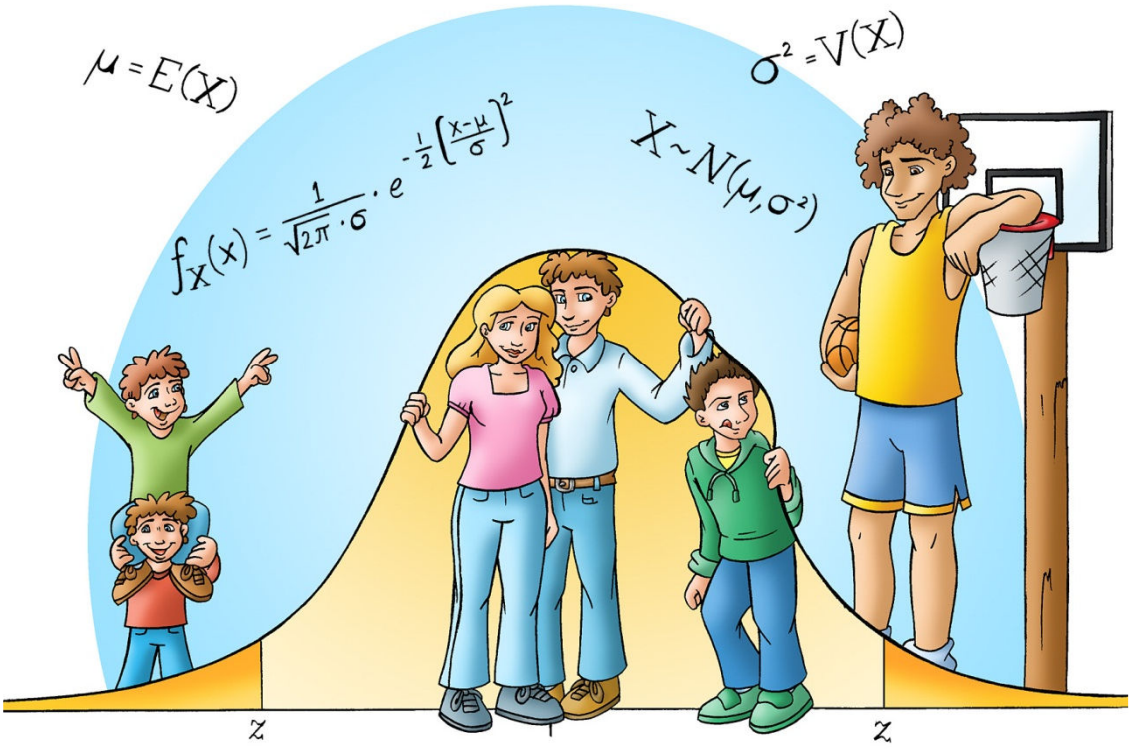
Normalfordelingen



Billedliste

Side 6: ©iStock.com/fizkes (Pige med mobiltelefon)

Side 8: Christian Albrecht Jensen [Public domain], via Wikimedia Commons (Carl Friedrich Gauss).

Side 8: ©iStock.com/traveler1116 (Pierre-Simon Laplace)

Side 14: ©iStock.com/Elenathewise (Mand ved maskine)

Side 18: ©iStock.com/gpointstudio (To fiskere)

Side 20: Pudelek [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)] (Den kongelige livgarde på Amalienborg)

1. Kontinuert stokastisk variabel

I en tidligere note har vi kigget på diskrete stokastiske variable, som kun kan antage *tælleligt* mange forskellige værdier, ofte endda kun endeligt mange forskellige værdier. Det er for eksempel tilfældet med en binomialfordelt stokastisk variabel, som kan antage værdierne 0, 1, … , n, hvor n er antal gange basiseksperimentet udføres. I denne note skal vi betragte stokastiske variable, som kan antage mere end tælleligt mange forskellige værdier, typisk hele intervaller af reelle tal eller hele R. Et godt eksempel er en stokastisk variabel, som angiver den tid, der går før man modtager det næste telefonopkald. Det er klart, at mængden af mulige værdier for X her er intervallet $[0,\infty[$. Det viser sig, at kontinuerte stokastiske variable skal behandles noget anderledes end de diskrete. Særligt bestemmes sandsynligheden for en hændelse forskelligt. Lad os definere:

Definition 1

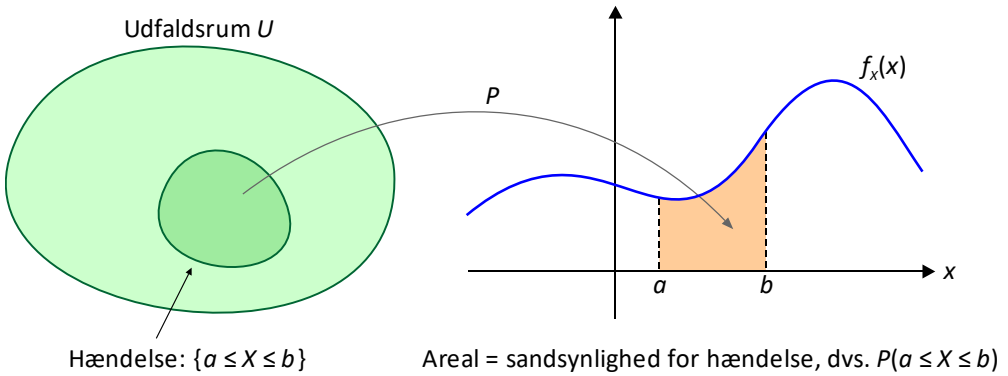
Lad X være en funktion fra et udfaldsrum U ind i mængden af reelle tal. Funktionen X siges da at være en *kontinuert stokastisk variabel*, hvis der findes en ikke-negativ integrabel funktion $f_X(x)$ med egenskaben

$$P(a \le X \le b) = \int_a^b f_X(x)\,dx \tag{1}$$

for alle reelle tal a og b med $a < b$. Funktionen $f_X(x)$ kaldes *tæthedsfunktionen* for X. På engelsk betegnes den *probability density function*, ofte forkortet *pdf*. Ligesom i det diskrete tilfælde er der en tilhørende *fordelingsfunktion* $F_X(x)$, som på engelsk hedder *cumulative distribution function* (forkortet *cdf*), og den er defineret ved:

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(z)\,dz \tag{2}$$

Her repræsenterer $\{a \le X \le b\}$, som er en forkortelse for $\{u \in U \mid a \le X(u) \le b\}$, en *hændelse*, nemlig mængden af de udfald fra udfaldsrummet U, som ved X afbildes i intervallet $[a,b]$. Ifølge definitionen er kravet til X altså, at der findes en fast tæthedsfunktion, så sandsynligheden $P(a \le X \le b)$ for den pågældende hændelse kan bestemmes som arealet under tæthedsfunktionen fra a til b.



I det følgende angives en række egenskaber, hvoraf nogle faktisk er lidt tekniske at bevise stringent, da man skal dykke helt ned i selve sandsynlighedsbegrebet. Det vil vi ikke gøre her. Egenskaberne virker desuden meget naturlige.

