Fordelinger

En oversigt over de vigtigste sandsynlighedsteoretiske fordelinger
Anden udgave
Udvidet version

Ulrich Fahrenberg
uli@math.auc.dk

Da denne fordelingsoversigt's første udgave så verdens lys i oktober 1998, var det egentlig kun meningen, at den skulle bruges som et hurtigt opslagsværk til vores projektarbejde i statistik. Tingene udviklede sig, og oversigten blev en del mere omfangsrig.

Denne anden udgave foreligger i to versioner, en normal og en udvidet. Den normale dækker alle de fordelinger, der almindeligvis bliver omtalt i et første kursus i sandsynlighedsteori. I den udvidede version er der et ekstra kapitel, hvor der omtales nogle af de mindre almindelige fordelinger.

Meningen med oversigten er stadig, at den skal være et hurtigt opslagsværk til fordelinger. Den går således ikke ret meget i dybden, og man vil heller ikke kunne finde definitioner på de bagvedliggende sandsynlighedsteoretiske begreber.

Fejl og unøjagtigheder bedes rapporteret til uli@math.auc.dk.

Tak til Bo Rosbjerg, Dennis Nilsson, Søren L. Buhl, Henrik V. Christensen, Tina Madsen og Bjarne Pedersen for råd, vejledning og kritik.

For en god ordens skyld skal det nævnes, at $\mathbb{N} = \{1, 2, \dots\}$ betegner mængden af de naturlige tal. $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ betegner mængden af de naturlige tal sammen med tallet 0. $\mathbb{R}$ betegner mængden af reelle tal og $\mathbb{R}^+$ mængden af positive reelle tal.

$[a, b]$ betegner det lukkede interval fra a til b og $]a, b[$ det åbne interval fra a til b.

Ulrich Fahrenberg, 1. august 2002

Indhold

Kapitel 1

Diskrete fordelinger

En stokastisk variabel X kaldes *diskret* fordelt, hvis den kan antage tælleligt mange værdier x_i, hver med en sandsynlighed $p_i = \mathrm{P}(X = x_i)$, så $\sum_i p_i = 1$.

1.1 Den diskrete ligefordeling

En diskret stokastisk variabel, der kan antage værdierne $x_1, x_2, \ldots, x_n$ ($n \in \mathbb{N}$), kaldes ligefordelt, hvis der gælder

$$\mathrm{P}(X = x_k) = \frac{1}{n} \qquad (k = 1, 2, \ldots, n)$$

Middelværdien er

$$\mathbf{E}X = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

Variansen beregnes direkte af formlen $\mathbf{Var}X = \mathbf{E}(X^2) - (\mathbf{E}X)^2$ og kan ikke forenkles:

$$\mathbf{Var}X = \frac{1}{n} \sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n} \sum_{k=1}^{n} x_k \right)^2$$

1.2 Binomialfordelingen

En binomialfordelt stokastisk variabel kan antage heltalsværdier mellem 0 og et fast tal $n \in \mathbb{N}$. Fordelingens anden parameter er en sandsynlighed p $(0 < p < 1)$, og der gælder:

$$\mathrm{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad (x = 0, 1, \ldots, n)$$

At en stokastisk variabel X er binomialfordelt med parametrene n og p skrives også

$$X \sim \mathrm{b}(n, p).$$

En model for en binomialfordelt stokastisk variabel er følgende:

Et forsøg med sandsynligheden p for at en hændelse A indtræffer gentages n gange. Er de enkelte forsøg uafhængige af hinanden, og angiver den stokastiske variabel X antallet af forsøg, der resulterer i hændelsen A, da er $X \sim \mathrm{b}(n, p)$.

I stedet for ovennævnte formel kan man for $n \geq 2$ også benytte en rekursionsformel, der har følgende form:

$$\begin{aligned}
&\mathrm{P}(X = 0) = (1-p)^n \\
&\mathrm{P}(X = x) = \frac{n-x}{x} \frac{p}{1-p} \mathrm{P}(X = x-1) \qquad x = 1, \ldots, n-1
\end{aligned}$$

Binomialfordelingens momentfrembringende funktion er

$$M(t) = (pe^t + 1 - p)^n,$$

og middelværdi og varians har formlerne

$$\mathbf{E}X = np \qquad \mathbf{Var}X = np(1-p).$$

For binomialfordelte stokastiske variable gælder følgende additionssætning:

Er X og Y uafhængige stokastiske variable med $X \sim \mathrm{b}(m, p)$ og $Y \sim \mathrm{b}(n, p)$, da er $X + Y \sim \mathrm{b}(m + n, p)$.

1.3 Multinomialfordelingen

Lad der blive udført et stokastisk eksperiment, der kan have r forskellige udfald $A_1, \ldots, A_r$ med respektive sandsynligheder $p_1, \ldots, p_r$ $(0 < p_i < 1$, $p_1 + \cdots + p_r = 1)$. Lad eksperimentet blive udført et antal n gange, og lad de stokastiske variable X_i $(i = 1, \ldots, r)$ betegne det antal forsøg, der resulterer i udfaldet A_i. Da siges $X_1, \ldots, X_r$ at være multi- eller polynomialfordelte, og der gælder

$$\mathrm{P}\left((X_1, \ldots, X_r) = (x_1, \ldots, x_r)\right) = \frac{n!}{\prod_{i=1}^{r} x_i!} \prod_{i=1}^{r} p_i^{x_i}$$

$$\left(i = 1, 2, \ldots, r,\ \sum_{i=1}^{r} x_i = n\right)$$

Der gælder $X_i \sim \mathrm{b}(n, p_i)$, hvilket giver følgende formler for middelværdierne og varianserne:

$$\mathbf{E}X_i = np_i \qquad \mathbf{Var}X_i = np_i(1 - p_i) \qquad (i = 1, \ldots, r)$$

Desuden gælder der for kovarianserne:

$$\mathbf{Cov}(X_i, X_j) = -np_ip_j \qquad i, j = 1, \ldots, r \quad i \neq j$$

1.4 Den geometriske fordeling

En stokastisk variabel X siges at være geometrisk fordelt med parameter p $(0 < p < 1)$, hvis der for sandsynlighederne gælder følgende formel:

$$\mathrm{P}(X = x) = p(1-p)^{x-1} \qquad (x = 1, 2, \dots)$$

Et eksperiment, hvori en hændelse A har en fast sandsynlighed p for at indtræffe, udføres flere gange efter hinanden, og de enkelte udførelser er uafhængige af hinanden. Den stokastiske variabel, der betegner det antal forsøg, der skal udføres, indtil hændelsen A indtræffer første gang, er da geometrisk fordelt. [1]

Den geometriske fordelings momentfrembringende funktion har følgende udseende:

$$M(t) = \frac{1}{1 + \frac{1}{p}(e^{-t} - 1)},$$

og formlerne for middelværdi og varians er

$$\mathbf{E}X = \frac{1}{p} \qquad \mathbf{Var}X = \frac{1-p}{p^2}.$$

[1] Bemærk dog den afvigende definition, der omtales i afsnit 1.5 på side 9.

