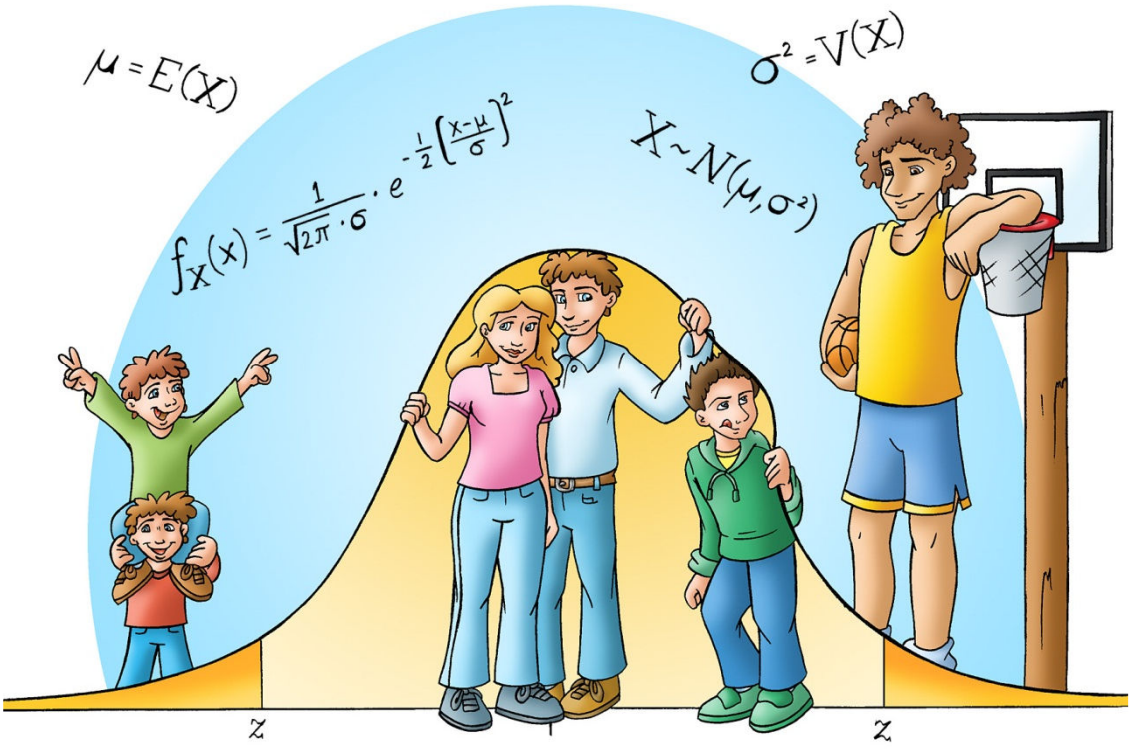
# Normalfordelingen



En ekstra opgave tilføjet 25.04.21

## Billedliste

Side 14: ©iStock.com/Jacob Wackerhausen (Kvinde_IQ)

Side 15: ©iStock.com/Elenathewise (Mand ved maskine)

Side 17: Christian Albrecht Jensen [Public domain], via Wikimedia Commons (Carl Friedrich Gauss).

Side 17: ©iStock.com/traveler1116 (Pierre-Simon Laplace)

Side 18: ©iStock.com/leonid_tit (Appelsin-plantage)

Side 22: ©iStock.com/KhotenkoVolodymyr (Vægtlodder)

Side 36: Pudelek [CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0)] (Den kongelige livgarde på Amalienborg)

Side 36: ©iStock.com/NiPlot (Glas med piller)

Side 37: ©iStock.com/pressdigital (Bananer)

Side 38: ©iStock.com/XiXinXing (To basketboldspillere)

# Forord

Denne note er beregnet til at dække den del om *normalfordelingen*, som mangler i forhold til matematik A STX i 2017-reformen, set i forhold til min e-bog *Sandsynlighedsregning og statistik, B-niveau STX* (se [1]). Vi forudsætter, at læseren allerede har stiftet bekendtskab med den indledende sandsynlighedsregning, begrebet *stokastisk variabel* og specielt arbejdet med *binomialfordelingen*.

# 1. Kontinuert stokastisk variabel

En *binomialfordelt* stokastisk variabel $X$ med antalsparameter $n$ kan som bekendt antage værdierne $0, 1, \ldots, n$ – altså *endeligt* mange forskellige værdier. En stokastisk variabel, som kun kan antage endeligt mange eller eventuelt tælleligt mange forskellige værdier kaldes for en *diskret* stokastisk variabel. I denne note skal vi betragte en stokastisk variabel, som kan antage alle værdier i $R$, altså mængden af reelle tal. Der er tale om den såkaldte *normalfordelte* stokastisk variabel, som er et eksempel på en *kontinuert* stokastisk variabel. Det viser sig, at kontinuerte stokastiske variable skal behandles noget anderledes end de diskrete. Særligt bestemmes sandsynligheden for en hændelse forskelligt.

**Definition 1**

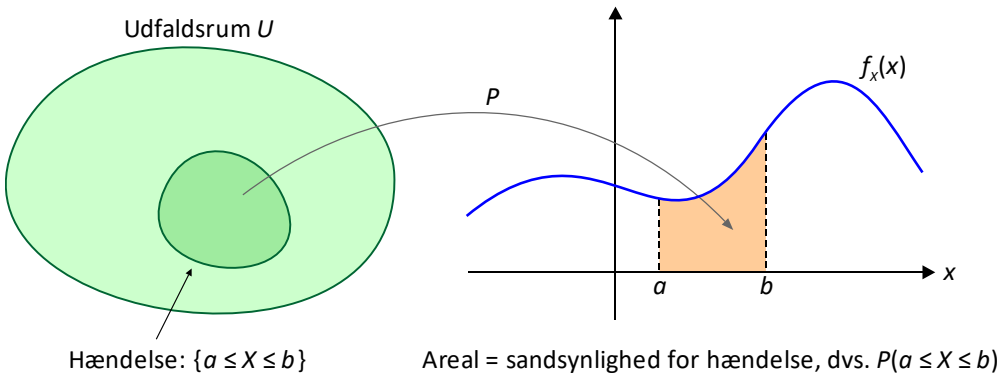
Lad $X$ være en funktion fra et udfaldsrum $U$ ind i mængden af reelle tal. Funktionen $X$ siges da at være en *kontinuert stokastisk variabel*, hvis der findes en ikke-negativ, integrabel funktion $f_X(x)$ med egenskaben

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

for alle reelle tal $a$ og $b$ med $a < b$, eventuelt også $\pm\infty$. Funktionen $f_X(x)$ kaldes *tæthedsfunktionen* eller *frekvensfunktionen* 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 betegnes *cumulative distribution function* (ofte 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 for en kontinuert stokastisk variabel.

**Sætning 2**

For en kontinuert stokastisk variabel $X$ med tæthedsfunktion $f_X$ gælder følgende:

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

*Bevis*: Følger direkte af (1) samt at $P(U) = 1$.

□

**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) $F_X$ er en voksende funktion.

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

*Bevis*: a), b) og c) følger umiddelbart af (1), (2) og (3). Hvad angår d), så følger den af integralregningens fundamentalsætning.

□

**Bemærkninger 4**

Det skal bemærkes, at *punktsandsynligheder* af formen $p(X = a)$ ikke er tilladte i forbindelse med kontinuerte stokastiske variable. Havde man tilladt det, skulle sandsynligheden ifølge (1) have været 0, da $a = b$ og integralet derfor giver 0. Men det er i princippet muligt, at $X$ kan antage værdien $a$, selv om sandsynligheden for det er "uendeligt lille". Man ser desuden, at det er ligegyldigt, om man bruger $\le$ eller $<$ i beregningen af sandsynligheder $(a < b)$:

(4) $$P(a \le X \le b) \;=\; P(a \le X < b) \;=\; P(a < X \le b) \;=\; P(a < X < b)$$

□

Man kan også tale om middelværdi, varians og spredning for en kontinuert fordelt stokastisk variabel. I tilfældet med en diskret stokastisk variabel som for eksempel i [1] side 30 er størrelserne defineret ved hjælp af *summer*. Når der er tale om kontinuerte stokastiske variable er summerne udskiftet med integraler, som vist i definition 5 på næste side.

**Definition 5**

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

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

og variansen og spredningen er defineret ved henholdsvis

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx \tag{6}$$

$$\sigma(X) = \sqrt{Var(X)} \tag{7}$$

forudsat at integralerne giver mening.

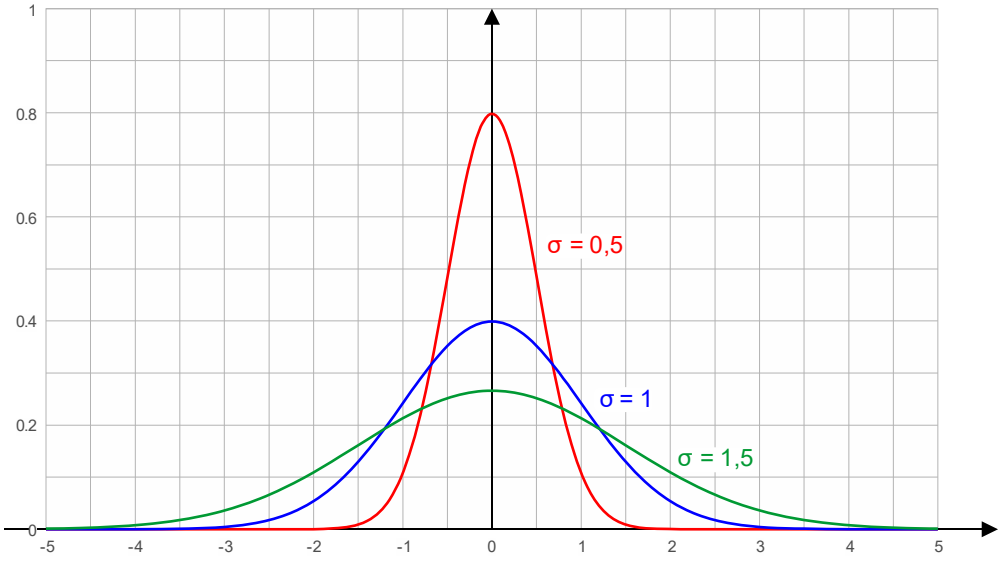
## 2. Normalfordelingen

Normalfordelingen er en kontinuert fordeling. Den har to parametre, nemlig $\mu$ og $\sigma$, som vi senere skal se er henholdsvis middelværdi og spredning for fordelingen (se sætning 6 og appendiks A). Med notationen $X \sim N(\mu,\sigma)$ vil vi mene, at $X$ er en normalfordelt stokastisk variabel med parametre $\mu$ og $\sigma$. Dens tæthedsfunktion ser således ud:

$$f_{\mu,\sigma}(x) = \frac{1}{\sigma\cdot\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{8}$$

Nedenfor er grafen for tæthedsfunktionen afbildet for tre forskellige værdier af $\sigma$, mens $\mu = 0$ i alle tre tilfælde.

Graferne for tre tæthedsfunktioner for normalfordelingen (alle $\mu = 0$)



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 (8). 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.