Sætning 2

For en kontinuert stokastisk variabel med tæthedsfunktion f_X gælder følgende:

a) $\int_{-\infty}^{\infty} f_X(z)\,dz = 1$

b) $P(X = a) = 0$ for ethvert $a \in R$

Sætning 3

Nedenstående egenskaber gælder for fordelingsfunktionen for en kontinuert stokastisk variabel X.

a) $P(a \le X \le b) = F_X(b) - F_X(a)$

b) $0 \le F_X(x) \le 1$ for alle $x \in R$

c) $x_1 \le x_2 \Rightarrow F_X(x_1) \le F_X(x_2)$, så F_X er en voksende funktion

d) $\lim_{x\to-\infty} F_X(x) = 0$ og $\lim_{x\to\infty} F_X(x) = 1$

e) Hvis f_X er kontinuert i x_0 gælder: $F_X'(x_0) = f_X(x_0)$.

Bevis: Lad os nøjes med at bevise b). Hvis $x_1 \le x_2$, så gælder $X \le x_1 \Rightarrow X \le x_2$. Om de tilhørende hændelser gælder da: $\{X \le x_1\} \subseteq \{X \le x_2\}$, hvormed $P(X \le x_1) \le P(X \le x_2)$. □

Bemærkninger 4

Egenskaben a) i sætning 2 udtrykker blot, at sandsynligheden for hele udfaldsrummet U er lig med 1: $P(U) = 1$. Ang. 2b): Hvis vi tillader at sætte $b = a$ i (1), fås $P(X = a) = 0$. Punktsandsynligheder er altså 0. Det kan godt virke underligt, for i princippet kan X jo godt antage værdien a. Men hvis vi tænker i et konkret eksempel, kan vi godt intuitivt forstå, at man ikke kan tildele en sandsynlighed større end 0 til hændelsen $X = a$. Tænk for eksempel på eksemplet med ventetid på et telefonopkald: Sandsynligheden for at opkaldet skulle ske præcist kl. 14.25 er nærmest usandsynlig. Den kunne lige så vel foregå kl. 14.250001, hvis vi regner i kommatal. Så det giver altså ikke mening at udregne punktsandsynligheder for kontinuerte stokastiske variable – i modsætning til hvad tilfældet er for diskrete stokastiske variable. Af samme grund er det ligegyldigt, om man benytter $\le$ eller blot $<$ i beregningen af sandsynligheder:

(3) $$P(a \le X \le b) = P(a \le X < b) = P(a < X \le b) = P(a < X < b)$$

Man kan også tale om middelværdi, varians og spredning for en kontinuert fordelt stokastisk variabel. Her er summer blot udskiftet med integraler.

Definition 5

Lad X være en kontinuert stokastisk variabel. Da er middelværdien, også kaldet den *forventede værdi* af X defineret ved

$$\mu = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx \tag{4}$$

og variansen og spredningen er defineret ved henholdsvis

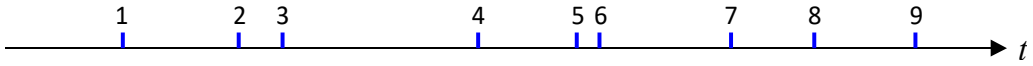
$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx \tag{5}$$

$$\sigma(X) = \sqrt{Var(X)} \tag{6}$$

forudsat at integralerne giver mening.

Eksempel 6 (Eksponentialfordelingen)

Hvis en række begivenheder indtræffer, kan man være interesseret i at studere fordelingen af tidsintervallerne mellem begivenhederne. Man indfører en stokastisk variabel X, som angiver tidsrummet mellem to på hinanden følgende begivenheder. Under nogle bestemte antagelser, kan man faktisk sige noget om fordelingen af X.



Antagelserne er:

1. Begivenhederne indtræffer uafhængig af hinanden (fordelingen er *glemsom*).
2. Begivenhederne indtræffer med en fast gennemsnitlig frekvens pr. tidsenhed.

Man kan vise, at under disse betingelser er X en kontinuert stokastiske variabel med denne tæthedsfunktion:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases} \tag{7}$$

hvor λ angiver den gennemsnitlige begivenhedsrate pr. tidsenhed. Den stokastiske variabel siges da at være *eksponentialfordelt*. Det kan nævnes, at eksponentialfordelingen er i familie med den velkendte Poisson-fordeling, som dog er en diskret fordeling. Ikke mere herom. Lad os beregne et udtryk for fordelingsfunktionen.

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(y)\,dy = \int_{0}^{x} \lambda \cdot e^{-\lambda \cdot y}\,dy = \lambda \cdot \int_{0}^{x} e^{-\lambda \cdot y}\,dy \\ &= \lambda \cdot \left[-\frac{1}{\lambda} \cdot e^{-\lambda \cdot y} \right]_0^x = -\left[e^{-\lambda \cdot y} \right]_0^x = -\left(e^{-\lambda x} - 1 \right) = 1 - e^{-\lambda x} \end{aligned} \tag{8}$$

Middelværdien af X kan bestemmes:

$$(9) \qquad \mu = E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\,dx = \int_{0}^{\infty} x \cdot \lambda \cdot e^{-\lambda x}\,dx = \frac{1}{\lambda}$$

Vi udelader detaljer, da partiel integration ikke længere er pensum i gymnasiet. Alternativt kan man med sit CAS-værktøj afprøve påstanden (hvis du gør det generelt med λ, skal du huske at "fortælle værktøjet", at $\lambda > 0$). Resultatet er ikke underligt, for hvis for eksempel den gennemsnitlige begivenhedsrate λ er 2 pr. time, så vil der i middel være 1/2 time mellem hver begivenhed. På tilsvarende vis kan variansen bestemmes:

$$(10) \qquad Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx = \int_{0}^{\infty} \left(x - \frac{1}{\lambda}\right)^2 \cdot \lambda \cdot e^{-\lambda x}\,dx = \frac{1}{\lambda^2}$$

Fordelingen finder anvendelse indenfor *køteori* og til at vurdere fejlrater ved produktioner (*Reliability Theory*), ligesom den også dukker op flere steder i fysik. Det næste eksempel er et mere jordnært eksempel.

□

Eksempel 7

Louise har lige afsluttet en samtale med sin kæreste, som er på konference i udlandet. Hun opdager, at hun har glemt at fortælle ham noget vigtigt. Desværre kan hun ikke ringe tilbage, da kæresten ikke har sin egen telefon med på konferencen. Hun må altså vente indtil han selv ringer tilbage til hende. Hun ved, at han i gennemsnit plejer at ringe hver ca. 2,5 time i dagtimerne. Nedenstående spørgsmål ønskes besvaret, idet det antages, at den måde kæresten generelt set ringer på overholder de to antagelser fra eksempel 6.

a) Hvad er sandsynligheden for, at han ringer indenfor 1 time?
b) Hvad er sandsynligheden for, at han ringer om mellem 1 og 2 timer?
c) Hvad er sandsynligheden for, at det tager mindst 6 timer, før han ringer?

