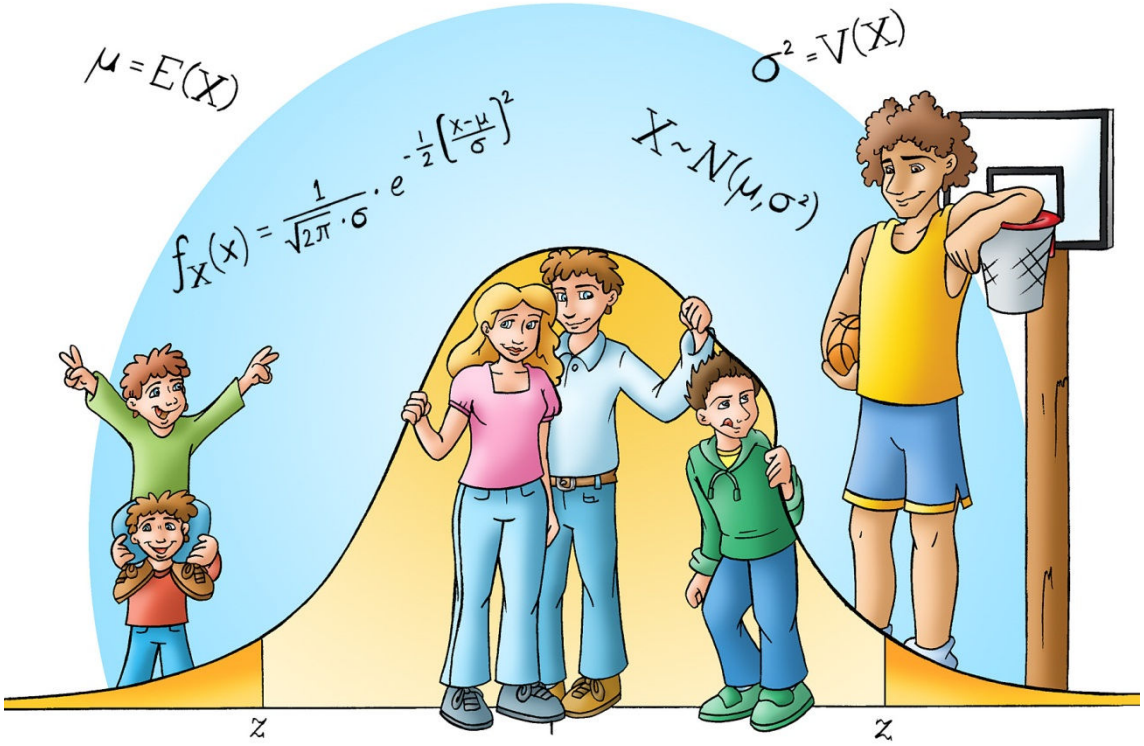
Normalfordelingen



Billeder:

Forside: jakobkramer.dk/Jakob Kramer
Side 17: ©iStock.com/Elenathewise
Side 28: ©iStock.com/jaroon
Side 29: ©iStock.com/3pod
Side 53: ©iStock.com/travellinglight
Desuden egne fotos og illustrationer.

Indholdsfortegnelse

Forsidebilledet udført af Jakob Kramer.

1. Indledning

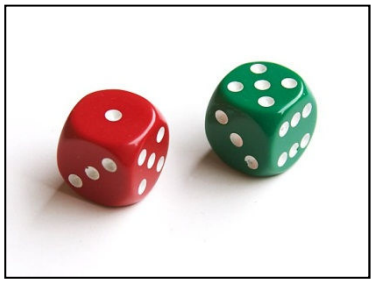
Statistik er den af de matematiske discipliner, som bliver anvendt mest intensivt i praksis, og i hjertet af statistikken og sandsynlighedsregningen ligger *normalfordelingen*. Fordelingen er på ingen måde selvindlysende. Selv om den efterhånden blev teoretisk underbygget, beror dens succes i høj grad på, at den er i stand til at beskrive mange forskellige slags data fra den virkelige verden. Mange praktiske data er med andre ord omtrent normalfordelte. Før vi kan beskrive normalfordelingen skal vi have nogle grundlæggende begreber på plads.

2. Stokastiske variable

Et *stokastisk* eksperiment er et forsøg, hvor man ikke på forhånd kan forudsige udfaldet. Der kan forekomme en række *udfald*, og mængden af disse udgør *udfaldsrummet* U. En *stokastisk variabel* X er en funktion, som til ethvert udfald $u \in U$ knytter et tal $X(u)$. Man er ofte interesseret i *sandsynlighedsfordelingen* for den stokastiske variabel, dvs. sandsynlighederne for de forskellige værdier, som X kan antage. Sandsynlighederne kan angives på lidt forskellig måde. Vi skal se på to eksempler.

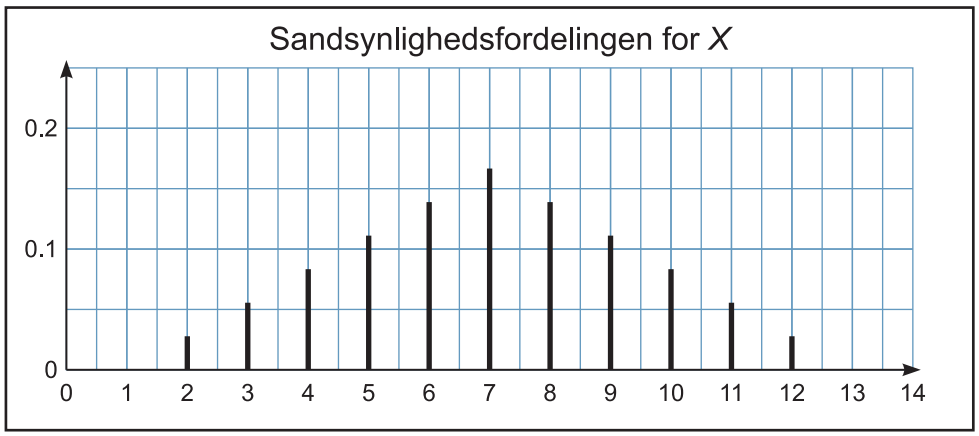
Eksempel 1

Eksperiment: Et kast med to terninger, en grøn og en rød. Antal øjne betragtes.
Udfaldsrum: $U = \{(1,1), (1,2), \ldots, (1,6), (2,1), (2,2), \ldots, (2,6), \ldots, (6,6)\}$, hvor første koordinaten i talparret angiver antal øjne for den grønne terning, mens andenkoordinaten angiver antal øjne for den røde terning.
Stokastisk variabel: X angiver summen af øjnene af de to terninger.



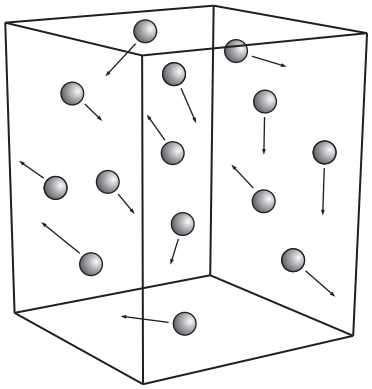
Den stokastiske variabel X kan antage værdierne fra 2 til 12. Da den kun kan antage endeligt (specielt tælleligt) mange værdier, kaldes X for en *diskret* stokastisk variabel. For at finde sandsynlighedsfordelingen for X skal vi beregne sandsynligheden for hver af dens mulige værdier. X giver 4 for følgende udfald: $(1,3), (2,2)$ og $(3,1)$. Sandsynligheden for hver af disse er som bekendt $\frac{1}{36}$. Derfor er den ønskede sandsynlighed lig med $\frac{3}{36}$. Vi skriver: $P(X=4) = \frac{3}{36}$. Sandsynlighederne for de øvrige værdier af X udregnes tilsvarende. Sandsynlighedsfordelingen kan passende beskrives i et stolpediagram, som vist på næste side. Ved hjælp af sandsynlighedsfordelingen kan man løse forskellige opgaver, for eksempel bestemme sandsynligheden for at summen af terningernes øjne *højst* er lig med 5:

$$P(X \le 5) = P(X=2) + P(X=3) + P(X=4) + P(X=5)$$
$$= \tfrac{1}{36} + \tfrac{2}{36} + \tfrac{3}{36} + \tfrac{4}{36} = \tfrac{10}{36} = \tfrac{5}{18}$$



□

Eksempel 2



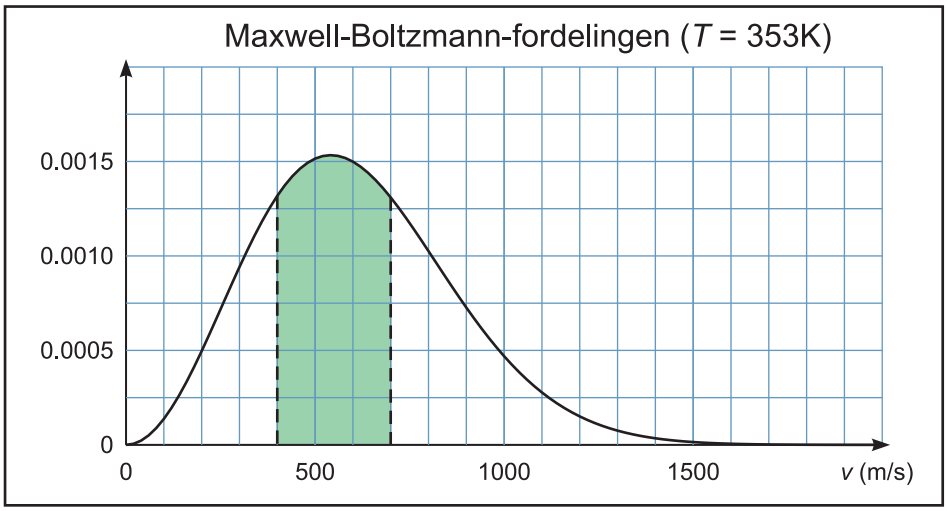
Det er velkendt, at jo højere temperaturen i en gas er, jo hurtigere bevæger gasmolekylerne sig. Her burde man egentligt sige: jo højere temperatur jo højere er *gennemsnitsfarten*, for molekylerne i en gas med en given temperatur har nemlig meget forskellig fart. Vi kan også i dette tilfælde betragte et eksperiment, et udfaldsrum og en stokastisk variabel:

Eksperiment: Et tilfældig molekyle udtages fra gassen.
Udfaldsrum: Mængden af alle molekyler i gassen med den faste temperatur T.
Stokastisk variabel: X angiver *farten* af det udtrukne molekyle.

Det viser sig, at de mulige værdier for X kan være alle tal fra 0 til uendelig, altså et helt interval. Da X antager værdier i et helt interval, har det ikke mening at forsøge at bestemme sandsynligheden for, at X er lig med en given værdi, fx 23,762 m/s. Hvis man skulle gøre det, måtte svaret være, at sandsynligheden er uendelig lille for, at X er eksakt lig 23,762 m/s, selv om det i princippet er muligt at molekylet netop har denne fart. I stedet tillader man kun at stille spørgsmål om hvad sandsynligheden er i visse fart-intervaller. Det viser sig, at vi kan beskrive sandsynlighedsfordelingen for X via en såkaldt *tæthedsfunktion*, som er defineret på et helt interval – i dette tilfælde defineret for alle hastigheder fra 0 til ∞. Hvis man er meget godt inde i den fysiske disciplin *termodynamik*, kan man vise, at gasmolekylernes fart v er beskrevet ved *Maxwell-Boltzmann*-fordelingen med følgende tæthedsfunktion:

(1) $$f(v) = 4\pi \cdot \left(\frac{m}{2\pi kT} \right)^{3/2} \cdot v^2 e^{-\frac{mv^2}{2kT}}$$ (Maxwell-Boltzmann-fordelingen)

Grafen for tæthedsfunktionen for gassen Ne-20 ved $T = 353$ K er afbildet på næste side. I forrige terningeeksempel var summen af sandsynlighederne i sandsynlighedsfordelingen for X lig med 1. I dette tilfælde er det *arealet* under grafen for tæthedsfunktionen, som er lig med 1. Det vil vi dog ikke vise her.



Hvis man bliver spurgt om sandsynligheden for at et gasmolekyles fart er mellem 400 m/s og 700 m/s, så er svaret, at det er arealet under grafen i netop dette interval. Arealet er skraveret på figuren. Med grafregnerens hjælp får vi:

$$(2) \qquad P(400 \le X \le 700) = \int_{400}^{700} f(v)dv = \int_{400}^{700} 4\pi \cdot \left(\frac{m}{2\pi kT}\right)^{3/2} \cdot v^2 e^{-\frac{mv^2}{2kT}} dv = 0,437$$

hvor $m = 20,0237\text{u} = 20,0237 \cdot 1,66054 \cdot 10^{-27}\ \text{kg} = 3,32502 \cdot 10^{-26}\ \text{kg}$ for gassen Ne-20, $k = 1,380658 \cdot 10^{-23}\ \text{J/K}$ for Boltzmanns konstant og 353 K for temperaturen i Kelvin. Vi ser, at 43,7% af molekylerne har en fart mellem 400 m/s og 700 m/s. En stokastisk variabel, der har et helt interval af værdier og hvor sandsynlighederne kan bestemmes via et tæthedsfunktion som her, betegnes en *kontinuert* stokastisk variabel.

□

3. Middelværdi, varians og spredning

Der er specielt to størrelser, som er med til at sige noget om en stokastisk variabel, og det er *middelværdien* og *variansen* eller *spredningen*. Middelværdien fås ved at tage det *vejede gennemsnit* af værdierne for den stokastiske variabel, mens variansen fås som det vejede gennemsnit af kvadratet på afvigelserne fra middelværdien. Spredningen er kvadratroden af variansen. Lad os se på de to eksempler ovenfor.

Eksempel 3

Middelværdien af den stokastiske variabel fra eksempel 1 er følgende sum:

$$(3) \qquad \mu = E(X) = \sum_{i=1}^{n} x_i P(X = x_i) = 2 \cdot \tfrac{1}{36} + 3 \cdot \tfrac{2}{36} + 4 \cdot \tfrac{3}{36} + \ldots + 12 \cdot \tfrac{1}{36} = 7$$

Dette resultat er ikke overraskende, da fordelingen er symmetrisk omkring 7. Bemærk, at middelværdien både betegnes med μ og $E(X)$. Sidstnævnte står for "expectation of X" – den forventede værdi af X. For at finde variansen skal vi udregne følgende sum:

$$\begin{aligned}\text{(4)}\qquad \mathrm{Var}(X) &= \sum_{i=1}^{n}(x_i-\mu)^2\cdot P(X=x_i)\\ &= (2-7)^2\cdot\tfrac{1}{36}+(3-7)^2\cdot\tfrac{2}{36}+(4-7)^2\cdot\tfrac{3}{36}+\ldots+(12-7)^2\cdot\tfrac{1}{36} = 5{,}83\end{aligned}$$

Vi tager altså forskellene mellem den stokastiske variabels værdier og middelværdien, opløfter til 2. potens og vejer med sandsynlighederne. Spredningen fås ved at tage kvadratroden af variansen:

$$\text{(5)}\qquad \sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{5{,}83} = 2{,}52$$

Selv om talværdien for spredningen ikke siger noget direkte om fordelingen, så kan man dog sige, at jo større tallet er, jo mere spredte er data.

□

Eksempel 4

Middelværdien for den kontinuerte stokastiske variabel i eksempel 2 fås ved:

$$\begin{aligned}\text{(6)}\qquad E(X) &= \int_0^\infty v\cdot f(v)dv = \int_0^\infty v\cdot 4\pi\cdot\left(\frac{m}{2\pi kT}\right)^{3/2}\cdot v^2 e^{-\frac{mv^2}{2kT}}dv\\ &= \int_0^\infty 4\pi\cdot\left(\frac{m}{2\pi kT}\right)^{3/2}\cdot v^3 e^{-\frac{mv^2}{2kT}}dv = \sqrt{\frac{8kT}{\pi m}}\end{aligned}$$

Du skal ikke forsøge at kontrollere sidste lighedstegn: det er en meget kompliceret sag. Med talværdierne fra eksempel 2 giver det en middelhastighed på 610,9 m/s. Man kan ivrigt vise (se opgave 2.6), at den mest sandsynlige hastighed er givet ved 541,4 m/s. Vi ser desuden, at fordelingen *ikke* er symmetrisk om nogen lodret akse. Variansen giver ikke noget pænt udtryk i denne opgave, så vi undlader at udregne den.

□

Lad os sammenfatte begreberne ovenfor:

Definition 5 (Middelværdi af stokastisk variabel)
Middelværdien for en diskret stokastisk variabel X er givet ved følgende udtryk:

$$\text{(7)}\qquad E(X) = \sum_i x_i P(X=x_i)$$

hvor der summeres over de mulige værdier $x_1, x_2, \ldots$, som X kan antage. I tilfældet med en kontinuert stokastisk variabel X ser udtrykket for middelværdien således ud:

$$\text{(8)}\qquad E(X) = \int_{-\infty}^{\infty} x\cdot f(x)dx$$

hvor f er tæthedsfunktionen. (8) er under forudsætning af, at integralet eksisterer.

Definition 6 (Varians af stokastisk variabel)
Variansen for en diskret stokastisk variabel X er givet ved følgende udtryk:

$$\text{Var}(X) = \sum_i (x_i - \mu)^2 P(X = x_i) \tag{9}$$

hvor der summeres over de mulige værdier $x_1, x_2, \ldots$, som X kan antage, og hvor μ angiver middelværdien af den stokastiske variabel. I tilfældet med en kontinuert stokastisk variabel X ser udtrykket for variansen således ud:

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx \tag{10}$$

hvor f er tæthedsfunktionen. (10) er under forudsætning af, at integralet eksisterer.

Man kan definere nye stokastiske variable ud fra allerede eksisterende. Specielt interessant er lineære transformationer $Y = aX + b$. Her gælder følgende sætning:

Sætning 7
Lad X være en stokastisk variabel, og lad $Y = aX + b$. Da er Y en stokastisk variabel med følgende middelværdi og varians:

$$E(Y) = a \cdot E(X) + b \tag{11}$$

$$\text{Var}(Y) = a^2 \cdot \text{Var}(X) \tag{12}$$

Bevis: Vi vil kun bevise (11), og kun i tilfældet med en diskret stokastisk variabel. For et bevis i tilfældet med en kontinuert stokastisk variabel: se opgave 3.1.

$$\begin{aligned}
E(Y) &= \sum_i y_i \cdot P(Y = y_i) = \sum_i (ax_i + b) \cdot P(X = x_i) \\
&= (ax_1 + b) \cdot P(X = x_1) + (ax_2 + b) \cdot P(X = x_2) + \ldots \\
&= ax_1 \cdot P(X = x_1) + b \cdot P(X = x_1) + ax_2 \cdot P(X = x_2) + b \cdot P(X = x_2) + \ldots \\
&= a \cdot \left(x_1 \cdot P(X = x_1) + x_2 \cdot P(X = x_2) + \ldots \right) + b \cdot \left(P(X = x_1) + P(X = x_2) + \ldots \right) \\
&= a \cdot \sum_i x_i \cdot P(X = x_i) + b \cdot \sum_i P(X = x_i) \\
&= a \cdot E(X) + b
\end{aligned}$$

I tredje lighedstegn er der blevet ganget ind i parenteser, i fjerde lighedstegn har vi sat a og b udenfor parentes. I sidste lighedstegn har vi udnyttet, at sandsynlighederne tilsammen giver 1: $\sum_i P(X = x_i) = 1$. Dette beviser (11).