1.5 Den negative binomialfordeling

Et eksperiment, hvori en hændelse A har den faste sandsynlighed p for at indtræffe, udføres flere gange efter hinanden. De enkelte forsøg er uafhængige, og den stokastiske variabel X betegner det antal forsøg, der udføres, indtil hændelsen A er indtruffet et antal k gange. Da siges X at have en negativ binomialfordeling.

For $k = 1$ er X dermed geometrisk fordelt, og for vilkårligt $k \in \mathbb{N}$ gælder

$$\mathrm{P}(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k} \qquad (x = k, k+1, \dots)$$

Fordelingens momentfrembringende funktion ser ud som følger:

$$M(t) = \left(1 + \frac{1}{p}(e^{-t} - 1)\right)^{-k},$$

og formlerne for middelværdi og varians er

$$\mathbf{E}X = \frac{k}{p} \qquad \mathbf{Var}X = k\frac{1-p}{p^2}.$$

Den negative binomialfordeling kan ses som et forbindelsesled mellem binomialfordelingen og den geometriske fordeling:

Er den diskrete stokastiske variabel X negativt binomialfordelt, udtrykker hændelsen $X = x$, at der måtte blive udført x forsøg, indtil en hændelse A var indtruffet k gange. Dette er det samme som at hændelsen A blandt de første $x - 1$ forsøg indtraf $k - 1$ gange, og at det x-te forsøg resulterede i hændelsen A. Det førstnævnte er en binomialfordelt hændelse, og af denne grund gælder der

$$\mathrm{P}(X = x) = \binom{x-1}{k-1} p^{k-1} (1-p)^{x-k}\, p = \binom{x-1}{k-1} p^k (1-p)^{x-k}.$$

Sammenhængen til den geometriske fordeling er givet ved, at det at vente på, at hændelsen A indtræffer k gange, er det samme som k gange at vente på, at A indtræffer én gang. Det vil sige, en negativt binomialfordelt stokastisk variabel X med parametrene k og p kan skrives som en sum af k stokastiske variable $X_1, \dots, X_k$, hvor X_1 angiver det antal forsøg, der skulle udføres indtil A indtraf første gang, X_2 angiver det antal forsøg, der herefter skulle udføres indtil A indtraf anden gang og så videre. Variablerne X_i er uafhængige og geometrisk fordelte med parameter p.

Dette giver for eksempel en nem måde at beregne middelværdien på, idet

$$\mathbf{E}X = \mathbf{E}\left(\sum_{i=1}^{k} X_i\right) = \sum_{i=1}^{k} \mathbf{E}X_i = \sum_{i=1}^{k} \frac{1}{p} = \frac{k}{p}$$

Der er i litteraturen delte meninger om definitionen af den geometriske og den negative binomialfordeling. Jeg har holdt mig til definitionerne i [Ross, 1998] og [Grinstead og Snell, 1997], mens [Johnson *et al.*, 1992] giver en anden definition, hvor man ikke tæller *alle* udførelser af forsøget, indtil hændelsen er indtruffet k gange, men kun de forsøg, hvor A *ikke* indtræffer.

Hvis X betegner det antal forsøg, der udføres indtil hændelsen A er indtruffet k gange, og Y antallet af forsøg, der ikke resulterede i A, gælder der således $Y = X - k$, hvilket fører frem til følgende udtryk for denne form af den negative binomialfordeling:

$$\mathrm{P}(Y = y) = \binom{y+k-1}{y} p^k (1-p)^y \qquad (y = 0, 1, \dots)$$

For Y gælder

$$\mathbf{E}Y = \mathbf{E}(X - k) = \mathbf{E}X - k = k\frac{1-p}{p}$$

og

$$\mathbf{Var}Y = \mathbf{Var}(X - k) = \mathbf{Var}X = k\frac{1-p}{p^2}.$$

Definitionen af denne form af den negative binomialfordeling udvides ofte til at omfatte vilkårlige reelle parametre k (i stedet for kun naturlige), idet binomialkoefficienten for $\alpha \in \mathbb{R}$ og $m \in \mathbb{N}$ defineres som

$$\binom{\alpha}{m} = \frac{1}{m!} \prod_{i=0}^{m-1} (\alpha - i).$$

Der gælder følgende sætning om sammenhængen mellem denne form af den negative binomialfordeling og Poisson- og gammafordelingen (se afsnittene 1.6 på den følgende side og 2.4 på side 17):

Hvis X og Λ er stokastiske variable med $(X|\Lambda = \lambda) \sim \mathrm{Poi}(\lambda)$ og $\Lambda \sim \mathrm{Gamma}(\alpha, \beta)$, da er X negativt binomialfordelt med $k = \alpha$ og $p = \frac{\beta}{1+\beta}$.

1.6 Poisson-fordelingen

En diskret stokastisk variabel X kaldes Poisson-fordelt med parameter $\lambda \in \mathbb{R}$, hvis der gælder

$$\mathrm{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad (x = 0, 1, \dots)$$

At en stokastisk variabel X er Poisson-fordelt med parameter λ skrives

$$X \sim \mathrm{Poi}(\lambda).$$

Poisson-fordelingens momentfrembringende funktion har formlen

$$M(t) = e^{\lambda(e^t - 1)},$$

og middelværdi og varians fås til

$$\mathbf{E}X = \lambda \qquad \mathbf{Var}X = \lambda.$$

Poisson-fordelingen bruges, hvis et forsøg med en lille sandsynlighed for, at en hændelse A indtræffer, udføres et stort antal gange, og de enkelte forsøg er uafhængige. X angiver da antallet af forsøg, der resulterer i hændelsen A. Af denne grund kan en binomialfordeling $\mathrm{b}(n, p)$ approksimeres med Poisson-fordelingen $\mathrm{Poi}(np)$ for store n og små p.

Som for binomialfordelingen, så findes der også en additionssætning for Poisson- fordelingen:

Er X og Y uafhængige stokastiske variable med $X \sim \mathrm{Poi}(\lambda_1)$ og $Y \sim \mathrm{Poi}(\lambda_2)$, da er $X + Y \sim \mathrm{Poi}(\lambda_1 + \lambda_2)$.

1.6.1 Poisson-processen

Poisson-fordelingens hyppigste anvendelse er Poisson-processen. Denne opstår, hvis man ser på stokastiske begivenheders fordeling over tid.

Hvis $N(t)$ betegner antallet af begivenheder i tidsrummet $[0, t]$, skal der være opfyldt følgende fem betingelser:

1. $N(0) = 0$
2. $\mathrm{P}(N(h) = 1) = \lambda h + o(h)$ for tilstrækkeligt lille h
3. Hændelser i disjunkte tidsrum er uafhængige
4. $N(t_0+t)-N(t_0) = N(t)$ – hvornår man starter med at tælle, har ingen betydning
5. $\mathrm{P}(N(h) \geq 2) = o(h)$ – i et tilstrækkeligt lille tidsrum h sker der højst én hændelse

(her er $o(h)$ et symbol for en (vilkårlig) funktion $f(h)$ med $\lim_{h\to 0} \frac{f(h)}{h} = 0$).

