angiver sandsynligheden for, at én cyklist kører uden lygte. X kan igen antage værdierne $0, 1, 2, \ldots, n$ dvs. heltallig, men *begrænset* til max. n, det samlede antal standsede cyklister.

Lægemidler/Kvalitetskontrol
Udtag n ampuller med et vist lægemiddel, og undersøg de enkelte ampullers indhold mht. koncentration/styrke af aktivt stof. X = antal ampuller af utilfredsstillende kvalitet (blandt de n undersøgte ampuller). p angiver fejlandelen af ampuller.

Fysik/Materialer
Udtag n prøver af et bestemt materiale. Mål på de n prøver, om de kan modstå en vis elektrisk spændingsbelastning. X = antal blandt de n prøver, som modstår belastningen. p angiver andelen af materialeprøver, som modstår belastningen.

I eksemplerne er X i alle tilfælde heltallig, men begrænset til max. n.

$$X \in Bin(n,p) \Longleftrightarrow P_r\{X = x\} = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \quad ; \quad x = 0, 1, 2, \ldots, n$$

$$P_r\{X \leq x\} = \sum_{i=0}^{x} \left(\begin{array}{c} n \\ i \end{array} \right) p^i (1-p)^{n-i} \quad ; \quad x = 0, 1, 2, \ldots, n$$

$$E\{X\} = np \quad \text{og} \quad V\{X\} = np(1-p)$$

Approximation for store n-værdier men små p-værdier:
$Bin(n,p) \Rightarrow Pois(np) \Rightarrow$

$$P_r\{X \leq x\} \simeq P_r\{Pois(np) \leq x\}$$

Anvendes fortrinsvis for $n \geq 20$ og $p \leq 0.05$. Andre brugbare regler er ($n \geq 50$ og $np \leq 5$) eller ($n \geq 100$ og $np \leq 10$) eller tilsvarende.

Approximation for store n-værdier og $np \geq 5$ og $n(1-p) \geq 5$:
$Bin(n,p) \rightarrow N(np, np(1-p)) \Rightarrow$

$$P_r\{X \leq x\} \simeq \Phi\left(\frac{x + 1/2 - np}{\sqrt{np(1-p)}} \right)$$

$$P_r\{x_1 \leq X \leq x_2\} \simeq \Phi\left(\frac{x_2 + 1/2 - np}{\sqrt{np(1-p)}} \right) - \Phi\left(\frac{x_1 - 1/2 - np}{\sqrt{np(1-p)}} \right)$$

Hvis $X_1 \in Bin(n_1, p)$ og $X_2 \in Bin(n_2, p) \Rightarrow X_1 + X_2 \in Bin(n_1 + n_2, p)$. Kan generaliseres til flere binomialfordelte variable.

2.3 Bernoulli-fordelingen: Bern(p)

Binomialfordelingen for $n = 1$:

$$X \in Bern(p) \Leftrightarrow f(x) = P_r\{X \leq x\} = p^x(1-p)^{1-x} \quad \text{for} \quad x = \{0, 1\}$$

$$E\{X\} = p \quad \text{og} \quad V\{X\} = p(1-p)$$

Hvis $X_1 \in Bern(p)$ og $X_2 \in Bern(p) \Rightarrow X_1 + X_2 \in Bin(2, p)$. Kan generaliseres til flere Bernoullifordelte variable.

3 Andre diskrete fordelinger

3.1 Hypergeometrisk fordeling: Hyp(n,M,N)

Udtagning af prøver uden tilbagelægning/Kvalitetskontrol
Sæt, at man har i alt N enheder og, at ud af disse er præcis M defekte. Udtager vi nu en stikprøve på n enheder blandt de N, vil antallet af defekte, X, i stikprøven følge en hypergeometrisk fordeling. X kan antage værdierne 0, 1, 2,...,M.

$$X \in Hyp(n, M, N) \Longleftrightarrow P_r\{X = x\} = \frac{\begin{pmatrix} M \\ x \end{pmatrix}\begin{pmatrix} N-M \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}$$

$$E\{X\} = n \cdot M/N \quad \text{og} \quad V\{X\} = n\frac{M}{N}\frac{(N-M)}{N}\frac{(N-n)}{(N-1)}$$

Approximation for $n << N$ og $n << (N - M)$: $Hyp(n, M, N) \Rightarrow Bin(n, M/N)$

3.2 Multinomialfordelingen: Multi$(n, p_1, p_2, \ldots, p_k)$

Kaldes i mange fremstillinger også for *polynomial-fordelingen*. Er en generalisering af binomialfordelingen.