Eksempel 8

For at få en bedre forståelse af begreberne middelværdi og varians samt af sætning 7 skal vi kigge på et eksempel. En bankør tilbyder et spil, hvor spilleren slår med to terninger, en rød og en grøn. Hvis der er en 1'er blandt de to terninger, så skal spilleren betale 4 kr. til bankøren. I alle andre tilfælde vinder spilleren det beløb i kroner, som svarer til forskellen mellem de to terningers visning. Hvis den ene terning viser 5 og den anden 2, så vinder spilleren altså $5-2=3$ kroner. Lad os være systematiske igen:

Eksperiment: Et kast med to terninger, en grøn og en rød. Antal øjne betragtes.
Udfaldsrum: $U=\{(1,1),(1,2),\ldots,(1,6),(2,1),(2,2),\ldots,(2,6),\ldots,(6,6)\}$, hvor første koordinaten i talparret angiver antal øjne for den grønne terning, mens anden koordinaten angiver antal øjne for den røde terning.
Stokastisk variabel: X angiver det beløb spilleren vinder i ét spil.

For at bestemme sandsynlighedsfordelingen for X skal vi først finde ud, hvad X giver for de enkelte udfald i udfaldsrummet. Det er gjort skematisk på figuren nedenfor til venstre. Situationen er set fra spillerens synspunkt, så et tab anføres som et negativt tal. Vi ser, at X kan antage 6 forskellige værdier: –4, 0, 1, 2, 3, og 4. Sandsynligheden for hver af disse værdier fås ved at tælle op, hvor ofte de forekommer i skemaet og så gange med 1/36, som er sandsynligheden for hvert enkelt udfald. Det giver tabellen til højre:

Rød terning

6	-4	4	3	2	1	0
5	-4	3	2	1	0	1
4	-4	2	1	0	1	2
3	-4	1	0	1	2	3
2	-4	0	1	2	3	4
1	-4	-4	-4	-4	-4	-4
	1	2	3	4	5	6

Grøn terning

Sandsynlighedsfordelingen for X:

x_i	–4	0	1	2	3	4
$P(X=x_i)$	$\frac{11}{36}$	$\frac{5}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Middelværdien og variansen af den stokastiske variabel X kan nu udregnes:

$$E(X)=\sum_i x_i P(X=x_i)=-4\cdot\tfrac{11}{36}+0\cdot\tfrac{5}{36}+1\cdot\tfrac{8}{36}+2\cdot\tfrac{6}{36}+3\cdot\tfrac{4}{36}+4\cdot\tfrac{2}{36}=-\tfrac{1}{9}$$

$$\begin{aligned}\mathrm{Var}(X) &= \sum_i (x_i-\mu)^2 P(X=x_i)\\ &= (-4-(-\tfrac{1}{9}))^2\cdot\tfrac{11}{36}+(0-(-\tfrac{1}{9}))^2\cdot\tfrac{5}{36}+(1-(-\tfrac{1}{9}))^2\cdot\tfrac{8}{36}\\ &\quad+(2-(-\tfrac{1}{9}))^2\cdot\tfrac{6}{36}+(3-(-\tfrac{1}{9}))^2\cdot\tfrac{4}{36}+(4-(-\tfrac{1}{9}))^2\cdot\tfrac{2}{36}\\ &= 7{,}65\end{aligned}$$

Vi ser, at middelværdien er negativ, hvilket ikke er så mærkeligt, da bankører har det med at sørge for, at de selv har de bedste odds! Helt præcist fortæller middelværdien, at spilleren i gennemsnit vil *tabe* 1/9 kr. i hvert spil. Talværdien for variansen er ikke nem at give en god fortolkning af, men vi kan lave lidt om på spilreglerne og studere den virkning, som det har på variansen. Lad os sige, at spilleren stadig taber 4 kr. hvis blot en af terningerne viser 1, at to ens giver summen af øjnene i kr., undtagen hvis de to ens er (1,1), mens alle andre kombinationer hverken giver tab eller gevinst. Situationen er vist på figuren nedenfor til venstre. Middelværdien viser sig at være nøjagtig den samme som i det oprindelige spil, men variansen kan vises at være vokset til 14,87. Det skyldes, at spillet er blevet mere chancebetonet. Der er større præmier, som er fordelt på færre udfald. Men bankøren vil altså i gennemsnit få den samme indtjening!

Rød terning						
6	-4	0	0	0	0	12
5	-4	0	0	0	10	0
4	-4	0	0	8	0	0
3	-4	0	6	0	0	0
2	-4	4	0	0	0	0
1	-4	-4	-4	-4	-4	-4
	1	2	3	4	5	6

Grøn terning

Rød terning						
6	-11	13	10	7	4	1
5	-11	10	7	4	1	4
4	-11	7	4	1	4	7
3	-11	4	1	4	7	10
2	-11	1	4	7	10	13
1	-11	-11	-11	-11	-11	-11
	1	2	3	4	5	6

Grøn terning

Spilleren foreslår nu nye spilleregler: Han foreslår, at alle satser tredobles og at han får 1 kr. forud i hvert spil. Skal bankøren, som er matematiker, acceptere disse betingelser? Situationen er beskrevet i skemaet ovenfor til højre. Her er alle værdier for den stokastiske variabel X ganget med 3 og 1 er lagt til i forhold til det tidligere skema på forrige side. Den nye stokastiske variabel er altså givet ved $Y = 3X + 1$. Man kunne selvfølgelig finde middelværdien for Y ved hjælp af formel (7), men da middelværdien for X allerede er kendt, er det nemmere at benytte sætning 7, som også giver variansen umiddelbart:

$$E(Y) = 3 \cdot E(X) + 1 = 3 \cdot (-\tfrac{1}{9}) + 1 = \tfrac{2}{3}$$
$$\mathrm{Var}(Y) = 3^2 \cdot \mathrm{Var}(X) = 3^2 \cdot 7{,}65 = 68{,}9$$

Da middelværdien af Y er positiv, afviser bankøren spilbetingelserne, da de i længden vil være til fordel for spilleren, eftersom middelværdien er positiv.

□

Bemærkning 9

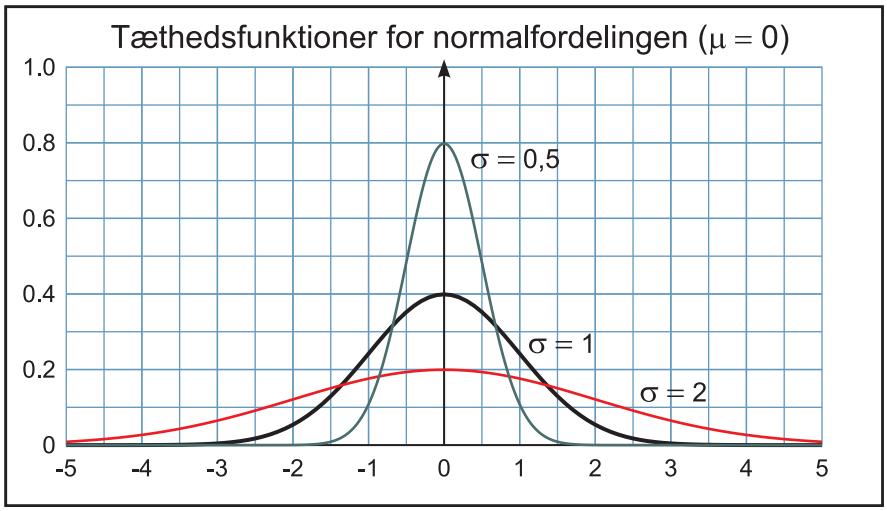
Hvis $Y = aX + b$, så fås ifølge (12) følgende sammenhæng mellem spredningerne for X og Y: $\sigma(Y) = \sqrt{\mathrm{Var}(Y)} = \sqrt{a^2 \cdot \mathrm{Var}(X)} = |a| \cdot \sqrt{\mathrm{Var}(X)} = |a| \cdot \sigma(X)$.

4. Normalfordelingen

Normalfordelingen er en kontinuert fordeling. Den har to parametre, nemlig μ og σ, som viser sig at være henholdsvis middelværdi og spredning for fordelingen (se sætning 10 på næste side). Normalfordelingens tæthedsfunktion er givet ved

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{13}$$

Nedenfor er graferne for tæthedsfunktionerne for tre forskellige værdier af σ afbildet, mens $\mu = 0$ i alle tre tilfælde. Vi ser, at det ikke er underligt, at tæthedsfunktionen for en normalfordeling ofte kaldes for en *klokkekurve*. Jo mindre σ er jo, smallere er klokkekurven, og jo større σ er, jo bredere er den.



Det ses direkte af udtrykket (13), at hvis man vælger en anden værdi af μ end 0, så parallelforskydes kurven blot med μ i x-aksens retning. Det betyder også, at den generelle klokkekurve er symmetrisk omkring $x = \mu$.

Den *kumulerede sandsynlighedsfunktion* eller *fordelingsfunktionen* for en vilkårlig stokastisk variabel X er defineret ved:

$$F(t) = P(X \le t) \tag{14}$$

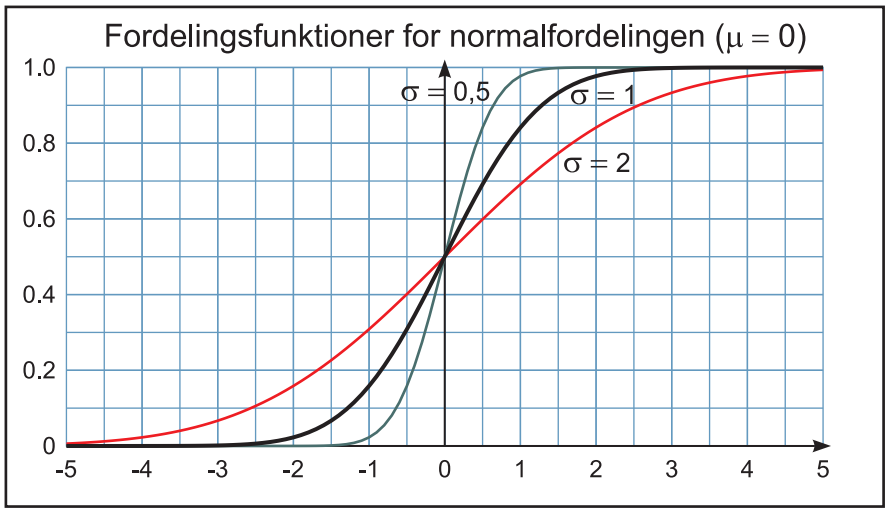
I tilfældet med en normalfordelt stokastisk variabel fås følgende fordelingsfunktion:

$$F_{\mu,\sigma}(t) = P(X \le t) = \int_{-\infty}^{t} f_{\mu,\sigma}(x)dx = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{15}$$

Ifølge en fundamental sætning fra integralregningen konkluderer vi, at fordelingsfunktionen er en stamfunktion til tæthedsfunktionen:

$$F'_{\mu,\sigma}(t) = f_{\mu,\sigma}(t) \tag{16}$$

De tre tæthedsfunktioner ovenfor har fordelingsfunktioner, hvis grafer er S-formede:



Som allerede illustreret i eksempel 2, kan sandsynligheden $P(a \le X \le b)$ bestemmes ved at finde arealet under grafen for tæthedsfunktionen fra a til b. Har man imidlertid fordelingsfunktionen til rådighed, er det meget nemmere idet:

(17) $$P(a \le X \le b) \;=\; F(b) - F(a)$$

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

(18) $$\Phi(t) \;=\; F_{0,1}(t) \;=\; \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{t} e^{-\frac{1}{2}\cdot x^2}\, dx$$

I dag kan grafregnere og softwareprogrammer såsom Microsoft Excel med numeriske metoder bestemme sandsynligheder for normalfordelingen. En del bøger med sandsynlighedsregning og statistik indeholder dog kun en tabel for standardnormalfordelingen. Heldigvis er der dog en forbindelse mellem en generel normalfordeling og standardnormalfordelingen, som vi skal se i det følgende.

I det følgende vil vi med notationen $X \sim N(\mu, \sigma^2)$ mene, at X er en normalfordelt stokastisk variabel med parametre μ og σ. Vi har en meget vigtig sætning:

Sætning 10

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

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

b) $P(X \le t) \;=\; \Phi\left(\dfrac{t - \mu}{\sigma}\right)$

c) Parametrene μ og σ i en normalfordeling angiver henholdsvis middelværdien 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.

$$\begin{aligned} P(X \le t) &= P(\sigma Z + \mu \le t) \\ &= P(Z \le (t-\mu)/\sigma) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{(t-\mu)/\sigma} e^{-\frac{1}{2}x^2}\, dx \\ &= \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot \int_{-\infty}^{t} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}\, dy \end{aligned}$$

Hvor vi for at redegøre for sidste lighedstegn har benyttet substitutionen: $y = \sigma x + \mu \Leftrightarrow x = (y-\mu)/\sigma$ dvs. $dx = 1/\sigma \cdot dy$. Sidstnævnte integral viser ifølge (15) netop, at X er en normalfordelt stokastisk variabel med parametre μ og σ.

b) $P(X \le t) = P(\sigma Z + \mu \le t) = P\left(Z \le \dfrac{t-\mu}{\sigma}\right) = \Phi\left(\dfrac{t-\mu}{\sigma}\right)$.

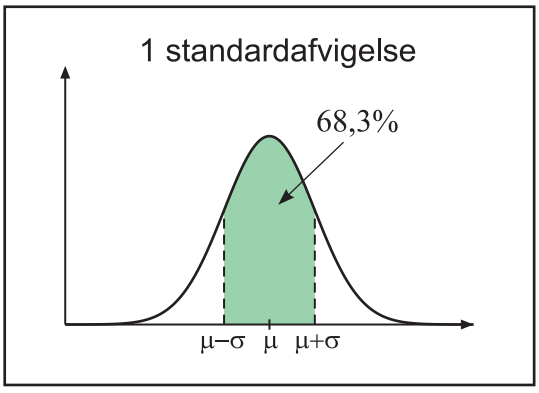
c) Beviset er ret teknisk. Der henvises til appendiks A.

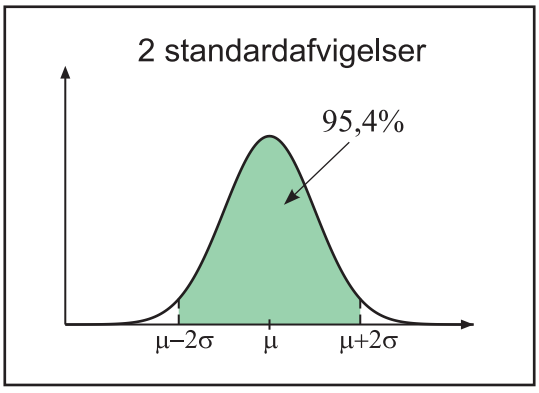
□

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 andet lighedstegn benyttes sætning 10b:

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

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





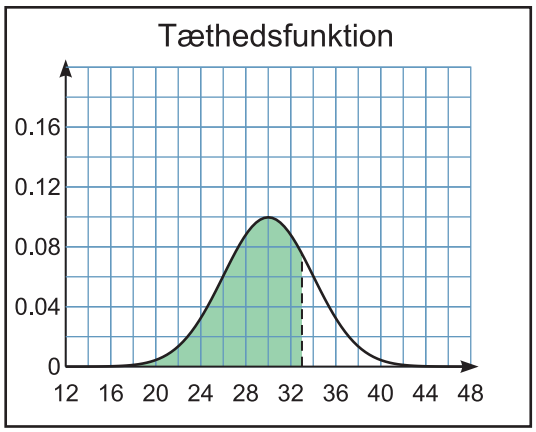
Sandsynligheden for, at X er højst 1 standardafvigelse fra middelværdien, er altså 68,3%, mens sandsynligheden for at X er højst 2 standardafvigelser fra μ er 95,4%. Vi ser, at svaret på spørgsmålene er uafhængige af hvilken normalfordeling, der er tale om. Det kan altså undertiden være fornuftigt at regne i enheder af standardafvigelsen σ fra middelværdien μ.

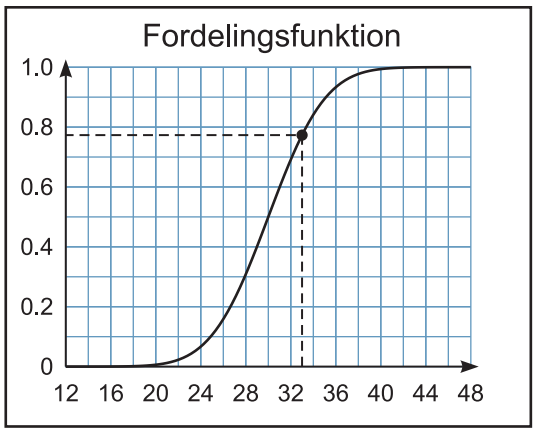
I det følgende skal vi se på eksempler på forskellige opgavetyper i forbindelse med normalfordelinger. Eksemplerne vil overvejende blive løst med grafregneren TI-89. Det er nødvendigt, at du på TI-89 har installeret `Stat/List Editor`. Andre hjælpemidler vil til opgaveløsning vil blive omtalt i opgavesektionen.

Eksempel 11

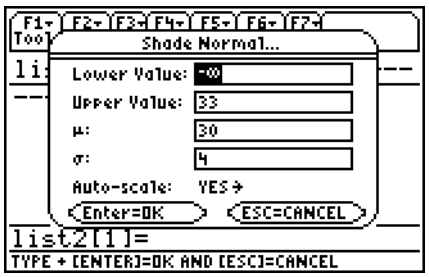
En stokastisk variabel X oplyses at være normalfordelt med middelværdi 30 og spredning 4. Bestem $P(X \le 33)$.

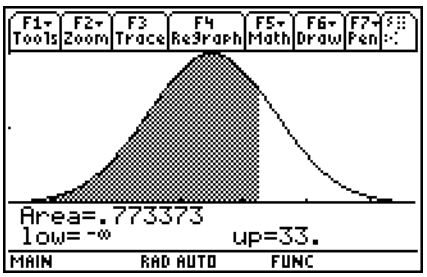
Løsning: Der er flere måder at løse denne opgave på. Enten kan man bestemme sandsynligheden som arealet under grafen for tæthedsfunktionen fra $-\infty$ til 33, eller også kan man blot bestemme fordelingsfunktionens værdi i 33.





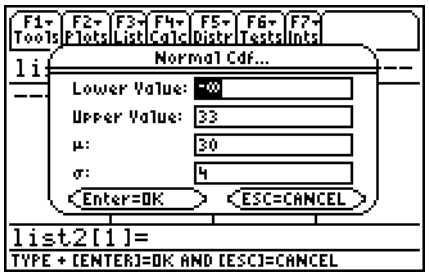
Tæthedsfunktionen med TI-89: I `Stat/List editor` vælges menuen `Distr` via □. Vælg menuen `1:Shade` og derefter `1:Shade Normal...`. Indtast de fire værdier, som afbildet på figuren nedenfor til venstre. Efter ÷. bliver tæthedsfunktionen tegnet og det relevante område skraveret. Dette ses på figuren til højre.