Er disse 5 aksiomer opfyldt, kan det vises, at $N(t) \sim \mathrm{Poi}(\lambda t)$. Det vil sige, der gælder

$$\mathrm{P}(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \qquad (k = 0, 1, \dots)$$

Der er en interessant sammenhæng mellem Poisson-processen og eksponentialfordelingen (se afsnit 2.2 på side 15):

Hvis den stokastiske variabel T betegner ventetiden fra Poisson-processens start til det første indtræffen af hændelsen (eller fra at hændelsen er indtruffet, til den indtræffer igen[2]), gælder der

$$F_T(t) = \mathrm{P}(T \leq t) = 1 - \mathrm{P}(T > t) = 1 - \mathrm{P}(N(t) = 0) = 1 - e^{-\lambda t}.$$

Ventetiden T er således eksponentialfordelt med parameter λ.

Det omvendte gælder ligeledes: Hvis ventetiden er exponentialfordelt, da er processen en Poisson-proces.

Ovenstående kan udvides på følgende måde: Hvis T_n betegner den tid i en Poisson-proces, der går, indtil hændelsen er indtruffet n gange, da er $T_n \sim \mathrm{Gamma}(n, \lambda)$. [3]

[2] Dette er efter aksiom 4 det samme.

[3] Gammafordelingen se afsnit 2.4 på side 17.

1.7 Den hypergeometriske fordeling

En hypergeometrisk fordelt stokastisk variabel med parametrene N, M og n er en diskret stokastisk variabel, der opfylder formlen

$$\mathrm{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}$$

$$(x = \max\{0, n-(N-M)\}, \ldots, \min\{n, M\})$$

At en stokastisk variabel X er hypergeometrisk fordelt med de ovennævnte parametre skrives som

$$X \sim \mathrm{hyp}(N, M, n),$$

og af ovenstående ses, at der gælder $\mathrm{hyp}(N, M, n) = \mathrm{hyp}(N, n, M)$.

En model for en hypergeometrisk fordelt stokastisk variabel er følgende:

I en urne er der N kugler, hvoraf M er sorte og resten hvide. Af urnen udtages i alt n kugler, og den stokastiske variabel X angiver antallet af sorte blandt de udtagne kugler. Da er $X \sim \mathrm{hyp}(N, M, n)$.

Formlerne for middelværdi og varians kan forenkles ved at indføre $p = \frac{M}{N}$. (Dermed er p sandsynligheden for ved ét forsøg at udtage en sort kugle.)

Middelværdien af $X \sim \mathrm{hyp}(N, M, n)$ er da

$$\mathbf{E}X = n\frac{M}{N} = np,$$

og variansens formel er

$$\mathbf{Var}X = \frac{nM(N-M)(N-n)}{N^2(N-1)} = np(1-p)\left(1 - \frac{n-1}{N-1}\right).$$

Ligheden med binomialfordelingen i de ovenstående formler er ikke tilfældig. For $n \ll N$ kan den hypergeometriske fordeling nemlig approksimeres med binomialfordelingen:

Den stokastiske variabel X er binomialfordelt, hvis kuglerne i det ovenstående eksempel udtages enkeltvis og lægges tilbage, før den næste trækkes. Jo større antallet af kugler er i forhold til antallet af udtagninger, jo mindre betydning har det, om man lægger kuglerne tilbage eller ej.

Kapitel 2

Kontinuerte fordelinger

En sandsynlighedsfordeling kaldes *kontinuert*, hvis den kan beskrives ved en kontinuert *tæthedsfunktion* $f_X : I \mapsto [0,1]$, så der gælder

$$\mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(x)$$

for alle $x \in I$. Her er I et generaliseret reelt interval – det vil sige, I er et interval mellem a og b, hvor $a, b \in \mathbb{R}$ eller $a = \infty$ eller $b = \infty$. Funktionen F_X med $F_X(x) = \mathrm{P}(X \leq x)$ kaldes fordelingens *fordelingsfunktion*.

Et ofte benyttet værktøj i undersøgelsen af kontinuerte stokastiske variable er *standardisering*: Er X en kontinuert stokastisk variabel med middelværdi $\mathbf{E}X$ og varians $\mathbf{Var}X$, da defineres den tilhørende standardiserede stokastiske variabel U ved

$$U = \frac{X - \mathbf{E}X}{\sqrt{\mathbf{Var}X}}.$$

Der gælder

$$F_X(x) = F_U\left(\frac{x - \mathbf{E}X}{\sqrt{\mathbf{Var}X}}\right) \qquad f_X(x) = \frac{1}{\sqrt{\mathbf{Var}X}} f_U\left(\frac{x - \mathbf{E}X}{\sqrt{\mathbf{Var}X}}\right).$$

U er ofte nemmere at håndtere end den oprindelige variabel X, og der gælder $\mathbf{E}U = 0$ og $\mathbf{Var}U = 1$.

2.1 Den kontinuerte ligefordeling

En kontinuert stokastisk variabel X siges at være ligefordelt på intervallet $]a,b[\subseteq \mathbb{R}$, hvis der for dens tæthed gælder

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{ellers.} \end{cases}$$

At en kontinuert stokastisk variabel X er ligefordelt på intervallet $]a,b[$, skrives også på følgende måde:

$$X \sim \mathrm{Unif}(a,b)$$

Fordelingsfunktionen er

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq a \\ \frac{x-a}{b-a} & \text{for } a < x < b \\ 1 & \text{for } x \geq b. \end{cases}$$

Der gælder

$$\mathbf{E}X = \frac{a+b}{2} \qquad \mathbf{Var}X = \frac{(b-a)^2}{12}.$$

En kontinuert *standard*-ligefordeling er en ligefordeling på intervallet $]0,1[$, og for en standard-ligefordelt stokastisk variabel Y gælder:

$$f_Y(y) = \begin{cases} 1 & \text{for } 0 < y < 1 \\ 0 & \text{ellers} \end{cases} \qquad F_Y(y) = \begin{cases} 0 & \text{for } y \leq 0 \\ y & \text{for } 0 < y < 1 \\ 1 & \text{for } y \geq 1. \end{cases}$$

Middelværdi og varians til Y er

$$\mathbf{E}Y = \frac{1}{2} \qquad \mathbf{Var}Y = \frac{1}{12}.$$

Bemærk, at en *standard*-ligefordeling ikke er en *standardiseret* fordeling.

2.2 Eksponentialfordelingen

En kontinuert stokastisk variabel X siges at være eksponentialfordelt med parameter $\lambda \in \mathbb{R}^+$, hvis der for dens tæthed gælder

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

At X er eksponentialfordelt med parameter λ kan forkortes ved

$$X \sim \mathrm{Exp}(\lambda).$$

Fordelingsfunktionen er

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

Fordelingens momentfrembringende funktion er

$$M(t) = \frac{\lambda}{\lambda - t}$$

og formlerne for middelværdi og varians ser således ud:

$$\mathbf{E}X = \frac{1}{\lambda} \qquad \mathbf{Var}X = \frac{1}{\lambda^2}$$

Eksponentialfordelingen "har ikke nogen hukommelse". Det vil sige, der gælder

$$\mathrm{P}(X \leq D + x \mid X > D) = \mathrm{P}(X \leq x) = F_X(x)$$

for alle $x, D \geq 0$.

Billedligt talt betyder dette, at ligegyldigt hvor længe man venter med at gå over vejen, da forbliver sandsynligheden for at blive kørt over den samme (forudsat at trafikkens intensitet forbliver den samme).