Løsninger:

For det første skal vi lige have bestemt begivenhedsraten: Her benytter vi den formel for middelværdien for en eksponentialfordelt stokastisk variabel med parameter λ, som vi udledte i eksempel 6. Vi kender jo middelværdien:

$$\mu = \frac{1}{\lambda} \Leftrightarrow \lambda = \frac{1}{\mu} = \frac{1}{2{,}5} = 0{,}4$$

så hans opkaldsrate er altså 0,4 opkald i timen.

a) Vi løser dette spørgsmål ved hjælp af fordelingsfunktionen:

$$P(X \le 1) = F_X(1) = 1 - e^{-0.4 \cdot 1} = 0{,}3297$$

Der er altså en sandsynlighed på ca. 33,0%, for at han ringer indenfor den første time.

b) Her vælger vi både at bestemme sandsynligheden ved hjælp af et areal under tæthedsfunktionen og ved hjælp af fordelingsfunktionen. Først bruger vi (1):

$$P(1 \le X \le 2) = \int_1^2 f_X(x)\,dx = \int_1^2 0.4 \cdot e^{-0.4 \cdot x}\,dx = 0{,}2210$$

og derefter ved hjælp af sætning 3a):

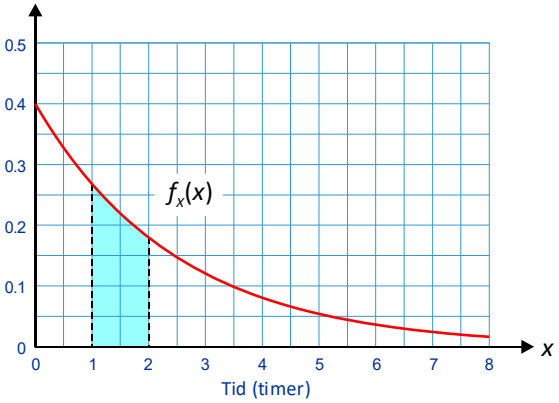
$$P(1 \le X \le 2) = F_X(2) - F_X(1) = 0{,}5507 - 0{,}3297 = 0{,}2210$$

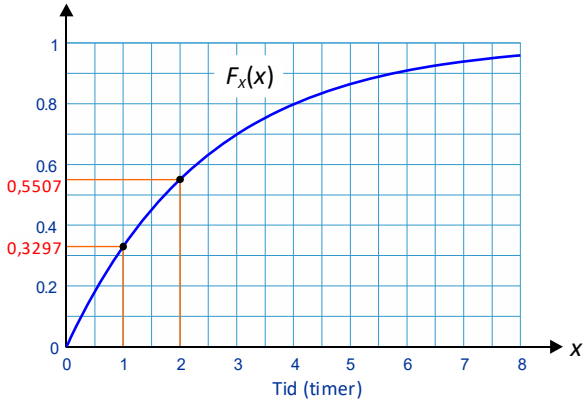
I begge tilfælde får vi altså en sandsynlighed på 22,1% for at kæresten ringer om mellem 1 og 2 timer.

c) Her bruger vi igen fordelingsfunktionen, idet vi indser, at den *komplementære* hændelse til at $X \ge 6$ er at $X < 6$. Summen af deres sandsynligheder er 1, da foreningsmængden af hændelserne er hele udfaldsrummet U og fordi der er tale om disjunkte hændelser. Vi får altså:

$$P(X \ge 6) = 1 - P(X < 6) = 1 - F_X(6) = 1 - 0{,}9093 = 0{,}0907$$

hvor vi har udnyttet bemærkning 4 med at punktsandsynligheder er 0. Altså er der omkring 9,1% sandsynlighed for, at han først ringer efter 6 timer.





2. Normalfordelingen

Som så mange andre at matematikkens teorier kom normalfordelingen til verden ad kringlede veje. Normalfordelingen blev, som vi skal se, dels set som en approksimation til binomialfordelingen, dels blev den betragtet som en fejlkurve, som kan beskrive fordelingen af fejl ved målinger af en fysisk størrelse. I dag er normalfordelingen den mest benyttede af alle de fordelinger, der findes indenfor sandsynlighedsregningen og statistikken. Det var oprindeligt dog på ingen måde selvindlysende, at denne fordeling skulle få den særstatus, som den har fået. En del af dens succes skyldes da også, at fordelingen med stor tilnærmelse kan benyttes til at beskrive eller forudsige så mange forhold fra den virkelige verden. Adskillige af de allerdygtigste matematikere var involveret i udviklingen af normalfordelingen, herunder Abraham De Moivre (1667-1754), Jacob Bernouilli (1654-1705), Pierre-Simon Laplace (1749-1823) og Carl Friedrich Gauss (1777-1855).

Carl Friedrich Gauss (1777-1855)

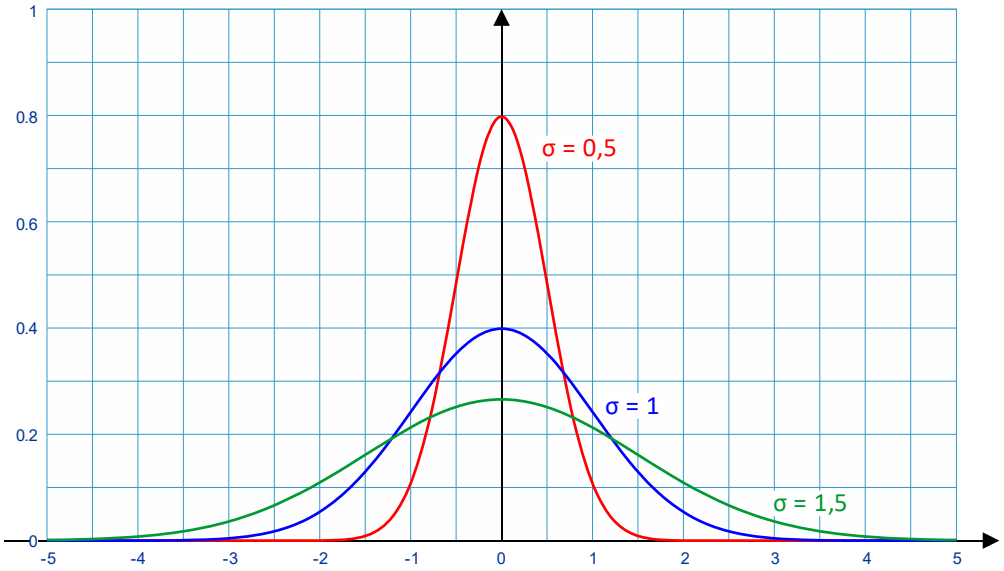
Pierre-Simon Laplace (1749-1827)

Normalfordelingen er en kontinuert fordeling. Den har to parametre, nemlig μ og σ, som vi senere skal se er henholdsvis middelværdi og spredning for fordelingen (se sætning 8c) side 10). Med notationen $X \sim N(\mu,\sigma^2)$ vil vi mene, at X er en normalfordelt stokastisk variabel med parametre μ og σ. Dens tæthedsfunktion ser således ud:

(11)
$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$$

På næste side er grafen for tæthedsfunktionen afbildet for tre forskellige værdier af σ, mens $\mu = 0$ i alle tre tilfælde. Det er ikke så interessant at variere μ, for det bevirker blot en parallelforskydning af grafen med μ i x-aksens retning, hvilket ses direkte af forskriften. Forskriften afslører desuden, at grafen er symmetrisk omkring den lodrette linje $x = \mu$. Det er ikke underligt, at grafen for tæthedsfunktionen for en normalfordeling ofte kaldes for en *klokkekurve*. Parameteren σ styrer, hvor bred klokkekurven er.