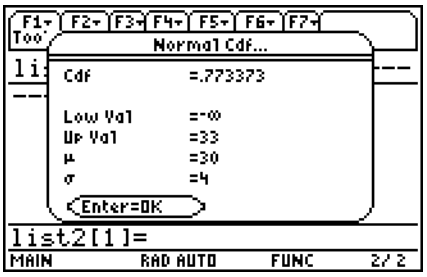




Vi har altså resultatet $P(X \le 33) = 0,773373$.

Fordelingsfunktionen med TI-89: I `Stat/List editor` vælges igen menuen `Distr` via ◻. Vælg menuen `4:Normal Cdf...` Indtast de fire værdier, som afbildet på figuren nedenfor til venstre. Efter ÷ fås resultatet til højre.





Igen har vi $P(X \le 33) = F(33) = 0,773373$.

Løsning med tabel: Her vil vi gøre kraftigt brug af sætning 10b for at knytte en sammenhæng mellem en generel normalfordeling og standardnormalfordelingen. Tabellerne indeholder nemlig data for standardnormalfordelingens fordelingsfunktion Φ.

(19) $$P(X \le 33) = \Phi\left(\frac{33-30}{4}\right) = \Phi(0,75) = 0,77337$$

◻

Eksempel 12

Intelligenskvotient scorer er normalfordelte med middelværdi 100 og en spredning på 15.

a) Hvor stor en del af befolkningen har en intelligenskvotient på under 80?

b) Hvor stor en andel af befolkningen 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 mindst have for at blive optaget?

Løsning:

a) Ved at bruge menuen `4:Normal Cdf...` ligesom i opgave 11, får vi svaret 9,1%:

$$P(X \le 80) = 0,091211$$

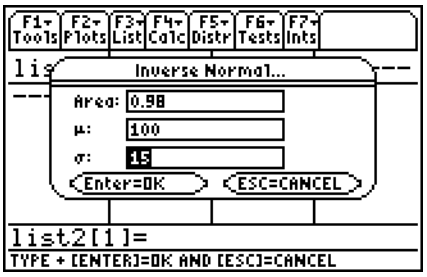
b) Samme menu benyttes her med `Lower Value` sat til 110 og `Upper Value` sat til 120. Det giver 16,1%:

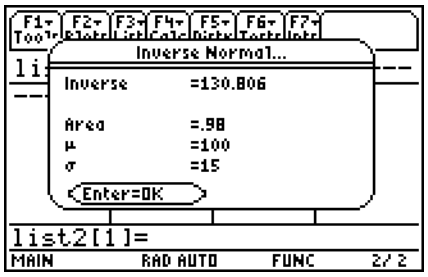
$$P(110 \le X \le 120) = 0,161281$$

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

$$F_{100,15}(t) \;=\; P(X \leq t) \;=\; 1 - P(X \geq t) \;=\; 1 - 0{,}02 \;=\; 0{,}98 \quad \Leftrightarrow \quad t = F_{100,15}^{-1}(0{,}98)$$

Værdien af den inverse funktion til fordelingsfunktionen i 0,98 kan findes direkte ved hjælp af TI-89: I `Stat/List editor` vælges menuen `Distr` via ◻. Vælg menuen `2:Inverse` og derefter `1:Inverse Normal...`. Indtast de tre værdier, som afbildet på figuren nedenfor til venstre. Efter ÷ fås svaret vist til højre. Vi ser, at $F_{100,15}^{-1}(0{,}98) = 130{,}806$. Man skal altså have en intelligenskvotient på 131 for at blive optaget i Mensa.





◻

Eksempel 13

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 en spredning på 0,1 mm. Fabrikanten kan acceptere en afvigelse på max. 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, således at kun 2% af cylindrene skal kasseres.

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

Opgaven ønskes løst med grafregner.

Løsning:

a) Vi skal altså finde sandsynligheden for at diameteren enten er over 20,2 mm eller under 19,8 mm. Det er lettere at regne sandsynligheden for det modsatte ud, dvs. at diameteren er imellem 19,8 mm og 20,2 mm, og trække resultatet fra 1:

$$P(X < 19{,}8) + P(X > 20{,}2) \;=\; 1 - P(19{,}8 \leq X \leq 20{,}2) \;=\; 1 - 0{,}9545 \;=\; 0{,}0455$$

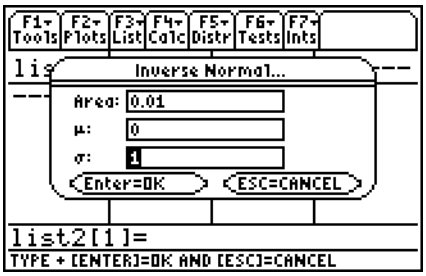
hvor sandsynligheden er udregnet med TI-89 som i eksempel 11, altså via menuen `Normal Cdf`. `Lower Value` sættes til 19,8, `Upper Value` til 20,2, μ til 20 og σ til 0,1. Vi ser, at 4,6% af cylindrene må kasseres.

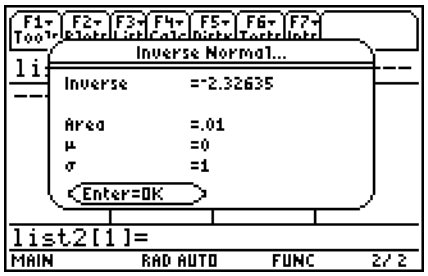
b) Vi skal bestemme σ, så $P(X<19,8)+P(X>20,2)=0,02$. Da normalfordelingen er symmetrisk, har vi $P(X\le 19,8)=0,01$. Ved brug af sætning 10b fås:

$$P(X\le 19,8)=0,01 \quad\Leftrightarrow\quad \Phi\left(\frac{19,8-20}{\sigma}\right)=0,01 \quad\Leftrightarrow\quad \frac{19,8-20}{\sigma}=\Phi^{-1}(0,01)$$

$$\frac{19,8-20}{\sigma}=-2,32635 \quad\Leftrightarrow\quad \sigma=0,086$$

Beregningen med den inverse funktion til fordelingsfunktionen for standardnormalfordelingen, dvs. $\Phi^{-1}(0,01)$, kan klares på samme måde, som vi gjorde i eksempel 12 spørgsmål c): I `Stat/List editor` vælges menuen `Distr` via ◻. Vælg menuen `2:Inverse` og derefter `1:Inverse Normal...`. Indtast de tre værdier, som afbildet på figuren nedenfor til venstre. Efter ÷ fås svaret vist til højre. Beregningerne ovenfor viser altså, at spredningen skal reduceres fra 0,1 til 0,086.





□

5. Normalfordelt data

Ikke sjældent er data fra den virkelige verden omtrent normalfordelt. Der er en række situationer, hvori man har erfaring for, at data approksimativt er normalfordelte. Det er for eksempel i forbindelse med måleusikkerheder i fysiske eksperimenter, variationerne i outputtet fra industrielle produkter (se eksempel 13) og biologiske variationer såsom højde og vægt (se eksempel med personhøjde nedenfor). Lidt abstrakt kan man sige, at hvis der er grund til at mistænke tilstedeværelsen af et stort antal mindre effekter, som er indbyrdes uafhængige og som lægger sammen til en samlet effekt, så kan man formode, at observationerne er normalfordelte. Dette er både verificeret empirisk og teoretisk. Sidstnævnte gennem den såkaldte *centrale grænseværdisætning*, som er en af sandsynlighedsregningens vigtige hjørnestene. Men alt er ikke normalfordelt. Vi har allerede i eksempel 2 set, at molekylers fart i en gas er fordelt efter en Maxwell-Boltzmann fordeling, som ikke er symmetrisk ligesom normalfordelingen. Hvis man kiggede på lønninger i Danmark, så ville man se, at fordelingen er noget højreskæv, idet der vil være en ret lang hale med store lønninger.

Men hvordan afgør man, om noget givet data er normalfordelt? Man kan selvfølgelig tegne et histogram over data og se, om det ligner en klokkekurve og om histogrammet er symmetrisk, men det kan være vanskeligt at afgøre, om den kurve, som tilnærmer histogrammet, krummer på den rigtige måde. Det samme vil være tilfældet med grafen for fordelingsfunktionen. Men hvis man tager den inverse funktion af fordelingsfunktionen til standardnormalfordelingen, dvs. Φ^{-1}, til de kumulerede frekvenser, så skal man teoretisk få en ret linje, hvis data er normalfordelt. Vi har nemlig ifølge sætning 10b:

$$(20)\quad F_{\mu,\sigma}(t) = P(X \le t) = \Phi\left(\frac{t-\mu}{\sigma}\right) \quad \Leftrightarrow \quad \Phi^{-1}(F_{\mu,\sigma}(t)) = \frac{t-\mu}{\sigma} = \frac{1}{\sigma}\cdot t - \frac{\mu}{\sigma}$$

Og det er meget nemmere at afgøre om nogle punkter ligger på en ret linje end at afgøre, om en kurve krummer på den rette måde. Samtidigt kan vi benytte linjen til at give et overslag over middelværdi og spredning. I det følgende skal vi med et velvalgt eksempel se metoden illustreret.

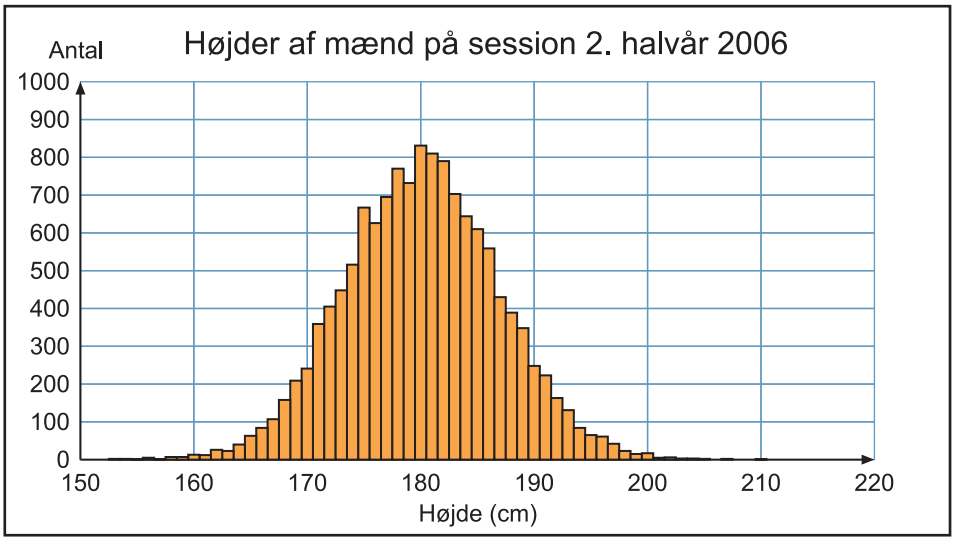
Eksempel 14 (Højden af værnepligtige på session)

På næste side er en tabel over højderne af 13427 danske værnepligtige mænd, der var på session i andet halvår af 2006. En tak til sessionslæge P. Winberg for at levere disse data og tillade brugen af dem i denne note. I tabellens søjle 1 er de værnepligtiges højde angivet i cm, i anden søjle hyppighederne, i tredje søjle frekvenserne og i fjerde søjle de kumulerede frekvenser. Datamaterialet tænkes grupperet, således, at alle personer med en højde på fx 175 cm, kommer til at tælles med i intervallet $]174,5\,\text{cm}, 175,5\,\text{cm}]$. Begrundelsen er åbenlys, hvis vi tænker på, at de angivne højder i cm er et resultat af en afrunding! Femte søjle i tabellen angiver så højre endepunkt af intervallerne. I sjette og sidste søjle er taget funktionen Φ^{-1} til de kumulerede frekvenser.

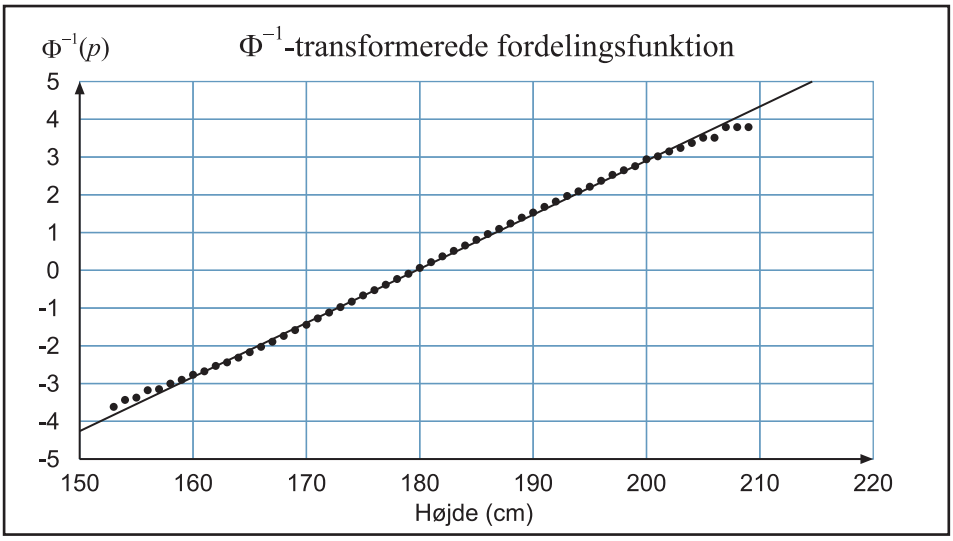
Højde (cm)	Antal	Frekvens	Kum. frekvens, p	Højre endepunkt	$\Phi^{-1}(p)$
153	2	0,000149	0,000149	153,5	-3,6170
154	2	0,000149	0,000298	154,5	-3,4337
155	1	0,000074	0,000372	155,5	-3,3726
156	5	0,000372	0,000745	156,5	-3,1767
157	1	0,000074	0,000819	157,5	-3,1490
158	7	0,000521	0,001341	158,5	-3,0021
159	7	0,000521	0,001862	159,5	-2,9006
160	13	0,000968	0,002830	160,5	-2,7668
161	12	0,000894	0,003724	161,5	-2,6761
162	26	0,001936	0,005660	162,5	-2,5327
163	23	0,001713	0,007373	163,5	-2,4385
164	40	0,002979	0,010352	164,5	-2,3133
165	63	0,004692	0,015044	165,5	-2,1689
166	84	0,006256	0,021300	166,5	-2,0276
167	107	0,007969	0,029269	167,5	-1,8916
168	158	0,011767	0,041037	168,5	-1,7388
169	209	0,015566	0,056602	169,5	-1,5840
170	241	0,017949	0,074551	170,5	-1,4427
171	359	0,026737	0,101288	171,5	-1,2742
172	405	0,030163	0,131452	172,5	-1,1196
173	448	0,033366	0,164817	173,5	-0,9749
174	516	0,038430	0,203247	174,5	-0,8301
175	667	0,049676	0,252923	175,5	-0,6653
176	626	0,046622	0,299546	176,5	-0,5257
177	695	0,051761	0,351307	177,5	-0,3818
178	770	0,057347	0,408654	178,5	-0,2310
179	732	0,054517	0,463171	179,5	-0,0924
180	831	0,061890	0,525061	180,5	0,0629
181	810	0,060326	0,585388	181,5	0,2157
182	790	0,058837	0,644224	182,5	0,3698
183	703	0,052357	0,696582	183,5	0,5146
184	644	0,047963	0,744545	184,5	0,6574
185	610	0,045431	0,789975	185,5	0,8063
186	559	0,041633	0,831608	186,5	0,9605
187	430	0,032025	0,863633	187,5	1,0968
188	389	0,028971	0,892604	188,5	1,2405
189	348	0,025918	0,918522	189,5	1,3952
190	248	0,018470	0,936993	190,5	1,5300
191	223	0,016608	0,953601	191,5	1,6808
192	163	0,012140	0,965741	192,5	1,8216
193	131	0,009756	0,975497	193,5	1,9685

194	84	0,006256	0,981753	194,5	2,0914
195	65	0,004841	0,986594	195,5	2,2143
196	61	0,004543	0,991137	196,5	2,3713
197	42	0,003128	0,994265	197,5	2,5281
198	23	0,001713	0,995978	198,5	2,6502
199	15	0,001117	0,997095	199,5	2,7584
200	17	0,001266	0,998362	200,5	2,9405
201	5	0,000372	0,998734	201,5	3,0195
202	6	0,000447	0,999181	202,5	3,1490
203	3	0,000223	0,999404	203,5	3,2410
204	3	0,000223	0,999628	204,5	3,3726
205	2	0,000149	0,999777	205,5	3,5111
206	0	0,000000	0,999777	206,5	3,5111
207	2	0,000149	0,999926	207,5	3,7928
208	0	0,000000	0,999926	208,5	3,7928
209	0	0,000000	0,999926	209,5	3,7928
210	1	0,000074	1,000000	210,5	Ikke defineret

Det kan være fornuftigt lige at tegne et histogram, for at se om data omtrent følger en klokkekurve. Det ses at være tilfældet, så der er håb for, at der kan være tale om en normalfordeling. Se i afsnit 9, hvordan dette gøres i regnearket Microsoft Excel.



På figuren på næste side er grafen for den Φ^{-1}-transformerede fordelingsfunktion. Vi ser, at punkterne ligger meget fint på linje, hvorfor vi slutter, at der er tale om normalfordelt data. For at bestemme den bedste rette linje kan man udføre lineær regression. I den forbindelse kan det være meget fornuftigt kun at benytte en vis mængde af de midterste datapunkter, for de yderste punkter er ret usikre – bemærk, at blot ganske få meget lave eller meget høje personer kan rykke betydeligt med punkterne i diagrammet. Punkterne i enderne ligger da heller ikke helt perfekt!