Biologi
Undersøg n individer af en bestemt art (mennesker f.eks.) og opdel dem efter fænotype (med et bestemt fremtoningspræg, øjenfarve f.eks.). X_1 er antal med blå øjne, X_2 er antal med brune øjne, X_3 er antal med grønne øjne, X_4 er antal med grå øjne, X_5 er antal med andre øjenfarver (f.eks. ét brunt og ét blåt øje). $p_1, p_2, \ldots, p_k$ angiver andelene af de k kategorier.

Sundhed/Sygdom
Undersøg n personer af en bestemt kategori (f.eks. mænd med en bestemt sygdomsdisposition) og optæl X = antal personer (blandt de n), som har en bestemt blodtype. X_1=antal blodtype 0, X_2=antal blodtype A, X_3=antal blodtype B, X_4=antal blodtype AB. (p_1, p_2, p_3, p_4) angiver blodtypefordelingen for den pågældende kategori af personer.

Kvalitetskontrol
Udtag n prøver af en vare (f.eks. en stor sending af et medicinsk præparat) og kategoriser de enkelte prøver efter X_1= antal OK, X_2= antal med små fejl, X_3= antal med betydelige fejl, X_4= antal med kritiske fejl.

$$f(x_1, x_2, \ldots, x_k) = P_r\{X_1 = x_1 \cap X_2 = x_2 \cap \ldots \cap X_k = x_k\} = \frac{n! \cdot p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}}{x_1! x_2! \cdots x_k!}$$

$$E\{X_1, X_2, \ldots, X_k\} = (np_1, np_2, \ldots, np_k)$$

Det gælder, at $X_1 + X_2 + \ldots + X_k = n$ og $p_1 + p_2 + \ldots + p_k = 1$.

3.3 Geometrisk fordeling: Geom(p)

Den geometriske fordeling hører sammen med binomialfordelingen, idet den er fordeling for *ventetiden* (målt i antal) ved gentagne binomialforsøg, indtil man møder den første "succes". X kan antage heltalsværdierne 1, 2, 3,..., i princippet ubegrænset.

Kvalitetskontrol
Udtag prøver af en vare (f.eks. en sending af et medicinsk præparat) og bliv ved, indtil du finder den første fejlbehæftede enhed. X angiver antal prøver, der er udtaget *ialt* (altså inklusiv den fejlbehæftede). X kan antage værdierne 1, 2, 3, ... (principielt ubegrænset).

Biologi
Sæt, at man vil udvikle en ny type plante, hvor man forsøger at krydse to beslægtede arter. I ét forsøg krydses de to arter, og X angiver antal forsøg, der skal gøres, indtil (og med) det første gang lykkes at få en levedygtig plante, der kan opformeres. X kan antage værdierne 1, 2, 3, ... (principielt ubegrænset).

Sundhed
Sæt, at man har en behandlingsmetode, som ikke virker hver gang, men kun med en vis sandsynlighed, f.eks. et bestemt træningprogram (rygeafvænning f.eks). Hvis sandsynligheden for, at ét behandlingsforsøg virker, er p, vil antallet af nødvendige behandlinger følge en geomentrisk fordeling med parameter p. X kan antage værdierne 1, 2, 3, ... (principielt ubegrænset).

$$X \in Geom(p) \Leftrightarrow f(x) = P_r\{X = x\} = p(1-p)^x \quad \text{for} \quad x = 1, 2, 3, \ldots$$

$$P_r\{X \leq x\} = \sum_{i=1}^{x} p(1-p)^i \quad \text{for}$$

$$E\{X\} = 1/p \quad \text{og} \quad V\{X\} = (1-p)/p^2$$

4 Kontinuerte fordelinger

Karakteristikum: Den stokastiske variabel, X er kendetegnet ved, at den kan antage ikke heltallige værdier, f.eks. *alle reelle tal* (normalfordelingen), *de positive tal* (log-normalfordelingen) eller eksponentialfordelingen. De kan også være begrænset til et *interval* som i den uniforme fordeling.

4.1 Exponentialfordelingen: Exp(β)

Hører naturligt sammen med Poissonfordelingen, idet den er ventetidsfordeling for den næste hændelse i en Poissonproces. X kan antage positive reelle værdier , dvs. alle $x > 0$.