Graferne for tre tæthedsfunktioner for normalfordelingen (alle μ = 0)



Ifølge definition 1 side 3 får vi fordelingsfunktionen $F_{\mu,\sigma}(x)$ ved at integrere tæthedsfunktionen fra $-\infty$ til x:

$$(12) \qquad F_{\mu,\sigma}(x) = P(X \le x) = \int_{-\infty}^{x} f_{\mu,\sigma}(z)dz = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{z-\mu}{\sigma}\right)^2} dz$$

Resultatet kan ikke udtrykkes ved hjælp af de sædvanlige matematiske funktioner, så man må ty til numerisk beregning af integralet for hver værdi af x. De fleste CAS-værktøjer har dog indbygget værktøjer til at håndtere både tæthedsfunktionen og fordelingsfunktionen for en normalfordelt stokastisk variabel. Husk den engelske forkortelse *pdf* for tæthedsfunktionen og *cdf* for fordelingsfunktionen for en generel kontinuert stokastisk variabel, så det er meget tænkeligt at værktøjerne indeholder disse tre bogstaver som en del af navnet. Sætning 3e) giver os straks, at hvis man differentierer fordelingsfunktionen, så fås tæthedsfunktionen:

$$(13) \qquad F'_{\mu,\sigma}(x) = f_{\mu,\sigma}(x)$$

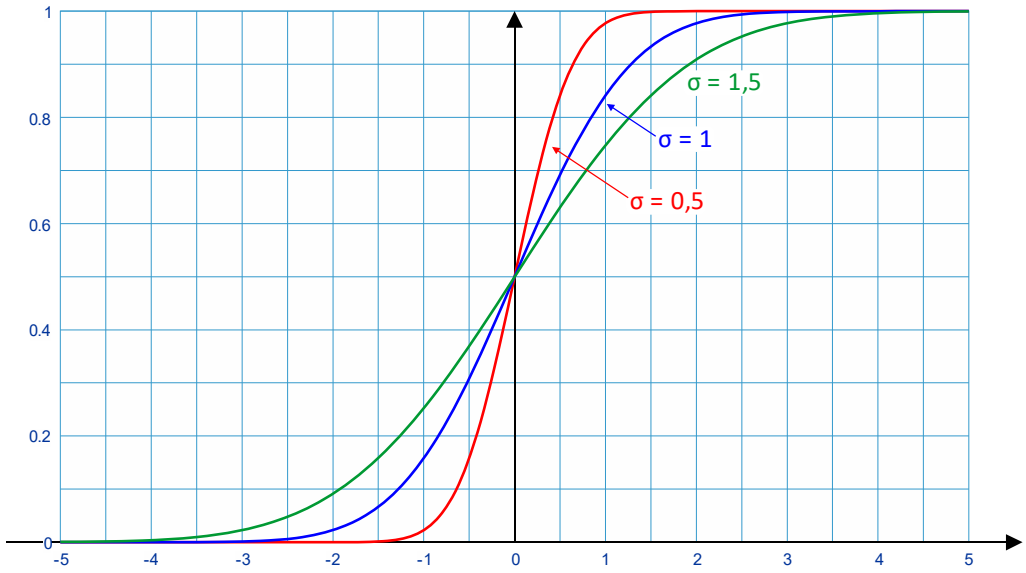
Det er egentligt bare en anvendelse af integralregningens fundamentalsætning på (12). De tre tæthedsfunktioner på figuren ovenfor har fordelingsfunktioner, hvis grafer er S-formede. De er afbildet på figuren på næste side.

Der er én af normalfordelingerne, som har en særstatus, nemlig den med $\mu = 0$ og $\sigma = 1$. Den har fået navnet *standardnormalfordelingen*, og dens fordelingsfunktion får sit eget specielle symbol, nemlig Φ:

$$(14) \qquad \Phi(x) = F_{0,1}(x) = \frac{1}{\sqrt{2\pi}}\cdot\int_{-\infty}^{x} e^{-\frac{1}{2}\cdot z^2} dz$$

Som sætning 8 på næste side viser, så er der en snæver sammenhæng mellem en normalfordeling med generelle parametre μ og σ og så standardnormalfordelingen. I tidligere tider var det meget vigtigt, for så behøvede man kun én tabel med de kumulerede sandsynligheder.

Graferne for tre fordelingsfunktioner for normalfordelingen (alle μ = 0)



Sætning 8

Lad X og Z være stokastiske variable med $X = \sigma \cdot Z + \mu$, dvs. $Z = \dfrac{X-\mu}{\sigma}$. Da gælder:

a) $Z \sim N(0,1) \Leftrightarrow X \sim N(\mu, \sigma^2)$

b) Hvis $X \sim N(\mu, \sigma^2)$ gælder: $P(X \le x) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$

c) Parametrene μ og σ i en normalfordeling angiver henholdsvis middelværdi og spredningen for fordelingen.

Bevis:

a) Lad os vise, at $Z \sim N(0,1) \Rightarrow X \sim N(\mu, \sigma^2)$. Den anden vej foregår analogt.

$$P(X \le x_0) = P(\sigma \cdot Z + \mu \le x_0) = P\left(Z \le \frac{x_0 - \mu}{\sigma}\right)$$

(15)

$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{(x_0-\mu)/\sigma} e^{-\frac{1}{2}\cdot z^2}\, dz = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{x_0} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}\, dx$$

hvor vi i tredje lighedstegn har benyttet antagelsen om at Z er standardnormalfordelt. For at få det fjerde og sidste lighedstegn har vi benyttet integration ved substitution: $x = \sigma \cdot z + \mu \Leftrightarrow z = (x-\mu)/\sigma$, hvoraf $dz = 1/\sigma \cdot dx$. Substitutionen giver desuden de nye y-grænser $-\infty$ og x_0. Sidstnævnte integral viser netop ifølge (12), at X er en normalfordelt stokastisk variabel med parametre μ og σ.

b) Fås blot ved at gentage de første dele af (15), blot med x i stedet for x_0.

$$(16) \qquad P(X \le x) \;=\; P(\sigma \cdot Z + \mu \le x) \;=\; P\left(Z \le \frac{x-\mu}{\sigma} \right) \;=\; \Phi\left(\frac{x-\mu}{\sigma} \right)$$

c) Vi antager igen, at $X \sim N(\mu, \sigma^2)$. Ifølge definition 5 og (11) fås middelværdien:

$$\begin{aligned}
E(X) \;&=\; \int_{-\infty}^{\infty} x \cdot f_{\mu,\sigma}(x)\, dx \\
&= \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot e^{-\frac{1}{2} \cdot \left(\frac{x-\mu}{\sigma} \right)^2} \, dx \\
&= \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\sigma \cdot t + \mu) \cdot e^{-\frac{1}{2} \cdot t^2} \cdot \sigma \cdot dt \\
&= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (\sigma \cdot t + \mu) \cdot e^{-\frac{1}{2} \cdot t^2} \, dt \\
&= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \sigma \cdot t \cdot \cdot e^{-\frac{1}{2} \cdot t^2} + \mu \cdot e^{-\frac{1}{2} \cdot t^2} \, dt \\
&= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \sigma \cdot t \cdot e^{-\frac{1}{2} \cdot t^2} \, dt \;+\; \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} \mu \cdot e^{-\frac{1}{2} \cdot t^2} \, dt \\
&= \frac{\sigma}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} t \cdot e^{-\frac{1}{2} \cdot t^2} \, dt \;+\; \mu \cdot \left(\frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2} \cdot t^2} \, dt \right) \\
&= 0 \;+\; \mu \\
&= \mu
\end{aligned}$$

hvor vi for at få andet lighedstegn har benyttet integration ved substitution. Vi har sat: $t = (x-\mu)/\sigma \Leftrightarrow x = \sigma \cdot t + \mu$, hvoraf $dx = \sigma \cdot dt$. 5. lighedstegn: Der er ganget ind i parentesen. 6. lighedstegn: Integralet er delt op i to integraler. 7. lighedstegn: konstanter er ganget ud foran integralet. 8. lighedstegn: I det første integral er integranden en *ulige* funktion. Når der integreres fra $-\infty$ til ∞, fås derfor 0. I parentesen i andet led integreres tæthedsfunktionen for standardnormalfordelingen fra $-\infty$ til ∞, og den ved vi giver 1. Heraf ses, at middelværdien for X er lig med μ.