Udføres lineær regression på målepunkterne fra 158 cm til 205 cm, fås følgende forskrift: $y = 0{,}1433 \cdot t - 25{,}755$. Sammenlignes med (20) fås

$$(21) \qquad \frac{1}{\sigma} \approx 0{,}1433 \ \wedge \ -\frac{\mu}{\sigma} \approx -25{,}755 \quad \Leftrightarrow \quad \sigma \approx 6{,}98 \ \wedge \ \mu \approx 179{,}7$$

begge i enheden cm. Bemærk, at dette er noget approksimative værdier for middelværdi og spredning. For at finde mere pålidelige og rimelige estimater for middelværdi og spredning bør man benytte formlerne for den *empiriske middelværdi*, den *empiriske varians* og den *empiriske spredning*:

$$(22) \qquad \overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \ , \quad s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \ , \quad s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Vi vil ikke argumentere for disse formler i detaljer her, blot nævne, at de er beslægtede med formlerne i definition 5 og 6. Det skal dog nævnes, at man med vilje kalder den empiriske middelværdi for $\overline{x}$, og *ikke* μ. Det skyldes, at man skal opfatte $\overline{x}$ som et *estimat* på den "rigtige middelværdi", μ. Førstnævnte afhænger jo af de forhåndenværende datapunkter. Tilsvarende med s og σ. Først den empiriske middelværdi:

$$(23) \qquad \overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i = \frac{1}{13427} \cdot \left(2 \cdot 153 + 2 \cdot 154 + 1 \cdot 155 + 5 \cdot 156 + \ldots + 1 \cdot 210 \right) = 180{,}1$$

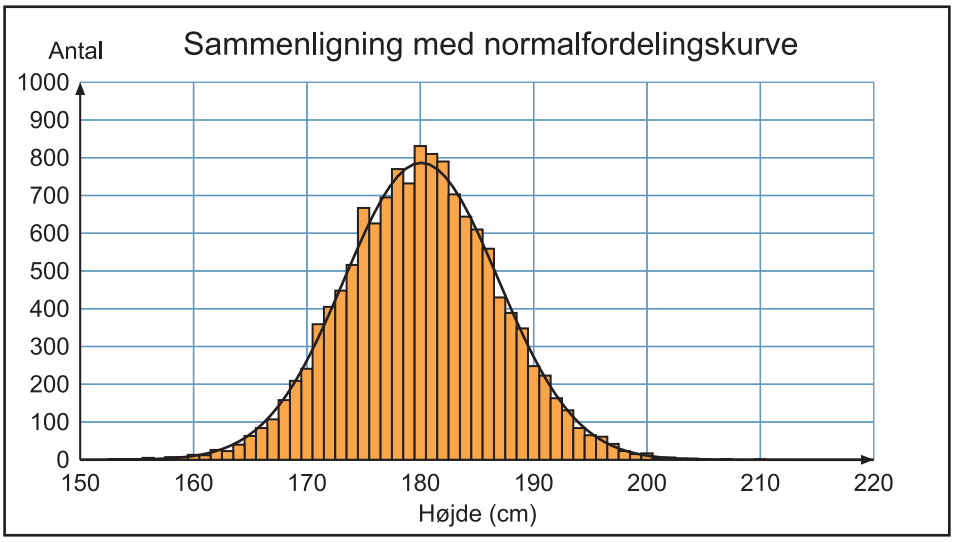
Bemærk, at formlen antager data skrevet på en lang række $x_1, x_2, x_3, \ldots$ Da vi har angivet hyppighederne for hver observation, så skriver vi $2 \cdot 153$ i stedet for $153 + 153$. Nu til den empiriske varians, hvori den empiriske middelværdi skal bruges:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2 \\ (24) \qquad &= \frac{1}{13427-1} \cdot \left(2 \cdot (153 - 180{,}1)^2 + 2 \cdot (154 - 180{,}1)^2 + \ldots + 1 \cdot (210 - 180{,}1)^2 \right) \\ &= 46{,}38 \end{aligned}$$

hvor vi igen har samlet enslydende led. Vi får nu straks den empiriske spredning:

(25) $$s = \sqrt{46{,}38} = 6{,}81$$

Ved at sammenligne (23) og (25) med (21), ser vi, at der er rimelig pæn overensstemmelse! Men det er altså værdierne i (23) og (25), som er de mest "retvisende" estimater på middelværdi og spredning. Hvis man tegner normalfordelingskurven svarende til parametrene 180,1 og 6,81 oveni i histogrammet, så ses det, at kurven meget fint følger histogrammet. Og hvis man havde målt mændenes højder i mm, så kunne intervalbredderne indskrænkes og histogrammet ville fremtone mere "glat".



□

6. Lidt kombinatorik

Som hjælperedskab til næste afsnit får vi brug for lidt *kombinatorik*. Kombinatorikken er den matematiske disciplin, som beskæftiger sig med at tælle antal kombinationer.

Definition 15

Antag givet n forskellige elementer opstillet på en række. Med betegnelsen en *permutation* af de n elementer menes en ombytning af elementerne, så de står i en ny (evt. samme) rækkefølge. Antallet af mulige permutationer af n forskellige elementer betegnes $n!$ og udtales "n fakultet".

På figuren nedenfor er vist permutationer af tallene fra 1 til 7. Da der er i alt 5040 forskellige permutationer, kan vi kun opskrive nogle af dem.

Permutationer $7! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 = 5040$

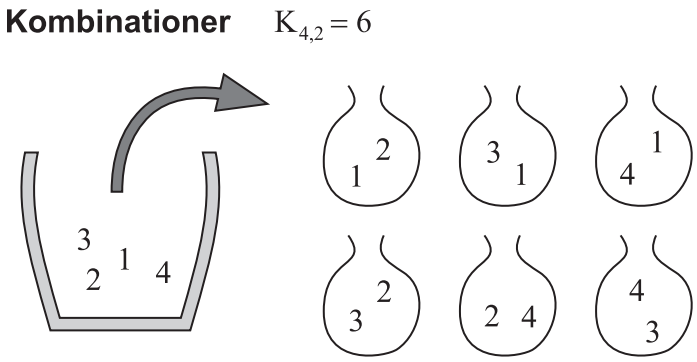
1	2	3	4	5	6	7	,	2	1	3	4	5	6	7	,	2	3	1	4	5	6	7	,

... , | 4 | 6 | 1 | 2 | 5 | 7 | 3 | , ... , | 7 | 6 | 5 | 4 | 3 | 2 | 1 | .

Definition 16

En anden vigtig størrelse er *kombinationer*. Med betegnelsen $K_{n,r}$ vil vi mene antallet af måder, hvorpå man kan udtage r elementer ud af en mængde på n forskellige elementer (antaget $0 \le r \le n$).

Situationen er illustreret nedenfor: Der skal udtages 2 tal fra en krukke med de fire tal 1, 2, 3 og 4. Det kan gøres på 6 forskellige måder som vist. Derfor er $K_{4,2} = 6$. De udtrukne tal er vist i "poser" for at indikere, at man ikke interesserer sig for, i hvilken rækkefølge tallene udtrækkes.

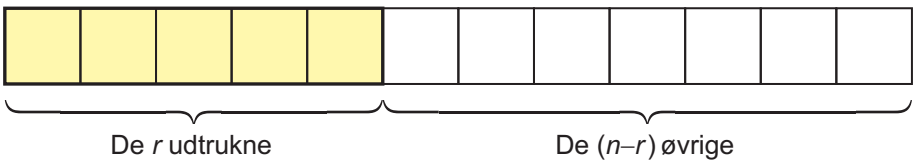


Sætning 17 (Permutationer og kombinationer)

a) $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$

b) $K_{n,r} = \dfrac{n!}{r!(n-r)!}$

Bevis: a) Givet n forskellige elementer. Den første plads kan besættes på n måder. For hver af disse tilfælde kan plads 2 besættes på $(n-1)$ måder. For hver af disse kan plads 3 besættes på $(n-2)$ måder, osv. … indtil den sidste plads, som kun kan besættes på 1 måde. Antallet af muligheder fås dermed ved at gange disse tal sammen.



b) Man kan betragte problemet på følgende måde: Vælg en tilfældig permutation af de n elementer og lad os vedtage, at de r elementer, som står først, skal være de udtrukne. Der er imidlertid en masse tilfælde, som tælles med flere gange. De første r elementer kan permuteres på $r!$ forskellige måder, men de repræsenterer den samme udtrækning.

Vi må derfor dividere med $r!$ for at tage høje for de permutationer, som tælles med flere gange. På samme måde vil permutationer af de $n-r$ øvrige elementer heller ikke resultere i nogen ny udtrækning, hvorfor vi må dividere med $(n-r)!$ for at tage højde for de tilfælde, som tælles med flere gange. Antallet af mulige udtrækninger fås derfor til:

$$K_{n,r} = \frac{n!}{r!(n-r)!}$$

□

Eksempel 18

I eksemplet på den første figur i afsnit 6 har vi, at antallet af permutationer af de 7 elementer er $7! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 = 5040$. I eksemplet på den anden figur har vi, at man kan udtage 2 elementer ud af en krukke med 4 forskellige elementer på 6 måder:

$$K_{4,2} = \frac{4!}{2!(4-2)!} = \frac{4!}{2! \cdot 2!} = \frac{1 \cdot 2 \cdot 3 \cdot 4}{1 \cdot 2 \cdot 1 \cdot 2} = 6$$

□

Eksempel 19 (Texas 89)

Fakultetsfunktionen kan i grafregneren Texas 89 aktiveres via ∀: 7! kan således indtastes ved μ∞ε÷, og det giver 5040.

I Texas 89 betegnes $K_{n,r}$ med `nCr` og funktionen kan i ∀ enten skrives direkte eller hentes i |: `nCr(4,2)` giver således 6.

7. Binomialfordelingen

Vi har hidtil behandlet normalfordelingen. I dette afsnit skal vi betragte en anden meget vigtig fordeling, nemlig den såkaldte *binomialfordeling*. At den anvendes så ofte, har sin årsag i, at den er defineret ud fra nogle abstrakte kriterier, som gør den så generel, at den passer på mange situationer. En binomialfordelt stokastisk variabel X er defineret ved følgende:

Eksperiment: Et *basiseksperiment* udføres n gange. n kaldes for *længden* af eksperimentet. Et basiseksperiment kan *lykkes* eller mislykkes. Udfaldene af hver af basiseksperimenterne skal være indbyrdes *uafhængige*. Sandsynligheden for, at et enkelt basiseksperiment lykkes betegnes p, hvormed sandsynligheden for at det mislykkes er $1-p$.
Udfaldsrum: Et udfald u kan beskrives ved en vektor med n koordinater. Den i'te koordinat er lig l, hvis det i'te basiseksperiment lykkes og m, hvis det mislykkes. Eksempel: Hvis $n=5$, så er (l, m, m, l, m) det udfald, at første og fjerde basiseksperiment lykkes, mens de øvrige tre mislykkes. Mængden af alle mulige udfald udgør udfaldsrummet.
Stokastisk variabel: X angiver det antal gange, som basiseksperimentet lykkes.

Det generelle i definitionen gør som sagt, at fordelingen finder anvendelse i vidt forskellige situationer: Det kan være, at man slår 8 gange med en terning og vil undersøge sandsynligheden for at få 3 femmere. Det kan være, at man ønsker at finde sandsynligheden for at få højst 2 piger ved 4 fødsler, eller måske vil man undersøge usikkerheden af stikprøver i forbindelse med folketingsvalg.

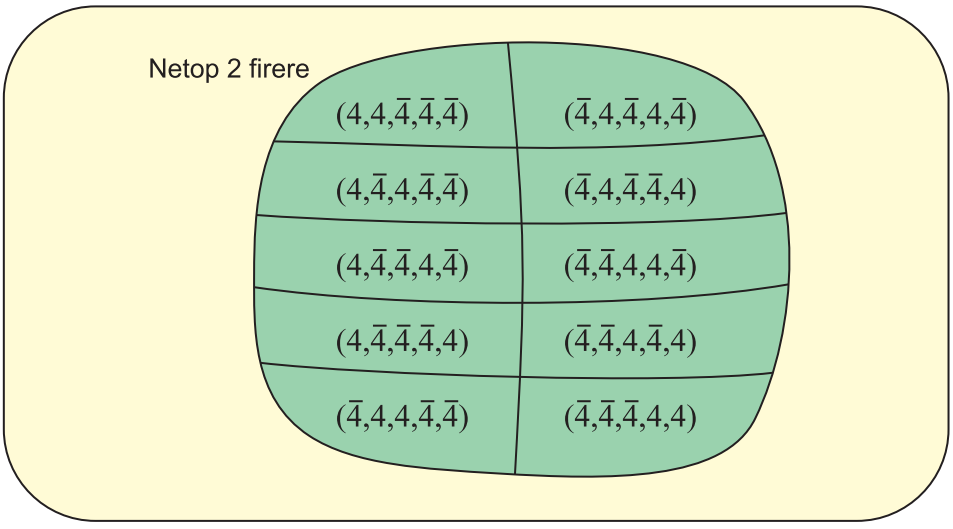
Vi skal senere se, hvorfor disse eksempler opfylder kriterierne. Først skal vi dog opstille en formel for sandsynligheden for, at basiseksperimentet lykkes *r* gange:

Sætning 20 (Binomialfordelingen)

Lad X være en binomialfordelt stokastisk variabel. Sandsynligheden for, at X antager værdien r er givet ved følgende udtryk:

(26) $$P(X=r)=K_{n,r}\cdot p^r\cdot(1-p)^{n-r}$$

Bevisskitse: Af hensyn til overskueligheden vil vi argumentere for ovenstående formel ved hjælp af et eksempel: Lad os antage, at vi slår 5 gange med en terning og ønsker at finde sandsynligheden for at få netop 2 firere. Basiseksperimentet er da at slå én gang med en terning. Vi vedtager, at basiseksperimentet lykkes, hvis man får en firer. Det giver en basissandsynlighed på $p=\frac{1}{6}$. Sandsynligheden for at basiseksperimentet mislykkes, dvs. at resultatet ikke er en firer, er dermed $1-p=\frac{5}{6}$. Den stokastiske variabel X skal angive antallet af gange basiseksperimentet lykkes, dvs. hvor mange gange man får en firer i $n=5$ kast. Netop 2 firere kan fremkomme på forskellig vis: Det er hensigtsmæssigt at dele op i en række hændelser. På figuren nedenfor symboliserer skrivemåden $(4,4,\overline{4},\overline{4},\overline{4})$, at man fik firere i de to første kast, og *ikke-firere* i de næste tre kast. Man kan også tænke, at man fik firere i første og tredje kast og ikke-firere i de øvrige, etc.



I alt er der 10 muligheder, kan vi se, og de udelukker indbyrdes hinanden samtidigt med, at de udgør alle tilfælde med netop 2 firere. Sandsynligheden for netop to firere kan da findes ved at lægge sandsynligheden for hver af de 10 muligheder sammen. Lad os først bestemme sandsynligheden for $(4,4,\overline{4},\overline{4},\overline{4})$:

$$P(4,4,\overline{4},\overline{4},\overline{4}) = \tfrac{1}{6}\cdot\tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6} = \left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3$$

Hvor vi har benyttet, at de enkelte kast er uafhængige af hinanden, så vi får sandsynligheden for hele kombinationen ved at gange deres indbyrdes sandsynligheder sammen. Vi kan gøre det samme med hændelsen $(4,\overline{4},4,\overline{4},\overline{4})$:

$$P(4,\overline{4},4,\overline{4},\overline{4}) = \tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{1}{6}\cdot\tfrac{5}{6}\cdot\tfrac{5}{6} = \left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3$$

Vi ser, at resultatet er det samme som før. Man overbeviser sig hurtigt om, at alle disse sandsynligheder er ens, nemlig $(\frac{1}{6})^2(\frac{5}{6})^3$, og at man dermed får sandsynligheden for netop 2 firere ved at gange $(\frac{1}{6})^2(\frac{5}{6})^3$ med antallet af kombinationer:

$$P(X=2) = 10\cdot\left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^3 = K_{5,2}\cdot\left(\tfrac{1}{6}\right)^2\cdot\left(\tfrac{5}{6}\right)^{5-2}$$

idet $K_{5,2}=10$ er antallet af kombinationer, svarende til antallet af måder, hvorpå man kan udpege de to pladser, hvorpå firerne skal stå. Resten af pladserne skal indeholde ikke-firere. Vi ser, at resultatet stemmer med (26) for $n=5,\ r=2,\ p=\frac{1}{6}$.

□

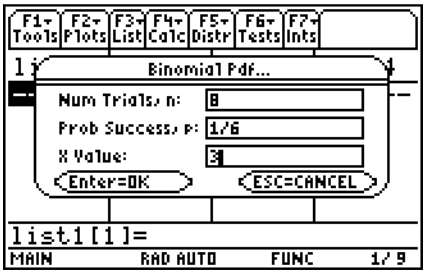
Eksempel 21

Bestem sandsynligheden for ved 8 kast med en terning at få netop 3 femmere. Angiv desuden hele sandsynlighedsfordelingen for den stokastiske variabel *X*, som angiver antal femmere ved de 8 kast.

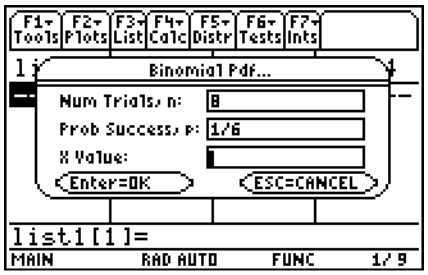
Løsning: Basiseksperimentet er ét kast med en terning, og det udføres $n = 8$ gange. Udfaldene af de enkelte basiseksperimenter er klart uafhængige. Dermed kan vi benytte binomialfordelingen. Basissandsynligheden er $p=\frac{1}{6}$. Ved brug af formel (26) fås:

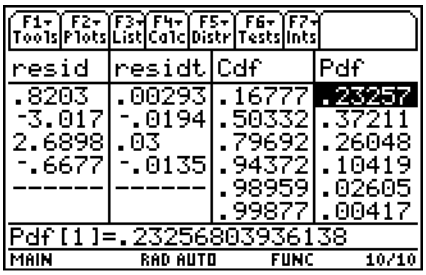
$$P(X=3) = K_{8,3}\cdot\left(\tfrac{1}{6}\right)^3\left(\tfrac{5}{6}\right)^5 = 56\cdot\left(\tfrac{1}{6}\right)^3\left(\tfrac{5}{6}\right)^5 = 0{,}1042$$

Hvis man har Stats/List-Editoren installeret på sin Texas 89, kan problemet løses hurtigere: Gå ind i editoren via O og vælg menuen `Distr` via □ og derefter undermenuen `B:Binomial Pdf...`. I den fremkomne boks indtastes de relevante værdier, som vist på figuren nedenfor.



Man kan få hele sandsynlighedsfordelingen for X ved at gentage proceduren ovenfor bortset fra, at man lader feltet "X Value" stå tomt. Efter tryk på ÷ fås en boks og efter ÷ igen, fås billedet nedenfor til højre. De ønskede sandsynligheder findes i kolonnen Pdf (Eng: *Probability Density Function*). De kumulerede sandsynligheder fås i øvrigt samtidigt i kolonnen Cdf (Eng: *Cumulative Distribution Function*).





x	0	1	2	3	4	5	6	7	8
$P(X=x)$	0,2326	0,3721	0,2605	0,1042	0,0261	0,0042	0,0004	0,0000	0,0000

Eksempel 22

Det er en udbredt opfattelse blandt folk, at hvis man får mange børn af samme køn, så er det en ekstrem hændelse. Men er det nu det? Lad os se på eksempler. Hvad er sandsynligheden for ved 4 fødsler at få: a) fire piger?, b) mindst en pige?, c) højst to piger?

Løsning: Basiseksperimentet er en fødsel. Det oplyses, at udfaldet pige og dreng i basiseksperimentet kan betragtes som uafhængige hændelser. Statistikker viser, at piger forekommer en smule sjældnere end drenge, nemlig i 49% af tilfældene. Lad os vedtage at basiseksperimentet lykkes, hvis det bliver en pige, dvs. basissandsynligheden er $p=0,49$. Lad X angive antallet af piger.

a) $P(X=4)=K_{4,4}\cdot 0,49^4\cdot(1-0,49)^{4-4}=K_{4,4}\cdot 0,49^4\cdot 0,51^0=0,49^4=0,0576$, så fire piger ved 4 fødsler sker altså trods alt i knap 6% af alle tilfælde!