Biologi/Trafik
Man måler tiden, der går, mellem en bestemt dyreart observeres på et bestemt sted eller tidsafstanden mellem cyklister uden lys på en bestemt vej. Tiden, T, mellem observationerne kan ofte beskrives ved en exponentialfordeling. Hvis *antallet* X pr tidsenhed er Poissonfordelt med parameter λ, er *tidsafstanden* T exponentialfordelt med middelværdi $\beta = 1/\lambda$ - og vice versa.

Fysik
Ventetiden, T, mellem to γ-partikler fra en radioaktiv kilde.

Kvalitetskontrol/Levetidsundersøgelser For mange komponenter i apparater kan tiden fra komponenten ibrugtages, til den fejler (f.eks. en el-pære, som brænder over) beskrives ved en exponentialfordeling. Den gennemsnitlige tid kaldes ofte komponentens middellevetid eller blot levetiden.

$$X \in \exp(\beta) \Leftrightarrow f(x) = \frac{1}{\beta}\exp(-x/\beta) \quad \text{for} \quad x > 0 \quad \text{og} \quad \beta > 0$$

$$F(x) = P_r\{X \leq x\} = \int_0^x f(t)dt = 1 - \exp(-x/\beta)$$

$$E\{X\} = \beta \quad \text{og} \quad V\{X\} = \beta^2$$

Hvis $X_1 \in \exp(\beta)$ og $X_2 \in \exp(\beta) \Rightarrow X_1 + X_2 \in Gam(2, \beta)$. Kan generaliseres til flere exponentialfordelte variable.

4.2 Gammafordelingen: Gam(k,β)

hører naturligt sammen med Poisson- og exponentialfordelingen. Den er nemlig ventetidsfordeling mellem k hændelser, hvor de enkelte hændelsers tidsafstande alle er exponentialfordelte $\exp(\beta)$ og ikke influerer på hinanden (er uafhængige, som man siger). X kan antage positive reelle værdier , dvs. alle $x > 0$.

Biologi/Trafik
Man måler tiden, der går, mellem en bestemt dyreart observeres på et bestemt sted eller tidsafstanden mellem cyklister uden lys på en bestemt vej. Tiden, T, fra man starter med at observere, til det k'te individ observeres, følger en Gam(k, β)-fordeling, hvis tidsafstandende fra ét individ til det næste er exponentialfordelt exp(β).

Køteori/Planlægning
Man kan forestiller sig, at det tager en bestemt tid at udføre en bestemt opgave (f.eks. en operation på et sygehus). Samtidig ankommer kunder (patienter), som skal behandles med en vis intensitet, som kunne svare til en Poissonfordeling. Man kan så være interesseret i at vurdere sandsynligheden for, at køen af kunder (patienter) ikke overskrider et vist antal, svarende til, at der ikke ankommer flere end et vist antal i et bestemt tidsrum.

Hvis T er ventetiden til den k'te kunde (patient) ankommer, er $P_r\{T \leq t_0\}$ lig med sandsynligheden for, at den k'te kunde (patient) ankommer inden tidspunktet t_0

$$X \in Gam(k,\beta) \Leftrightarrow f(x) = \frac{x^{k-1}}{\beta^k \Gamma(k)} \exp(-x/\beta) \quad \text{for} \quad x > 0 \quad , \quad k > 0 \quad \text{og} \quad \beta > 0$$

$$F(x) = P_r\{X \leq x\} = \int_0^x f(t)dt \quad \text{(beregnes på computer f.eks.)}$$

$$E\{X\} = k \cdot \beta \quad \text{og} \quad V\{X\} = k \cdot \beta^2$$

Hvis $X_1 \in Gam(k_1, \beta)$ og $X_2 \in Gam(k_2, \beta) \Rightarrow X_1 + X_2 \in Gam(k_1 + k_2, \beta)$. Kan generaliseres til flere gammafordelte variable.

4.3 Uniformfordelingen: U(α,β)

Afrundingsfejl/Måling
Når man anfører et måleresultat afrundet til et begrænset antal cifre, vil der være en afvigelse fra den faktiske værdi. På et pH-meter udlæses f.eks. værdien pH=7.42. Den faktiske pH-værdi ligger et sted i intervallet $[7.415 - 7.425]$ og med lige stor sandsynlighed over hele intervallet.

Afrundingsfejlen X er altså mellem -0.005 og +0.005, dvs $X \in U(-0.005, +0.005)$.

Fysik/Materialer
Hvis man undersøger et stykke metaltråd af en bestemt længde, f.eks. b, og finder dens svageste sted, vil man i reglen antage, at dette optræder et helt tilfældigt sted over trådens længde. Kaldes stedet X, vil $X \in U(0, b)$.