Ifølge definition 1 side 5 får vi fordelingsfunktionen $F_{\mu,\sigma}(x)$ ved at integrere tæthedsfunktionen fra $-\infty$ til $x$:

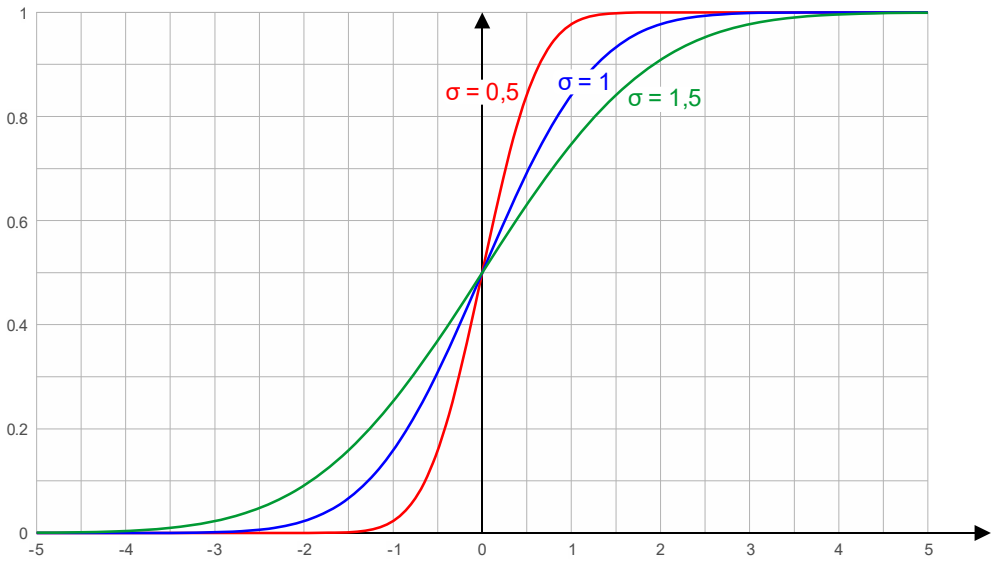
$$(9) \qquad F_{\mu,\sigma}(x) = P(X \le x) = \int_{-\infty}^{x} f_{\mu,\sigma}(z)dz = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-\frac{1}{2}\cdot\left(\frac{z-\mu}{\sigma}\right)^2} dz$$

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. Sætning 3d) giver os straks, at hvis man differentierer fordelingsfunktionen, så fås tæthedsfunktionen:

$$(10) \qquad F'_{\mu,\sigma}(x) = f_{\mu,\sigma}(x)$$

Det er egentligt bare en anvendelse af integralregningens fundamentalsætning på (9). De tre tæthedsfunktioner på figuren på forrige side har fordelingsfunktioner, hvis grafer er S-formede. De ser således ud:

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




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 $\Phi$:

$$(11) \qquad \Phi(x) = F_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{1}{2}\cdot z^2} dz$$

Som sætning 6 herunder viser, så er der en snæver sammenhæng mellem en normalfordeling med generelle parametre $\mu$ og $\sigma$ og så standardnormalfordelingen. I tidligere tider var det meget vigtigt, for så behøvede man kun én tabel med de kumulerede sandsynligheder for standardnormalfordelingen. Sammenhængen er dog stadig væsentlig, både af teoretiske grunde og til opgaver.

**Sætning 6**

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

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

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

c) Parametrene $\mu$ og $\sigma$ i en normalfordeling angiver henholdsvis middelværdi og spredning for fordelingen.

*Bevis*:

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

(12)
$$P(X \le x_0) = P(\sigma \cdot Z + \mu \le x_0) = P\left(Z \le \frac{x_0-\mu}{\sigma}\right)$$
$$= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{(x_0-\mu)/\sigma} e^{-\frac{1}{2}\cdot z^2}\,dz = \frac{1}{\sigma\cdot\sqrt{2\pi}} \cdot \int_{-\infty}^{x_0} e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}\,dx$$

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 $x$-grænser $-\infty$ og $x_0$. Sidstnævnte integral i (12) viser netop, at $X$ er en normalfordelt stokastisk variabel med parametre $\mu$ og $\sigma$.

b) Fås blot ved at gentage de første dele af (12), blot med $x$ i stedet for $x_0$.

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

c) I appendiks A kan ses et bevis for påstanden om middelværdien. Et bevis for, at $\sigma$ er spredningen for den stokastiske variabel $X$, undlader vi.

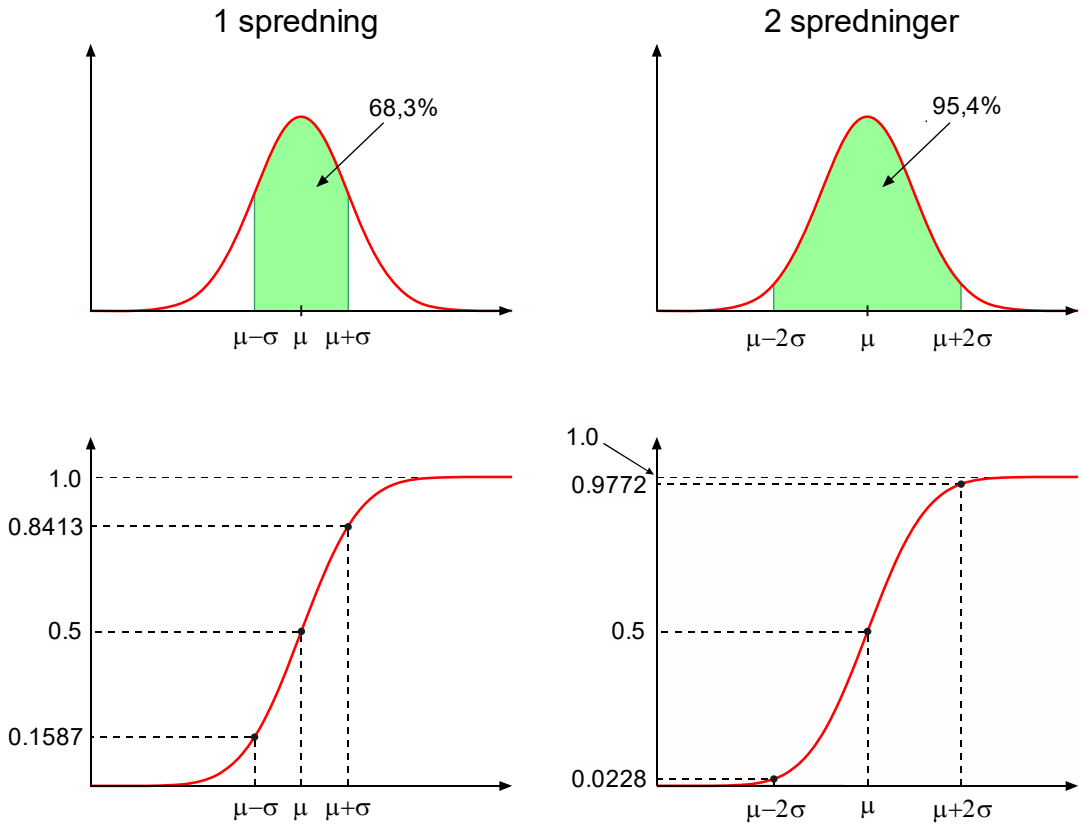
□

Spredningen $\sigma$ kaldes også for *standardafvigelsen*. Man kan stille sig selv 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 6b):

$$
\begin{aligned}
P(\mu-\sigma \le X \le \mu+\sigma) &= P(X \le \mu+\sigma) - P(X \le \mu-\sigma) \\
&= \Phi\left(\frac{\mu+\sigma-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-\sigma-\mu}{\sigma}\right) \\
&= \Phi(1) - \Phi(-1) \qquad (14) \\
&= 0{,}841345 - 0{,}158655 \\
&= 0{,}682690
\end{aligned}
$$

$$
\begin{aligned}
P(\mu-2\sigma \le X \le \mu+2\sigma) &= P(X \le \mu+2\sigma) - P(X \le \mu-2\sigma) \\
&= \Phi\left(\frac{\mu+2\sigma-\mu}{\sigma}\right) - \Phi\left(\frac{\mu-2\sigma-\mu}{\sigma}\right) \\
&= \Phi(2) - \Phi(-2) \qquad (15) \\
&= 0{,}97725 - 0{,}02275 \\
&= 0{,}95450
\end{aligned}
$$

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 $\mu$ 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 $\sigma$ fra middelværdien $\mu$.



Værdier $x$ af $X$, som ligger indenfor maksimalt 2 standardafvigelser (spredninger) fra middelværdien $\mu$ kaldes *normale*. Vi ser, at sandsynligheden for et *normalt udfald* er lig med $P(\mu-2\sigma \le X \le \mu+2\sigma)=95{,}4\%$. Sandsynligheden for at få en værdi, som højst er 3 spredninger fra middelværdien, kan vi udregne på lignende vis:

$$\begin{aligned} P(\mu-3\sigma \le X \le \mu+3\sigma) &= P(X \le \mu+3\sigma)-P(X \le \mu-3\sigma) \\ &= \Phi\left(\frac{\mu+3\sigma-\mu}{\sigma}\right)-\Phi\left(\frac{\mu-3\sigma-\mu}{\sigma}\right) \\ &= \Phi(3)-\Phi(-3) \\ &= 0{,}99865-0{,}00135 \\ &= 0{,}99730 \end{aligned} \tag{16}$$

Værdier af $X$, som ligger *mere end* 3 spredninger fra middelværdien kaldes *exceptionelle* og sandsynligheden for en sådant er $1-0{,}99730=0{,}00270=0{,}27\%$.

### Eksempel 7

En producent af müsli sælger pakker med påtrykt 1000 g. Maskinerne, der påfylder indholdet, er imidlertid ikke helt nøjagtige. Nettovægten af müsli kan beskrives ved en normalfordelt stokastisk variabel $X$ med middelværdi $\mu=1000\text{ g}$ og spredning $\sigma=12\text{ g}$.

a) Angiv vægtintervallet for de normale udfald for nettovægten af müslipakker.
b) Hvad kan man sige om nettovægten for de exceptionelle udfald.
c) Bestem sandsynligheden for, at indholdet af müsli i en pakke ligger et sted mellem 988 gram og 1012 gram.

*Løsning*: Vi kan besvare opgaven uden hjælp fra et CAS-værktøj:

a) Vægtintervallet for de normale udfald er følgende, underforstået i gram:

$$\left[\mu-2\cdot\sigma,\ \mu+2\cdot\sigma\right] = \left[1000-2\cdot12,\ 1000+2\cdot12\right] = \left[976,\ 1024\right]$$

b) Vi leder her efter udfald, som er mere end 3 spredninger fra middelværdien:

$$\mu-3\cdot\sigma = 1000-3\cdot12 = 964 \quad \text{og} \quad \mu+3\cdot\sigma = 1000+3\cdot12 = 1036$$

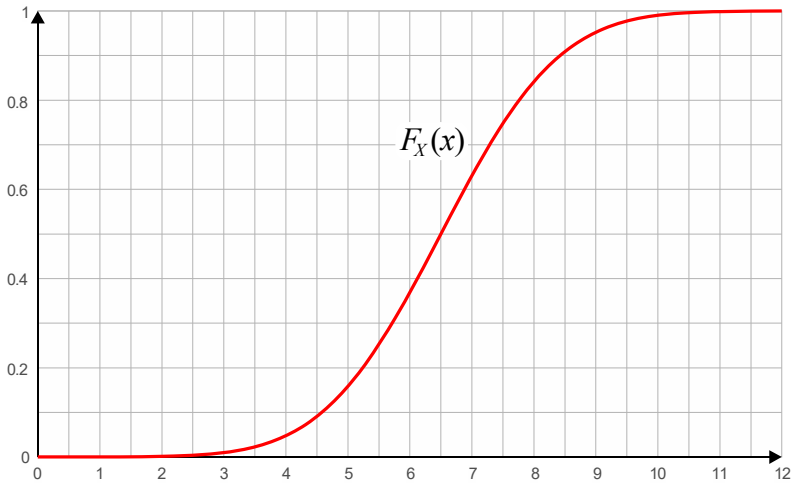
Exceptionelle udfald er altså karakteriseret ved, at nettovægten er under 964 gram eller over 1036 gram.

c) Vi ser, at grænserne netop er 1 spredning fra middelværdien. Ifølge (14) fås derfor en sandsynlighed på 68,3%.

## Eksempel 8

På figuren nedenfor er afbildet grafen for fordelingsfunktionen for en normalfordelt stokastisk variabel $X$.

a) Bestem middelværdien og spredningen for $X$.
b) Bestem grafisk sandsynlighederne $P(X \le 5{,}5)$, $P(X \ge 8{,}75)$ samt $P(5{,}5 \le X \le 6{,}5)$.

