Kontinuerlige sannsynlighetsfordelinger.

Dekkes av kap. 6 og deler av kap. 8.5 i boka.

Husk: $f(x)$ er sannsynlighetstettheten til en kontinuerlig X dersom:

1. $f(x) \geq 0$ for alle $x \in \mathcal{R}$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$
3. $P(a < X < b) = \int_a^b f(x)dx$

Vi skal nå se på noen av de mest brukte kontinuerlige sannsynlighetsfordelingene. (For hver slik fordeling er $f(x)$ gitt ved en bestemt parametrisk funksjon).

Kontinuerlig uniformfordeling (6.1)

Dersom X kun kan anta verdier i et intervall $[a, b]$, og det er like sannsynlig at X faller i et hvilket som helst område av samme utstrekning innen dette intervallet er X uniformfordelt (f.eks. hvor på en wire et brudd oppstår). Da er

$$f(x) = \frac{1}{b-a}\ , \quad a \leq x \leq b.$$

og det kan vises at $\mathrm{E}(X) = \frac{(a+b)}{2}$ og $\mathrm{Var}(X) = \frac{(b-a)^2}{12}$.

Kontinuerlig uniformfordeling brukes mye i forbindelse med simulering. De fleste randomgeneratorer genererer uniformfordelte tall mellom 0 og 1. Ved hjelp av transformasjoner o.l. kan man ut fra dette få generert tall fra andre sannsynlighetsfordelinger.

Normalfordeling (6.2-6.4)

Normalfordelingen er den viktigste og mest brukte av alle sannsynlighetsfordelinger. En lang rekke fenomen i natur, teknologi, økomoni, etc kan modelleres ved en normalfordeling. Normalfordeling er også svært sentral i statistisk teori pga at mange fordelinger under visse betingelser kan tilnærmes til normalfordeling.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right), \quad -\infty < x < \infty,$$

$\mathrm{E}(X) = \mu$ og $\mathrm{Var}(X) = \sigma^2$. Vi skriver at $X \sim N(\mu, \sigma^2)$.

Siden $f(x)$ er komplisert å integrere, regner vi ut normalfordelingssannsynligheter ved å gjøre transformasjonen

$$Z = \frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} = \frac{X - \mu}{\sigma}$$

der Z er *standard normalfordelt* (sjekk selv at $E(Z) = 0$ og $\mathrm{Var}(Z) = 1$) og så bruke tabell

over standard normalfordelingen. F.eks.

$$P(X < a) = P(\frac{X-\mu}{\sigma} < \frac{a-\mu}{\sigma}) = P(Z < \frac{a-\mu}{\sigma})$$

der $P(Z < \frac{a-\mu}{\sigma})$ finnes fra tabellen. (Går ikke nærmere inn på dette da det antas kjent fra før.)

Sentralgrenseteoremet (8.5)

Summer/gjennomsnitt av **mange** uavh. stok. var. (fra en hvilken som helst fordeling!) vil være tilnærmet normalfordelt!

La $X_1, \ldots, X_n$ være uavh. variable fra en fordeling med forventning μ og varians σ^2. Vi har da (sjekk selv!) at $\mathrm{E}(\bar{X}) = \mu$ og $\mathrm{Var}(\bar{X}) = \sigma^2/n$.

Sentralgrenseteoremet (SGT) sier nå at

$$Z = \frac{\bar{X} - \mathrm{E}(\bar{X})}{\sqrt{\mathrm{Var}(\bar{X})}} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \approx N(0,1)$$

Generelt OK tilnærming når $n \geq 30$.

Vi ser at dette også gir at

$$Z = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} = \frac{\sum_{i=1}^{n} X_i - \mathrm{E}(\sum_{i=1}^{n} X_i)}{\sqrt{\mathrm{Var}(\sum_{i=1}^{n} X_i)}} \approx N(0,1)$$

Dvs, generelt

$$Z = \frac{V - \mathrm{E}(V)}{\sqrt{\mathrm{Var}(V)}} \approx N(0,1)$$

der $V = \bar{X}$ eller $V = \sum_{i=1}^{n} X_i$.