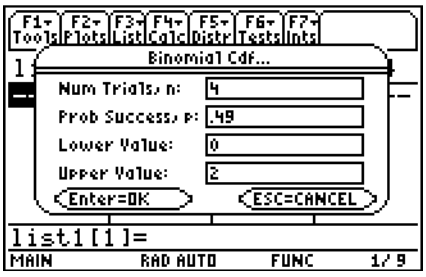
b) Hvis der skal være mindst én pige, så skal X altså være lig med 1, 2, 3 eller 4. Problemet kan altså løses ved at udregne $P(X=1)+P(X=2)+P(X=3)+P(X=4)$.

Der er imidlertid en nemmere måde: Det modsatte af at få mindst 1 pige er, at man ingen piger får. Derfor kan opgaven også løses ved:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - K_{4,0} \cdot 0,49^0 \cdot 0,51^4 = 1 - 0,51^4 = 0,932$$

så sandsynligheden for at få mindst én pige ved 4 fødsler er altså 93,2%.

c) Hvis man skal have højst 2 piger, så skal X være lig med 0, 1 eller 2. Opgaven kan altså løses ved at udregne $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$. Her er det dog hurtigere at benytte Texas 89's facilitet for *kumulerede binomial-sandsynligheder*: Gå ind i Stats/List-editoren via O og vælg menuen `Distr` via □ og derefter undermenuen `C:Binomial Cdf...`. I den fremkomne boks indtastes de relevante værdier, som vist på figuren nedenfor.



Svaret er $P(X \leq 2) = 0,7023$. Så man vil få højst to piger i 70,23% af tilfældene.

□

Eksempel 23

Op til folketingsvalg foretages mange valgprognoser. Typisk udspørges mellem 1000 og 1500 personer om, hvilket parti de vil stemme på. I dette eksempel skal vi forsøge at få en fornemmelse for, hvor usikre sådanne stikprøver er. Vi vil antage, at man udspørger 1200 helt tilfældigt udvalgte personer. Lad os antage, at et parti A på landsplan har 30% af stemmerne, uden at man kender tallet. Hvad er da sandsynligheden for, at en stikprøve på 1200 vil give et resultat, som afviger med mere end 2 procentpoint fra det rigtige tal?

Løsning: Vi kan benytte binomialfordelingen: Basiseksperimentet er, at man udspørger én person. Vi kan antage, at personerne stemmer uafhængigt, og basissandsynligheden er $p = 0,30$. Den stokastiske variabel X angiver, hvor mange af de 1200 personer, som stemmer på partiet. Vi skal bestemme sandsynligheden for, at resultatet af stikprøven ligger under 28% eller over 32%. Ud af 1200 svarer det til, at vi skal finde sandsynligheden for, at der er færre end 336 eller over 384, som stemmer på partiet A. Det modsatte er, at der er mindst 336 og højst 384, som stemmer på partiet. Vi kan da skrive:

$$P(X<336)+P(X>384) \;=\; 1-P(336\le X\le 384) \;=\; 1-0{,}877 \;=\; 0{,}123$$

hvor $P(336\le X\le 384)$ kan udregnes med Texas 89 på samme måde, som vi gjorde i eksempel 22 c): Her med *n* sat til 1200, *p* til 0,30, `Lower Value` sat til 336 og `Upper Value` sat til 384. Vi ser, at chancen for at stikprøven viser mere end 2 procentpoint galt, er større end 12%! Man skal altså være forsigtig med at konkludere for håndfast ud fra stikprøver. Her gik vi endda ud fra, at gruppen på 1200 var helt tilfældigt udvalgt. Er der en skævhed i sammensætningen, vil det forøge usikkerheden.

Bemærkning 24

Til den opmærksomme læser skal det lige tilføjes, at man egentligt principielt ikke kan anvende binomialformlen i eksempel 23, hvis man skal være nøjeregnende: Når en person har afgivet en stemme, så er sandsynligheden for, at den næste person stemmer på partiet A ikke helt 30%, for en person er jo fjernet fra gruppen! Ud af en stor population, som den danske befolkning er, har dette dog ingen praktisk betydning. Denne problematik er årsagen til, at man undertiden siger, at binomialfordelingen gælder for situationer *med tilbagelægning*.

I det følgende skal vi kigge på middelværdien og variansen for en binomialfordelt stokastisk variabel. Netop i dette tilfælde gælder der nogle ekstra simple og smukke formler, så vi undgår at skulle gå tilbage til de generelle definitioner 5 og 6:

Sætning 25

Lad X være en binomial-fordelt stokastisk variabel med antalsparameter n og basissandsynlighed p. Da er middelværdien og variansen givet ved følgende formler:

a) $E(X)=n\cdot p$

b) $\mathrm{Var}(X)=n\cdot p\cdot(1-p)$

Bevis: Overspringes, da de er lidt tekniske.

□

Eksempel 26

I eksempel 21, hvor der kastes 8 gange med en terning fås $E(X)=n\cdot p=8\cdot\frac{1}{6}=1\frac{1}{3}$. Der vil altså i snit forekomme $1\frac{1}{3}$ femmere ved 8 kast med en terning. Variansen udregnes til $\mathrm{Var}(X)=n\cdot p\cdot(1-p)=8\cdot\frac{1}{6}\cdot\frac{5}{6}=\frac{10}{9}$.

Eksempel 27

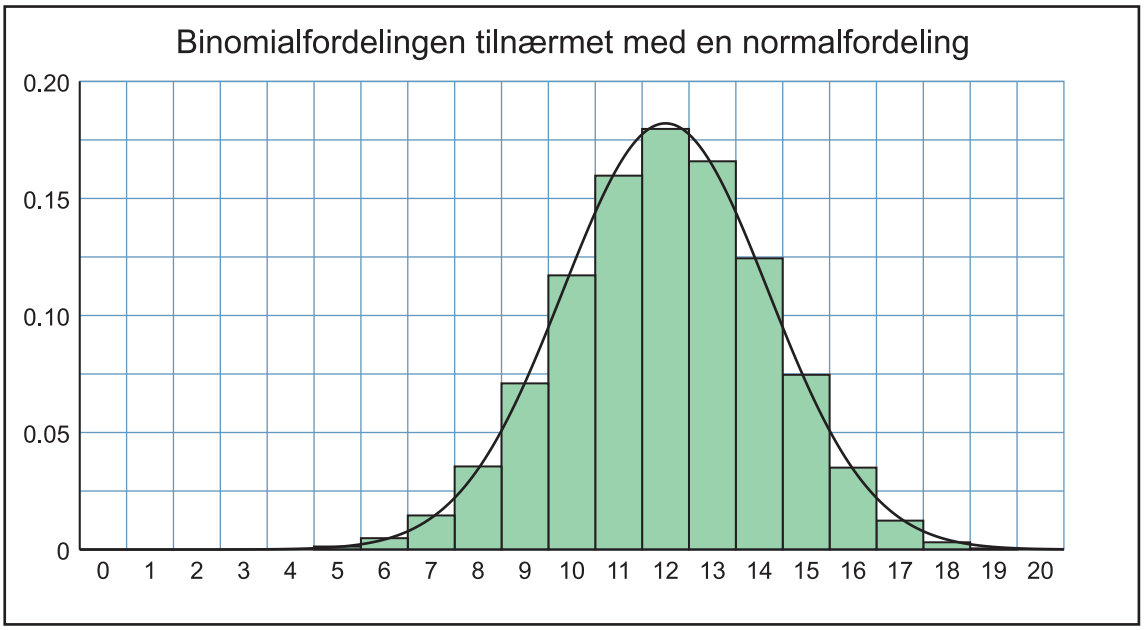
I eksempel 23 er middeltallet for antallet af stemmer på parti A ikke overraskende givet ved: $E(X)=n\cdot p=1200\cdot 0{,}30=360$, svarende til 30% af 1200!

Bemærkning 28 (Uafhængighed)

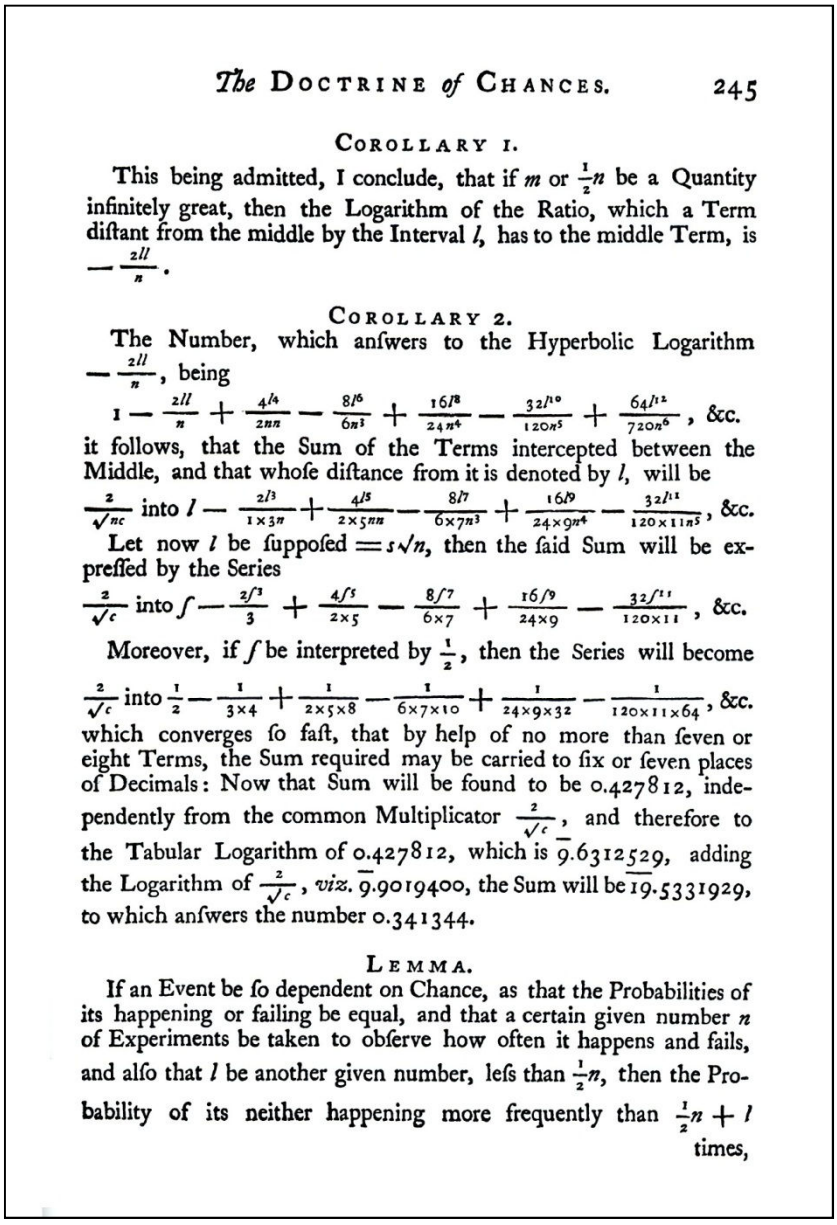
Som nævnt er uafhængigheden af de enkelte basiseksperimenter en vigtig forudsætning for, at vi har at gøre med en binomialfordeling. Derfor skal man være påpasselig med at overveje, om denne forudsætning kan siges at være opfyldt i den enkelte situation. Ved man statistisk over en lang periode, at Brøndbys fodboldhold plejer at vinde 70% af deres kampe, og man vil vurdere sandsynligheden for, at holdet vinder to af de næste tre kampe, så skal man passe på. For er holdet inde i en god stime, så ved man godt, at sandsynligheden for gevinst øges i næste kamp, fordi selvtilliden er i top. Udfaldene er med andre ord ikke uafhængige!

8. Normalfordelingens forbindelse til binomialfordelingen

Når binomialfordelingen finder anvendelse i så mange tilfælde, skyldes det som nævnt, at dens definition indeholder generelle elementer, der passer på så mange situationer. Binomialfordelingen er en såkaldt *diskret fordeling*, fordi den stokastiske variabel kun kan antage endeligt mange værdier. Normalfordelingen, derimod, er en *kontinuert* fordeling, fordi den stokastiske variabel kan antage et helt interval af værdier. Det er derfor meget overraskende, at binomialfordelingen kan tilnærmes med normalfordelingen. Som et eksempel kan vi se på en binomialfordelt stokastisk variabel X med antalsparameter $n = 20$ og basissandsynlighed $p = 0,6$. På figuren nedenfor er sandsynlighedsfordelingen for X afbildet, så søjlerne har centrum i de værdier, de hører til. Ifølge forrige afsnit er middelværdien for X givet ved $E(X) = n \cdot p = 20 \cdot 0,6 = 12$ og variansen er givet ved $\text{Var}(X) = n \cdot p \cdot (1-p) = 4,8$. På figuren er desuden indtegnet tæthedsfunktionen for normalfordelingen med de samme værdier for middelværdi og varians: $N(\mu, \sigma^2) = N(12; 4,8)$. Vi ser, at approksimationen er overraskende god!



Som en tommefingerregel 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, som at han tilnærmede binomialfordelingen med en normalfordeling. Han kan samtidigt siges at være den person, som gav den første formulering af den sætning, som går under navnet *den centrale grænseværdisætning* (*The Central Limit Theorem*) – en dyb sætning, der står som en hjørnesten i sandsynlighedsregningen.



The DOCTRINE *of* CHANCES. 245

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 $\bar{9}.6312529$, adding the Logarithm of $\frac{2}{\sqrt{c}}$, *viz.* $\bar{9}.9019400$, the Sum will be $\bar{19}.5331929$, 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.

9. Excel-tutorial

I denne tutorial vil det blive demonstreret, hvordan man kan lave histogrammer, grafer for fordelingsfunktioner m.m. i regnearket Microsoft Excel. Det antages, at læseren har et basalt kendskab til Excel, dvs. kan lave formler, nedkopiere, etc. Derimod vil det grafiske blive forklaret i detaljer. Som materiale benytter vi data hentet på Danmarks Statistik (www.dst.dk) fra deres *Statistikbank*. Kig under *Befolkning og valg* > *Vielser og skilsmisser*. Vi trækker nu data ud for alderen på de kvinder med dansk statsborgerskab, som er blevet viet til en 30-årig mand med dansk statsborgerskab fra 2006. Hensigten er blandt andet at undersøge om kvindernes aldre fordeler sig som en normalfordeling, der er symmetrisk omkring 30 år.

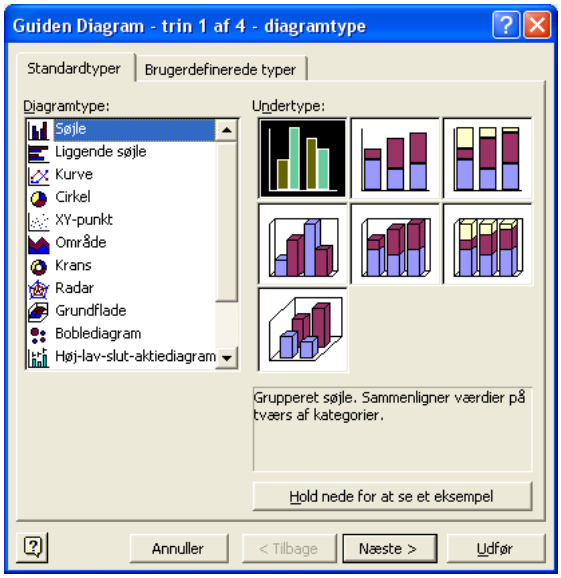
Du skal starte med at lave nedenstående skema. På næste side følger en kortfattet instruktion hertil.

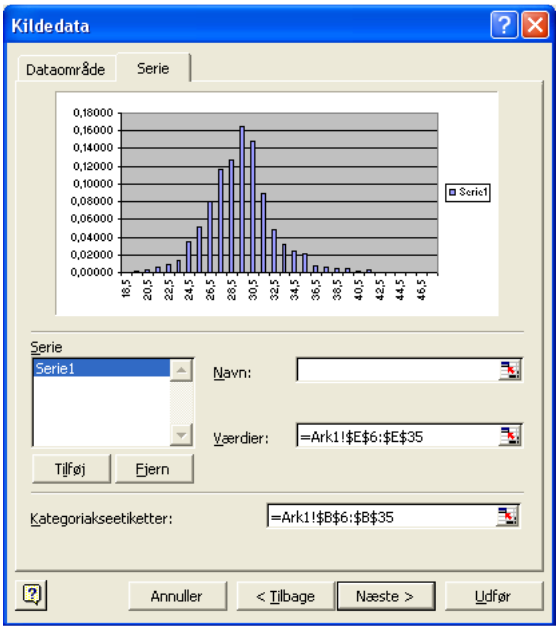
	A	B	C	D	E	F	G
1	**Alder for kvinder viet til 30 årige mænd 2006**						
2							
3							
4							
5	Alder	Midtpunkt	Højre endepunkt	Hyppighed	Frekvens	Kum. Frekvens, p	$\Phi^{-1}(p)$
6	18	18,5	19	1	0,00057	0,00057	-3,2544
7	19	19,5	20	2	0,00114	0,00171	-2,9281
8	20	20,5	21	4	0,00227	0,00398	-2,6538
9	21	21,5	22	10	0,00569	0,00966	-2,3391
10	22	22,5	23	15	0,00853	0,01819	-2,0926
11	23	23,5	24	25	0,01421	0,03240	-1,8466
12	24	24,5	25	61	0,03468	0,06708	-1,4979
13	25	25,5	26	90	0,05117	0,11825	-1,1838
14	26	26,5	27	142	0,08073	0,19898	-0,8453
15	27	27,5	28	205	0,11654	0,31552	-0,4803
16	28	28,5	29	224	0,12735	0,44287	-0,1437
17	29	29,5	30	290	0,16487	0,60773	0,2734
18	30	30,5	31	260	0,14781	0,75554	0,6920
19	31	31,5	32	157	0,08926	0,84480	1,0144
20	32	32,5	33	85	0,04832	0,89312	1,2433
21	33	33,5	34	57	0,03240	0,92553	1,4433
22	34	34,5	35	42	0,02388	0,94940	1,6391
23	35	35,5	36	37	0,02103	0,97044	1,8873
24	36	36,5	37	14	0,00796	0,97840	2,0217
25	37	37,5	38	11	0,00625	0,98465	2,1609
26	38	38,5	39	9	0,00512	0,98977	2,3177
27	39	39,5	40	7	0,00398	0,99375	2,4975
28	40	40,5	41	3	0,00171	0,99545	2,6084
29	41	41,5	42	5	0,00284	0,99829	2,9281
30	42	42,5	43	0	0,00000	0,99829	2,9281
31	43	43,5	44	1	0,00057	0,99886	3,0520
32	44	44,5	45	0	0,00000	0,99886	3,0520
33	45	45,5	46	0	0,00000	0,99886	3,0520
34	46	46,5	47	1	0,00057	0,99943	3,2544
35	47	47,5	48	1	0,00057	1,00000	
36			Sum:	1759			

1. Overfør kvindernes aldre og de tilhørende hyppigheder til søjle A og D.
2. Vi skal have grupperet data. Normalt siger man, at en person er 25 år gammel helt indtil denne fylder 26 år. Derfor skal personen figurere i intervallet fra 25 til 26. Midtpunkterne samt de højre endepunkter af intervallerne udregnes i søjle B og C.
3. Summen af hyppighederne udregnes med formlen `=SUM(D6:D35)` i celle D36.
4. Frekvenserne i søjle E udregnes ved i celle E6 at skrive formlen `=D6/$D$36` efterfulgt af nedkopiering.
5. De kumulerede frekvenser i søjle F klares på følgende måde: I feltet F6 skrives formlen `=E6` og i celle F7 skrive formlen `=E7+F6`, hvorefter sidstnævnte celle nedkopieres.
6. De Φ^{-1} – transformerede kumulerede frekvenser i søjle G frembringes ved i celle G6 at skrive formlen `=NORMINV(F6;0;1)` og nedkopiere cellen. Vi er nu klar til den grafiske side af sagen. Vi starter med histogrammet:

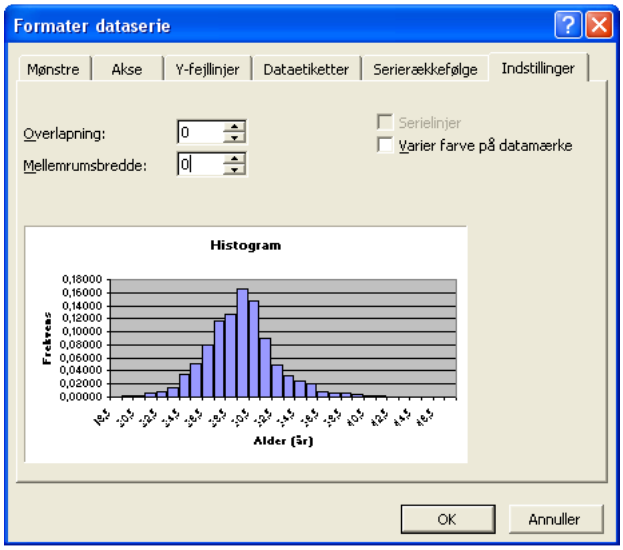
Histogram

1. Klik på knappen *Guiden Diagram* i værktøjslinjen. Ikonen forestiller nogle søjler.
2. I dialogboksen vælges diagramtypen *Søjle* og undertypen øverst til venstre – se figuren nederst til venstre. Tryk på *Næste*.
3. I trin 2 af diagramguiden skal du med fanebladet *Dataområde* stille cursoren i feltet *Dataområde* og derefter gå over og markere søjle F med frekvenserne. Når du slipper venstre musetast efter markeringen, vil søjlereferencerne automatisk blive skrevet i feltet, og du vil endda kunne se et preview.
4. Klik på fanebladet Serie. Vi skal have specificeret det, som skal stå på 1. aksen. I Excel hedder det *Kategoriakseetiketter*. Det er desværre ikke muligt i Excel at anføre intervalendepunkterne ud for overgangen fra en søjle til den næste. Som alternativ kan man passende skrive midtpunktet af hver søjle ud for hver søjle. Det gøres ved at sætte cursoren i feltet med kategoriakseetiketter og markere søjle B med midtpunkterne. Resultatet vises på figuren nedenfor til højre. Klik *Næste*.