Biologi
Under et eksperiment har man på et bestemt tidspunkt t_0 et individ, som lever. Til tiden t_1 konstaterer man, at individet er dødt. Hvis man ikke ved andet, vil

man kunne benytte uniformfordelingen $U(t_0, t_1)$ som model for tidspunktet, hvor individet døde.

$$X \in U(\alpha, \beta) \Leftrightarrow f(x) = \frac{1}{\beta - \alpha} \quad \text{for} \quad \alpha < x < \beta$$

$$P_r\{X \leq x\} = \int_\alpha^x f(t)dt = \frac{x - \alpha}{\beta - \alpha}$$

$$E\{X\} = (\alpha + \beta)/2 \quad \text{og} \quad V\{X\} = (\beta - \alpha)^2/12$$

4.4 Normalfordelingen: N(μ,σ^2)

er statistikkens (uden sammenligning) mest betydningsfulde fordeling, og den benyttes i et utal af sammenhænge, dels som model for et stort antal naturligt forekommende (mere eller mindre) tilfældige fænomener. Desuden er den grænsefordeling for mange andre fordelinger. X kan antage alle reelle værdier.

Den centrale grænseværdisætning
Denne sætning findes i flere varianter. Den enkleste er: Antag , at X_i , $i = \{1, 2, \ldots, n\}$, er uafhængige stokastiske variable med samme fordeling, der har middelværdi μ og varians σ^2. Da vil summen

$$\sum_{i=1}^{n} X_i \quad \text{asymptotisk} \quad \in N(n\mu, n\sigma^2)$$

Sætningen findes i andre varianter, hvor det f.eks.ikke kræves, at alle X'er har helt samme fordeling.

Hovedbudskabet er, at summer af mange stokastiske variable, som er af nogenlunde samme størrelsesorden, (eventuelt tilnærmelsesvist) vil følge en normalfordeling.

Approximation
Mange approksimationer udspringer af den centrale grænseværdisætning. Hvis f.eks. $E\{X_i\} = \mu$ og $V\{X_i\} = \sigma^2$ vil gennemsnit $\overline{X} = (\sum X_i)/n$ ikke alene have en middelværdi μ og varians σ^2/n, den vil også (evt tilnærmelsesvist) følge en normalfordeling.

Poisson- og binomialfordelingen tilnærmes for store n ved normalfordelingen (se under disse).

Fysik/Kemi/Målefejl
En fysisk eller kemisk størrelse har en bestemt (men ukendt) værdi f.eks. θ, som kunne være pH=6.4987856432765712456....=θ, dvs en reel værdi. Med en målemetode forsøger man at bestemme θ, og man udlæser værdien Y. Målefejlen er $X = Y - \theta$.

Hvis man ikke kender noget til sit apparat, kan man ikke sige meget om, hvor den faktiske værdi θ ligger i forhold til måleresultatet Y. Dvs. man kan ikke sige

præcist hvor stor X i det konkrete tilfælde er, men en god model er (meget ofte), at X er normalfordelt med en vis middelværdi, μ, og en vis varians, σ^2 (hvor μ gerne skulle være nær nul og σ^2 så lille som muligt).

Biologi
Hvis man har en population, hvor der er en vis variation mellem individerne (det kan f.eks. være deres vægt, længde, eller andre fysisk-kemiske egenskaber), vil normalfordelingen ofte være velegnet til at beskrive disse. Typisk vil der være to fordelinger for en given aldersgruppe, nemlig én for hun-individer og én for hanindivider.

Sundhed
Når man bestemmer blodsukker hos en patient, vil dette variere omkring en vis værdi, hvis man måler flere gange. Variationen omkring patientens middelniveau, μ, for indhold af sukker i blod beskrives godt med en normalfordeling med middelværdi μ (for raske personer mellem ca. 4 og 7 mmol/liter).

Kvalitetskontrol/Lægemidler
Ved fremstilling medicinske præparater tilstræbes, at deres styrke (koncentration) ligger i et bestemt (snævert) interval omkring den tilstræbte værdi μ. Der vil altid være større eller mindre variationer for denne koncentration i forhold til det tilstræbte. Sådanne variationer bekrives i reglen godt med normalfordelingen.