2.3 Normalfordelingen

En kontinuert stokastisk variabel X kaldes normalfordelt med parametrene μ og σ^2 ($\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}^+$), hvis dens tæthedsfunktion har formlen

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (x \in \mathbb{R})$$

En sådan stokastisk variabel siges at have fordelingen

$$\mathrm{N}(\mu, \sigma^2)$$

(betegnelsen $\mathrm{N}(\mu, \sigma)$ bruges dog også). Normalfordelingen kaldes også Gauss-fordelingen.

Formlerne for den momentfrembringende funktion, middelværdi og varians er

$$M(t) = e^{\mu t + \frac{\sigma^2}{2} t^2}$$
$$\mathbf{E}X = \mu \qquad \mathbf{Var}X = \sigma^2.$$

2.3.1 Standardnormalfordelingen

Er den kontinuerte stokastiske variabel $X \sim \mathrm{N}(\mu, \sigma^2)$, da er $U = \frac{X-\mu}{\sigma} \sim \mathrm{N}(0, 1)$.

Den standardiserede stokastiske variabel U har middelværdi 0 og varians 1, og for dens tætheds- og fordelingsfunktion gælder

$$f_U(u) = \phi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$$
$$F_U(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{t^2}{2}} dt$$

Funktionerne ϕ og Φ er tabellagte[1], og tætheds- og fordelingsfunktionen til en vilkårlig stokastisk variabel $X \sim \mathrm{N}(\mu, \sigma^2)$ kan således beregnes som

$$f_X(x) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma}\right) \qquad F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

For normalfordelingen gælder der en vigtig additionssætning:

Er $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$ og $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$ med $\mathbf{Cov}(X, Y) = \sigma_{12}$, da er

$$X + Y \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + 2\sigma_{12} + \sigma_2^2)$$

[1] Bemærk, at der gælder $\phi(-u) = \phi(u)$ og $\Phi(-u) = 1 - \Phi(u)$.

2.4 Gammafordelingen

Gammafunktionen er defineret på følgende måde:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt \qquad (\alpha \in \mathbb{R}^+)$$

Der gælder

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

og

$$\Gamma(n) = (n-1)! \qquad (n \in \mathbb{N})$$

samt $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ og $\Gamma\left(\frac{3}{2}\right) = \sqrt{\frac{\pi}{4}}$.

En kontinuert stokastisk variabel X siges at være gammafordelt med parametre $\alpha, \lambda \in \mathbb{R}^+$, hvis dens tæthedsfunktion har formlen

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x^{\alpha-1} e^{-\lambda x} \qquad (x > 0)$$

med $f_X(x) = 0$ for $x \leq 0$.

Man skriver i så fald

$$X \sim \text{Gamma}(\alpha, \lambda),$$

betegnelsen $X \sim \text{Gamma}(\alpha, \frac{1}{\lambda})$ bruges dog også.

α kaldes fordelingens *formparameter*, og λ kaldes fordelingens *skalaparameter*. $\frac{1}{\lambda}$ kaldes fordelingens *intensitet*; for $\lambda = 1$ kaldes fordelingen også en standard-gammafordeling.

Middelværdi og varians til $X \sim \text{Gamma}(\alpha, \lambda)$ har formlerne

$$\mathbf{E}X = \frac{\alpha}{\lambda} \qquad \mathbf{Var}X = \frac{\alpha}{\lambda^2}$$

Der gælder

$$\text{Gamma}(1, \lambda) = \text{Exp}(\lambda),$$

og en gammafordeling $\text{Gamma}(\alpha, \lambda)$, hvor α er et positivt heltal, kaldes også en Erlang-fordeling. Desuden har også gammafordelingen sin additionssætning:

Er X og Y uafhængige stokastiske variable med $X \sim \text{Gamma}(\alpha, \lambda)$ og $Y \sim \text{Gamma}(\beta, \lambda)$, da er $X + Y \sim \text{Gamma}(\alpha + \beta, \lambda)$.

2.5 χ^2-fordelingen

Lad $U_1, \ldots, U_n$ være stokastiske variable med $U_i \sim \mathrm{N}(0,1)$, og lad X betegne deres kvadratsum:

$$X = \sum_{i=1}^{n} U_i^2$$

Da siges X at være χ^2-fordelt med n frihedsgrader.[2]

Man kommer frem til, at

$$\chi^2(n) = \mathrm{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right),$$

hvormed χ^2-fordelingens tæthedsfunktion har følgende udseende:

$$f_X(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x}{2}} x^{\frac{n}{2}-1} \qquad (x > 0)$$

For $x \leq 0$ er $f_X(x) = 0$.

Specielt gælder altså

$$\chi^2(2) = \mathrm{Gamma}\left(1, \frac{1}{2}\right) = \mathrm{Exp}\left(\frac{1}{2}\right).$$

Fordelingens momentfrembringende funktion, middelværdi og varians har følgende formler:

$$M(t) = (1-2t)^{-\frac{n}{2}}$$
$$\mathbf{E}X = n \qquad \mathbf{Var}X = 2n$$

Som specialtilfælde af en Gammafordeling har også χ^2-fordelingen sin additionssætning:

Hvis de to stokastiske variable $X \sim \chi^2(m)$ og $Y \sim \chi^2(n)$ er uafhængige, da er $X + Y \sim \chi^2(m+n)$.

Desuden gælder følgende sætning:

Givet n uafhængige stokastiske variable $X_1, X_2, \ldots, X_n$ med $X_i \sim \mathrm{N}(\mu, \sigma^2)$, lad $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ og $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Da er $\overline{X}$ og S^2 uafhængige, og der gælder $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ samt $S^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$.

[2] Umiddelbart giver udtrykket “n frihedsgrader” således kun mening for $n \in \mathbb{N}$, men der viser sig også at være brug for χ^2-fordelinger med et ikke-naturligt antal frihedsgrader.

2.6 t-fordelingen

Lad U og Z være to uafhængige kontinuerte stokastiske variable, og lad $U \sim \mathrm{N}(0,1)$ og $Z \sim \chi^2(n)$ for $n \in \mathbb{R}^+$. Da siges den stokastiske variabel

$$X = \frac{U}{\sqrt{\frac{Z}{n}}}$$

at være t-fordelt med n frihedsgrader.

Det udledes, at X har følgende tæthedsfunktion:

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{n} B\left(\frac{n}{2}, \frac{1}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \\ &= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\, \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad (x \in \mathbb{R}) \end{aligned}$$

For store n kan t-fordelingen med n frihedsgrader approksimeres med standard-normalfordelingen.

t-fordelingen med 1 frihedsgrad har ingen middelværdi. Det kan imidlertid vises, at integralet

$$\int_{-\infty}^{\infty} |t| f_X(t) dt$$

eksisterer for $n > 1$, og siden f_X er en lige funktion, gælder der

$$\mathbf{E}X = 0 \qquad (n > 1)$$

Ligeledes kan det vises, at

$$\mathbf{Var}X = \frac{n}{n-2} \qquad (n > 2)$$

2.7 F-fordelingen

Lad Y og Z være uafhængige kontinuerte stokastiske variable, og lad $Y \sim \chi^2(m)$ og $Z \sim \chi^2(n)$ for $m, n \in \mathbb{R}^+$. Da siges den stokastiske variabel

$$X = \frac{\frac{Y}{m}}{\frac{Z}{n}} = \frac{nY}{mZ}$$

at være F-fordelt med tællerfrihedsgrad m og nævnerfrihedsgrad n.