Jeg skal spare læseren for at gå igennem udregningerne for at finde variansen af X. Den er nemlig endnu mere teknisk! Via definition 5 postulerer vi blot, at:

$$Var(X) \;=\; \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_{\mu,\sigma}(x)\, dx \;=\; \sigma^2$$

og dermed fås følgende værdi for spredningen for X: $\sigma(X) = \sqrt{Var(x)} = \sqrt{\sigma^2} = \sigma$.

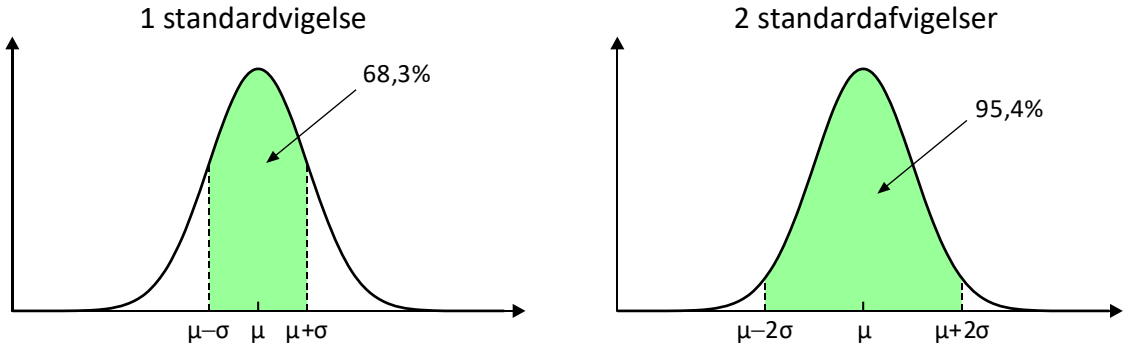
□

Spredningen σ kaldes også ofte for *standardafvigelsen*. Man kan stille sig det spørgsmål, hvad sandsynligheden er for, at en normalfordelt stokastisk variabel X højst ligger henholdsvis én eller to standardafvigelser fra middelværdien. I det følgende bruger vi både sætning 3a) og sætning 8b):

$$\begin{aligned} P(\mu-\sigma\le X\le\mu+\sigma) &= P(X\le\mu+\sigma)-P(X\le\mu-\sigma)\\ &= \Phi\left(\frac{\mu+\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-\sigma-\mu}{\sigma}\right)\\ &= \Phi(1)-\Phi(-1)\\ &= 0{,}841345-0{,}150655\\ &= 0{,}682690 \end{aligned}$$

$$\begin{aligned} P(\mu-2\sigma\le X\le\mu+2\sigma) &= P(X\le\mu+2\sigma)-P(X\le\mu-2\sigma)\\ &= \Phi\left(\frac{\mu+2\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-2\sigma-\mu}{\sigma}\right)\\ &= \Phi(2)-\Phi(-2)\\ &= 0{,}97725-0{,}02275\\ &= 0{,}95450 \end{aligned}$$

Man bruger selvfølgelig sit CAS-værktøj til at bestemme værdierne for fordelingsfunktionen for standardnormalfordelingen. Vi konkluderer, at sandsynligheden for, at X er højst én standardafvigelse fra middelværdien, er 68,3%, mens sandsynligheden for at X er højst to standardafvigelser fra μ er 95,4%. Svarene på spørgsmålene er åbenlyst uafhængige af hvilken normalfordeling, der er tale om. Det kan altså undertiden være fornuftigt at regne i enheder af standardafvigelsen σ fra middelværdien μ.



I det følgende skal vi se på eksempler på forskellige opgavetyper i forbindelse med normalfordelinger. Eksemplerne vil overvejende blive løst med CAS-værktøj.

Eksempel 9

En stokastisk variabel X oplyses at være normalfordelt med middelværdi 30 og spredning 4. Bestem $P(X\le 33)$.

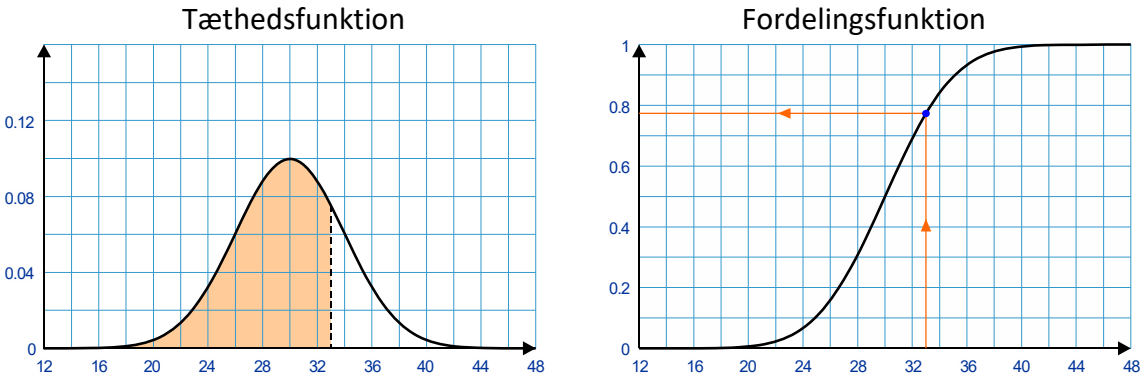
Løsning: Der er to måder at løse denne opgave på. Enten kan man bestemme sandsynligheden som arealet under grafen for tæthedsfunktionen fra $-\infty$ til 33, eller også kan man blot bestemme fordelingsfunktionens værdi i 33. Vil vi bruge tæthedsfunktionen, fås ifølge (12) følgende for $\mu=30$ og $\sigma=4$:

$$P(X \le 33) = \int_{-\infty}^{33} f_{30,4}(y)dy = \frac{1}{4\cdot\sqrt{2\pi}} \cdot \int_{-\infty}^{33} e^{-\frac{1}{2}\cdot\left(\frac{y-30}{4}\right)^2} dy = 0{,}7734$$

men man kan også undgå at skulle sætte ind i udtrykket for tæthedsfunktionen efter 2. lighedstegn, for de fleste CAS-værktøj har en indbygget tæthedsfunktionen $f_{\mu,\sigma}(x)$ for den generelle normalfordeling. På tilsvarende vis har de fleste CAS-værktøjer indbygget fordelingsfunktionen $F_{\mu,\sigma}(x)$ for den generelle normalfordeling. Så her skal man bare indsætte 33 på x's plads:

$$P(X \le 33) = F_{30,4}(33) = 0{,}7734$$

Har man derimod kun en tabel med værdier for fordelingsfunktionen for standardnormalfordelingen, så kan man gøre brug af sætning 8b). Det vil vi dog ikke vise her. Rent grafisk kan situationen med de to løsninger afbildes således:



□

Eksempel 10 (IQ-skala)

Skalaen for intelligenskvotienter er således bygget op, at menneskehedens IQ-værdier fordeler sig som en normalfordeling med middelværdi 100 og spredning 15.

a) Hvor stor en del af populationen har en intelligenskvotient på under 80?
b) Hvor stor en andel af populationen har en intelligenskvotient mellem 110 og 120?
c) En betingelse for at blive optaget i organisationen *Mensa* er, at man hører til de 2% mest intelligente personer. Hvor høj en score skal man have for at blive optaget?