Miljø
pH-værdien i almindelig nedbør bør være gennemsnitligt nær 7, dvs neutral. I det daglige vejr vil den faktiske nedbørs pH-værdi variere omkring en vis værdi, som kan være $\mu < 7$ (sur regn). En egnet model for disse *variationer* er normalfordelingen (hvis ikke pludselige begivenheder gør billedet atypisk - vulkanudbrud f.eks.).

Folkesundhed
Når man måler en egenskab hos et stort antal personer (en normalbefolkning f.eks.) finder man hyppigt, at værdierne fordeler sig som fra en normalfordeling. Det gælder f.eks. længde og vægt af spædbørn med en vis alder, og på samme måde vægt og længde for voksne (mænd og kvinder har typisk hver sin fordeling). Et mål for fedme er det såkaldte *body mass index*, som modelleres udmærket med en normalfordeling. Den enkelte persons værdi sammenholdes med normalpopulationens fordeling (den må ikke ligge for langt væk fra midten).

Intelligens
Intelligenskvotienten hos normale mennesker har et normalområde, og en fordeling, som intelligensforskere med et poppet udtryk kalder 'the bell curve', der altså blot er normalfordelingen.

Sport/Doping
Når sportsfolk indtager midlet epo (og lignende præparater) øges antal røde blodlegener, dvs blodets evne til at optage ilt. Iltoptagelsesevnen måles med den såkaldte hematocritværdi (koncentrationsmål for røde blodlegemer). Hos normale unge personer, varierer denne omkring 45 for unge mænd og 41 for unge kvinder, i begge tilfælde med en spredning på ca 3. Normalfordelingen kunne

muligvis anvendes, men i praksis ligger der relativt mange omkring fordelingens midte, og fordelingens haler er hurtigere aftagende end i normalfordelingen. Eksemplet er et eksempel på, at selv i tilfælde, hvor normalfordelingen synes et oplagt valg, kan der være problemer.

Dagligvarer
Når man køber en dagligvare, f.eks. en pakke fødevarer med påskriften '1000 g' vil det faktiske indhold afvige herfra. Ved påfyldning skal fabrikanten sikre, at sandsynligheden for, at der faktisk er *mindre* end de 1000 g, er lille. Det er ikke tilstrækkeligt, at indholdet i middel for mange pakker er 1000 g (det er det, 'e'-mærkningen handler om).

Som model for den faktisk påfyldte mængde benyttes gerne normalfordelingen, der følgelig skal have en middelværdi $\mu >$1000 g, for at sikre, at kun få pakker er undervægtige.

$$X \in N(\mu, \sigma^2) \Leftrightarrow f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$P_r\{X \leq x\} = \int_{-\infty}^{x} f(t)dt = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

hvor

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad \text{og} \quad \Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

$$E\{X\} = \mu \quad \text{og} \quad V\{X\} = \sigma^2$$

Hvis $X_1 \in N(\mu_1, \sigma_1^2)$ og $X_2 \in N(\mu_2, \sigma_2^2)$ og $Y = a + b \cdot X_1 + c \cdot X_2$, vil for alle reelle a, b og c

$$Y \in N(a + b \cdot \mu_1 + c \cdot \mu_2 \ , \ b^2\sigma_1^2 + c^2\sigma_2^2)$$

Resultatet kan generaliseres til alle linearkombinationer af vilkårligt mange *uafhængige* normalfordelte variable.

4.5 Log-normalfordelingen: LN(α, β^2)

Generelt, såfremt $Y \in N(\alpha, \beta^2) \Leftrightarrow X = e^Y \in LN(\alpha, \beta^2)$ (definition). Symbolet log(.) betegner *den naturlige logaritme*, dvs med grundtal e.

Det betyder altså, at såfremt $X \in LN(\alpha, \beta^2)$, vil den naturlige logaritme af X følge en $N(\alpha, \beta^2)$-fordeling. X kan antage positive reelle værdier, dvs. alle $x > 0$.

Når man arbejder med log-normalfordelte værdier, logaritmetransformeres først, og derefter benyttes normalfordelingen på sædvanlig måde.

Pålidelighed
Sæt, at et system består af et stort antal komponenter, og at $Q_1, Q_2, \ldots, Q_N$

angiver sandsynlighederne for, at de enkelte komponenter virker til et bestemt tidspunkt. Q'erne er stokastiske variable, $0 << Q_i \leq 1$, dvs at sandsynligheden for, at ét eksemplar af systemet som helhed fungerer, er:

$$Q_{system} = Q_1 \cdot Q_2 \cdots Q_N$$

Tager vi nu den naturlige logaritme:

$$\log(Q_{system}) = \log(Q_1) + \log(Q_2) + \cdots \log(Q_N)$$

Ifølge *den centrale grænseværdisætning* vil $\log(Q_{system})$ (tilnærmelsesvist) følge en normalfordeling. Q_{system} selv vil følge en lognormalfordeling.