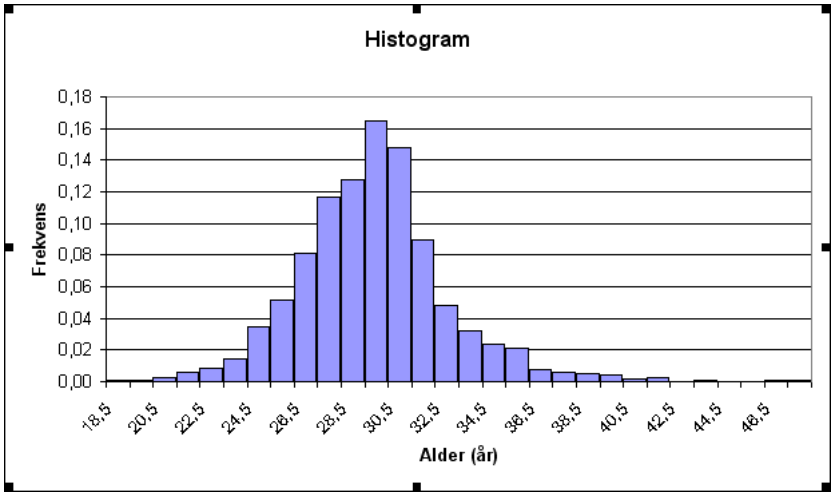




5. I trin 3 kan du anføre titler til diagram, kategoriakse og værdiakse m.m. Afslut med *Næste*.
6. I trin 4 klikker du *Udfør*. Du har nu et færdigt diagram.
7. Søjlerne i diagrammet har desværre mellemrum imellem søjlerne, og det duer ikke. Heldigvis kan det ændres ved at højreklikke på en af søjlerne og vælge *Formater dataserie…* Vælg fanebladet *Indstillinger*. Sæt *Mellemrumsbredde* til 0, og klik *Ok*.



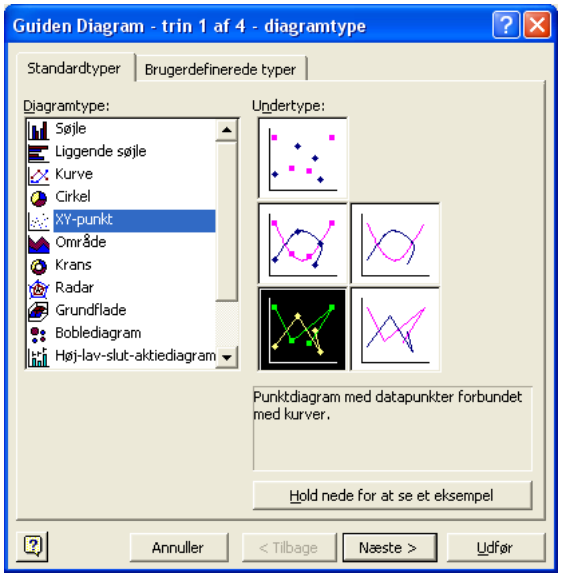
Efter andre små tilretninger, ser histogrammet således ud:

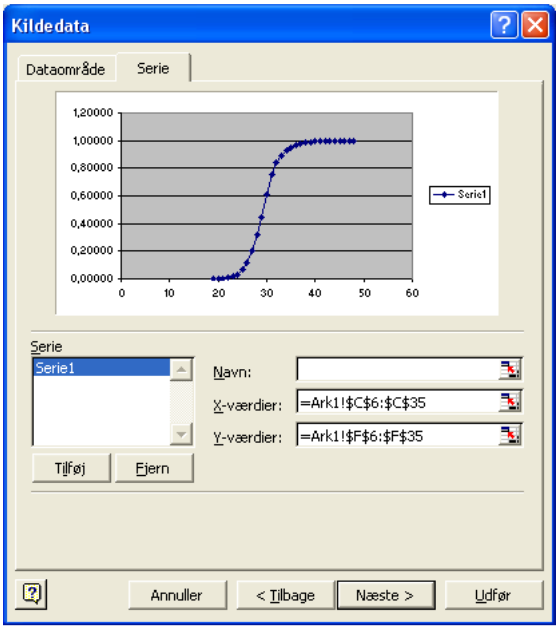


Vi ser, at histogrammet specielt til højre for toppen mangler lidt i at ligne en rigtig klokkekurve. Men for mere sikkert at kunne afvise, at der er tale om en normalfordeling, skal vi både tegne grafen for fordelingsfunktionen og dens Φ^{-1}-transformerede.

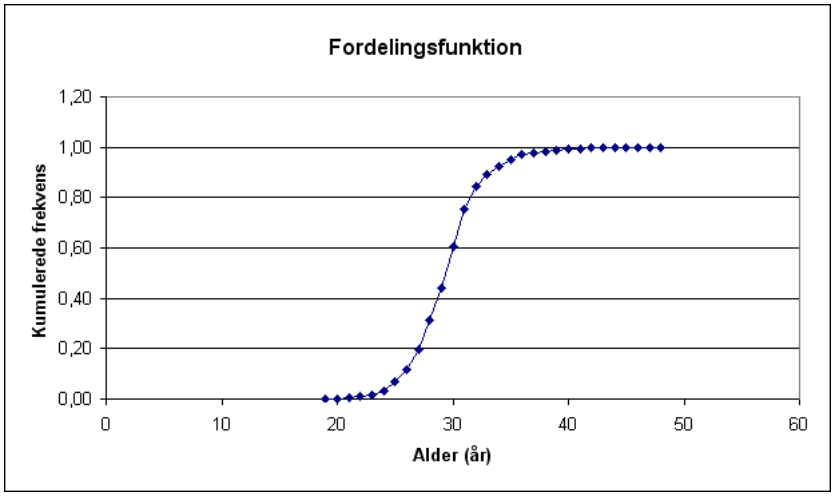
Fordelingsfunktionens graf

1. Vælg igen diagramværktøjet, men som *Diagramtype* skal du nu vælge *XY-punkt*, og derefter undertypen nederst til venstre – se figuren til venstre på næste side. Klik herefter på *Næste*.
2. Under fanen *Dataområde* sætter du cursoren i feltet Dataområde og markerer herefter søjle F med de kumulerede frekvenser. Et preview bliver vist.
3. Under fanebladet *Serie* sætter du cursoren i feltet *X-værdier* og markerer søjle D med de højre endepunkter. Resultatet ser ud som på figuren nedenfor til højre. Afslut med *Næste*.

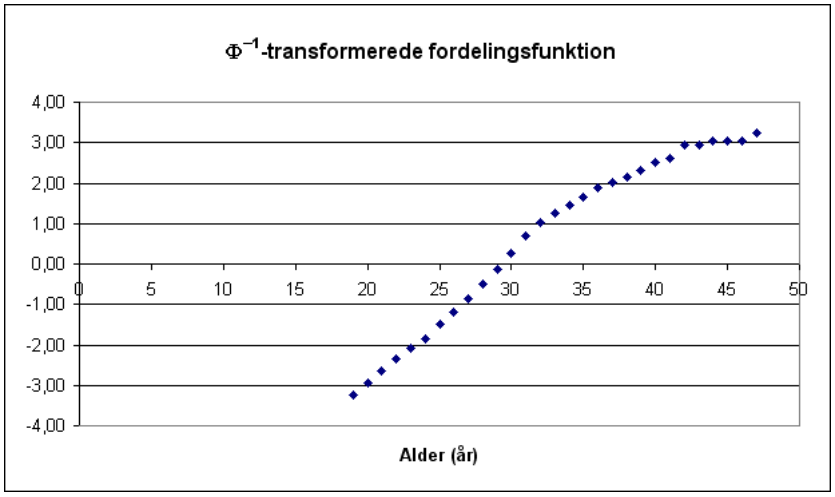




4. I de næste trin skal du skrive titler og rette til. Resultatet kan være således:



Grafen er S-formet, så det udelukker ikke en normalfordeling. For at afgøre dette skal vi se på den Φ^{-1}-transformerede graf, som i Excel laves på samme måde som grafen for fordelingsfunktionen, blot skal du som undertype til datatypen *XY-punkt* ikke vælge linjeforbundne punkter, men blot punkter. Resultatet kan ses på figuren på næste side. Endelig ses det nu, at det ikke er en normalfordeling. Punkterne bøjer systematisk af i højre ende. Der er tilsyneladende færre ældre kvinder, som gifter sig med en mand, end yngre kvinder!



Selvom der ikke er tale om en normalfordeling, kan det godt være interessant også at udregne den empiriske middelværdi, varians og spredning. Formlerne ses i (22). For at beregne middelværdien kan man selvfølgelig lave en ekstra søjle i regnearket, hvor produkterne af frekvenserne og midtpunktsværdierne står og derefter nedenfor søjlen anvende sum-funktionen til at udregne summen, men der er en mere elegant metode, hvorved hele beregningen kan foretages i en celle. Idéen er at anvende en såkaldt *matrix-formel*. Man gør som følger:

1. I en selvvalgt celle skrives `=SUM(B6:B35*E6:E35)`. Afslut med at trykke tastekombinationen Shift+Ctrl+Enter. Sidstnævnte fortæller Excel, at der er tale om en matrix-formel. Prøv at markere cellen, hvori du skrev matrix-formlen og kig i formelindtastningslinjen: Der er kommet {} om formlen! Vi ser, at den empiriske middelværdi er 29,41 for de kvinder, som gifter sig med en 30,5 år gammel mand. Det passer meget godt med den almindelige opfattelse af, at manden i et ægteskab ofte er lidt ældre end kvinden!
2. Den empiriske varians kan også beregnes med en matrixformel. Lad os sige, at du anbragte den empiriske middelværdi i celle C39. Følgende matrixformel løser da problemet: `=sum(D6:D35*(B6:B35-C39)^2)/(D36-1)`. Husk at afslutte med tastekombinationen Shift+Ctrl+Enter.
3. Lad os sige, at du anbragte den empiriske varians i celle C40. Den empiriske spredning klares med `=KVROD(C40)`.

Appendiks A

Bevis for sætning 10c: Ifølge definition 5 har vi for standardnormalfordelingen:

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_{0,1}(x)dx = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot e^{-\frac{1}{2} \cdot x^2} dx = \frac{1}{\sqrt{2\pi}} \cdot \left[-e^{-\frac{1}{2} \cdot x^2} \right]_{-\infty}^{\infty} = 0$$

Hvilket man i øvrigt straks kunne have sagt, da integranden er en ulige funktion! Variansen for standardnormalfordelingen findes af definition 6:

$$\begin{aligned} \mathrm{Var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_{0,1}(x)dx = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} (x-0)^2 \cdot e^{-\frac{1}{2} \cdot x^2} dx \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x^2 \cdot e^{-\frac{1}{2} \cdot x^2} dx = -\frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} x \cdot \left(-x \cdot e^{-\frac{1}{2} \cdot x^2} \right) dx \\ &= -\frac{1}{\sqrt{2\pi}} \left[x \cdot e^{-\frac{1}{2} \cdot x^2} \right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} 1 \cdot e^{-\frac{1}{2} \cdot x^2} dx = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\infty} e^{-\frac{1}{2} \cdot x^2} dx = 1 \end{aligned}$$

hvor vi i fjerde lighedstegn har omskrevet integralet med henblik på at udføre partiel integration med de to funktioner x og $-x \cdot e^{-\frac{1}{2} \cdot x^2}$. Sidstnævnte har $e^{-\frac{1}{2} \cdot x^2}$ som stamfunktion. Det sidste integrale giver 1. Det er overordentligt kompliceret at vise dette, hvorfor vi undlader det. Lad os blot nævne, at da integralet er hele arealet under grafen for tæthedsfunktionen, skal integralet give 1.

Vi skal vise det ønskede for enhver normalfordeling Ifølge sætning10a) kan en normalfordelt stokastisk variabel med parametre μ og σ, skrives på formen $X = \sigma Z + \mu$, hvor Z er en standardnormalfordelt stokastisk variabel. Ifølge sætning 7 fås nu:

$$\begin{aligned} E(X) &= E(\sigma Z + \mu) = \sigma \cdot E(Z) + \mu = \sigma \cdot 0 + \mu = \mu \\ \mathrm{Var}(X) &= \mathrm{Var}(\sigma Z + \mu) = \sigma^2 \cdot \mathrm{Var}(Z) = \sigma^2 \cdot 1 = \sigma^2 \end{aligned}$$

□

Opgaver

Nedenstående opgaver er nummereret, så tallet foran punktummet angiver det afsnit, som opgaven hører til. Således er opgave 3.2 opgave 2 i afsnit 3.

Opgave 2.0

Betragt følgende eksperiment: Et kast med to terninger, en grøn og en rød. Antal øjne betragtes. Lad den stokastiske variabel X angive *forskellen* på øjnene af de to terninger. Benyt idéerne i eksempel 1 til at løse følgende opgaver:

a) Bestem $P(X=3)$.
b) Hvilke værdier kan X antage? Bestem sandsynlighedsfordelingen for X.
c) Bestem sandsynligheden for at terningernes differens højst er 2, dvs. $P(X \le 2)$.

Opgave 2.1

Eksperimentet er et kast med tre mønter. Man interesserer sig for, om en mønt viser plat eller krone. Det er en pædagogisk hjælp at tænke på, at mønterne er nummererede. Lad for eksempel (k,p,k) betyde, at mønt 1 viser krone, mønt 2 viser plat og mønt 3 krone.

a) Opskriv de 8 mulige udfald. Overvej hvorfor de er lige mulige?

I det følgende lader vi X være den stokastiske variabel, som angivet antallet af plat ved det pågældende kast med tre mønter.

b) Angiv de mulige værdier for X og bestem sandsynlighedsfordelingen for X.

Opgave 2.2

Husk Maxwell-Boltzmann fordeling for molekylers fart i en gas, omtalt i eksempel 2. Det er et eksempel på en kontinuert fordeling. Vi skal undersøge molekylernes fart i en Argon-40 gas ved stuetemperatur. Argon-40 har $m = 39{,}9624\,\text{u} = 6{,}6359 \cdot 10^{-26}\,\text{kg}$ som atommasse og $T = 293\,\text{K}$.

a) Tegn grafen for fordelingens tæthedsfunktion på din grafregner. Benyt som vindue $x \in [0;1000]$ og $y \in [0;0{,}004]$. Beskriv kurvens form med ord.
b) Bestem sandsynligheden for at et molekyles fart er mindre end 300 m/s.
c) Bestem sandsynligheden for at farten af et molekyle er over 500 m/s.
d) Hvad er sandsynligheden for, at farten af et molekyle er imellem 100 og 400 m/s?
e) Hvad er den mest sandsynlige fart for et molekyle i gassen?

Opgave 2.3

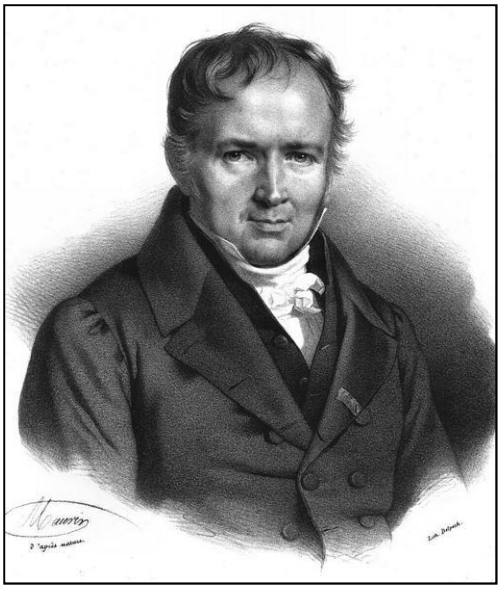
Bestem sandsynligheden for en sum på mindst 4 og højst 6 ved ét kast med to terninger?

Opgave 2.4 (Poissonfordelingen)

Den såkaldte Poissonfordeling er en diskret fordeling med følgende sandsynlighedsfordeling:

$$P(X=k) = \frac{e^{-\lambda}\cdot\lambda^k}{k!};\quad k=0,1,2,3,\ldots$$

Hvor $k!=1\cdot 2\cdot 3\cdot\ldots\cdot k$ og λ er en parameter. I Texas 89' $\forall$-menu kan man frembringe udråbstegnet med kombinationen ∞ε÷.