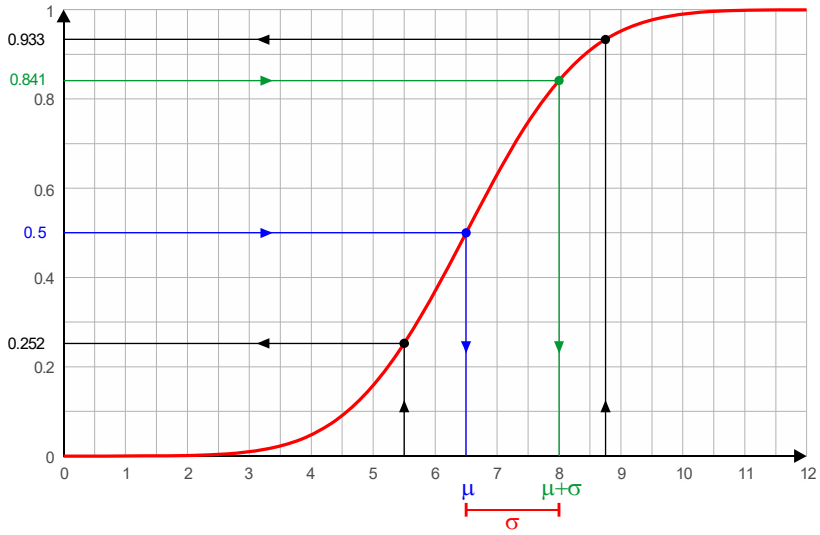



*Løsning*: Igen kan opgaven løses uden brug af CAS-værktøj.

a) På grund af symmetrien af klokkekurven omkring middelværdien $\mu$ er $F_X(\mu) = 0{,}5$, så vi får middelværdien til 6,5 ved at aflæse ud for 0,5 på 2. aksen (blå linje). Ifølge (14) eller figurerne nederst side 10 har vi:

$$F_X(\mu+\sigma) = P(X \le \mu+\sigma) = 0{,}841345\,.$$

Derfor aflæser vi ud for 0,841 på 2.aksen (grøn linje), hvilket giver $\mu+\sigma = 8$, hvorved spredningen fås til $\sigma = (\mu+\sigma)-\mu = 8-6{,}5 = 1{,}5$.




b) $P(X \le 5{,}5) = F_X(5{,}5) = 0{,}252$

$P(X \ge 8{,}75) = 1-P(X \le 8{,}75) = 1-F_X(8{,}75) = 1-0{,}933 = 0{,}067$

hvor vi har benyttet aflæsningerne markeret med sorte linjer.

$$\begin{aligned} P(5{,}5 \le X \le 6{,}5) &= P(X \le 6{,}5) - P(X \le 5{,}5) \\ &= F_X(6{,}5) - F_X(5{,}5) \\ &= 0{,}5 - 0{,}252 \\ &= 0{,}248 \end{aligned}$$

□

### Eksempel 9

En stokastisk variabel $X$ er normalfordelt $X \sim N(20,3)$.

a) Bestem intervallet for de normale udfald.
b) Bestem $P(X \le 17)$.
c) Opskriv det integral, som bestemmer sandsynligheden $P(X \le 15)$.

*Løsning*: Igen kan opgaven løses uden det store regneværktøj.

a) Vi har $\mu = 20$ og $\sigma = 3$. Intervallet for de normale udfald er dem, som holder sig indenfor 2 spredninger til hver side af middelværdien:

$$\left[\mu - 2\cdot\sigma,\, \mu + 2\cdot\sigma\right] = \left[20 - 2\cdot 3,\, 20 + 2\cdot 3\right] = \left[14,\, 26\right]$$

b) Vi ser, at 17 svarer til $\mu - \sigma$. Derfor fås ifølge (14) eller figurerne nederst side 10:

$$P(X \le 17) = 0{,}1587$$

c) Ifølge (9) haves:

$$P(X \le 15) = \int_{-\infty}^{15} f_{20,3}(z)\,dz = \frac{1}{3\cdot\sqrt{2\pi}} \cdot \int_{-\infty}^{15} e^{-\frac{1}{2}\cdot\left(\frac{z-20}{3}\right)^2}\,dz$$

hvorved det ønskede integral er opstillet.

□

I det følgende skal vi se på eksempler på forskellige opgavetyper i forbindelse med normalfordelinger, hvor et CAS-værktøj er nødvendigt.

### Eksempel 10 (Tæthedsfunktion og fordelingsfunktion)

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

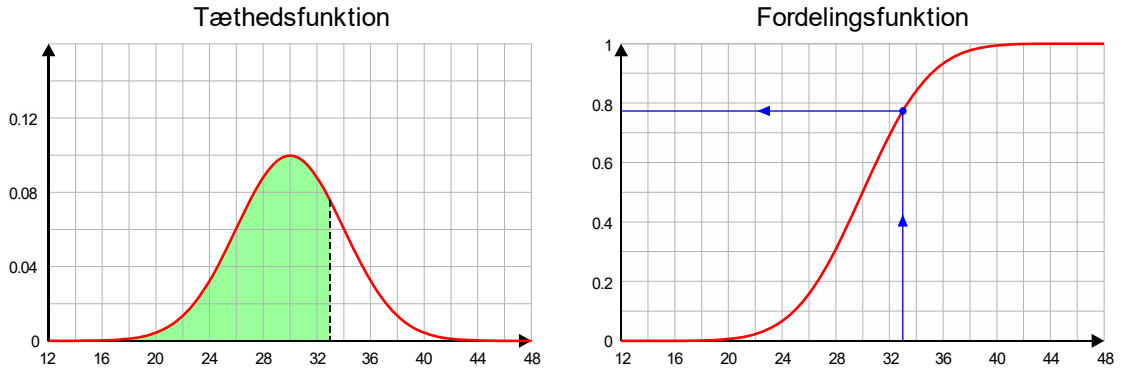
*Løsning*: Der er to måder at løse denne opgave på. Enten kan man bestemme sandsynligheden som arealet under grafen for tæthedsfunktionen fra $-\infty$ til 33, eller også kan man blot bestemme fordelingsfunktionens værdi i 33. Vil vi bruge tæthedsfunktionen, fås ifølge (9) følgende for $\mu = 30$ og $\sigma = 4$:

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

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:

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

Har man derimod kun en tabel med værdier for fordelingsfunktionen for standardnormalfordelingen, så kan man gøre brug af sætning 6b). Det vil vi dog ikke vise her. Rent grafisk kan situationen med de to løsninger afbildes således:




□

**Eksempel 11** (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:

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

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.

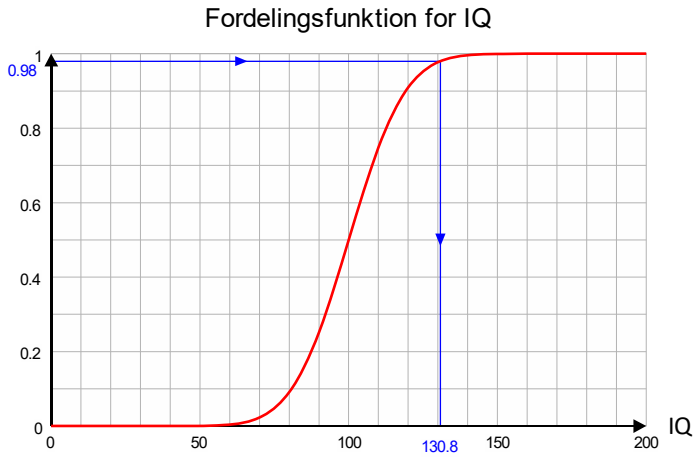
c) Vi skal bestemme $x$ således, at $P(X \ge x) = 0{,}02$. Heraf får vi, at værdien af fordelingsfunktionen i $x$ er lig med 0,98:

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

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:

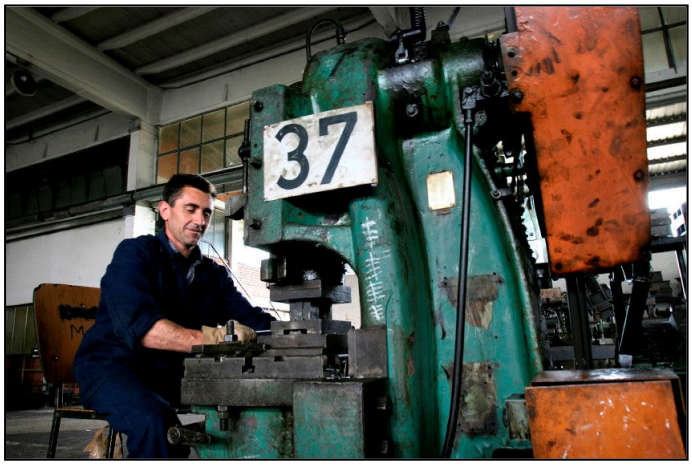
$$x \;=\; F^{-1}_{100,15}(0{,}98) \;=\; 130{,}81$$

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




## Eksempel 12 (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 0,10 mm. Fabrikanten kan acceptere en afvigelse på maksimalt 0,18 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?

*Løsning*:

a) Vi skal altså finde sandsynligheden for at diameteren enten er over 20,18 mm eller under 19,82 mm. Igen bruger vi tricket med den modsatte hændelse til $X > 20{,}18$.

$$\begin{aligned} P(X<19{,}82)+P(X>20{,}18) &= P(X<19{,}82)+1-P(X\le 20{,}18) \\ &= F_{20,0.10}(19{,}82)+1-F_{20,0.10}(20{,}18) \\ &= 0{,}0359+1-0{,}9641 \\ &= 0{,}0719 \end{aligned}$$

Vi ser, at 7,2% af cylindrene må kasseres.

b) Vi skal bestemme $\sigma$, så $P(X<19{,}82)+P(X>20{,}18)=0{,}02$. Da normalfordelingen er symmetrisk, har vi $P(X\le 19{,}82)=0{,}01$. Igen udregner vi fraktiler ud, ligesom i eksempel 11. Man må løse en ligning med hensyn til den ubekendte $\sigma$.

$$P(X\le 19{,}82)=0{,}01 \quad\Leftrightarrow\quad F_{20,\sigma}(19{,}82)=0{,}01 \quad\Leftrightarrow\quad \sigma=0{,}0774$$

Man skal altså reducere spredningen i produktionen fra 0,10 til 0,077, 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 6b) samt gå omvendt ind i en tabel, som indeholder værdier for fordelingsfunktionen for standardnormalfordelingen $\Phi$. Det svarer til at løse følgende ligning, idet $\mu=20$ og $x=19{,}82$:

$$\Phi\left(\frac{19{,}82-20}{\sigma}\right)=0{,}01 \quad\Leftrightarrow\quad \frac{19{,}82-20}{\sigma}=\Phi^{-1}(0{,}01)=-2{,}3263 \quad\Leftrightarrow\quad \sigma=0{,}0774$$

□

## Eksempel 13

Om en normalfordelt stokastisk variabel $X$ gælder $P(X\le 6)=0{,}46$ og $P(X\le 9)=0{,}88$. Bestem middelværdi og spredning for $X$.

*Løsning*: Hvis man forsøger at løse to ligninger med de to ubekendte $\mu$ og $\sigma$, direkte via fordelingsfunktionerne i et CAS-værktøj, kan man nemt rende ind i problemer. Her kommer sætning 6b) derimod til sin ret. Med den kan man nemlig simplificere ligningerne.

$$P(X\le 6)=0{,}67 \;\Leftrightarrow\; \Phi\left(\frac{6-\mu}{\sigma}\right)=0{,}67 \;\Leftrightarrow\; \frac{6-\mu}{\sigma}=\Phi^{-1}(0{,}67) \;\Leftrightarrow\; \frac{6-\mu}{\sigma}=-0{,}100434$$

$$P(X\le 9)=0{,}88 \;\Leftrightarrow\; \Phi\left(\frac{9-\mu}{\sigma}\right)=0{,}88 \;\Leftrightarrow\; \frac{9-\mu}{\sigma}=\Phi^{-1}(0{,}88) \;\Leftrightarrow\; \frac{9-\mu}{\sigma}=1{,}1749868$$

hvor $\Phi^{-1}$ er den inverse funktion til fordelingsfunktionen for standardnormalfordelingen. Løses de to ligninger med to ubekendte i et CAS-værktøj fås $\mu=6{,}24$ og $\sigma=2{,}35$.