Biologi/Stokastisk vækst
I mange biologiske systemer er væksten i et kort tidsrum (med en vis variation) proportional med den tilstedeværende mængde biomasse (bakterier, alger, gærceller). Til tidspunktet i kaldes biomassen X_i og vækstkoefficienten er C_i.

Som model for væksten fra tiden i til tiden $i+1$ benyttes da, at

$$X_{i+1} = X_i + C_i \cdot X_i = (1 + C_i)X_i$$

Kaldes biomassen til tiden '0' for x_0, vil den til tiden n være

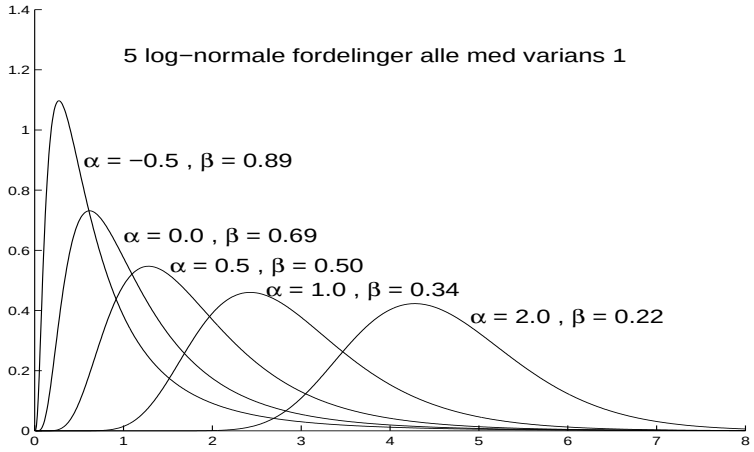
$$X_n = x_0 \cdot (1 + C_1)(1 + C_2) \ldots (1 + C_n) = x_0 \cdot \Pi_{i=1}^{n}(1 + C_i)$$

Hvis nu C_i'erne er stokastiske variable med (nogenlunde) samme fordeling, vil log-normalfordelingen (igen ifølge *den centrale grænseværdisætning*) være en naturlig fordeling for mængden af biomasse til et bestemt tidspunkt efter igangsætningen af væksten.

Pulverteknologi/Tabletter
Som model for f.eks. kornstørrelsen af de enkelte korn i et pulver, som f.eks. presses til tabletter, benyttes log-normalfordelingen ofte.

På samme måde benyttes log-normalfordelingen som model for såvel vægtfordelingen som antalsfordelingen for forskellige fraktioner ved sigtning af pulvere.

Skæve fordelinger/Transformation
I mange måletekniske situationer kan værdierne udvise en skæv fordeling, samtidig med, at det formelt set er umuligt at opnå negative værdier (mål og vægt f.eks). Når værdierne er langt fra 0, vil fordelingen ofte synes normal, men for små værdier finder man hyppigt fordelinger, som ligner log-normalfordelinger. Følgende figur illustrerer dette:



Man noterer, at fordelingen længst til højre meget ligner en normalfordeling, medens fordelingen for de små værdier er meget (højre-)skæv.

$$X \in LN(\alpha, \beta^2) \Leftrightarrow f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left(-\frac{(\log(x)-\alpha)^2}{2\beta^2}\right)$$

$$P_r\{X \le x\} = \int_0^x f(t)dt = \Phi\left(\frac{\log(x)-\alpha}{\beta}\right)$$

hvor

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \quad \text{og} \quad \Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

$$E\{X\} = \exp(\alpha + \beta^2/2) \quad \text{og} \quad V\{X\} = \exp(2\alpha + \beta^2)(\exp(\beta^2) - 1)$$

Hvis $X_1 \in LN(\alpha_1, \beta_1^2)$ og $X_2 \in LN(\alpha_2, \beta_2^2)$ og $Y = a \cdot X_1^b \cdot X_2^c$, vil for $a > 0$ og reelle b og c :

$$Y \in LN(\log(a) + b \cdot \alpha_1 + c \cdot \alpha_2 \ , \ b^2\beta_1^2 + c^2\beta_2^2)$$

Resultatet kan generaliseres til alle produktkombinationer af vilkårligt mange *uafhængige* log-normalfordelte variable.