Løsning:

a) Vi benytter et CAS-værktøj til at udregne værdier for fordelingsfunktionen:

$$P(X \le 80) \;=\; F_{100,15}(80) \;=\; 0{,}0912$$

så omkring 9,1% af populationen har en IQ på under 80. NB! Husk at det for kontinuerte fordelinger er ligegyldigt, om man spørger om "mindre end" eller "mindre end eller lig med".

b) Vi benytter sætning 3a):

$$P(110 \le X \le 120) \;=\; F_{100,15}(120) - F_{100,15}(110) \;=\; 0{,}1613$$

så omkring 16,1% af populationen har en IQ på mellem 110 og 120.

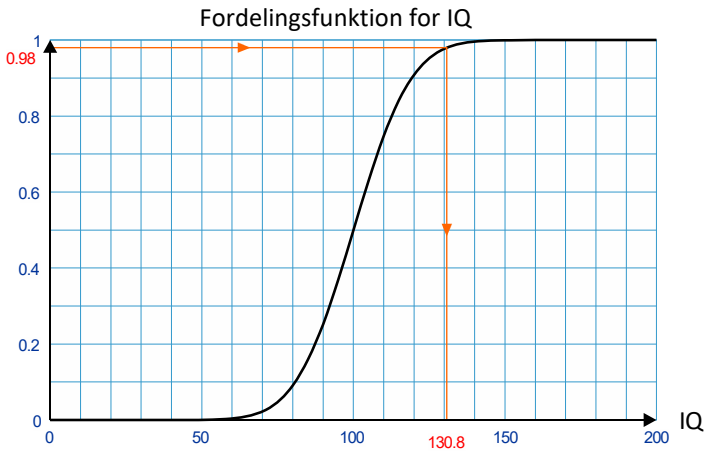
c) Vi skal bestemme x således, at $P(X \ge x) = 0{,}02$. Heraf får vi, at værdien af fordelingsfunktionen i x er lig med 0,98:

$$F_{100,15}(x) \;=\; P(X \le x) \;=\; 1 - P(X > x) \;=\; 1 - 0{,}02 \;=\; 0{,}98 \quad\Leftrightarrow\quad x = F^{-1}_{100,15}(0{,}98)$$

Vi skal altså bestemme 0,98-fraktilen i vores normalfordeling. Det kan enten løses som en ligning med fordelingsfunktionen, eller ved hjælp af en invers fordelingsfunktion, som de fleste CAS-værktøjer også har. Her giver svaret:

$$x \;=\; F^{-1}_{100,15}(0{,}98) \;=\; 130{,}81$$

Man skal altså have en intelligenskvotient på 131 for at blive optaget i Mensa.



Eksempel 11 (Variation i produktionen)

En maskine på en fabrik skal fremstille cylindre med en diameter på 20 mm. Imidlertid falder resultatet ikke altid helt nøjagtigt ud. Det viser sig, at diametrene er normalfordelte med middelværdi 20 mm og med en spredning på 0,1 mm. Fabrikanten kan acceptere en afvigelse på maksimalt 0,2 mm fra det ønskede.

a) Hvor stor en del af cylindrene må kasseres?

En ingeniør mener at kunne forbedre maskinen, så den bliver mere nøjagtig og det kun er nødvendigt at kassere 2% af cylindrene.

b) Hvor meget skal spredningen reduceres til, hvis målet skal nås?

Opgaven ønskes løst med CAS-værktøj.

Løsning:

a) Vi skal altså finde sandsynligheden for at diameteren enten er over 20,2 mm eller under 19,8 mm. Igen bruger vi tricket med den modsatte hændelse til af $X > 20{,}2$:

$$\begin{aligned} P(X<19{,}8)+P(X>20{,}2) &= P(X<19{,}8)+1-P(X\le 20{,}2) \\ &= F_{20,0.1}(19{,}8)+1-F_{20,0.1}(20{,}2) \\ &= 0{,}0228+1-0{,}9773 \\ &= 0{,}0455 \end{aligned}$$

Vi ser, at 4,6% af cylindrene må kasseres.

b) Vi skal bestemme σ, så $P(X<19{,}8)+P(X>20{,}2)=0{,}02$. Da normalfordelingen er symmetrisk, har vi $P(X\le 19{,}8)=0{,}01$. Igen udregner vi fraktiler ud, ligesom i eksempel 10. Man må løse en ligning med hensyn til den ubekendte σ.

$$P(X\le 19{,}8)=0{,}01 \quad\Leftrightarrow\quad F_{20,\sigma}(19{,}8)=0{,}01 \quad\Leftrightarrow\quad \sigma=0{,}0860$$

Man kan altså reducere spredningen i produktionen fra 0,1 til 0,086, for at der kun skal kasseres 2% af cylindrene.

Bemærkning! Hvis man ikke har et værktøj, der kan løse ligninger med fraktiler, så kan man alternativt udnytte sætning 8b) samt gå omvendt ind i en tabel, som indeholder værdier for fordelingsfunktionen for standardnormalfordelingen Φ. Det svarer til at løse følgende ligning, idet $\mu=20$ og $x=19{,}8$:

$$\Phi\left(\frac{19{,}8-20}{\sigma}\right)=0{,}01 \quad\Leftrightarrow\quad \frac{19{,}8-20}{\sigma}=\Phi^{-1}(0{,}01)=-2{,}3263 \quad\Leftrightarrow\quad \sigma=0{,}0860$$

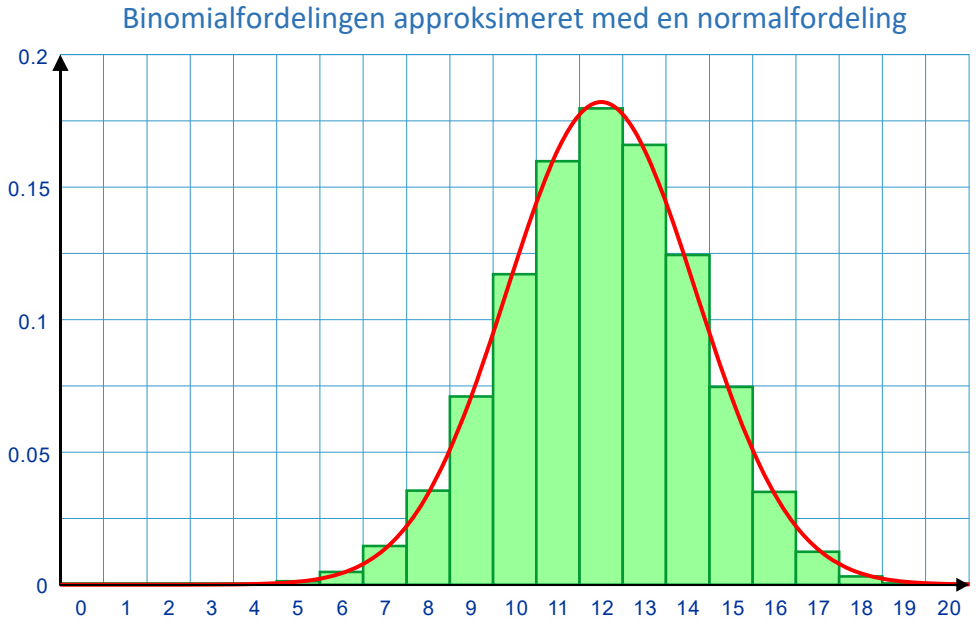
□

Bemærkning 12

Der er en lang række eksempler fra det virkelige liv, hvor man har erfaringer for at data tilnærmelsesvist er normalfordelte. Når man producerer komponenter i industrien (jf. eksempel 11), så falder komponenterne ikke altid ens ud, selv om det er intentionen. De kan måske have lidt forskellig vægt eller lidt forskellig længde. Når flere små tilfældige og indbyrdes uafhængige effekter er til stede i en produktionsproces, så har man praktisk erfaring for, at produkterne tilnærmelsesvist følger en normalfordeling på en eller flere punkter. Dette er også underbygget teoretisk gennem den såkaldte *centrale grænseværdisætning* (*The Central Limit Theorem*) – en dyb sætning, der står som en hjørnesten i sandsynlighedsregningen. Desværre alt for kompliceret til at blive behandlet nærmere her.