### Bemærkning 14

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 12), 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 et 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.

## Historie

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 udviklinen 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)

# 3. Normalfordelt data

I dette afsnit skal vi se, hvordan man kan vurdere, om et sæt data er normalfordelt eller ej. Først ser vi på tilfældet med *grupperet data*, derefter tilfældet med *rådata*. Vi tager udgangspunkt i eksempler.

### Eksempel 15 (Grupperet data)

En appelsinavler i Spanien ønsker at undersøge, om appelsinernes vægt er normalfordelt. Derfor udtrækker en assistent på tilfældig vis en stikprøve på 120 appelsiner. Avleren modtager data grupperet i vægtintervaller på 5 gram, som tabellen nedenfor viser. Frekvenserne og de kumulerede frekvenser er udregnet. Den sidste kolonne kommer vi ind på lidt senere. Intervallerne er underforstået at være i enheden gram.

| Intervaller | Hyppighed | Frekvens $f$ | Kumuleret frekvens $p$ | $\Phi^{-1}(p)$ |
|---|---|---|---|---|
| ]125, 130] | 1 | 0,00833 | 0,00833 | -2,3940 |
| ]130, 135] | 2 | 0,01667 | 0,02500 | -1,9600 |
| ]135, 140] | 2 | 0,01667 | 0,04167 | -1,7317 |
| ]140, 145] | 8 | 0,06667 | 0,10833 | -1,2354 |
| ]145, 150] | 12 | 0,10000 | 0,20833 | -0,8122 |
| ]150, 155] | 17 | 0,14167 | 0,35000 | -0,3853 |
| ]155, 160] | 26 | 0,21667 | 0,56667 | 0,1679 |
| ]160, 165] | 17 | 0,14167 | 0,70833 | 0,5485 |
| ]165, 170] | 15 | 0,12500 | 0,83333 | 0,9674 |
| ]170, 175] | 10 | 0,08333 | 0,91667 | 1,3830 |
| ]175, 180] | 5 | 0,04167 | 0,95833 | 1,7317 |
| ]180, 185] | 2 | 0,01667 | 0,97500 | 1,9600 |
| ]185, 190] | 3 | 0,02500 | 1,00000 | |
| | 120 | | | |

Ifølge den deskriptive statistik kan *middelværdien* udregnes som vist her:

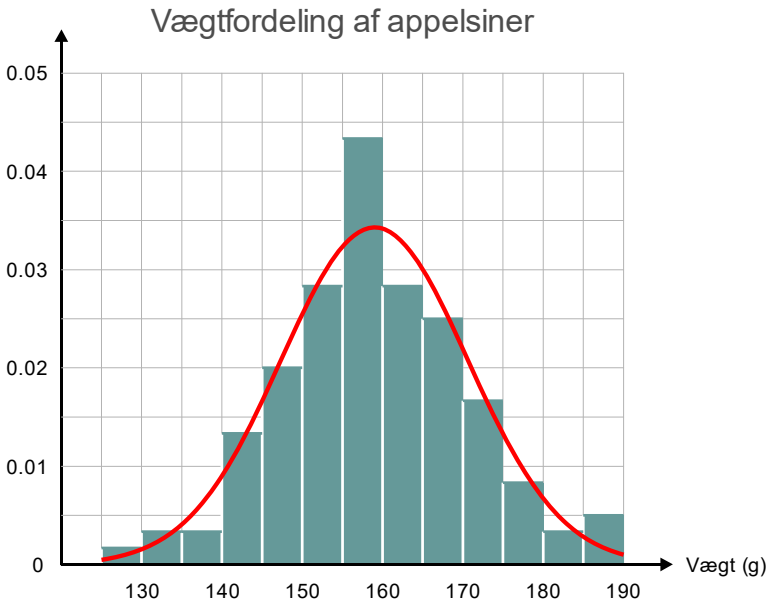
$$\begin{aligned}\mu &= m_1 \cdot f_1 + m_2 \cdot f_2 + \ldots + m_n \cdot f_n \; = \\ &= 127{,}5 \cdot 0{,}00833 + 132{,}5 \cdot 0{,}01667 + \ldots + 187{,}5 \cdot 0{,}02500 \\ &= 159{,}000\end{aligned}$$

og *spredningen* udregnes således:

$$\begin{aligned}\sigma &= \sqrt{(m_1-\mu)^2 \cdot f_1 + (m_2-\mu)^2 \cdot f_2 + \ldots + (m_n-\mu)^2 \cdot f_n} \\ &= \sqrt{(127{,}5-159)^2 \cdot 0{,}00833 + (132{,}5-159)^2 \cdot 0{,}01667 + \ldots + (187{,}5-159)^2 \cdot 0{,}02500} \\ &= 11{,}630\end{aligned}$$

hvor $m_i$ er midtpunktet af det *i*'te observationsinterval og $f_i$ er den *i*'te frekvens. En mulighed er at sammenligne *histogrammet* hørende til data i tabellen med grafen for tæthedsfunktionen for normalfordelingen med parametre $\mu = 159$ og $\sigma = 11{,}630$.

**Figur 1**




Det ser fornuftigt ud i dette tilfælde, men i andre kan det være svært at vurdere, om klokkekurven tilnærmer histogrammet tilstrækkeligt præcist. Man kunne også plotte *sumkurven* hørende til tabellens data med grafen for normalfordelingens fordelingsfunktion. Vi vil ikke gøre det her. Det er lidt bedre, men allerbedst er det at "rette sumkurven og den S-formede graf for fordelingsfunktionen ud", så man kommer til at afgøre, om noget ligger tæt på en ret linje eller ej. Ifølge sætning 6b) gælder der nemlig, at

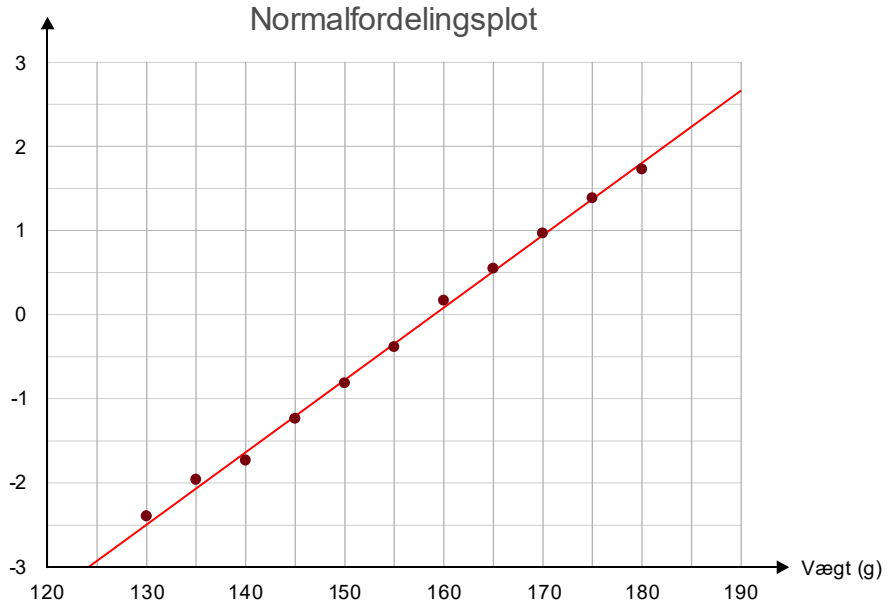
$$F_{\mu,\sigma}(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

hvor $\Phi$ er fordelingsfunktionen for standardnormalfordelingen. Tager vi den inverse til den funktion på begge sider, får vi:

(17) $$\Phi^{-1}\left(F_{\mu,\sigma}(x)\right) = \frac{x-\mu}{\sigma} = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$$

Så tager vi den inverse funktion til standardnormalfordelingen til alle de kumulerede frekvenser – det der er gjort i sidste kolonne i tabellen på forrige side – og plotter dem op imod de højre endepunkter af observationsintervallerne, så skal punkterne altså ligge på en ret linje, hvis data er normalfordelt! Dette er gjort på figur 2 herunder, hvor linjen med ligningen $y = 1/\sigma \cdot x - \mu/\sigma$ også er tilføjet. Vi ser, at punkterne ligger tæt og pænt fordelt omkring linjen. Det betyder, at vi godtager, at appelsinerne er normalfordelte.

**Figur 2**




Operationerne kan gennemføres i de fleste CAS-værktøj eller for eksempel i regnearksprogrammet Excel.

□

### Eksempel 16 (Rådata)

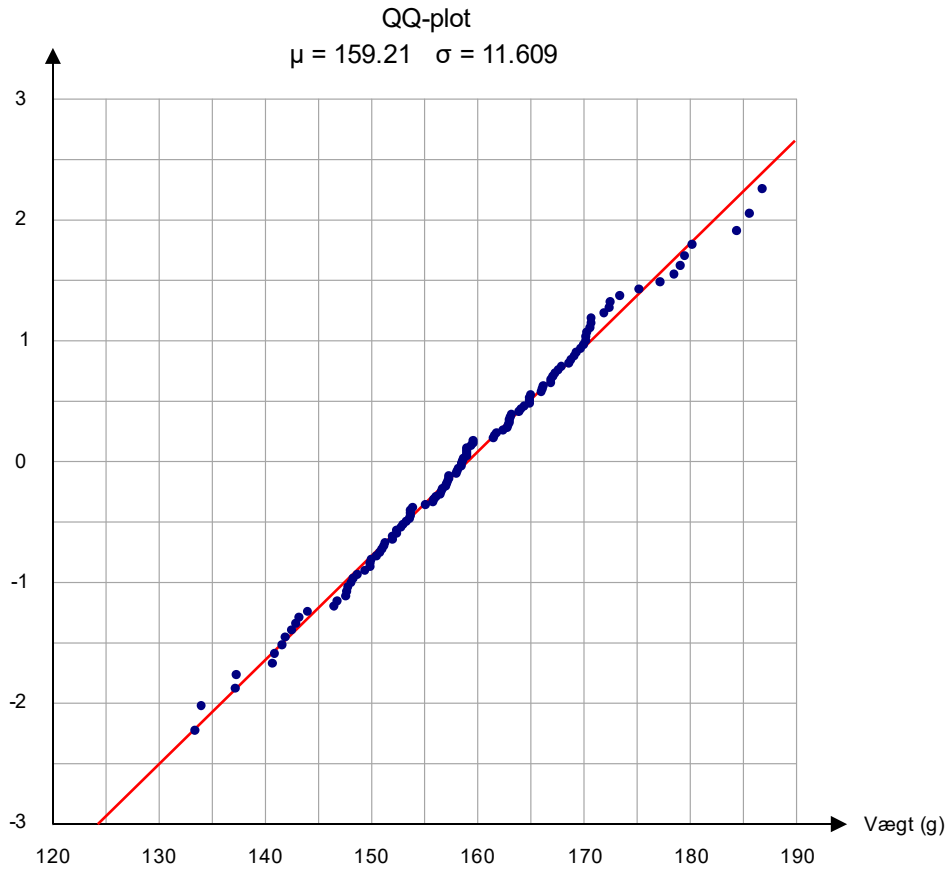
Lad os sige, at appelsinavleren ikke blot har de grupperede data til rådighed, men endda *rådata* i form af samtlige 120 appelsinvægte, som vist herunder i gram med 1 decimals nøjagtighed.