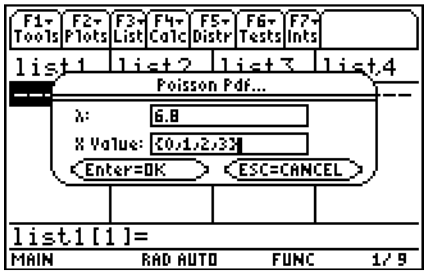
Den store franske matematiker *Simeon Denis Poisson* (1781 – 1840) lægger navn til fordelingen, fordi han var den første, der fandt frem til den som en approksimation til binomialfordelingen (afsnit 7) for $\lambda = n\cdot p$, hvor p er lille og n er stor. Vi skal ikke gå i detaljer hermed, blot nævne, at en tysk professor, *Ladislaus von Bortkiewicz*, i 1898 skrev en artikel, som hurtigt transformerede Poissons grænseformel til en helt ny sandsynlighedsfordeling. Bortkiewicz studerede blandt andet data for antallet af preussiske kavalerisoldater, som blev sparket til døde af deres heste. Ved at analysere disse data, var han i stand til at vise, at ovenstående formel er en brugbar sandsynlighedsmodel, helt uden reference til binomialfordelingen. Andre forskere fulgte hurtigt trop, og der viste sig en hel sværm af anvendelser af fordelingen. I dag betragtes fordelingen som værende blandt de 3-4 mest betydende fordelinger i sandsynlighedsregningen og statistikken.

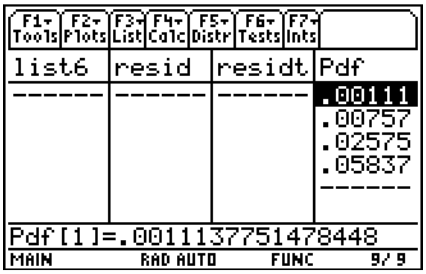
Man kan vise, at middelværdien og variansen af en Poissonfordelt stokastisk variabel begge er lig med parameteren λ: $E(X)=\lambda$, $\mathrm{Var}(X)=\lambda$.

I det følgende skal vi se på en vigtig anvendelse af Poissonfordelingen. I kernefysikken er det en velkendt sag, at man ikke kan forudsige, hvornår en given radioaktiv partikel henfalder, blot angive en sandsynlighed derfor. Dermed vil det være forskelligt, hvor mange kerner, der henfalder i forskellige tidsintervaller af samme tidslængde. Vi antager i det følgende, at den radioaktive kilde har en så stor halveringstid, at vi kan betragte kildens styrke som værende konstant. Lad os sige, at vi har givet en α-radioaktiv Thorium-kilde, som i gennemsnit henfalder med en hastighed af 3,4 kerner pr. minut. Lad os sige, at man betragter tidsrum af en længde på 2 minutter. Dermed vil der pr. tidsrum i gennemsnit henfalde 6,8 kerner. Ifølge formlen for middelværdien haves derfor $\lambda = 6{,}8$.

a) Benyt formlen ovenfor til at bestemme sandsynligheden for, at der forekommer netop 4 henfald i løbet af 2 minutter.

Du kan også benytte Texas 89 til at bestemme Poisson-sandsynligheder: I Stats/List-editoren vælges menuen `Distr` via ◻ og derefter undermenuen `D:Poisson Pdf ...`. I den fremkomne boks indtastes de relevante værdier: Dels parameteren λ og dernæst de *k*-værdier, for hvilket man ønsker $P(X=k)$ bestemt. På figuren til venstre er *k*-værdierne fra 0 til 3 skrevet med komma imellem og krøllede parenteser om. Resultatet ses på figuren til højre.



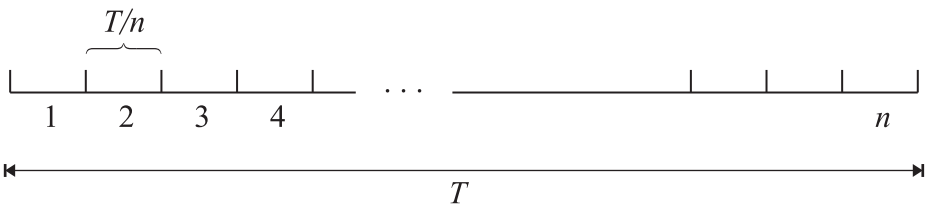


b) Bestem $P(X=k)$ for $k=0,1,2,\ldots,15$ ved hjælp af Texas 89.

c) Bestem sandsynligheden for, at der henfalder mere end 10 kerner i løbet af 2 minutter, dvs. bestem $P(X>10)$. *Hjælp*: Bemærk, at $P(X>10)=1-P(X\le 10)$. Benyt dernæst Texas 89' undermenu undermenuen `E:Poisson Cdf ...` til at bestemme den kumulerede sandsynlighed $P(X\le 10)$.

d) Med et Geiger-Müller-rør og en tæller kan du måle tællinger fra en radioaktiv kilde i en række tidsintervaller af fast længde og optælle hyppighederne af tælletallene. Sammenlign derefter fordelingen af tælletal med en Poissonfordeling med parameter λ, hvor du vælger λ til at være gennemsnittet af samtlige tælletal i serien.

Kommentar: Et spørgsmål, som uvægerligt dukker op er: "Hvorfor kan Poissonfordelingen benyttes i så forskelligartede eksempler som fordelingen af dødsfald efter hestespark i en preussisk krig og fordelingen af henfaldet af en α-radioaktiv kilde"? Forklaringen er, at hvis man piller de uvæsentlige detaljer fra, så ender man op med en basal model, som ligger til grund for begge situationer. Netop denne skelnen mellem væsentlige og uvæsentlige aspekter er selve kernen i anvendelsen af mange matematiske modeller i praksis. Poisson-modellen kan beskrives på følgende måde: Forestil dig, at et tidsinterval *T* inddeles i *n* lige store delintervaller af længden T/n, hvor *n* er stor.



Poisson-modellen

1. Sandsynligheden for to eller flere forekomster i et delinterval er praktisk talt 0.
2. De enkelte forekomster er uafhængige hændelser.
3. Sandsynligheden for en forekomst i et givet delinterval er konstant i hele tidsintervallet på *T*.

Man kan nu benytte binomialfordelingen fra afsnit 7 på situationen: Enten er der en forekomst i et tidsinterval eller ej. Vi skal ikke gå i detaljer med udledningen, men blot nævne, at i grænsen for $n \to \infty$ fås Poissonfordelingen.

Opgave 2.5

Eksponentialfordelingen er et eksempel på en kontinuert fordeling. Tæthedsfunktionen er givet ved følgende udtryk, hvor *c* er en parameter:

$$f(t) = \begin{cases} 0 & \text{for } t < 0 \\ c \cdot e^{-ct} & \text{for } t \geq 0 \end{cases}$$

Fordelingens vigtighed består i, at den forekommer i en række anvendelser, som har med *levetider* eller *ventetider* at gøre. Det kan være levetiden for en radioaktive partikel før den henfalder, ventetiden før en ansat i en virksomhed tager telefonen og reagerer på ens opkald eller som i dette eksempel ventetiden før en elektrisk pære springer.

NB! For at ovenstående tæthedsfunktion beskriver tiden *mellem* to begivenheder, skal nogle krav være opfyldt: Begivenhederne skal indtræffe tilfældigt i tid på en måde så sandsynligheden for, at der fremover er en bestemt ventetid, er *uafhængig* af, hvor lang tid der allerede er gået forud! Man siger også, at fordelingen er *glemsom*. Hvis begivenhederne optræder, således at antallet af begivenheder pr. tidsenhed er Poissonfordelt (se opgave 2.4), så er tiden *mellem* begivenhederne eksponentialfordelt.

I denne opgave skal vi se på levetiden for en elektrisk pære før den springer. En bestemt type pære har en gennemsnitlig levetid på 1000 timer, hvilket man kan vise giver anledning til følgende tæthedsfunktion:

$$f(t) = \begin{cases} 0 & \text{for } t < 0 \\ 0{,}001 \cdot e^{-0{,}001t} & \text{for } t \geq 0 \end{cases}$$

a) Tegn grafen for fordelingens tæthedsfunktion på din grafregner. Indtast kun den del af funktionen, som vedrører $t \geq 0$. Som vindue kan $x \in [0;5000]$ og $y \in [0;0{,}001]$ benyttes. Beskriv kurvens form med ord.
b) Bestem sandsynligheden for, at pæren lever under 200 timer.
c) Bestem sandsynligheden for, at pæren lever mere end 2000 timer.
d) (Svær) Bestem fordelingens *median*, dvs. den tid, for hvilket netop 50% af pærerne vil have en mindre levetid. *Hjælp*: Du skal have fat i en stamfunktion til tæthedsfunktionen.

Opgave 2.6

Betragt eksempel 2: Vis, ved at foretage en funktionsundersøgelse af tæthedsfunktionen i (1), at den mest sandsynlige molekylfart er givet ved udtrykket $v_p = \sqrt{2kT/m}$. Benyt udtrykket til at vise, at den mest sandsynlige molekylefart med de angivne talværdier er lig med 541,4 m/s - som også nævnt i eksempel 4.

Opgave 3.0

En stokastisk variabel X kan antage værdierne 0, 1, 2, 3, 4 og 5 og har følgende sandsynlighedsfordeling:

x_i	0	1	2	3	4	5
$P(X = x_i)$	0,15	0,10	0,30	0,10	0,20	0,15

a) Bestem middelværdien $E(X)$.
b) Bestem variansen $Var(X)$ og spredningen $\sigma(X)$.

Opgave 3.1

Bevis sætning 7 i tilfældet med en kontinuert stokastisk variabel. Du skal benytte nogle simple integrationsregler.

Opgave 3.2

Vi skal se på en kontinuert fordeling (se også opgave 2.5). En stokastisk variabel X siges at være *eksponentialfordelt*, hvis tæthedsfunktionen ser således ud:

$$f(x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ a \cdot e^{-ax} & \text{for} \quad x \geq 0 \end{cases}$$

hvor a er en positiv konstant. Lad os i det følgende sige, at $a = 2$.

a) Vis ved brug af Texas 89, at arealet under grafen for tæthedsfunktionen virkeligt er lig med 1, som det skal være. Benyt desuden grafregneren til at skitsere grafen for tæthedsfunktionen.
b) Eftervis ved brug af Texas 89, at $E(X) = \frac{1}{2}$ og at $Var(X) = \frac{1}{4}$.
c) Bestem middelværdi og varians for eksponentialfordelingen for et vilkårligt a.

Opgave 3.3

En bookmaker foreslår følgende spil til en spiller: Ét spil består i at kaste tre gange med en mønt. Hvis spilleren udelukkende får plat, skal han betale 20 kr. Hvis spilleren får krone to gange lige efter hinanden, vinder han 5 kr. I alle andre tilfælde vinder spilleren 1 kr. Er det et fornuftigt spil for spilleren i det lange løb?

Hjælp: Opskriv først de 8 mulige udfald ved ét spil, dvs. tre kast – husk at rækkefølgen er væsentlig! Indfør en stokastisk variabel X, som skal være gevinsten ved ét spil. Hvilke værdier kan X antage? Bestem sandsynligheden for hver af disse værdier ved at betragte listen med de 8 mulige udfald. Da du således har sandsynlighedsfordelingen for X, kan du afgøre spørgsmålet ved at udregne middelværdien $E(X)$.

Opgave 3.4

Nogle børn laver en spillebule. Spillet består i at kaste med to terninger som i eksempel 8. Det vedtages, at boden vil udbetale 8 kr. hvis summen af øjnene er mindst 8, 3 kr., hvis summen er 6 eller 7, mens spilleren må punge ud med 15 kr., hvis summen af øjnene er 5 og derunder.

a) Udregn en sandsynlighedsfordeling for X.
b) Bestem middelværdien for X, dvs. $E(X)$.
c) Er spillet fordelagtig for boden eller spilleren?

Opgave 3.5

Betragt den stokastiske variabel X i opgave 3.0. Vi indfører en ny stokastisk variabel Y ved $Y = 2X - 3$.

a) Opskriv sandsynlighedsfordelingen for Y.
b) Bestem middelværdien $E(Y)$ og variansen $Var(Y)$ for Y ved at benytte sandsynlighedsfordelingen fra spørgsmål a).
c) Stemmer resultatet fra spørgsmål b) med det du får ved at benytte sætning 7?

Opgave 3.6

Lad den stokastiske variabel X have følgende sandsynlighedsfordeling:

x_i	–2	–1	0	1	2	3
$P(X = x_i)$	0,1	0,2	0,1	0,2	0,1	0,3

a) Bestem middelværdi og varians for X.

Betragt en ny stokastisk variabel Y givet ved $Y = X^2$.

b) Angiv sandsynlighedsfordelingen for Y. *Hjælp*: Hvilke værdier y_i kan Y antage og hvad er sandsynlighederne for disse værdier?
c) Bestem middelværdien for Y: $E(Y) = E(X^2)$.
d) Kontroller ved hjælp af resultaterne i a) og c), at $\text{Var}(X) = E(X^2) - E(X)^2$.
e) (Svær). Formlen for variansen i spørgsmål d) gælder generelt. Prøv at bevise sætningen ved hjælp af definitionerne 5 og 6.

Opgave 4.0

Lad X være en stokastisk variabel, som er normalfordelt med middelværdi $\mu = 5,0$ og spredning $\sigma = 1,7$.

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

Opgave 4.1

Når der fyldes mælk på en liter karton, så fyldes der ikke altid præcist 1 liter mælk i. Maskinen, der foretager påfyldningen, vil udvise små variationer fra gang til gang, og vi kan antage, at der er tale om en normalfordeling. For at være lidt på den sikre side indstilles maskinen til i middel af påfylde 1,015 liter. Antag, at spredningen er 0,008 liter.

a) Hvad er sandsynligheden for at et karton indeholder mindre end 1 liter?
b) Man kan indstille maskinen, så middelværdien øges. Hvad skal den indstilles til for at sikre, at kun 0,5% indeholder mindre end 1 liter? *Hjælp*: Benyt sætning 10b).

Opgave 4.2

Et værksted producerer små metalplader. Det antages, at tykkelsen af de producerede plader er normalfordelt med en middelværdi på 2,50 mm og en standardafvigelse (= spredning) på 0,10 mm. Køberen kræver, at pladernes tykkelse højst må afvige med 0,15 mm fra middelværdien.

a) Hvor mange procent af metalpladerne må kasseres?
b) Firmaets kvalitetskontrol ønsker at forbedre nøjagtigheden, så kun 2% af pladerne skal kasseres. Hvad skal spredningen reduceres til?

Opgave 4.3

Som vist i afsnit 5 er voksne mænds højde med meget stor tilnærmelse normalfordelt. Løs nedenstående opgaver, idet det antages, at gennemsnitshøjden for danske mænd er 180 cm og spredningen er 6,8 cm.

a) Hvor mange procent af mændene har en højde på under 165 cm?
b) Hvor mange mænd på 195 cm og derover kan man forvente at finde på Handelshøjskolen i København med 1000 mandlige elever?

Opgave 4.4

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 4.5

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?

Opgave 4.6 (Svær)

Lad X være en normalfordelt stokastisk variabel. Det oplyses, at $P(X \le 3) = 0,32$ og at $P(X \le 7) = 0,79$. Bestem middelværdien og spredningen for normalfordelingen.

Hjælp: De to oplysninger giver to punkter på fordelingsfunktionen og det specificerer normalfordelingen fuldstændigt. Hvis man benytter sætning 10b) på det første punkt fås følgende:

$$0,32 \;=\; P(X \le 3) \;=\; \Phi\left(\frac{3-\mu}{\sigma}\right) \;\Leftrightarrow\; \frac{3-\mu}{\sigma} \;=\; \Phi^{-1}(0,32) \;=\; -0,467699$$

hvor den inverse funktion til fordelingsfunktionen for standardnormalfordelingen fås frem på Texas 89 som anført i eksempel 13. Det andet punkt behandles på samme måde. Herved fås to ligninger med to ubekendte, nemlig μ og σ. Herved kan de ubekendte bestemmes, evt. via Texas 89.

Opgave 4.7

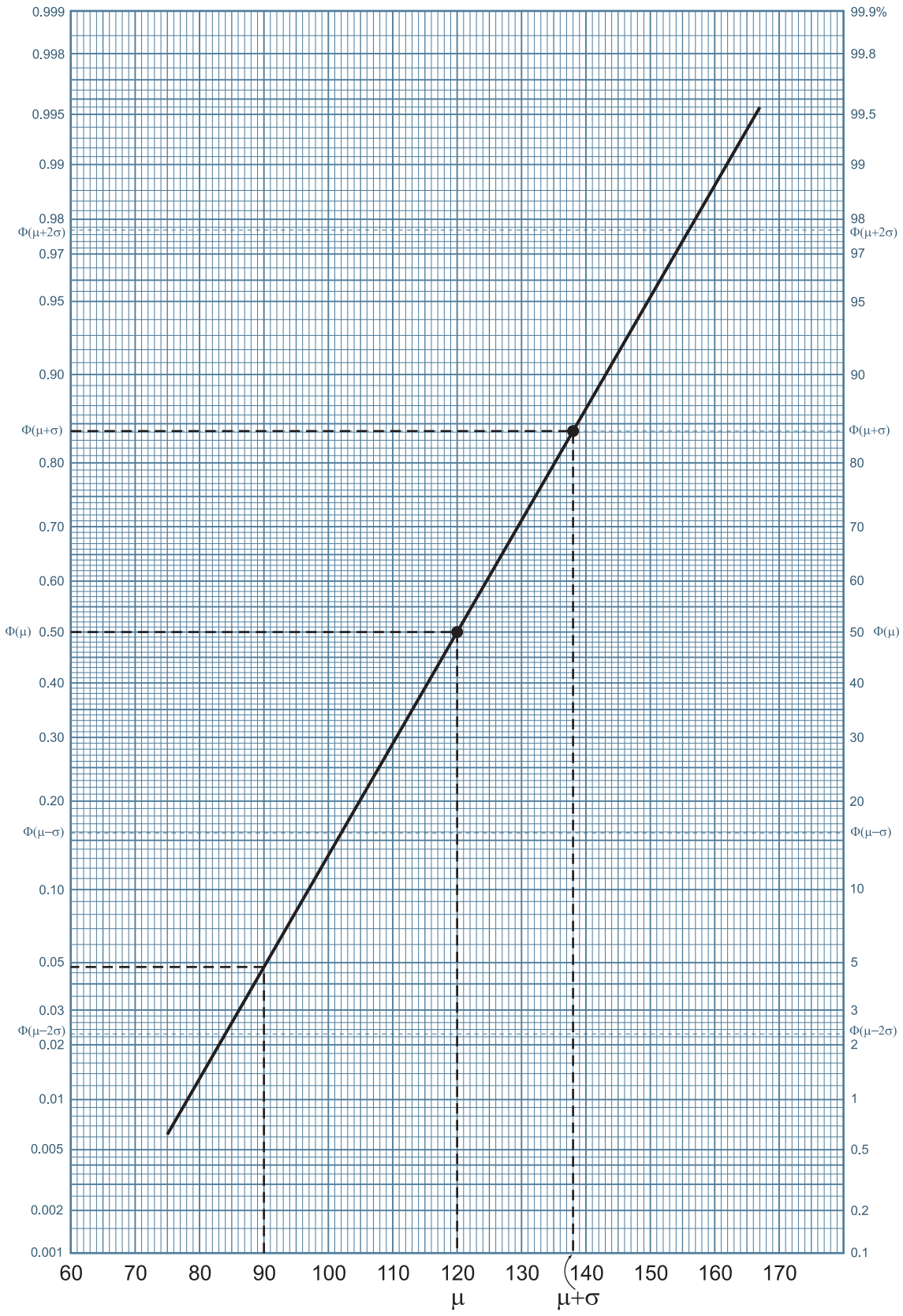
British Medical Journal, Vol. 307, 24. juli 1993, side 234, rapporterer om en undersøgelse af 5459 gravide kvinder, som benyttede Aarhus Universitets Hospital. Middeltallet for graviditetsperioden var 281,9 dage med en spredning på 11,4 dage. Bestem den procentdel af fødslerne, som resulterede i for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