3. Normalfordelingens forbindelse til binomialfordelingen

Binomialfordelingen er som bekendt en diskret fordeling, fordi en binomialfordelt stokastisk variabel kun kan antage endeligt mange værdier, nemlig 0, 1, … , n. Det er måske derfor overraskende, at den kan tilnærmes med normalfordelingen, som jo er en kontinuert fordeling. Som et eksempel kan vi se på en binomialfordelt stokastisk variabel X med antalsparameter $n = 20$ og basissandsynlighed $p = 0,6$. På figuren på næste side er sandsynlighedsfordelingen for X afbildet, så søjlerne har centrum i de værdier, de hører til. Fra teorien om binomialfordelingen ved vi, at der for en binomialfordelt stokastisk variabel X gælder, at middelværdien er givet ved $E(X) = n \cdot p = 20 \cdot 0,6 = 12$ og variansen er givet ved $\mathrm{Var}(X) = n \cdot p \cdot (1-p) = 4,8$. På nævnte figur er desuden indtegnet tæthedsfunktionen for normalfordelingen med de samme værdier for middelværdi og varians, altså henholdsvis 12 og 4,8. Vi ser, at approksimationen er overraskende god!



Som en tommefingeregel kan man sige, at hvis $n > 9 \cdot p/(1-p)$ og $n > 9 \cdot (1-p)/p$, så er normalfordelingen en rimelig god approksimation til binomialfordelingen. Hvorfra disse postulater kommer, er en længere historie, som vi absolut ikke skal gå i detaljer med her. Kort fortalt skal det dog nævnes, at den geniale matematiker *Abraham de Moivre* (1667-1754) søgte en måde at simplificere udregningerne af binomialsandsynligheder på – husk på, at alt måtte regnes i hånden dengang! På den tid var normalfordelingen endnu ikke opdaget, men de resultater de Moivre kom frem til, kan tolkes på den måde, at han tilnærmede binomialfordelingen med en normalfordeling. Han kan samtidigt siges at være den person, som gav den første formulering af den centrale grænseværdisætning omtalt i bemærkning 12. Du kan se en figur fra bogen af Abraham de Moivre på næste side.

COROLLARY 1.

This being admitted, I conclude, that if m or $\frac{1}{2}n$ be a Quantity infinitely great, then the Logarithm of the Ratio, which a Term distant from the middle by the Interval l, has to the middle Term, is $-\frac{2ll}{n}$.

COROLLARY 2.

The Number, which answers to the Hyperbolic Logarithm $-\frac{2ll}{n}$, being

$$1 - \frac{2ll}{n} + \frac{4l^4}{2nn} - \frac{8l^6}{6n^3} + \frac{16l^8}{24n^4} - \frac{32l^{10}}{120n^5} + \frac{64l^{12}}{720n^6}, \text{ \&c.}$$

it follows, that the Sum of the Terms intercepted between the Middle, and that whose distance from it is denoted by l, will be

$$\frac{2}{\sqrt{nc}} \text{ into } l - \frac{2l^3}{1\times 3n} + \frac{4l^5}{2\times 5nn} - \frac{8l^7}{6\times 7n^3} + \frac{16l^9}{24\times 9n^4} - \frac{32l^{11}}{120\times 11n^5}, \text{ \&c.}$$

Let now l be supposed $= s\sqrt{n}$, then the said Sum will be expressed by the Series

$$\frac{2}{\sqrt{c}} \text{ into } f - \frac{2f^3}{3} + \frac{4f^5}{2\times 5} - \frac{8f^7}{6\times 7} + \frac{16f^9}{24\times 9} - \frac{32f^{11}}{120\times 11}, \text{ \&c.}$$

Moreover, if f be interpreted by $\frac{1}{2}$, then the Series will become

$$\frac{2}{\sqrt{c}} \text{ into } \frac{1}{2} - \frac{1}{3\times 4} + \frac{1}{2\times 5\times 8} - \frac{1}{6\times 7\times 10} + \frac{1}{24\times 9\times 32} - \frac{1}{120\times 11\times 64}, \text{ \&c.}$$

which converges so fast, that by help of no more than seven or eight Terms, the Sum required may be carried to six or seven places of Decimals: Now that Sum will be found to be 0.427812, independently from the common Multiplicator $\frac{2}{\sqrt{c}}$, and therefore to the Tabular Logarithm of 0.427812, which is $\overline{9}.6312529$, adding the Logarithm of $\frac{2}{\sqrt{c}}$, *viz.* $\overline{9}.9019400$, the Sum will be $\overline{19}.5331929$, to which answers the number 0.341344.

LEMMA.

If an Event be so dependent on Chance, as that the Probabilities of its happening or failing be equal, and that a certain given number n of Experiments be taken to observe how often it happens and fails, and also that l be another given number, less than $\frac{1}{2}n$, then the Probability of its neither happening more frequently than $\frac{1}{2}n + l$ times,

Side 245 i Abraham De Moivres værk *The Doctrine of Chances*, 3. udgave 1756. Corollary 2 indeholder hovedresultatet.

Opgaver

Opgave 1 (eksponentialfordelingen)

Betragt eksponentialfordelingen fra eksempel 6.

a) Vis ved at regne i hånden, at arealet under grafen for tæthedsfunktionen for en eksponentialfordelt stokastisk variabel virkeligt er lig med 1, som det skal være.

b) Benyt desuden et CAS-værktøj til at tegne graferne for såvel tæthedsfunktionen som fordelingsfunktionen for tilfældet $\lambda = 2$.

Opgave 2 (Ventetider for fiskefangst)

Bo og hans far står hver weekend og fisker efter ørreder i den lokale sø. De har lige fanget en ørred. Spørgsmålet er hvor lang tid, der går, før de igen får bid. Antag i det følgende, at tiderne imellem fangsterne kan beskrives ved en eksponentialfordelt stokastisk variabel af den type, som er beskrevet i eksempel 6. De to har en fangstrate på 1,5 ørreder i timen.

a) Bestem sandsynligheden for at de fanger den næste fisk inden, der er gået 45 minutter, altså inden 0,75 time.

b) Hvad er sandsynligheden for, at de to må vente mere end 2 timer på at fange den næste fisk.

c) Hvor lang tid må de i gennemsnit vente for at fange den næste fisk?