148.3, 153.6, 157.1, 167.3, 152.4, 170.0, 149.9, 171.9, 147.7, 170.2, 148.1, 146.8, 164.9, 166.0, 155.9, 159.0, 162.4, 141.6, 150.0, 169.7, 159.6, 153.0, 168.8, 150.8, 147.8, 163.2, 152.4, 148.7, 166.9, 156.6, 127.8, 180.2, 177.2, 165.0, 170.6, 159.0, 157.0, 161.5, 167.9, 153.7, 158.1, 140.7, 179.5, 164.9, 164.9, 159.0, 157.2, 155.8, 163.0, 167.6, 169.3, 163.1, 166.1, 170.3, 156.7, 185.6, 149.9, 186.8, 158.6, 184.4, 133.4, 167.1, 164.4, 161.8, 172.5, 159.6, 166.9, 153.3, 170.7, 189.6, 179.1, 173.4, 152.0, 158.5, 156.1, 151.3, 163.0, 151.2, 163.9, 161.6, 150.5, 143.2, 151.0, 140.9, 156.5, 166.2, 158.5, 162.9, 147.6, 164.1, 170.2, 159.0, 153.7, 134.0, 137.3, 155.1, 149.4, 153.9, 144.0, 146.5, 141.9, 142.9, 158.7, 137.2, 168.6, 152.0, 153.7, 162.8, 169.1, 158.0, 170.7, 157.3, 142.5, 159.4, 172.4, 152.8, 178.5, 158.2, 175.2, 157.3.

For at gøre noget lignende som i eksempel 12, skal vi altså have fat i fordelingsfunktionen for de ugrupperede data – også kaldet den *empiriske fordelingsfunktion*. Dens graf er en trappefunktion. Det gør situationen anderledes i forhold til det grupperede tilfælde. Der kommer noget valgfrihed ind. Det viser sig fornuftigt her at lave et såkaldt *QQ-plot*, som

undertiden også betegnes et *fraktil-fraktil-plot* eller *probit-plot* på dansk. Vi skal ikke gå i detaljer med sagen her, blot nævne, at de fleste CAS-værktøj har en facilitet eller kommando, som kan udføre det, når der fodres med en liste af målinger. Igen handler det om at vurdere, om et sæt af punkter (ét punkt for hver måling) ligger tilstrækkeligt tæt omkring en linje.

Figur 3




Det skal lige nævnes, at man som estimater for μ og σ bruger henholdsvis den *empiriske middelværdi* $\bar{x}$ og *stikprøvespredningen* $s$, givet ved

$$\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \quad , \quad s = \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2} \tag{18}$$

CAS-værktøjet vil som regel også angive værdierne for disse. Punkterne ligger tæt omkring linjen i figur 3, så rådata tyder snildt på, at appelsinernes vægt er normalfordelt. Den interesserede læser kan se flere detaljer om QQ-plots i Appendiks B.

□

**Eksempel 17** (Normalfordelte unøjagtigheder i produktionen)

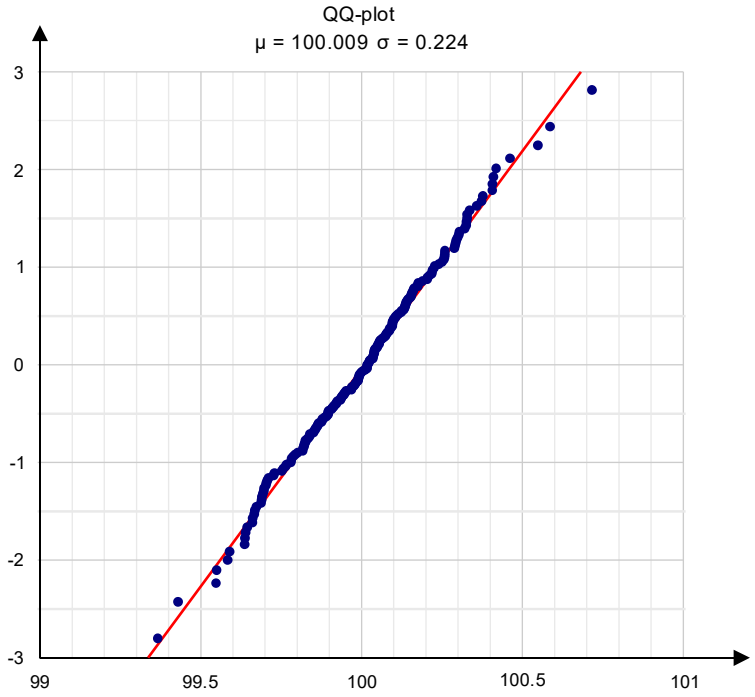
Det er tidligere omtalt, at komponenter i produktionen ofte har unøjagtigheder, og at fejlene ofte er normalfordelte. I dette tilfælde er der tale om et firma, som producerer lodder til undervisningsbrug. Hensigten er at producere lodder med vægten 100 gram. I en kontrol bliver der udtrukket en stikprøve på 200 lodder med henblik på at undersøge varia-

tionen i loddernes vægt og om de er normalfordelte. En Excel fil med de 200 værdier for loddernes vægte sendes til en statistiker, som importerer data i sit CAS-værktøj.

| Lodvægt (g) | 99.824 | 100.006 | ... | 100.094 | 99.695 | 99.927 |
|---|---|---|---|---|---|---|

Dele af data fra filen eks17_lodvaegte.xlsx

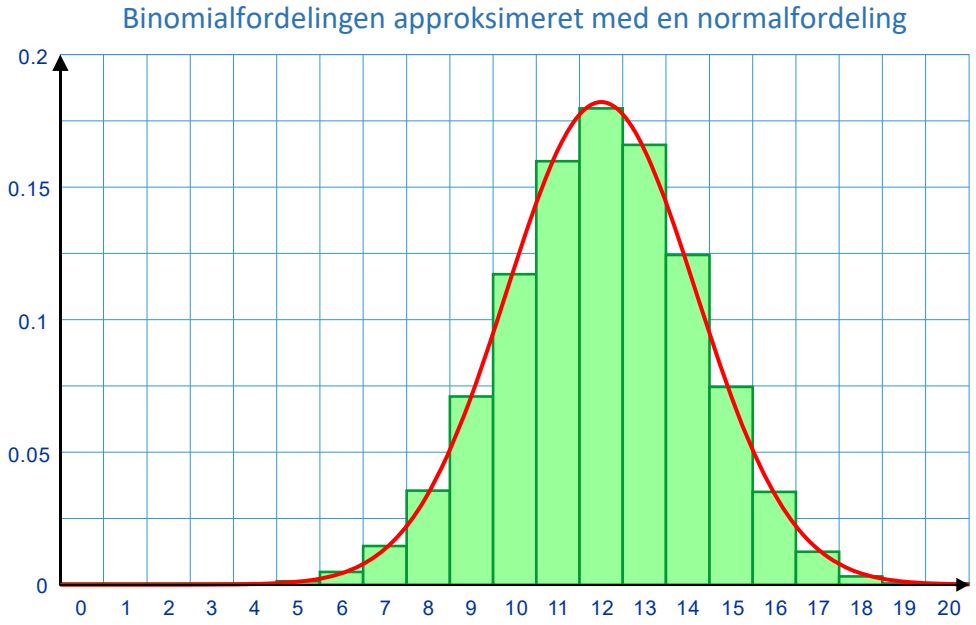
Heri laver personen et QQ-plot af lodvægtene, som giver resultatet herunder. Da punkterne ligger meget pænt på linje, godtager vi, at loddernes vægte er normalfordelte. Desuden får vi oplyst, at den empiriske middelværdi er 100,009 gram og at stikprøvespredningen er 0,224 gram.




□

# 4. 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 14. 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

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

Moreover, if $s$ 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.

# Appendiks A. Middelværdien

Vi har udskudt det tekniske bevis for første del af sætning 6c) til dette appendiks.

**Sætning A1**

Lad $X$ være en normalfordelt stokastisk variabel med parametrene $\mu$ og $\sigma$. Parameteren $\mu$ angiver middelværdien for $X$, altså $E(X)=\mu$.

*Bevis*: Ifølge definition 5 og (8) fås følgende:

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

hvor vi for at få andet lighedstegn har benyttet integration ved substitution. Vi har sat: $z=(x-\mu)/\sigma \Leftrightarrow x=\sigma\cdot z+\mu$, hvoraf $dx=\sigma\cdot dz$. Det femte lighedstegn: Der er ganget ind i parentesen. Det sjette lighedstegn: Integralet er delt op i to integraler. Det syvende lighedstegn: konstanter er ganget ud foran integralet. Det ottende lighedstegn: I det første integral er integranden en *ulige* funktion. Når der integreres fra $-\infty$ til $\infty$, fås derfor 0. I parentesen i det andet led integreres tæthedsfunktionen for standardnormalfordelingen fra $-\infty$ til $\infty$, og den ved vi giver 1 (sætning 2). Heraf ses, at middelværdien for $X$ er lig med $\mu$.

□

# Appendiks B. Forstå QQ-plot

Af og til får man brug for at afgøre, om noget givent datamateriale i form af en stikprøve er normalfordelt. Man kunne selvfølgelig tænke sig at gruppere data i intervaller og derefter se, om det tilhørende histogram kan fittes med en "klokkekurve", altså grafen for tæthedsfunktionen for normalfordelingen med middelværdi $\overline{x}$ og spredning $s$, jf. betegnelserne i Appendiks A. Det er imidlertid ikke altid nemt at afgøre, om et sådant fit er tilstrækkelig godt. Det samme ville være tilfældet, hvis man havde anvendt de kumulerede frekvenser og forsøgt at tilnærme sumkurven med grafen for fordelingsfunktionen for normalfordelingen. Det er nemmere at afgøre, om noget er omtrent *lineært*. Derfor anvender man ofte et såkaldt *QQ-plot*, eller som det også kaldes, et *fraktilplot*. Man ser undertiden også anvendt betegnelsen *probit-plot* på dansk. Vi holder os dog til betegnelsen QQ-plot. Q er en forkortelse for *Quantile* på engelsk. Det betyder *fraktil*.

Lad os kigge på et eksempel med henblik på at forstå, hvad et QQ-plot er. Vi tænker os foretaget en stikprøve med 10 elementer:

25,8 20,8 35,4 23,3 38,0 25,8 31,8 28,8 30,1 42,0

Vi kan eventuelt starte med at bestemme stikprøvens *empiriske middelværdi* $\overline{x}$ samt *stikprøvespredningen* $s$, som er omtalt i Appendiks A:

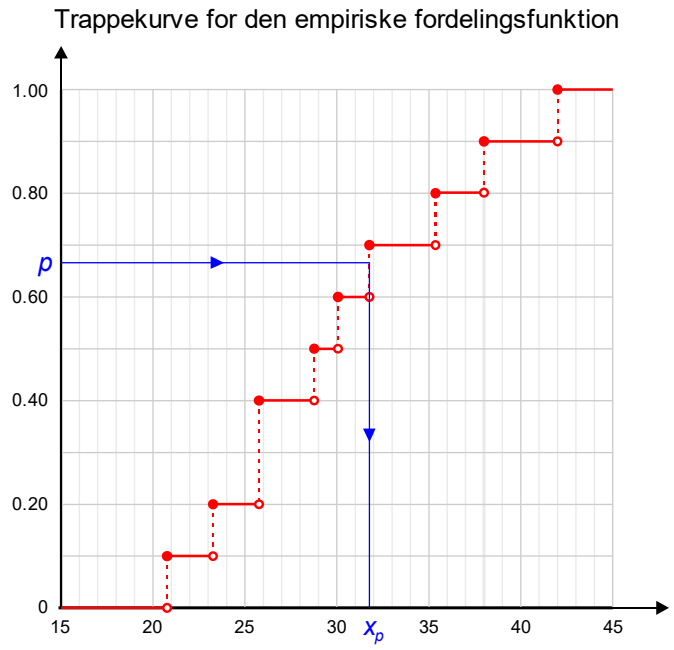
$$\begin{aligned}\overline{x} &= \frac{1}{n}\cdot\sum_{i=1}^{n} x_i \\ &= \frac{25{,}8+20{,}8+35{,}4+23{,}3+38{,}0+25{,}8+31{,}8+28{,}8+30{,}1+42{,}0}{10} \\ &= 30{,}180\end{aligned}$$

$$\begin{aligned}s &= \sqrt{\frac{1}{n-1}\cdot\sum_{i=1}^{n}(x_i-\overline{x})^2} \\ &= \sqrt{\frac{(25{,}8-30{,}180)^2+(20{,}8-30{,}180)^2+(35{,}4-30{,}180)^2+\ldots+(42{,}0-30{,}180)^2}{10-1}} \\ &= 6{,}7193\end{aligned}$$