X har følgende tæthedsfunktion:

$$\begin{aligned} f_X(x) &= \frac{\Gamma\left(\frac{m+n}{2}\right) m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} \\ &= \frac{m^{\frac{m}{2}} n^{\frac{n}{2}}}{B\left(\frac{m}{2}, \frac{n}{2}\right)} \frac{x^{\frac{m}{2}-1}}{(mx+n)^{\frac{m+n}{2}}} \qquad (x > 0) \end{aligned}$$

For $x \leq 0$ er $f_X(x) = 0$.

Middelværdi og varians har følgende formler:

$$\begin{aligned} \mathbf{E}X &= \frac{n}{n-2} \qquad (n > 2) \\ \mathbf{Var}X &= \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad (n > 4) \end{aligned}$$

Det udledes umiddelbart af definitionen, at hvis X er F-fordelt med frihedsgrader m og n, da er $\frac{1}{X}$ F-fordelt med frihedsgrader n og m.

Desuden gælder der, at hvis X er t-fordelt med n frihedsgrader, da er X^2 F-fordelt med 1 tæller- og n nævnerfrihedsgrader.

2.8 Ikke-centrale χ^2-, t- og F-fordelinger

2.8.1 Den ikke-centrale χ^2-fordeling

Normalfordelte variable, der ikke er standardnormalfordelte, har i almindelighed ikke en kvadratsum, der har central χ^2-fordeling. Af denne grund defineres den ikke-centrale χ^2-fordeling med n frihedsgrader og parameter λ som

$$\chi^2(n, \lambda) = \sum_{i=1}^{n} X_i^2,$$

hvor $X_i \sim \mathrm{N}(\mu_i, 1)$ $(i = 1, \ldots, n)$ og $\lambda = \sum_{i=1}^{n} \mu_i^2$.

Som det implicit allerede er indeholdt i ovenstående formel, er fordelingen $\chi^2(n, \lambda)$ ikke afhængig af de enkelte middelværdier μ_i, men kun af deres kvadratsum λ.

Fordelingens momentfrembringende funktion samt middelværdi og varians har følgende udseende:

$$M(t) = (1 - 2t)^{-\frac{n}{2}} e^{\frac{\lambda t}{1-2t}}$$
$$\mathbf{E}X = n + \lambda \qquad \mathbf{Var}X = 2n + 4\lambda$$

Endvidere gælder der, at hvis $Y_1, \ldots, Y_n$ er uafhængige stokastiske variable med $Y_i \sim \mathrm{N}(\mu_i, \sigma^2)$, og S^2 defineres som

$$S^2 = \tfrac{1}{n-1} \textstyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2 \quad (\text{med } \overline{Y} = \tfrac{1}{n} \textstyle\sum_{i=1}^{n} Y_i),$$

da er

$$S^2 \sim \frac{\sigma^2}{n-1} \chi^2(n-1, \lambda)$$

med $\lambda = \frac{1}{\sigma^2} \sum_{i=1}^{n} (\mu_i - \overline{\mu})^2$ (hvor $\overline{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i$).

2.8.2 Den ikke-centrale t-fordeling

Hvor en stokastisk variabel Y siges at have *central* t-fordeling (med n frihedsgrader), hvis den kan skrives som $Y = U(Z/n)^{-1/2}$, hvor $U \sim \mathrm{N}(0, 1)$ og $Z \sim \chi^2(n)$, da siges en stokastisk variabel X at have *ikke-central* t-fordeling med n frihedsgrader, hvis

$$X = \frac{V}{\sqrt{\frac{Z}{n}}},$$

hvor $V \sim \mathrm{N}(\mu, 1)$ og $Z \sim \chi^2(n)$. Den ikke-centrale t-fordeling har dermed en ekstra parameter μ, der er middelværdi til normalfordelingen i nævneren.

Middelværdi og varians til den stokastiske variabel X har formlerne

$$\mathbf{E}X = \sqrt{\frac{n}{2}}\,\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\mu \qquad (n > 1)$$

$$\mathbf{Var}X = \frac{n}{n-2}(1+\mu^2) - (\mathbf{E}X)^2 \qquad (n > 2),$$

der for store n kan approksimeres ved

$$\mathbf{E}X \approx \mu \qquad \mathbf{Var}X \approx 1 + \frac{\mu^2}{2n}$$

2.8.3 Den ikke-centrale F-fordeling

Den centrale F-fordeling er defineret som kvotient af to uafhængige χ^2-fordelinger, hver divideret med antallet af dens frihedsgrader. Erstattes de centrale χ^2-fordelinger med ikke-centrale χ^2-fordelinger, da kaldes fordelingen en *dobbelt ikke-central* F-fordeling.

Således siges en stokastisk variabel X at have dobbelt ikke-central F-fordeling med m og n frihedsgrader og parametrene λ og ξ, hvis den kan skrives som

$$X = \frac{Yn}{Zm},$$

hvor $Y \sim \chi^2(m, \lambda)$ og $Z \sim \chi^2(n, \xi)$.

Er $\xi = 0$, er altså χ^2-fordelingen i nævneren central, da kaldes fordelingen *enkelt ikke-central*, og i dette tilfælde beregnes middelværdi og varians til

$$\mathbf{E}X = \frac{n}{n-2}\,\frac{m+\lambda}{m} \qquad (n > 2)$$

$$\mathbf{Var}X = 2\left(\frac{n}{m}\right)^2 \frac{(m+\lambda)^2 + (m+2\lambda)(n-2)}{(n-2)^2(n-4)} \qquad (n > 4)$$

2.9 Betafordelingen

Betafunktionen $B(a,b)$ defineres på følgende måde:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt \qquad (a, b \in \mathbb{R}^+)$$

Man kommer rimeligt hurtigt frem til, at der gælder $B(a,b) = B(b,a)$, og desuden er $B(a,1) = \frac{1}{a}$ samt

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

En kontinuert stokastisk variabel X siges at være betafordelt med parametre $a, b \in \mathbb{R}^+$, hvis dens tæthedsfunktion er

$$f_X(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad (0 < x < 1)$$

I så fald skriver man også

$$X \sim \text{Beta}(a,b).$$

Middelværdien er

$$\mathbf{E}X = \frac{a}{a+b},$$

og variansen har formlen

$$\mathbf{Var}X = \frac{ab}{(a+b)^2(a+b+1)}.$$

Desuden gælder der, at $\text{Beta}(1,1) = \text{Unif}(0,1)$.

2.10 Den logaritmiske normalfordeling

Hvis der til en stokastisk variabel X findes et tal $\theta \in \mathbb{R}$, så $\log(X-\theta)$ er normalfordelt, da siges X at have logaritmisk normalfordeling. Den almindelige definition er, at X er logaritmisk normalfordelt med parametrene μ, σ og θ, hvis den stokastiske variabel

$$U = \frac{\log(X-\theta)-\mu}{\sigma}$$

er standardnormalfordelt.

Det udledes, at X har følgende tæthedsfunktion:

$$f_X(x) = \frac{1}{(x-\theta)\sqrt{2\pi}\sigma} e^{-\frac{(\log(x-\theta)-\mu)^2}{2\sigma^2}} \qquad (x > \theta)$$

For $x \leq \theta$ er $f_X(x) = 0$.