Opgave 3

Lad X være en stokastisk variabel, som er normalfordelt med middelværdi $\mu = 15$ og spredning $\sigma = 3$.

a) Tegn grafen for tæthedsfunktionen i intervallet $[0,30]$.

b) Bestem sandsynligheden $P(X \le 18)$ ved at benytte tæthedsfunktionen som i eksempel 9. Du må gerne benytte CAS-værktøjets indbyggede tæthedsfunktion. Forsøg eventuelt om muligt at tegne grafen med skravering af det areal, som svarer til sandsynligheden.

c) Gentag b) med sandsynligheden $P(10 \le X \le 18)$.

Opgave 4

Lad X være en stokastisk variabel, som er normalfordelt med middelværdi $\mu = 5{,}0$ og spredning $\sigma = 1{,}7$. Benyt normalfordelingens fordelingsfunktion til at udregne følgende:

a) Bestem $P(X \le 4)$
b) Bestem $P(2 \le X \le 7)$
c) Bestem $P(X \ge 6{,}2)$

Opgave 5 (svær)

Det oplyses, at der for en normalfordelt stokastisk variabel X gælder, at $P(X \le 34) = 0{,}38$ og $P(X \le 51) = 0{,}87$. Bestem middelværdi og spredning.

Hjælp: Benyt sætning 8b) til at vise: $(34-\mu)/\sigma = \Phi^{-1}(0{,}38)$ og $(51-\mu)/\sigma = \Phi^{-1}(0{,}87)$. Bestem højresiderne med dit CAS-værktøj og løs derefter to ligninger med to ubekendte.

Opgave 6 (Variation i produktionen)

Vi har tidligere i eksempel 11 set, at der uvægerligt vil være små variationer i produktionen på industrivirksomheder. Et eksempel kan være påfyldning af mælk på mælkekartoner. Det viser sig, at mængden af mælk i et karton kan beskrives ved en normalfordelt stokastisk variabel X. Det skal gerne være sådan, at forbrugeren med stor sandsynlighed får den mængde mælk, som vedkommende har betalt for, fx 1 liter eller 1000 ml. Lad os i det følgende antage, at tapningen af mælk foregår på en maskine, som er indstillet til at fylde 1015 ml mælk på hvert karton i gennemsnit (dvs. $\mu = 1015$), mens spredningen er 10 ml. Besvar da følgende spørgsmål:

a) Hvad er sandsynligheden for, at der er mindre end 1000 ml mælk i en given karton?
b) Bestem sandsynligheden for, at der er mere end 1020 ml mælk i en given karton.
c) Bestem sandsynligheden for, at der er mellem 995 ml og 1008 ml mælk i kartonen?

Afdelingslederen på mejeriet er ikke helt tilfreds med den procentdel af mælkekartonerne, som indeholder for lidt mælk. Han kan ikke indføre en mere nøjagtig maskine, for det er der ikke råd til. Derimod kan man på simpel vis justere aftapningsmaskinen, så den gennemsnitligt kommer lidt mere mælk i hvert karton.

d) Hvor meget skal man justere μ, for at mængden af kartoner med for lidt mælk i bliver færre end 3%? (*Hjælp*: Opstil en ligning indeholdende μ som ubekendt og løs den. Rent praktisk kan det være hensigtsmæssigt at løse ligningen numerisk, evt. med et gæt som input.
e) Prøv at løse samme spørgsmål som i d) ved hjælp af sætning 8b).

Opgave 7 (Værnepligtiges højde)

Ved at analysere højderne af 13427 værnepligtige i Danmark fra andet halvår af 2006 kan man konkludere, at soldaternes højde med meget stor nøjagtighed følger en normalfordeling med middelværdi 180,1 cm og spredning 6,81 cm. Med disse oplysninger skal følgende spørgsmål besvares.

a) Hvor mange procent af de værnepligtige mænd har en højde på under 170 cm?
b) Hvor mange procent har en højde på mellem 175 cm og 180 cm?
c) Hvor mange procent har en højde på mindst 200 cm?
d) Hvad kan man sige om de 10% højeste værnepligtige mænd?
e) Hvor mange procent har en højde på eksakt 180 cm?

Opgave 8

Lad os sige, at en professor efter flere års erfaring har observeret, at point-scorerne i en bestemt test er normalfordelte med middelværdi 70 og spredning 15. Hvor skal han lægge bestået-grænsen, hvis han ønsker at 80% af eleverne skal bestå?

Opgave 9 (Gravide kvinder)

British Medical Journal, Vol. 307, 24. juli 1993, side 234, rapporterer om en undersøgelse af 5459 gravide kvinder, som benyttede Aarhus Universitets Hospital. Middeltallet for graviditetsperioden var 281,9 dage med en spredning på 11,4 dage. Bestem den procentdel af fødslerne, som resulterede i for tidligt fødte børn (< 258 dage), under forudsætning af, at graviditetsperiodens længde er normalfordelt.

Opgave 10 (Chi-i-anden fordelingen)

Måske har du tidligere arbejdet med Chi-i-anden fordelingen. I så fald kan det oplyses, at denne fordeling er nært beslægtet med normalfordelingen. For tilfældet med 1 frihedsgrad er Chi-i-anden fordelingen simpelthen bare lig med den stokastiske variabel $Y = X^2$, hvor X er standardnormalfordelt. Man kan vise, at Y har følgende tæthedsfunktion:

$$f_Y(x) = \begin{cases} \dfrac{1}{\sqrt{2\pi}} \cdot x^{-1/2} \cdot e^{-x/2} & \text{for } x > 0 \\ 0 & \text{for } x \le 0 \end{cases}$$

a) Tegn grafen for denne tæthedsfunktion for $0 < x \le 10$.
b) Bestem sandsynligheden $P(Y \le 1{,}5)$.
c) Bestem 0,5-fraktilen for Chi-i-anden fordelingen. *Hjælp*: Bestem ved løsning af en ligning den værdi for x, som opfylder $P(Y \le x) = 0{,}5$.

Bemærkning: Chi-i-anden fordelingen benyttes blandt andet til at undersøge for statistisk uafhængighed, og spiller en vigtig rolle i statistikken.

Opgave 11 (Egenskaber for normalfordelingens tæthedsfunktion)

Side 8 blev betydningen af parameteren μ omtalt.

a) Vis at grafen for tæthedsfunktionen for en normalfordeling med $\mu = 0$ og vilkårlig spredning σ er symmetrisk omkring y-aksen.
b) Vis, at hvis man parallelforskyder grafen for tæthedsfordelingen for den normalfordeling, som er omtalt i a), med μ i x-aksens retning, så fås grafen for tæthedsfunktionen for den generelle normalfordeling med parametre μ og σ.
c) Benyt a) og b) til at argumentere for, hvorfor grafen for tæthedsfunktionen for den generelle normalfordeling med parametre μ og σ. er symmetrisk omkring den lodrette linje $x = \mu$.

Hjælp: a) vis at $f(x) = f(-x)$ for alle $x \in R$. Hvilken grafisk betydning har det? b) Vis, at hvis man udskifter x med $x - \mu$ i forskriften for tæthedsfunktionen, så bevirker det en parallelforskydning af grafen med μ i x-aksens retning …

Opgave 12

Massen af det aktive stof i nogle piller fremstillet på en fabrik viser sig at være normalfordelt med middelværdi 6,00 g og spredning 0,045 g.

a) Hvor mange procent af pillerne har en vægt på under 5,92 gram?
b) Hvor stor en del har en vægt på mellem 5,95 og 6,02 gram?
c) Hvor stor en del af pillerne har en vægt på over 6,15 gram?