Men vi mangler at vurdere, om stikprøven med rimelighed kan tænkes at komme fra en normalfordeling. For at vurdere det, vil vi først tage fat i stikprøvens *empiriske fordelingsfunktion*. Den er defineret meget formelt i en ramme på næste side. Lad os se på stikprøven med de $n=10$ elementer ovenfor og se, hvordan fordelingsfunktionen ser ud i dette tilfælde. For ethvert $x$ skal vi finde ud af, hvor mange af stikprøvens elementer eller observationer, som er mindre end eller lig med $x$. Derefter skal vi dividere med antal elementer, her 10. Så har vi $F_{emp}(x)$. For at klare dette er det hensigtsmæssigt at sætte elementerne i voksende rækkefølge:

20,8 23,3 25,8 25,8 28,8 30,1 31,8 35,4 38,0 42,0

Med denne sortering er det nemmere at overskue. Lad os se på $x = 33$. Vi ser hurtigt, at der i den sorterede række af elementer 7 elementer, som er mindre end eller lig med 33. Derfor haves: $F_{emp}(33) = \frac{7}{10}$. Sådan skal vi i princippet gøre for alle $x$-værdier. Man indser hurtigt, at det giver anledning til en *trappefunktion* (*stykvis lineær funktion*). Hver gang man kommer til en af de nye sorterede observationer, vil funktionsværdien vokse med et helt multiplum af $1/n$. Vi ser, at observationen 25,8 står anført to gange i den sorterede række. Derfor vil fordelingsfunktionen her vokse med $2/10 = 0{,}2$. Alle de øvrige observationer forekommer kun én gang, så her vil fordelingsfunktionen kun vokse med $1/10 = 0{,}1$. Grafen for fordelingsfunktionen bliver derfor en *trappekurve*:




**Definition B1** (En stikprøves empiriske fordelingsfunktion)

Den *empiriske fordelingsfunktion* hørende til stikprøven bestående af observationerne $x_1, x_2, \ldots, x_n$ er defineret ved

(B1) $$F_{emp}(x) = \frac{\#\{i \mid x_i \le x\}}{n}$$

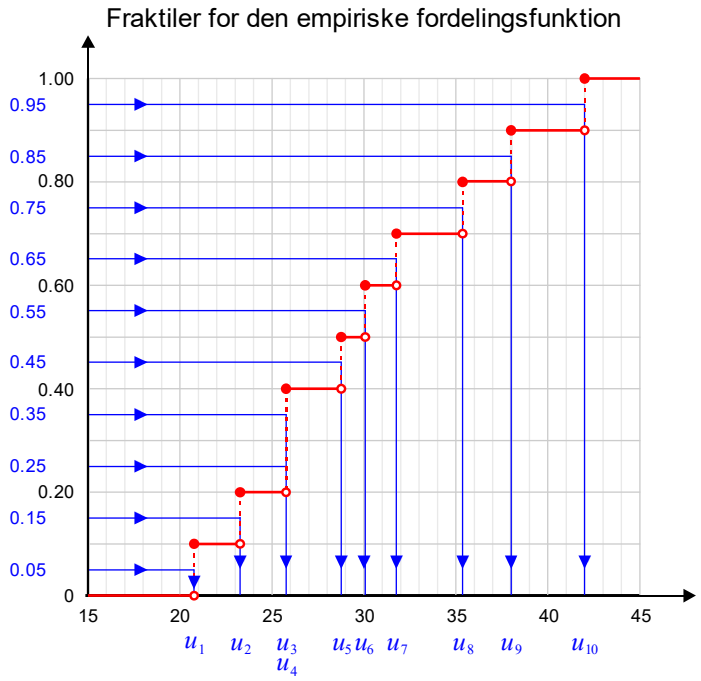
hvor # betyder "antallet af".

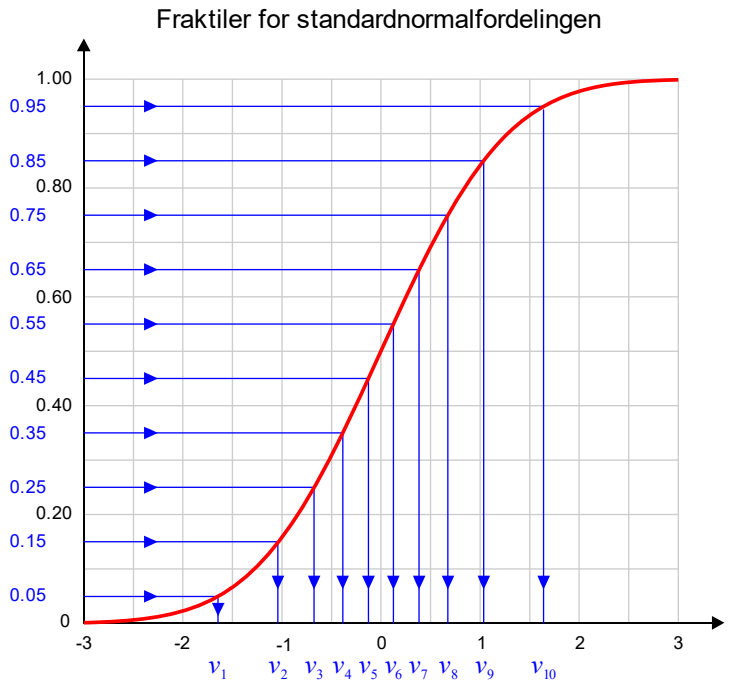
For ethvert $p$ opfyldende $0 < p < 1$ definerer vi $p$-fraktilen for den empiriske fordelingsfunktion som den værdi $x_p$, som fremkommer ved at gå op til $p$ på 2. aksen i plottet ovenfor, vandret hen til trappekurven og lodret ned på 1. aksen. Hvis den vandrette linje rammer den stiplede linje mellem to trin, så skal $x_n$ være den værdi, som ligger umiddelbart under denne lodrette stiplede linje. Det er vist med et eksempel med blå linjer på figuren ovenfor. Hvis den vandrette linje derimod rammer direkte ind på et trin, er der tvetydighed. Så kan man vælge $x$-koordinaten for ethvert punkt på det vandrette trin, fx midtpunktet. Da vi imidlertid ikke får brug for det i det følgende, behøver vi ikke foretage

et valg her. For at afgøre, om stikprøven har en normalfordeling, vil vi afbilde et fraktilsæt for standardnormalfordelingen som funktion af det tilsvarende fraktilsæt for den empiriske fordelingsfunktion. Her skal man vælge sig en række $p$-værdier. Det er almindeligt at vælge lige så mange $p$-værdier som antallet af elementer i stikprøven, dvs. $n$, og man bruger ofte følgende:

(B2) $$p_i = \frac{i-\frac{1}{2}}{n}, \quad i = 1, 2, \ldots, n$$

I vores eksempel giver det følgende værdier: $p_1 = 0{,}05;\ p_2 = 0{,}15; \ldots;\ p_n = 0{,}95$.



På de to grafer på forrige side, er aflæst kvartilerne for $p_1, p_2, \ldots, p_{10}$ for henholdsvis den empiriske fordelingsfunktion og standardnormalfordelingen. Det giver de markerede fraktilsæt, henholdsvis $u_1, u_2, \ldots, u_{10}$ og $v_1, v_2, \ldots, v_{10}$. Bemærk i øvrigt, at sidstnævnte serie kan beregnes ved at anvende den inverse funktion til fordelingsfunktionen for standardnormalfordelingen på sandsynlighederne:

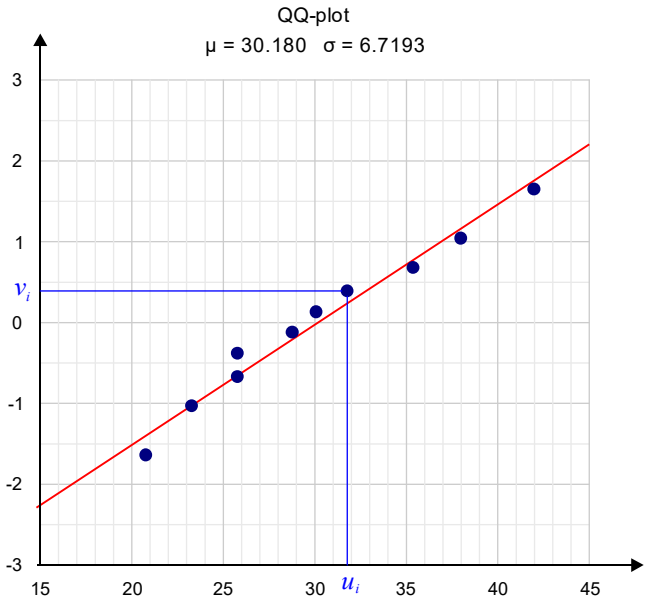
(B3) $$v_i = \Phi^{-1}(p_i)$$

QQ-plottet består da af punkterne $(u_1, v_1), (u_2, v_2), \ldots, (u_{10}, v_{10})$.

Lad os et øjeblik antage, at observationerne i stikprøven virkelig var omtrent normalfordelte med middelværdi μ og spredning σ, og at vi i stedet for standardnormalfordelingen havde anvendt den generelle normalfordeling med middelværdi μ og spredning σ i definitionen af QQ-plottet. Da ville punkterne i QQ-plottet ligge tæt på diagonalen $y = x$ i koordinatsystemet! Men heldigvis er der en snæver lineær sammenhæng mellem standardnormalfordelingen og den generelle normalfordeling. Vi har nemlig

(B4) $$X \sim N(\mu, \sigma) \Leftrightarrow Z \sim N(0,1) \quad \text{hvor } Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma} \cdot X - \frac{\mu}{\sigma}$$

Derfor kan vi nøjes med at anvende standardnormalfordelingen i definitionen af QQ-plottet. Med dette valg vil man normalt ikke få punkter omkring diagonalen, men om en anden linje. Det vil ofte være ret nemt visuelt at afgøre, om punkterne ligger omtrent på linje, og i det tilfælde er det en god indikation af, at man har at gøre med en normalfordeling. I øvrigt vil et QQ-plot også typisk indeholde linjen $y = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$, hvor man har valgt den empiriske middelværdi $\bar{x}$ og stikprøvespredningen $s$ som estimater for henh. μ og σ. Det er bedre end at udføre lineær regression på punkterne i QQ-plottet!




## Bemærkning B2

Vores stikprøve var i det anvendte eksempel kun på 10 elementer – for at gøre situationen overskuelig. Det vil almindeligvis være for få observationer til at kunne afgøre om opsamlet data er normalfordelt. Tilfældigheder vil for nemt kunne spille ind.

# Opgaver

### Opgave 1 (Tæthedsfunktionens graf)

a) Tegn grafen for tæthedsfunktionen $f_{20,5}$ for normalfordelingen med $\mu = 20$ og $\sigma = 5$ i vinduet $0 \le x \le 40,\ 0 \le y \le 0{,}1$. Benyt udtrykket (8) for tæthedsfunktionen.

b) Gentag a), hvor du bruger CAS-værktøjets kommando for tæthedsfunktionen.

c) Vis at arealet under grafen fra $-\infty$ til $\infty$ er lig med 1, som skal være opfyldt ifølge sætning 2.

### Opgave 2 (Grafer for tæthedsfunktion og fordelingsfunktion med mere)

Betragt normalfordelingen med middelværdi $\mu = 60$ og spredning $\sigma = 2{,}5$.

a) Tegn grafen for tæthedsfunktionen $f_{60,2.5}$ i vinduet $50 \le x \le 70,\ 0 \le y \le 0{,}2$.

b) Tegn grafen for fordelingsfunktionen $F_{60,2.5}$ i vinduet $50 \le x \le 70,\ 0 \le y \le 1$.

c) Bestem $F_{60,2.5}(63)$ og giv en fortolkning af dette tal. NB! Passer værdien med det, du kan aflæse på grafen i b)?

d) Udregn fraktilen $F_{60,2.5}^{-1}(0,75)$ og giv en fortolkning af tallet. NB! Passer værdien med det, du kan aflæse på grafen i b)?