For middelværdi og varians gælder

$$\mathbf{E}X = \theta + e^{\mu + \frac{1}{2}\sigma^2} \qquad \mathbf{Var}X = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1).$$

Af normalfordelingens additionssætning udledes, at produktet af to logaritmisk normalfordelte stokastiske variable med parametrene $(\mu_1, \sigma_1, 0)$ og $(\mu_2, \sigma_2, 0)$ $(\theta_1 = \theta_2 = 0)$ igen er logaritmisk normalfordelt, med parametrene $(\mu_1 + \mu_2, \sigma_1 + \sigma_2, 0)$.

Kapitel 3

Vigtige sætninger

- *De store tals svage lov*:

  $X_1, X_2, \ldots$ uafhængige og ensfordelte med middelværdi $\mathbf{E}X_i = \mu$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$:

  $$\underset{\epsilon\in\mathbb{R}^+}{\forall} \lim_{n\to\infty} \mathrm{P}(|\overline{X}_n - \mu| < \epsilon) = 1$$

- *De store tals stærke lov*:

  $X_1, X_2, \ldots$ uafhængige og ensfordelte med middelværdi $\mathbf{E}X_i = \mu$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$:

  $$\mathrm{P}(\lim_{n\to\infty} \overline{X}_n = \mu) = 1$$

- *Den centrale grænseværdisætning*:

  $X_1, X_2, \ldots$ uafhængige og ensfordelte med middelværdi $\mathbf{E}X_i = \mu$ og varians $\mathbf{Var}X_i = \sigma^2$:

  $$\underset{a\in\mathbb{R}}{\forall} \lim_{n\to\infty} \mathrm{P}\left(\frac{1}{\sigma\sqrt{n}}\left[\sum_{i=1}^{n} x_i - n\mu\right] \leq a\right) = \Phi(a)$$

Kapitel 4

Flere fordelinger

Dette kapitel omhandler sandsynlighedsfordelinger, der er mindre benyttede end de i kapitlerne 1 og 2 opførte. Kapitlet omfatter både diskrete og kontinuerte fordelinger.

4.1 Den logaritmiske fordeling

En diskret stokastisk variabel X, der kan antage positive heltalsværdier, siges at have logaritmisk fordeling[1] med en parameter $\theta \in \mathbb{R}$, $0 < \theta < 1$, hvis der gælder

$$\mathrm{P}(X = x) = a_\theta \frac{\theta^x}{x} \qquad (x = 1, 2, \dots)$$

med $a_\theta = -\frac{1}{\log(1-\theta)}$.

Dens momentfrembringende funktion er da

$$M(t) = -a_\theta \log(1 - \theta e^t),$$

og middelværdi og varians fås ved formlerne

$$\mathbf{E}X = \frac{a_\theta \theta}{1-\theta} \qquad \mathbf{Var}X = \frac{a_\theta \theta(1 - a_\theta \theta)}{(1-\theta)^2}.$$

4.2 Zeta- eller Zipf-fordelingen

En diskret stokastisk variabel X, der kan antage positive heltalsværdier, siges at være zeta- eller Zipf-fordelt med parameter $\alpha \in \mathbb{R}^+$, hvis der gælder

$$\mathrm{P}(X = x) = \frac{C_{\alpha+1}}{x^{\alpha+1}} \qquad (x = 1, 2, \dots)$$

[1]Denne fordeling må ikke forveksles med den logaritmiske *normal*fordeling (se afsnit 2.10)

$C_{\alpha+1}$ er da en konstant med formlen

$$C_{\alpha+1} = \frac{1}{\sum_{k=1}^{\infty} \frac{1}{k^{\alpha+1}}}.$$

Fordelingen kaldes zeta-fordelingen, da funktionen

$$\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$$

for $s \in \mathbb{R}$, $s > 1$, er kendt som Riemanns zeta-funktion.

Fordelingen anvendes i linguistikken, da man har fundet frem til, at antallet n_x af ord, der i en længere tekst forekommer x gange, nogenlunde kan udtrykkes ved formlen $n_x = cx^{-(\alpha+1)}$, med c som proportionalitetsfaktor og $\alpha > 0$. Antallet af ord, der forekommer x gange, er dermed zeta-fordelt.

Zeta-fordelingens middelværdi er

$$\mathbf{E}X = \frac{C_{\alpha+1}}{C_\alpha} \qquad (\alpha > 1)$$

og variansen har formlen

$$\mathbf{Var}X = \frac{C_{\alpha+1}}{C_{\alpha-1}} - \frac{C_{\alpha+1}^2}{C_\alpha^2} \qquad (\alpha > 2)$$

Zeta-fordelingen kaldes også den diskrete Pareto-fordeling (se afsnit 4.5 på side 30) på grund af ligheden i tæthederne.

4.3 Naor's fordeling

Naor's fordeling fremkommer ved eksperimenter, der kan beskrives ved følgende urnemodel:

En urne indeholder n kugler, deraf 1 rød og $n-1$ hvide. En kugle bliver taget ud. Er denne hvid, erstattes den i urnen med en rød kugle, hvorefter der trækkes igen. Eksperimentet slutter, når der bliver trukket en rød kugle. Den stokastiske variabel Y betegner da antallet af kugler, der i alt blev taget ud af urnen.

Der gælder

$$\mathrm{P}(Y=y)=\frac{(n-1)!y}{(n-y)!n^y} \qquad (y=1,2,\dots)$$

I praksis har det vist sig at være nemmere at arbejde med en stokastisk variabel $X=n-Y$. X angiver da antallet af kugler, der er tilbage i urnen ved eksperimentets afslutning, og der gælder

$$\mathrm{P}(X=x)=\frac{(n-1)!(n-x)}{x!n^{n-x}} \qquad (x=0,1,\dots,n-1)$$

Middelværdi og varians kan for store n approksimeres ved

$$\mathbf{E}X \approx n+\frac{1}{3}-\sqrt{\frac{n\pi}{2}} \qquad \mathbf{E}Y \approx \sqrt{\frac{n\pi}{2}}-\frac{1}{3}$$

$$\mathbf{Var}X=\mathbf{Var}Y \approx n^2+\frac{8}{3}n+\frac{1}{3}-(2n+1)\sqrt{\frac{n\pi}{2}}$$

En anden anvendelse af Naor's fordeling er følgende:

Man har et nøglebundt med et antal nøgler på samt en lås, som der er en og kun en af nøglerne på bundtet, der passer til. Man vælger en nøgle tilfældigt og afprøver denne, hvis den ikke passer, tager man en anden, og så videre. Den mest efficiente måde at afprøve nøglerne på (den måde, man hurtigst finder frem til den rigtige nøgle på) er uden tilbagelægning, det vil sige, man husker, hvilke nøgler man allerede har afprøvet. Hvis man afprøver nøglerne med tilbagelægning, vil der gå længere tid, indtil man finder den rette nøgle, da man risikerer at afprøve nogle af nøglerne to gange. Antallet af forsøg, der udføres, indtil den første nøgle bliver afprøvet to gange, er da Naor-fordelt. På denne måde giver Naor's fordeling et mål for, hvornår tilbagelægning begynder at få indflydelse på antallet af nødvendige afprøvninger.