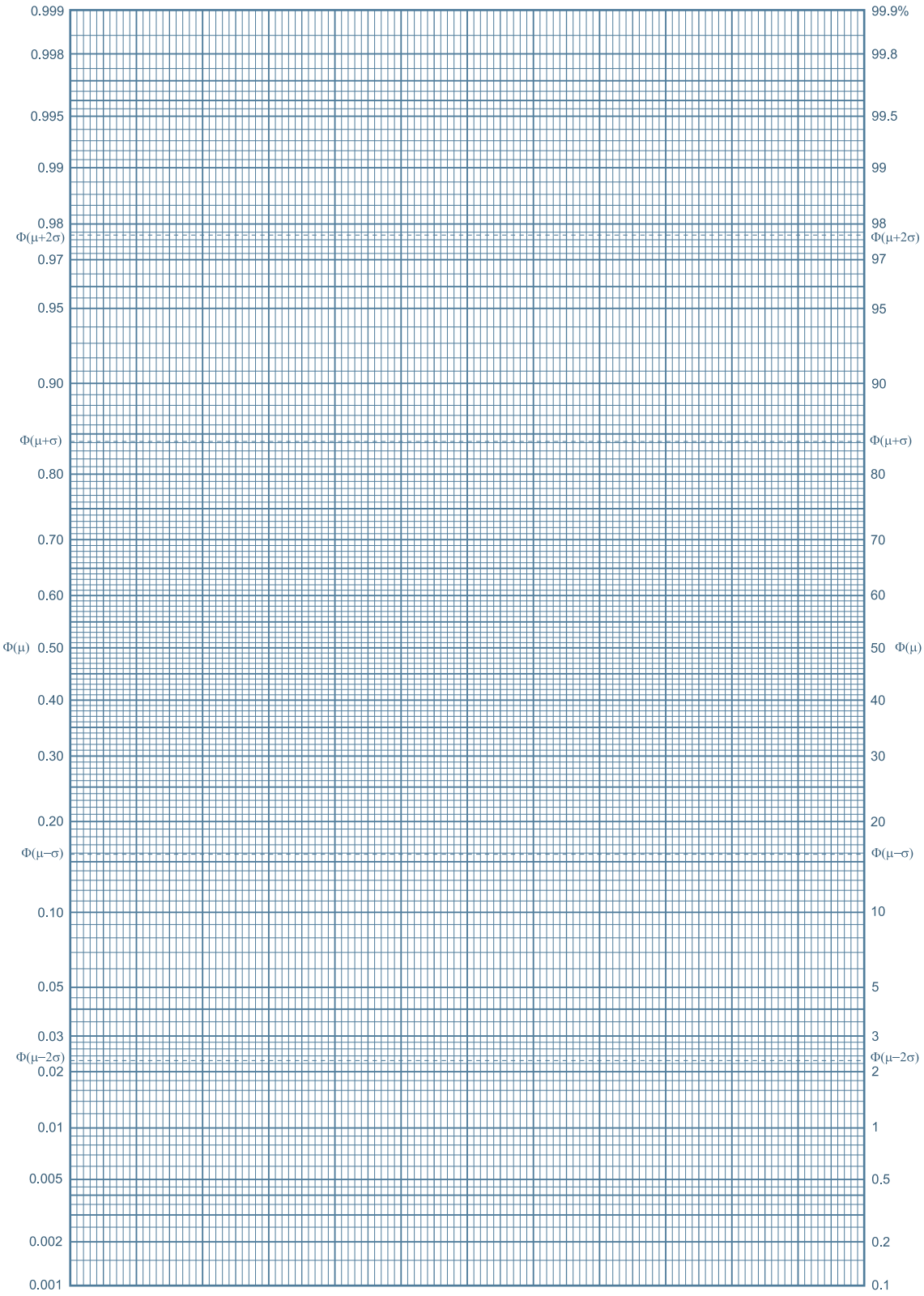
Opgave 4.8 (Normalfordelingspapir)

Der er et alternativ til at løse opgaver med normalfordelingen ved hjælp at grafregner eller IT-hjælpemidler, og det er ved at bruge *normalfordelingspapir*. I et almindeligt koordinatsystem vil fordelingsfunktionen for en normalfordeling være en S-formet kurve, som omtalt i afsnit 4. På normalfordelingspapir er *y*-aksen blevet deformeret, så fordelingskurven i stedet bliver til en ret linje. Pointen er, at det er nemmere at afgøre, om en kurve er en ret linje end at afgøre om den S-formede kurve krummer på den helt rigtige måde. Ved hjælp at papiret kan man løse mange opgaver. Normalt skal man have mindst to oplysninger for at kunne specificere situationen fuldstændigt.

Som et eksempel har vi her fået opgivet, at vi har en normalfordelt stokastisk variabel X med parametre $\mu = 120$ og $\sigma = 18$. Bestem sandsynligheden $P(X \ge 90)$. Den første oplysning med middelværdien giver os et punkt på grafen: (120; 0,50). Husk, at middelværdien samtidig er median, da normalfordelingen er symmetrisk! Den anden oplysning giver os $(\mu + \sigma; \Phi(\mu + \sigma)) = (120 + 18; 0,8413) = (138; 0,8413)$ som det andet punkt. Bemærk, at $\Phi(\mu + \sigma)$ er markeret på *y*-aksen på papiret. Vi ved, at der er tale om en normalfordeling og at en sådan skal give en ret linje på papiret. Derfor tegnes en ret linje gennem de nævnte to punkter. Herefter kan $P(X \le 90) = 0,048$ aflæses på grafen. Dermed haves: $P(X \ge 90) \;=\; 1 - P(X \le 90) \;=\; 1 - 0,048 \;=\; 0,952$. I det følgende skal du selv løse følgende opgaver på samme graf:

a) $P(X \le 150)$ b) $P(95 \le X \le 140)$


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
99.9%
99.8
99.5
99
98
Φ(μ+2σ)
97
95
90
Φ(μ+σ)
80
70
60
50 Φ(μ)
40
30
20
Φ(μ−σ)
10
5
3
Φ(μ−2σ)
2
1
0.5
0.2
0.1
60 70 80 90 100 110 120 130 140 150 160 170
μ
μ+σ


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
99.9%
99.8
99.5
99
98
Φ(μ+2σ)
97
95
90
Φ(μ+σ)
80
70
60
50 Φ(μ)
40
30
20
Φ(μ−σ)
10
5
3
Φ(μ−2σ)
2
1
0.5
0.2
0.1

Opgave 4.9

De følgende blandede opgaver skal du løse med normalfordelingspapir. Hvis du mangler et tomt papir, så kopier/print normalfordelingspapiret på forrige side. Antag, at X er normalfordelt. På 1. aksen kan enheder vælges vilkårlige. Det er fornuftigt at sørge for, at medianen ligger tæt på midten af aksen!

a) Det oplyses, at $\mu = 0,7$ og $\sigma = 0,3$. Bestem $P(X \le 0,4)$.
b) Givet $P(X \le 32) = 0,42$ og $P(X \ge 60) = 0,26$. Bestem μ og σ.
c) Lad $P(X \le 15) = 0,10$ og $\mu = 2$. Bestem spredningen σ.
d) Lad $P(X \ge 13,5) = 0,7$ og $\sigma = 2,2$. Bestem middelværdien μ.

Opgave 4.10

Man kan vise, at hvis X_1 og X_2 er to *uafhængige* normalfordelte stokastiske variable, så er linearkombinationen $X = a_1X_1 + a_2X_2,\ a_1, a_2 \in R$ igen en normalfordelt stokastisk variabel med $E(X) = a_1E(X_1) + a_2E(X_2)$ og $\text{Var}(X) = a_1^2\,\text{Var}(X_1) + a_2^2\,\text{Var}(X_2)$. For drenge i alderen 18-20 år gælder der ifølge eksempel 14, at deres højde er normalfordelt med middelværdi 180,1 cm og spredning 6,81 cm. Der findes ikke tilsvarende data for kvinder, men vi gætter på, at middelværdien er 168,0 cm og spredningen 6,0 cm. En dreng og en pige i alderen 18-20 år vælges tilfældigt ud. Hvad er sandsynligheden for, at drengen er højere end pigen? *Hjælp*: Lad X_1 angive drengens højde og lad X_2 angive pigens højde. Disse kan antages uafhængige. Sæt $X = X_1 - X_2$ og udnyt ovenstående sætning.

Opgave 4.11 (Abstrakt svær – måleusikkerhed i fysik)

I det følgende skal vi gøre brug af en sætning, som anføres uden bevis:

Lad $X_1, X_2, \ldots, X_n$ være normalfordelte stokastiske variable med middelværdier henholdsvis $\mu_1, \mu_2, \ldots, \mu_n$ og varianser henholdsvis $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Enhver linearkombination $Y = a_1 \cdot X_1 + a_2 \cdot X_2 + ,\ldots, + a_n \cdot X_n$, hvor $a_i \in R$, er da igen en normalfordelt stokastisk variabel med middelværdi $\mu_Y = a_1 \cdot \mu_1 + a_2 \cdot \mu_2 + \ldots + a_n \cdot \mu_n$ og varians lig med $\sigma_Y^2 = a_1{}^2 \cdot \sigma_1{}^2 + a_2{}^2 \cdot \sigma_2{}^2 + \ldots + a_n{}^2 \cdot \sigma_n{}^2$.

Erfaringer har vist, at normalfordelingen ofte er brugbar til at beskrive de variationer, som man oplever ved målinger af en given størrelse i fysiske eksperimenter. Lad os i det følgende gå ud fra, at den stokastiske variabel X_i, som angiver resultatet af den i'te måling, er normalfordelt med middelværdi μ og spredning σ.

a) Vis at den stokastiske variabel Y givet ved $Y = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$ har middelværdi $\mu_Y = \mu$ og varians lig med $\sigma_Y^2 = \frac{1}{n} \cdot \sigma^2$.
b) Forklar, hvorfor spørgsmål a) bekræfter den velkendte kendsgerning, at man i fysiske eksperimenter får et mere sikkert resultat ved at gentage forsøget flere gange og tage gennemsnittet af måleværdierne.

Opgave 5.0 (Projekt med normalfordelt data)

Fra *Danmarks Statistik*s hjemmeside www.dst.dk kan man finde følgende indekstal for kvinders *fertilitetskvotienter* i 2005 som funktion af alderen i år:

Alder	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Indeks	0,5	1,4	3,7	7,4	15,9	24,3	32,2	41,4	52,1	67,0
Alder	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**
Indeks	91,4	108,7	128,7	143,5	153,9	151,4	144,0	133,1	112,6	94,5
Alder	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**
Indeks	80,6	64,4	47,2	35,5	24,8	16,9	11,8	6,6	3,3	1,7
Alder	**45**	**46**	**47**	**48**	**49**					
Indeks	0,9	0,4	0,1	0,1	0,0					

Behandl data på samme måde som i afsnit 5, enten ved hjælp af Excel eller Texas 89. Afgør om data er normalfordelte? Bestem i så fald middelværdi og spredning. Hvad fortæller dette om kvinders fertilitet?

Opgave 5.1

Undersøg om lønninger er normalfordelte via Danmarks statistiks hjemmeside.

Opgave 6.0

I en fodboldtrup er der 15 spillere. Der ses i det følgende bort fra de specifikke pladser, som spillerne kan spille.

a) På hvor mange måder kan man udtage de 11 spillere, som skal spille den dag? (man tager ikke hensyn til spillernes placering på banen).
b) Samme spørgsmål, hvor man tager hensyn til spillernes placering på banen.

Opgave 6.1

Lotto er et spil, hvor en spiller kan afkrydse 7 tal på en liste med tallene fra 1 til 36. Ved udtrækning udtrækkes 7 numre tilfældigt fra en tromle.

a) På hvor mange måder kan man udtrække 7 numre ud af 36? Hvad er sandsynligheden dermed for at få 7 rigtige?
b) Hvor mange måder er der at få 6 rigtige på? (Benyt multiplikationsprincippet!)

Opgave 6.2

Hvor mange måder kan man udtage 5 kort af et spil på 52 kort?

Opgave 7.0

a) Bestem sandsynligheden for at få netop 4 toere ved 20 slag med en terning.
b) Hvad er sandsynligheden for at få mindst 11 rigtige på en tipskupon, hvis man benytter "sypigetips"? *Hjælp*: Hvad er basiseksperimentet og basissandsynligheden?

Opgave 7.1

Man regner med, at 5% af drengene i Danmark er farveblinde. Benyt binomialfordelingen til at løse nedenstående opgaver. Redegør samtidigt for, hvorfor fordelingen kan benyttes i den aktuelle situation. Vi betragter en klasse med 25 drenge.

a) Hvad er sandsynligheden for, at der netop er 1 farveblind dreng i klassen?
b) Hvad er sandsynligheden for, at der i klassen findes mindst én dreng, som er farveblind?
c) På hele skolen er der 300 drenge. Hvad er sandsynligheden for, at der er mere end 25 farveblinde drenge på skolen?
d) Hvad er det gennemsnitlige antal farveblinde drenge på en skole af den nævnte størrelse?

Opgave 7.2

I den danske befolkning er der følgende fordeling af blodtyper:

	Rhesus positiv	Rhesus negativ
A	37%	7%
B	8%	2%
0	35%	6%
AB	4%	1%

Der udtages en tilfældig gruppe på 50 personer fra den danske befolkning.
a) Hvad er sandsynligheden for, at der netop er fem B Rhesus positive i gruppen?
b) Hvad er middelværdien for antallet af B Rhesus positive i gruppen?
c) Bestem sandsynligheden for, at der er mindst seks med blodtype 0 Rhesus negativ.

Opgave 7.3 (svær)

Hvor mange gange skal man kaste en mønt for at være 99% sikker på at få mindst én krone? *Hjælp*: Hvad er det modsatte af at få mindst én krone? Hvad skal sandsynligheden for at dette forekommer da være?

Opgave 7.4 (Chevalier de Mérés problem)

Den franske adelsmand og gambler *Chevalier de Méré* studerede følgende to hændelser:

1) At få mindst én sekser ved fire kast med *en* terning.
2) Mindst én gang at få en dobbelt-sekser ved 24 kast med *to* terninger.

Selv om han havde en anelse om, hvilken hændelse, der var den mest sandsynlige, kunne de Méré ikke redegøre for det. Derfor henvendte han sig til den store franske matematiker *Blaise Pascal* (1623 – 1662). Pascals svar bekræftede Mérés formodning. I det følgende skal du prøve selv at bestemme de to sandsynligheder. *Hjælp*: Se på den omvendte eller modsatte hændelse og benyt: $P(X \geq 1) = 1 - P(X = 0)$.

Opgave 7.5

På en fabrik produceres der en komponent til en bil. Det vides erfaringsmæssigt, at 12% af komponenterne har en defekt. Der udtages på tilfældig vis 50 komponenter.

a) Hvad er sandsynligheden for, at netop 7 komponenter har defekten?
b) Hvad er sandsynligheden for at få mindst 6 og højst 8 defekte komponenter?
c) Hvad er sandsynligheden for at mindst 4 komponenter er defekte?
d) Hvad er det gennemsnitlige antal defekte komponenter i stikprøven?
e) Hvor stor er spredningen?

Opgave 7.6

I en multiple-choice prøve er der 25 spørgsmål med hver 5 mulige svar. Antag, at du ikke ved et klap om emnet og svarer helt tilfældigt. Hvad er da sandsynligheden for at få mindst 10 rigtige i prøven? Hvad er middeltallet for antal rigtige svar?

Opgave 8.0

Bestem middelværdi, varians og spredning for den normalfordeling, som tilnærmer binomialfordelingen med antalsparameter $n = 75$ og basissandsynlighed $p = 0,2$.

Opgave 8.1

Boing 757 fly er på visse destinationer beregnet til at flyve med 168 passagerer på økonomiklasse. Imidlertid viser erfaringer, at kun 90% af dem, der reserverer billet til turen, dukker op. For at udnytte kapaciteten bedst muligt gør man ofte det, at man *overbooker* flyet. Hvis for mange bookede personer skulle møde op, tilbydes nogle at vente til det næste fly, mod at

modtage en pæn kompensation. Lad X være den stokastiske variabel, som angiver antallet af bookede personer, som dukker op. Begivenheden at to bookede personer dukker op eller ej kan ikke helt siges at være *uafhængige* hændelser, for undertiden ankommer personer i familier! Alligevel er det en brugbar approksimation at antage, at de er uafhængige. Lad os sige, at man overbooker flyet ved at booke 178 personer.

a) Benyt binomialfordelingen til at bestemme sandsynligheden for, at der ankommer for mange passagerer til flyafgangen.

I afsnit 8 blev det antydet, at binomialfordelingen kan tilnærmes med en normalfordeling under visse forudsætninger. Figuren i afsnit 8 indikerer, at der gælder følgende approksimation:

$$P(a \le X \le b) \approx \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot\int_{a-0.5}^{b+0.5} e^{-\frac{1}{2}\cdot\left(\frac{t-\mu}{\sigma}\right)^2} dt = F_{\mu,\sigma}(b+0.5) - F_{\mu,\sigma}(a-0.5)$$

hvor $\mu = n\cdot p$ og $\sigma = \sqrt{n\cdot p\cdot(1-p)}$. Som en tommefingerregel kan det tilføjes, at approksimationen er god, når $n > 9\cdot p/(1-p)$ og $n > 9\cdot(1-p)/p$.

b) (Svær). Benyt approksimationen med normalfordelingen samt Texas 89 til at besvare samme spørgsmål som i a). Sammenlign værdierne udregnet i a) og b).

Litteratur

Jesper Blom-Hanssen. *Statistik for praktikere*. Ingeniøren|bøger, 2002.

Abraham de Moivre. *The Doctrine of Chances or A Method of Calculating the Probabilities of Events in Play*. The Third Edition, 1756. Genoptryk af The American Mathematical Society, 2000.

Preben Blæsild, Jørgen Granfeldt. *Idrætsstatistik*, bind 1 og 2, Det naturvidenskabelige fakultet, Aarhus Universitet, 2001.

Claus Jessen, Peter Møller, Flemming Mørk. *Tal, statistik og sandsynligheder*. Gyldendal 1994.

Ole Groth Jørsboe. *Sandsynlighedsregning*. Matematisk Institut, Danmarks tekniske højskole, 1984.

Richard J. Larsen, Morris L. Marx: *An Introduction to Mathematical Statistics and Its Applications*. Fourth Edition. Pearson Prentice Hall, 2006.

Nikolaj Malchow-Møller, Allan Würtz. *Indblik i statistik – en grundbog for videregående uddannelser*. Gyldendal, 2003.