Eksempel: Du kaster en terning 50 ganger. Hva er sannsynligheten for at summen av alle terningkastene blir større enn 200?

La X_i være resultatet av det $i$te terningkast og $Y = \sum_{i=1}^{50} X_i$. Vi vet at (se diskret uniform fordeling) $\mathrm{E}(X_i) = 3.5$ og $\mathrm{Var}(X_i) = 2.92$.

$$\begin{aligned} P(Y > 200) &= 1 - P(Y \leq 200) \\ &= 1 - P(\frac{Y - \mathrm{E}(Y)}{\sqrt{\mathrm{Var}(Y)}} \leq \frac{200 - \mathrm{E}(Y)}{\sqrt{\mathrm{Var}(Y)}}) \\ &= 1 - P(Z \leq \frac{200 - 50 \cdot 3.5}{\sqrt{50 \cdot 2.92}}) \\ &= 1 - P(Z \leq 2.07) \\ &\overset{SGT}{\approx} 1 - 0.981 = \underline{\underline{0.019}} \quad \square \end{aligned}$$

Binomisk fordeling tilnærmet med normalfordeling (6.5)

X = "antall suksesser i n forsøk" er en sum (la "suksess"=1 og "ikke suksess"=0) og kan dermed i følge SGT tilnærmes med normalfordeling. Når $X \sim B(n,p)$ husker vi at $\mathrm{E}(X) = np$ og $\mathrm{Var}(X) = np(1-p)$, og vi får dermed at

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1)$$

OK tilnærming når $np > 5$ og $n(1-p) > 5$. Se figurene 6.22-6.25 i boka.

$$\begin{aligned} &P(a \leq X \leq b) \\ &\approx P(\frac{a - 0.5 - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}) \\ &= P(Z \leq \frac{b + 0.5 - np}{\sqrt{np(1-p)}}) - P(Z \leq \frac{a - 0.5 - np}{\sqrt{np(1-p)}}) \end{aligned}$$

Her er -0.5 i nedre grense og +0.5 i øvre grense korreksjoner vi gjør for å få en bedre tilnærming (se fig. 6.23).

Poissonfordeling tilnærmet med normalfordeling

Også Poisson-fordelte variable kan tilnærmes med normalfordeling. Dersom $\lambda t > 15$ er

$$\begin{aligned} & P(a \leq X \leq b) \\ & \approx \quad P(\frac{a - 0.5 - \lambda t}{\sqrt{\lambda t}} \leq Z \leq \frac{b + 0.5 - \lambda t}{\sqrt{\lambda t}}) \end{aligned}$$

en god tilnærming som kan brukes til å forenkle beregninger. (Husk at $\mathrm{E}(X) = \mathrm{Var}(X) = \mu = \lambda t$ i Poisson-fordeling.)

Eksempel: Antall visninger av en webside per dag er Poisson-fordelt med forventing 92. Hva er sannsynligheten for at siden har mer enn 100 visninger i løpet av en dag? Hva er sannsynligheten for at siden har mer enn 700 visninger i løpet av en uke?

La X="antall visninger per dag" og Y="antall visninger per uke". Vi har da at $X \sim$Poisson(92) og $Y \sim$Poisson$(92 \cdot 7)$.

$$\begin{aligned} P(X > 100) &= 1 - P(X \leq 100) \\ &\approx 1 - P(Z \leq \frac{100 + 0.5 - 92}{\sqrt{92}}) \\ &= 1 - P(Z \leq 0.89) \\ &= 1 - 0.8133 = \underline{\underline{0.1867}} \\ P(Y > 700) &= 1 - P(Y \leq 700) \\ &\approx 1 - P(Z \leq \frac{700 + 0.5 - 92 \cdot 7}{\sqrt{92 \cdot 7}}) \\ &= 1 - P(Z \leq 2.23) \\ &= 1 - 0.9871 = \underline{\underline{0.0129}} \quad \square \end{aligned}$$

Gammafordeling (6.6-6.7)

Gammafordelingen er en fleksibel fordelingsklasse for $x \geq 0$. Gammafordeling brukes bl.a. til å beskrive levetider/funksjonstider både for tekniske systemer og i biologi. Flere mye brukte fordelinger er spesialtilfeller av gammafordelingen.

$$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}\,, \quad x \geq 0$$

$\mathrm{E}(X) = \alpha\beta$ og $\mathrm{Var}(X) = \alpha\beta^2$. $\Gamma(\alpha)$ er gammafunksjonen. Se formelsamlingen side 33. Merk spesielt at $\Gamma(n+1) = n!$ dersom n er et heltall.