*Hjælp*: Det er meget tænkeligt, at kommandoen til tæthedsfunktionen hedder noget med *pdf* (for Probability Density Funktion), og at fordelingsfunktionen hedder noget med *cdf* (Cumulative Distribution Function) i dit CAS-værktøj. Led også efter den inverse funktion til fordelingsfunktionen, som eventuelt kan hedde noget med *inv*.

### Opgave 3 (Forstå tæthedsfunktionen og fordelingsfunktionen)

Betragt normalfordelingen med middelværdi $\mu = 8{,}8$ og spredning $\sigma = 1{,}9$. Du skal bestemme sandsynligheden $P(8 \le X \le 11)$ på to måder: dels som et integral af tæthedsfunktionen, dels ved hjælp af fordelingsfunktionen.

a) Benyt (1) i definition 1 til at bestemme sandsynligheden ved at integrere tæthedsfunktionen fra 8 til 11. Sandsynligheden svarer til arealet under grafen i det omtalte interval. Tegn om muligt samtidigt grafen for tæthedsfunktionen med det relevante område under grafen skraveret.

b) Bestem den omtalte sandsynlighed via fordelingsfunktionen samt sætning 3a). Tegn desuden grafen for fordelingsfunktionen og vis, hvad det svarer til grafisk.

### Opgave 4 (Standardnormalfordelingen vs den generelle normalfordeling)

Sætning 6b) giver en sammenhæng mellem standardnormalfordelingens fordelingsfunktion $\Phi$ og fordelingsfunktionen $F_{\mu,\sigma}$ for den generelle normalfordeling med middelværdi $\mu$ og spredning $\sigma$. Tag eksemplet $\mu = 14$ og $\sigma = 3$. Vis at $F_{\mu,\sigma}(x)$ og $\Phi((x-\mu)/\sigma)$ giver det samme for to selvvalgte værdier af $x$.

### Opgave 5 (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 tæthedsfunktionen for den i spørgsmål a) omtalte normalfordeling med μ i *x*-aksens retning, så fås grafen for tæthedsfunktionen for den generelle normalfordeling med parametre μ og σ.

*Hjælp*: a) vis at $f_{0,\sigma}(x) = f_{0,\sigma}(-x)$ 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 6 (Opgave uden hjælpemidler)

En stokastisk variabel *X* er normalfordelt $X \sim N(20, 4)$. Besvar nedenstående uden brug af CAS-værktøj. Formelsamlingen eller siderne 10-11 i denne note er nok.

a) Angiv intervallet for de normal udfald.

b) Angiv intervallerne for de exceptionelle udfald.

c) Bestem $P(12 \le X \le 28)$.

d) Bestem $P(X \le 24)$.

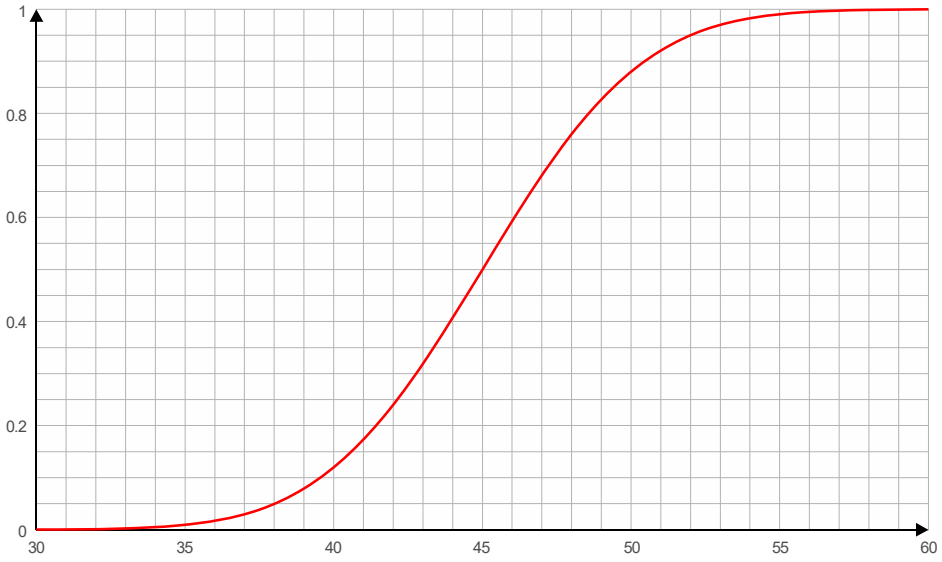
### Opgave 7 (Opgave uden hjælpemidler)

På figuren nedenfor er afbildet grafen for fordelingsfunktionen for en normalfordelt stokastisk variabel *X*.

a) Bestem middelværdien og spredningen for *X*.

b) Bestem grafisk sandsynligheden $P(X \le 43)$.

c) Bestem $P(44 \le X \le 52)$.



## Opgave 8 (Opgave uden hjælpemidler)

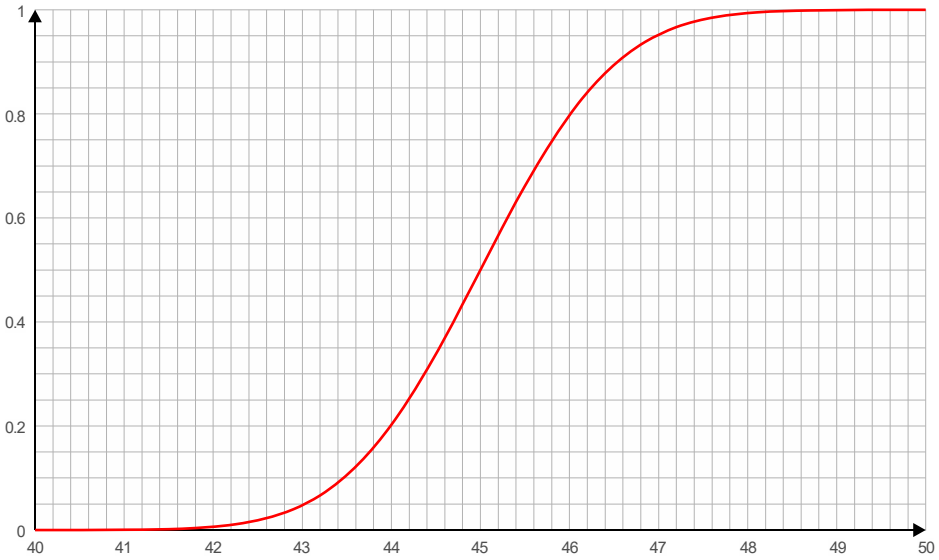
På figuren nedenfor er afbildet grafen for fordelingsfunktionen for en normalfordelt stokastisk variabel $X$.

a) Bestem middelværdien og spredningen for $X$.

b) Bestem grafisk følgende sandsynligheder:

$$P(X \le 44,2),\ P(X \ge 43,0) \text{ og } P(43,0 \le X \le 45,8)$$

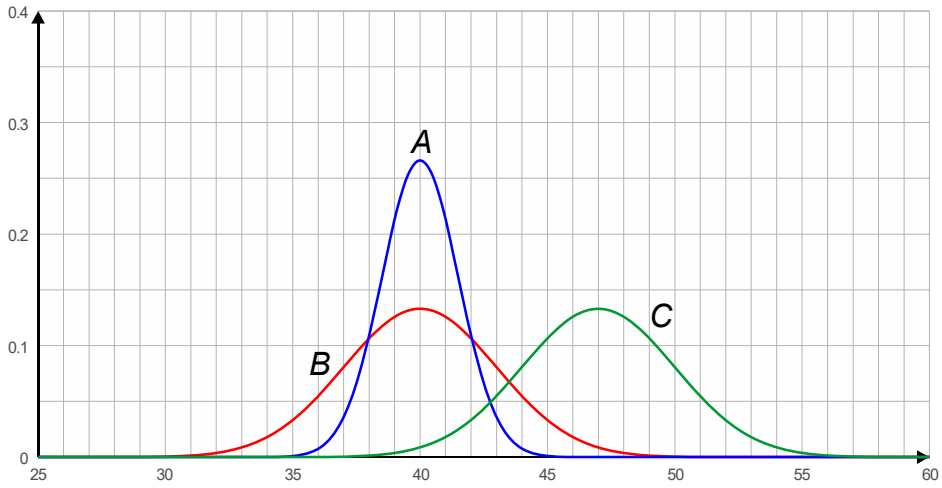
c) Bestem $k$ således, at $P(X = k) = 0,65$. Værdien af $k$ kaldes for 65%-fraktilen.




## Opgave 9 (Opgave uden hjælpemidler)

Figuren nedenfor viser graferne for tæthedsfunktionerne for tre normalfordelte stokastiske variable.

a) Gør rede for hvilken en af de tre grafer, som hører til tæthedsfunktionen for den normalfordelte stokastiske variabel med middelværdi 40 og spredning 1,5.



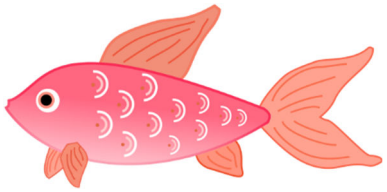
### Opgave 10 (Opgave uden hjælpemidler)

En stokastisk variabel $X$ er normalfordelt $X \sim N(47,5)$. Besvar nedenstående uden brug af CAS-værktøj. Formelsamlingen eller siderne 10-11 i denne note er nok.

a) Angiv intervallet for de normal udfald.
b) Angiv intervallerne for de exceptionelle udfald.
c) Bestem $P(42 \le X \le 52)$.

### Opgave 11 (Opgave uden hjælpemidler)

Fisk af en bestemt art har en længde, som er normalfordelt med middelværdi 35 cm og spredning 4,5 cm.



a) Angiv intervallet for de normale længder for fisk af den pågældende art.

En fisk af den pågældende art har længden 48 cm, mens en anden har længden 20 cm.

b) Afgør, for hver af de to fisks vedkommende, om der er tale om et normalt eller exceptionelt udfald eller ingen af delene.
c) Bestem sandsynligheden for, at en tilfældigt udtrukket fisk har en længde mellem 26 cm og 44 cm.

### Opgave 12

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 $P(X \le 19)$ ved at benytte tæthedsfunktionen som i eksempel 10. 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 13

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 14

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*: Se eksempel 13.

### Opgave 15 (Variation i produktionen)

Vi har tidligere i eksempel 12 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 $\mu$, for at mængden af kartoner med for lidt mælk i bliver færre end 3%? *Hjælp*: Opstil en ligning indeholdende $\mu$ som ubekendt og løs den.
e) Prøv at løse samme spørgsmål som i d) ved hjælp af sætning 6b).

### Opgave 16 (Points i test)

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

### Opgave 17 (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 18** (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 ($\leq 257$ dage), under forudsætning af, at graviditetsperiodens længde er normalfordelt.

**Opgave 19** (Medicin)

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

**Opgave 20** (Variation i industrien)

En maskine, som producerer komponenter til industrien, leverer metalskiver, hvis diameter er normalfordelt med en middeldiameter på 72,0 mm og en spredning på 1.2 mm.

a) Hvor mange procent af skiverne har en diameter, som afviger med mere end 2 mm fra middeldiameteren?

Producenten vil justere maskinens nøjagtighed, så procenten fra a) reduceres til 2%. Det indebærer en reduktion i spredningen i produktionen af metalskiver.

b) Hvad skal spredningen $\sigma$ reduceres til?