4.4 Den inverse Gauss-fordeling

En kontinuert stokastisk variabel X siges at have invers Gauss-fordeling med parametrene μ og λ, hvis dens tæthedsfunktion er givet ved

$$f_X(x) = \sqrt{\frac{\lambda}{2\pi x^3}} e^{-\frac{\lambda}{2\mu^2 x}(x-\mu)^2} \qquad (x > 0)$$

For $x \leq 0$ er $f_X(x) = 0$.

Fordelingsfunktionen beregnes til

$$F_X(x) = \Phi\left(\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} - 1\right)\right) + e^{\frac{2\lambda}{\mu}} \Phi\left(-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu} + 1\right)\right),$$

med Φ som standardnormalfordelingens fordelingsfunktion.

Fordelingens momentfrembringende funktion har formlen

$$M(t) = e^{\frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2 t}{\lambda}}\right)},$$

og middelværdi og varians fås til

$$\mathbf{E}X = \mu \qquad \mathbf{Var}X = \frac{\mu^3}{\lambda}.$$

Den inverse Gauss-fordeling med $\mu = 1$ kaldes også Wald-fordeling, og en model for denne er følgende:

Lad $Z_1, Z_2, \ldots$ være uafhængige stokastiske variable, der er ensfordelte med middelværdi $\mathbf{E}Z_i > 0$ og varians $\mathbf{Var}Z_i > 0$. For et fast tal $k > 0$, lad da den stokastiske variabel N være defineret ved, at følgende skal være opfyldt: $\sum_{i=1}^{j} Z_i < k$ for alle $j = 1, \ldots, N-1$ samt $\sum_{i=1}^{N} Z_i \geq k$. N er da Wald-fordelt med $\lambda = k\frac{\mathbf{E}Z_i}{\mathbf{Var}Z_i}$.

4.5 Pareto-fordelingen

Pareto-fordelingen udspringer af Paretos lov, der konstruerer en sammenhæng mellem antallet af personer med en vis indkomst og denne indkomst. Hvis N er antallet af personer med en indkomst, der er højere end eller lig med en givet indkomst x, gælder der efter Paretos lov

$$N = Ax^{-a},$$

hvor $A, a > 0$ er parametre.

En kontinuert stokastisk variabel X siges at være Pareto-fordelt med parametre $a, k > 0$, hvis dens fordelingsfunktion er

$$F_X(x) = \begin{cases} 1 - \left(\frac{k}{x}\right)^a & \text{for } x \geq k \\ 0 & \text{for } x < k \end{cases}$$

Tæthedsfunktionens formel er da

$$f_X(x) = \begin{cases} ak^a x^{-(a+1)} & \text{for } x \geq k \\ 0 & \text{for } x < k \end{cases}$$

Fordelingens middelværdi og varians eksisterer kun for henholdsvis $a > 1$ og $a > 2$ og har formlerne

$$\mathbf{E}X = k\frac{a}{a-1} \qquad (a > 1)$$
$$\mathbf{Var}X = k^2\frac{a}{(a-1)^2(a-2)} \qquad (a > 2)$$

4.6 Den logistiske fordeling

En kontinuert stokastisk variabel X siges at være logistisk fordelt med parametrene $\alpha \in \mathbb{R}$ og $\beta \in \mathbb{R}^+$, hvis der for dens fordelingsfunktion gælder

$$F_X(x) = \frac{1}{1 + e^{-\frac{x-\alpha}{\beta}}} \qquad (x \in \mathbb{R})$$

Tæthedsfunktionen har da formlen

$$f_X(x) = \frac{1}{\beta}\frac{e^{-\frac{x-\alpha}{\beta}}}{\left(1 + e^{-\frac{x-\alpha}{\beta}}\right)^2} \qquad (x \in \mathbb{R})$$

For middelværdi og varians gælder der

$$\mathbf{E}X = \alpha \qquad \mathbf{Var}X = \frac{\pi^2}{3}\beta^2.$$

Er X logistisk fordelt med parametrene α og β, da kan det skrives som

$$X = \alpha + \beta \log\left(\frac{U}{1-U}\right),$$

hvor $U \sim \mathrm{Unif}(0,1)$.

4.7 Ekstremværdi-fordelingen

En kontinuert stokastisk variabel X siges at være ekstremværdi-fordelt med parametrene $\xi \in \mathbb{R}$ og $\theta \in \mathbb{R}^+$, hvis dens fordelingsfunktion er

$$F_X(x) = e^{-e^{-\frac{x-\xi}{\theta}}} \qquad (x \in \mathbb{R})$$

På grund af denne formel kaldes fordelingen også dobbelt-eksponential-fordelingen, hvilket er lidt uheldigt, da dette navn ligeledes bruges til Laplace-fordelingen (se afsnit 4.9 på næste side). Et tredje navn til fordelingen er log-Weibull, da $\log X$ er Weibull-fordelt (se afsnit 4.8 på den følgende side; parametrene er $v = \xi$, $a = \theta$ og $b = 1$).

Fordelingen kaldes ekstremværdi-fordelingen, da den opstår som grænsefordeling for $n \to \infty$, når det er fordelingen af den største værdi blandt n uafhængige stokastiske variable med samme individuelle fordeling, der efterspørges.

Fordelingens tæthedsfunktion har formlen

$$f_X(x) = \frac{1}{\theta} e^{-\frac{x-\xi}{\theta}} e^{-e^{-\frac{x-\xi}{\theta}}} \qquad (x \in \mathbb{R})$$

Dens momentfrembringende funktion er

$$M(t) = e^{t\xi}\Gamma(1-\theta t),$$

og middelværdi og varians fås til

$$\mathbf{E}X = \xi + \gamma\theta \qquad \mathbf{Var}X = \frac{\pi^2}{6}\theta^2,$$

hvor γ er Euler's konstant:

$$\gamma = \lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k} - \log n\right) \approx 0.5772156649$$

4.8 Weibull-fordelingen

En kontinuert stokastisk variabel X siges at være Weibull-fordelt med parametrene $v \in \mathbb{R}$ og $a, b \in \mathbb{R}^+$, hvis dens tæthedsfunktion er givet ved

$$f_X(x) = \begin{cases} \frac{b}{a}\left(\frac{x-v}{a}\right)^{b-1} e^{-\left(\frac{x-v}{a}\right)^b} & \text{for } x > v \\ 0 & \text{for } x \leq v \end{cases}$$

Er X Weibull-fordelt med parametre v, a, b, da gælder der

$$\left(\frac{X-v}{a}\right)^b \sim \text{Exp}(1).$$

Fordelingsfunktionen til Weibull-fordelingen er

$$F_X(x) = \begin{cases} 1 - e^{-\left(\frac{x-v}{a}\right)^b} & \text{for } x > v \\ 0 & \text{for } x \leq v \end{cases}$$

4.8.1 Standard-Weibullfordelingen

Weibull-fordelingens standardform fås med $v = 0$ og $a = 1$. Dens fordelingsfunktion er

$$f_X(x) = \begin{cases} bx^{b-1}e^{-x^b} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

og middelværdi og varians har formlerne

$$\mathbf{E}X = \Gamma\left(1 + \frac{1}{b}\right) \qquad \mathbf{Var}X = \Gamma\left(1 + \frac{2}{b}\right) - \left(\Gamma\left(1 + \frac{1}{b}\right)\right)^2.$$

For store b kan dette approksimeres ved

$$\mathbf{E}X \approx 1 - \frac{\gamma}{b} + \frac{1}{2b^2}\left(\frac{\pi^2}{6} + \gamma^2\right) \qquad \mathbf{Var}X \approx \frac{\pi^2}{6b^2},$$

hvor γ er Euler's konstant.