Eksempel: Levetiden til en bestemt type insekt i måneder er beskrevet ved en gammafordeling med parametre $\alpha = 2$ og $\beta = 0.5$. Hva er sannsynligheten for at et insekt levere i mer enn 2 måneder?

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) \\ &= 1 - \int_0^2 \frac{1}{0.5^2\Gamma(2)} x^{2-1} e^{-x/0.5} dx \\ &= 1 - 4\int_0^2 x e^{-2x} dx \\ &= 1 - 4 \cdot 0.227 = \underline{\underline{0.092}} \quad \square \end{aligned}$$

Eksponensialfordeling (6.6-6.7)

Et viktig spesialtilfelle av gammafordelingen er eksponensialfordelingen som vi får ved å sette $\alpha = 1$ i gammafordelingen:

$$f(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x \geq 0,$$

$\mathrm{E}(X) = \beta$ og $\mathrm{Var}(X) = \beta^2$ (se øving 2).

Merk: Ofte brukes parameteren $\lambda = 1/\beta$.

Eksponensialfordelingen brukes bl.a. til å modellere funksjonstider (bl.a. til elektroniske komponenter), tider til og mellom hendelser, etc.

Videre har eksponensialfordelingen en del viktige relasjoner til Poisson-prosser.

La T være tid til første hendelse i en Poisson-prosess. Vi har da at:

$$\begin{aligned} P(T > t) &= P(\text{ingen hendelser i } [0,t]) \\ &= p(0;\lambda t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \end{aligned}$$

Dvs $F(t) = P(T \leq t) = 1 - e^{-\lambda t}$, og dermed blir tettheten for tid til første feil $f(t) = F'(t) = \lambda e^{-\lambda t}$ som er en eksponsialfordeling (med $\lambda = 1/\beta$).

Dvs tid til første hendelse i en Poisson-prosess er eksponensialfordelt. Det kan på tilsvarende måte også vises at tid *mellom* hendelser i en Poisson-prosess er eksponensialfordelt (f.eks. tid mellom visninger av en webside, tid mellom utrykninger fra en brannstasjon osv.)

Det kan også vises at tid til $k$te hendelse i en Poisson-prosess er gammafordelt med parametre $\alpha = k$ og $\beta = 1/\lambda$. Vi skal se litt mer på disse tingene senere.

Eksempel: Bussproblemet. Anta at ankomsten av busser til en bussholdeplass kan modelleres som en Poisson-prosess med parameter λ. Tiden mellom hver gang det kommer en buss er da eksponensialfordelt med forventningsverdi $1/\lambda$.

Du ankommer bussholdeplassen. Hva er forventet tid du må vente før det kommer en buss? La U være ventetiden fra du ankommer til det kommer en buss.

$$\begin{aligned} P(U > t) &= P(\text{ingen busser i } [0,t]) \\ &= \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \end{aligned}$$

dvs $\mathrm{E}(U) = 1/\lambda$ - samme som for tid mellom busser!!

Hvorfor? Når vi ankommer til bussholdeplassen har vi større sannsynlighet for å treffe et stort tidsintervall mellom busser enn et lite. □

χ^2-fordeling (6.8)

Viktig spesialtilfelle av gammafordelingen der $\alpha = \nu/2$ og $\beta = 2$ der ν er et positivt heltall. Brukes mye i statistisk teori, bl.a. ifbm tester/konfidensintervall for varians.

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} \ , \quad x \geq 0 \ ,$$

$$\nu = 1, 2, \ldots$$

$\mathrm{E}(X) = \nu$ og $\mathrm{Var}(X) = 2\nu$.