**Opgave 21** (Er det grupperede data normalfordelt?)

På en bananplantage ønsker man at undersøge, om bananernes vægt er normalfordelt og hvad middelvægten og spredningen i vægt er. På tilfældigvis udtrækkes 300 bananer, som vejes. Statistikeren modtager data grupperet i vægtintervaller på 10 gram.

| Intervaller | Hyppighed | Frekvens | Kumuleret frekvens $p$ | $\Phi^{-1}(p)$ |
|---|---|---|---|---|
| ]60, 70] | 2 | | | |
| ]70, 80] | 1 | | | |
| ]80, 90] | 4 | | | |
| ]90, 100] | 5 | | | |
| ]100, 110] | 10 | | | |
| ]110, 120] | 13 | | | |
| ]120, 130] | 17 | | | |
| ]130, 140] | 32 | | | |
| ]140, 150] | 26 | | | |
| ]150, 160] | 35 | | | |
| ]160, 170] | 33 | | | |
| ]170, 180] | 20 | | | |
| ]180, 190] | 24 | | | |
| ]190, 200] | 31 | | | |
| ]200, 210] | 15 | | | |
| ]210, 220] | 11 | | | |
| ]220, 230] | 10 | | | |
| ]230, 240] | 2 | | | |
| ]240, 250] | 5 | | | |
| ]250, 260] | 1 | | | |
| ]260, 270] | 2 | | | |
| ]270, 280] | 0 | | | |
| ]280, 290] | 1 | | | |

a) Udfyld de resterende kolonner i tabellen med banan data. Data er desuden vedhæftet i en Excel fil til denne pdf-fil, hvor venstre og højre endepunkt i intervallerne er angivet i hver sin søjle. Data kan enten behandles direkte i Excel regnearket, eller det kan indskrives/importeres i et CAS-værktøj og behandles derfra. Bemærk, at hvis man benytter Excel, så hedder den inverse funktion til standardnormalfordelingen på dansk `NORM.INV(p;0;1)`, hvor $p$ er den kumulerede frekvens og 0 og 1 er henholdsvis middelværdi og spredning – for standardnormalfordelingen.

b) Bestem middelværdien μ og spredningen σ for det grupperede data, eventuelt ved at bruge definitionerne, som beskrevet i eksempel 15.

c) Afbild i det samme koordinatsystem linjen med ligning

$$y = \frac{1}{\sigma} \cdot x - \frac{\mu}{\sigma}$$

sammen med datapunkterne $(h, \Phi^{-1}(p))$, hvor $h$ er intervallets højre ende og $\Phi^{-1}(p)$ er den inverse funktion til fordelingsfunktionen for standardnormalfordelingen evalueret i den tilhørende kumulerede frekvens $p$. Hvor godt ligger data i forhold til linjen? Hvis svaret er, at de ligger godt, kan vi acceptere hypotesen om, at bananernes vægt er normalfordelte.

## Opgave 22 (Er rådata normalfordelt?)

I en dansk by er 300 unge mænd indkaldt til session. Som et biprodukt af undersøgelserne er man interesseret i at undersøge, om højden af voksne mænd mon er normalfordelt, og hvad gennemsnitshøjden og variationen i højde er.

Excel-filen "opg22_hoejder.xlsx" indeholder rådata for højderne af de 300 mænd og er vedhæftet denne pdf-fil, hvorfra den kan downloades.

a) Importer data fra Excel-filen I dit CAS-værktøj og lav et QQ-plot. Afgør, om man ud fra data med rimelighed kan sige, at mændenes højder er normalfordelte?

b) Bestem den empiriske middelværdi og stikprøvespredningen for data.

I det følgende antager vi, at mændene i byen har en middelhøjde og en spredning i højde, som bestemt i b). Der udtrækkes nu på tilfældig vis en ung mand fra byen.

c) Hvad er sandsynligheden for at personen er over 190 cm høj?

d) Bestem højdeintervallet for *normale* udfald.

**Opgave 23** (To punkter på grafen for fordelingsfunktionen)

Om en normalfordelt stokastisk variabel $X$ gælder følgende:

$$P(X \le 30) = 0{,}22 \text{ og } P(X \le 55) = 0{,}78$$

Bestem middelværdi og spredning for $X$. *Hjælp*: Se eksempel 13.

**Opgave 24** (To punkter på grafen for fordelingsfunktionen)

Om en normalfordelt stokastisk variabel $X$ gælder følgende:

$$P(X \le 8{,}5) = 0{,}125 \text{ og } P(X \ge 13{,}1) = 0{,}075$$

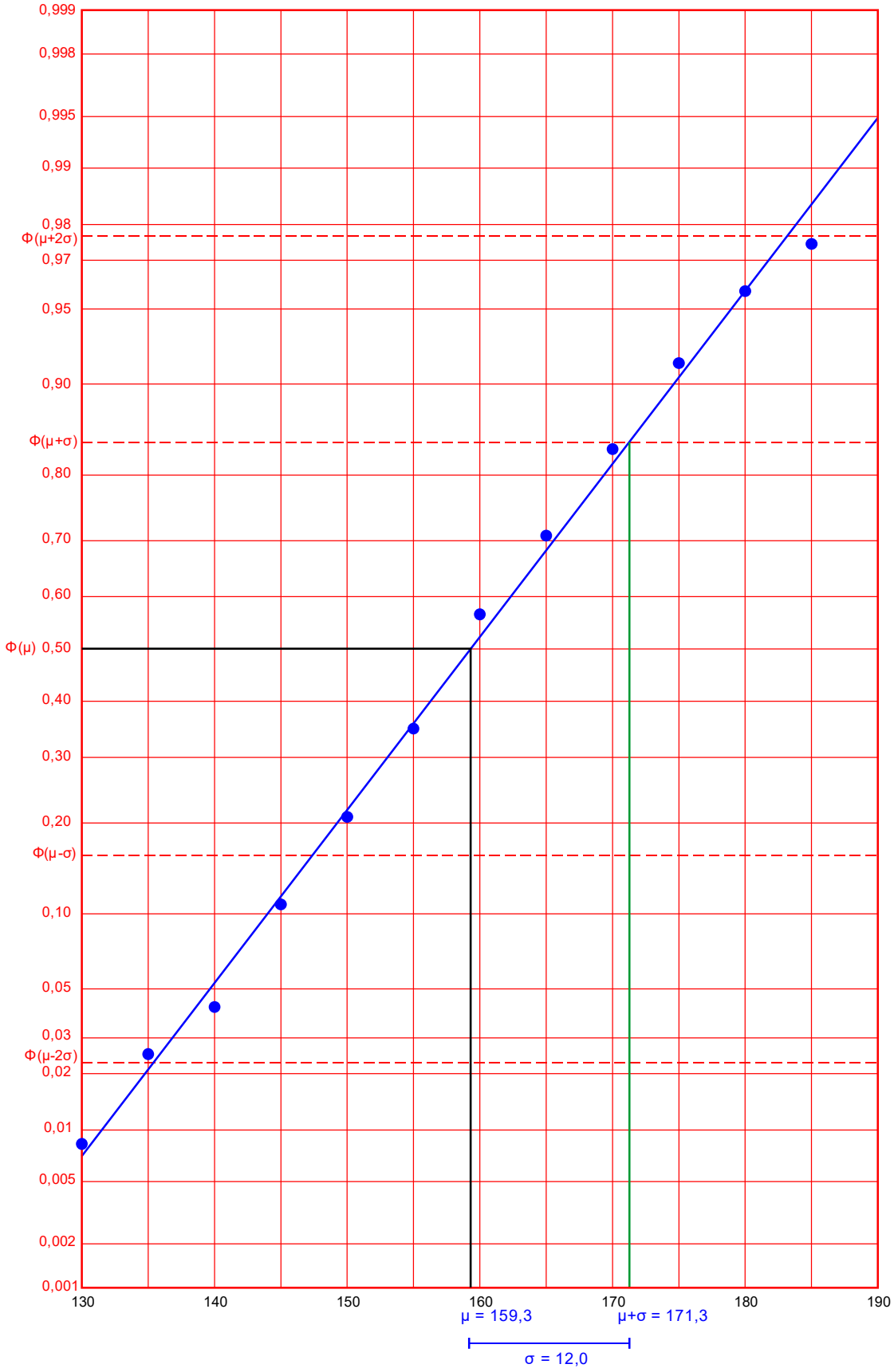
Bestem middelværdi og spredning for $X$. *Hjælp*: Se eksempel 13.

**Opgave 25** (Normalfordelingspapir)

I eksempel 15 så vi på tilfældet med grupperet data. Ved hjælp af regnearket Excel udregnede vi de kumulerede frekvenser. Vi så, at hvis man tager den inverse fordelingsfunktion for standardnormalfordelingen, $\Phi^{-1}$, til disse kumulerede frekvenser og afbilder disse værdier som funktion af de højre endepunkter i observationsintervallerne, så vil data være normalfordelt, hvis punkterne ligger på en ret linje. En anden måde at gøre det samme på, er ved at anvende et såkaldt *normalfordelingspapir*, hvor skalaen på 2. aksen er "deformeret", så det svarer til, at $\Phi^{-1}$ på forhånd er anvendt. Man skal dermed ikke selv anvende denne funktion, men kan straks indtegne punkterne.

På næste side er data fra eksempel 15 indtegnet på et simplificeret normalfordelingspapir. Et rigtigt normalfordelingspapir har underinddelinger, så man kan afsætte punkterne mere nøjagtigt. Bemærk at 1. aksen er en almindelig akse! Vi ser, at punkterne omtrent ligger på en ret linje i det pågældende papir. Derfor konkluderer vi, at data med god tilnærmelse er normalfordelt. Vi tegner efter bedste evne den bedste rette linje i forhold til datapunkterne. En anden behændighed ved dette papir er, at man nemt kan aflæse middelværdien $\mu$: Den fås ved at gå op til 0.50 på 2. aksen, vandret hen til linjen og derefter lodret ned på 1. aksen. Vi ser, at det giver $\mu = 59{,}3$. Desuden kan vi bestemme $\mu + \sigma$ på følgende måde: Gå op til $\Phi(\mu + \sigma) \approx 0{,}8413$ på 2.aksen, hen langs den stiplede vandrette linje til linjen og ned på 1. aksen. Vi ser, at det giver $\mu + \sigma = 171{,}3$. Trækker vi de to tal fra hinanden, får vi $\sigma = 171{,}3 - 159{,}3 = 12{,}0$. Altså en hurtig måde til at vurdere, om data er normalfordelt og få værdier for middelværdi og spredning. Men metoden er selvfølgelig lidt mere unøjagtig, da den beror på aflæsninger og det faktum, at man selv tegner den bedste rette linje …

En vigtig pointe: Indtegner man for normalfordelt data de kumulerede frekvenser som funktion af de højre endepunkter i observationsintervallerne i et almindeligt koordinatsystem, vil det give anledning til punkter på en S-formet kurve. som vi kender fra side 8. Men indtegnes de samme data på et normalfordelingspapir, ligger punkterne på en ret linje. Det er nemmere af afgøre, om noget ligger på en ret linje end en bestemt kurve!


0,999
0,998
0,995
0,99
0,98
Φ(μ+2σ)
0,97
0,95
0,90
Φ(μ+σ)
0,80
0,70
0,60
Φ(μ) 0,50
0,40
0,30
0,20
Φ(μ-σ)
0,10
0,05
0,03
Φ(μ-2σ)
0,02
0,01
0,005
0,002
0,001
130
140
150
160
170
180
190
μ = 159,3
μ+σ = 171,3
σ = 12,0

a) Benyt data for bananers vægt fra opgave 21. Udregn frekvenserne og de kumulerede frekvenser og afbild derefter de kumulerede frekvenser som funktion af de højre endepunkter i observationsintervallerne på et normalfordelingspapir. Kan du sige god for, at bananernes vægt er normalfordelt? Bestem desuden middelværdi og spredning. NB! Data i opgave 21 er i Excel-filen "bananer" vedhæftet denne pdf-fil.

---

# Litteratur

[1] Erik Vestergaard. *Sandsynlighedsregning og statistik, B-niveau STX*. e-bog 2020. Direkte link: https://www.matematikfysik.dk/mat/noter_tillaeg/SANDSYNLIGHEDSREGNING%20OG%20STATISTIK%20B-niveau%20stx.pdf