4.9 Laplace-fordelingen

En kontinuert stokastisk variabel X siges at være Laplace-fordelt med parametrene $\theta \in \mathbb{R}$ og $\phi \in \mathbb{R}^+$, hvis dens tæthedsfunktion har formlen

$$f_X(x) = \frac{1}{2\phi} e^{-\frac{|x-\theta|}{\phi}} \qquad (x \in \mathbb{R})$$

Sættes $\theta = 0$ og $\phi = 1$, fås en standardform med tæthedsfunktionen

$$f_Y(y) = \frac{1}{2}e^{-|x|} \qquad (x \in \mathbb{R})$$

Dennes momentfrembringende funktion er

$$M_Y(t) = \frac{1}{1-t^2},$$

og middelværdi og varians fås til

$$\mathbf{E}Y = 0 \qquad \mathbf{Var}Y = 2.$$

Dermed gælder der for den oprindelige variabel X

$$\mathbf{E}X = \theta \qquad \mathbf{Var}X = 2\phi^2.$$

Der gælder, at hvis $V_1, V_2 \sim \mathrm{Exp}(\phi)$ er uafhængige, da er $V_1 - V_2$ Laplace-fordelt med parameter ϕ (og $\theta = 0$).

En sammenhæng mellem Laplace- og standardnormalfordelingen er følgende:

Hvis $U_1, U_2, U_3, U_4 \sim \mathrm{N}(0,1)$ er uafhængige, da er

$$\begin{vmatrix} U_1 & U_2 \\ U_3 & U_4 \end{vmatrix} = U_1U_4 - U_2U_3$$

Laplace-fordelt med parametrene $\theta = 0$ og $\phi = 2$.

4.10 χ-fordelingen

Hvis en kontinuert stokastisk variabel Y er (centralt) χ^2-fordelt med n frihedsgrader (se afsnit 2.5), da siges den stokastiske variabel $X = \sqrt{Y}$ at være χ-fordelt med n frihedsgrader.

Tæthedsfunktionen har følgende formel:

$$f_X(x) = \frac{1}{2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}\right)} e^{-\frac{x^2}{2}} x^{n-1} \qquad (x > 0)$$

For $x \leq 0$ er $f_X(x) = 0$.

For $n = 1$ kaldes denne fordeling også halvnormalfordelingen, for $n = 2$ kaldes den Rayleigh-fordelingen og for $n = 3$ Maxwell-Boltzmann-fordelingen.

Fordelingsfunktionen har følgende udseende:

$$F_X(x) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{x^2}{2}} e^{-t} t^{\frac{n}{2}-1} dt \qquad (x > 0)$$

(for $x \leq 0$ er igen $F_X(x) = 0$).

For middelværdi og varians gælder der

$$\mathbf{E}X = \sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \frac{\sqrt{\pi}}{B\left(n, \frac{1}{2}\right)}$$

$$\mathbf{Var}X = n - 2\left(\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\right)^2 = n - \frac{2\pi}{\left(B\left(n, \frac{1}{2}\right)\right)^2}$$

Da $\frac{\Gamma\left(z+\frac{1}{2}\right)}{\Gamma(z)} \approx \sqrt{z}$ for store z, kan $\mathbf{E}X$ approksimeres med $\sqrt{n}$ for store n.

4.10.1 Rayleigh-fordelingen

Rayleigh-fordelingen er som sagt det samme som χ-fordelingen med 2 frihedsgrader, og dens tætheds- samt fordelingsfunktion har dermed formlerne

$$f_X(x) = xe^{-\frac{x^2}{2}} \qquad F_X(x) = 1 - e^{-\frac{x^2}{2}} \qquad (x > 0)$$

For $x \leq 0$ er $f_X(x) = F_X(x) = 0$.

Middelværdi og varians fås til følgende:

$$\mathbf{E}X = \sqrt{\frac{\pi}{2}} \qquad \mathbf{Var}X = 2 - \frac{\pi}{2}$$

Hvis X er Rayleigh-fordelt, da er $X^2 \sim \text{Exp}\left(\frac{1}{2}\right)$, og desuden gælder der følgende sætning:

Lad X, Y være uafhængige standardnormalfordelte stokastiske variable, der betragtes som koordinater i et retvinklet koordinatsystem. Da fås ved at gå over til polære koordinater to nye stokastiske koordinater, radius $R \in \mathbb{R}$ ($R \geq 0$) og vinklen $\Theta \in [0, 2\pi[$. Der gælder $X = R\cos\Theta$ og $Y = R\sin\Theta$, og R og Θ er uafhængige med $\Theta \sim \text{Unif}(0, 2\pi)$ og $R \sim$ Rayleigh.

4.11 Cauchy-fordelingen

En kontinuert stokastisk variabel X siges at være Cauchy-fordelt med parameter $\theta \in \mathbb{R}$, hvis dens tæthed er givet ved

$$f_X(x) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} \qquad (x \in \mathbb{R})$$

Fordelingsfunktionen er da

$$F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x - \theta) \qquad (x \in \mathbb{R}$$

Det specielle ved Cauchy-fordelingen er, at den ikke har nogen middelværdi. For at have en sådan, kræves det, at integralet

$$\int_{-\infty}^{\infty} |t| f_X(t) dt$$

eksisterer, hvilket for Cauchy-fordelingen ikke er tilfældet.

En momentfrembringende funktion eksisterer af samme årsager heller ikke, men man kan dog finde en formel for fordelingens *karakteristiske funktion* $\varphi(t)$:

$$\varphi(t) = \mathbf{E}\left(e^{itX}\right) = e^{it\theta - |t|}$$

Cauchy-fordelingen med $\theta = 0$ er det samme som t-fordelingen med 1 frihedsgrad.

Litteratur

[Grinstead og Snell, 1997] Charles M. Grinstead og J. Laurie Snell. *Introduction to Probability.* American Mathematical Society, 1997. ISBN 0-8218-0749-8 – se også http://dartmouth.edu/∼chance/JLSnell.html.

[Johnson *et al.*, 1992] Norman L. Johnson, Samuel Kotz, og Adrienne W. Kemp. *Univariate Discrete Distributions.* Wiley Interscience, 2. edition, 1992. ISBN 0-471-58494-0.

[Johnson *et al.*, 1994] Norman L. Johnson, Samuel Kotz, og N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley Interscience, 2. edition, 1994. ISBN 0-471-58495-9.

[Johnson *et al.*, 1995] Norman L. Johnson, Samuel Kotz, og N. Balakrishnan. *Continuous Univariate Distributions*, volume 2. Wiley Interscience, 2. edition, 1995. ISBN 0-471-58494-0.

[Ross, 1998] S. Ross. *A First Course in Probability.* Prentice Hall International, 1998. ISBN 0-13-896523-4.