Lognormal-fordeling (6.9)

X er lognormal-fordelt dersom $\ln(X)$ er normalfordelt. Dersom i tillegg $\mathrm{E}(\ln(X)) = \mu$ og $\mathrm{Var}(\ln(X)) = \sigma^2$ er

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\frac{(\ln(x)-\mu)^2}{\sigma^2}\right), \quad x > 0,$$

$\mathrm{E}(X) = e^{\mu+\sigma^2/2}$ og $\mathrm{Var}(X) = e^{2(\mu+\sigma^2)} - e^{2\mu+\sigma^2}$.
Lognormalfordeling brukes bl.a. som modell for konsentrasjoner, levetider, utmatting, etc

Eksempel: Konsentrasjonen (målt i p.p.m.) av et forurensende stoff i utslippsvannet fra en kjemisk fabrikk er lognormalfordelt med parametre $\mu = 3.2$ og $\sigma = 1$. Hva er sannsynligheten for at konsentrasjonen overstiger 8?

$$\begin{aligned} P(X > 8) &= 1 - P(X \le 8) \\ &= 1 - P(\ln(X) \le \ln(8)) \\ &= 1 - P(\frac{\ln(X) - 3.2}{1} \le \frac{\ln(8) - 3.2}{1}) \\ &= 1 - P(Z \le -1.12) \\ &= 1 - 0.1314 = \underline{\underline{0.8686}} \quad \square \end{aligned}$$

Weibull-fordeling (6.10)

Weibull-fordeling blir bl.a. brukt til å modellere levetider/funksjonstider til komponenter og system. Enklere å regne med en gamma-fordelingen.

$$f(x) = \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}, \quad x \ge 0,$$

$\mathrm{E}(X) = \alpha^{-1/\beta}\Gamma(1 + \frac{1}{\beta})$ og
$\mathrm{Var}(X) = \alpha^{-2/\beta}\left[\Gamma(1 + \frac{2}{\beta}) - \Gamma(1 + \frac{1}{\beta})^2\right]$.

Merk: $\beta = 1$ gir eksponensialfordeling.

Eksempel: Oppetiden (i dager) for en server kan modelleres ved en Weibull-fordeling med $\alpha = 0.25$ og $\beta = 0.5$. Hva er sannsynligheten for at serveren har en oppetid på mindre enn 30 dager?

$$\begin{aligned} P(X < 30) &= \int_0^{30} 0.25 \cdot 0.5 x^{0.5-1} e^{-0.25x^{0.5}} dx \\ &= [-e^{-0.25x^{0.5}}]_0^{30} \\ &= -e^{-0.25\cdot 30^{0.5}} - (-1) = \underline{\underline{0.746}} \quad \square \end{aligned}$$

Til slutt

(Litt om tolkning etc...)

En sannsynlighetsfordeling angir hvordan utfallene vil fordele seg dersom man gjentar et forsøk mange ganger. Man kan tenke på en sannsynlighetsfordeling som et histogram over uendelig mange gjentak av forsøket.

Ulike typer sannsynlighetsfordelinger er egnet til å beskrive ulike fenomen. Parametrene i en sannsynlighetsfordeling bestemmer den nøyaktige form og beliggenhet på fordelingen (se f.eks. figur 6.3, 6.4, 6.5, 6.28 og 6.30 i læreboka). Egenskaper ved en fordeling som forventningsverdi og varians vil alltid være en funksjon av parametrene.

Når type fordeling og parameterverdier er kjente kan man regne ut alt av interesse (forventning, varians, sannsynligheter, osv).

I praksis vil ofte parameterverdiene være ukjente, og da ønsker man ut fra tilgjengelig informasjon å finne et best mulig anslag på verdiene. Estimering, som vi kommer til senere i kurset, handler om dette.

Ved diskrete fordelinger kan man ofte avgjøre ut fra informasjon om fenomenet hvilken type fordeling som er den rette. Ved kontinuerlige fordelinger er det ikke alltid like lett å avgjøre hvilken type fordeling som er best egnet til å beskrive fenomenet. I tillegg til fenomenkunnskap etc finnes det bl.a. ulike plott som man kan bruke til å avgjøre ut fra observerte data om en antatt fordelingstype ser ut til å være rimelig (f.eks. normalplott omtalt i avsnitt 8.3).