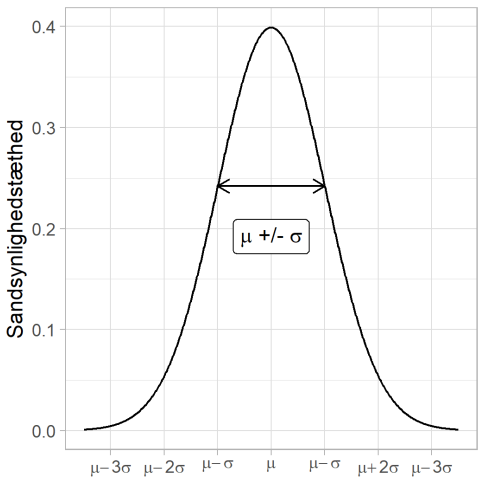
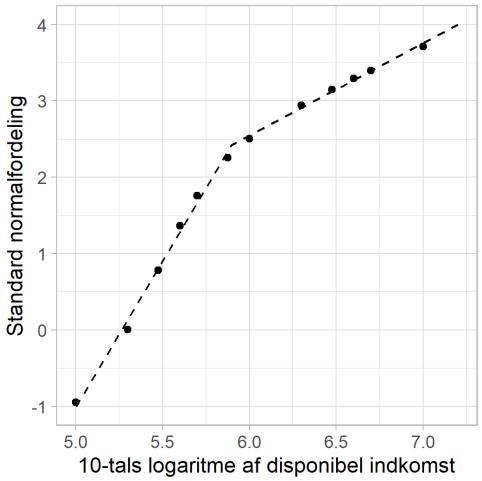
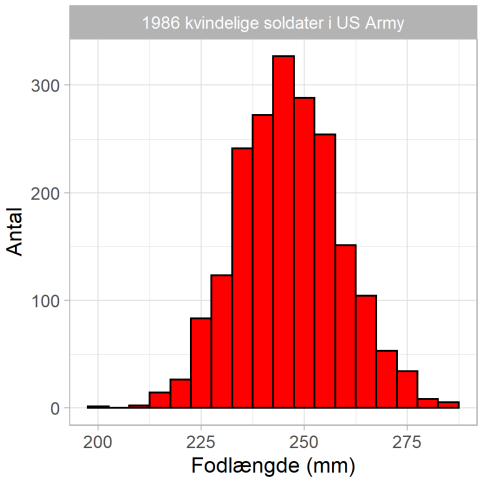
Normalfordelingen

Bo Markussen
Københavns Universitet

11. marts, 2020







1 Indledning

Forfatteren til dette notat er 186 cm høj. 186 er bare et tal, som her beskriver højden (målt i cm) på en konkret person. Højdemålinger siges ofte at være *normalfordelte*, men når man betragter højden (målt i cm) på en konkret person, så er det altså bare et tal. Højderne i en given gymnasieklasse, som vi for eksemplets skyld vil antage har 24 elever, kan tilsvarende beskrives med 24 tal. Nogle gange er sådan en nøjagtig beskrivelse på sin plads, f.eks. hvis der skal bestilles 24 ens kostumer i forskellige størrelser til en teaterforestilling gymnasieklassen skal opføre. Andre gange er en beskrivelse med de enkelte højder alt for detaljeret. Jo flere mennesker vi måler højden på, jo større synes behovet for en mere simpel beskrivelse at blive. I første halvår af 2018 var der f.eks. 18237 unge mennesker (primært mænd) der deltog i *Forsvarets Dag*, hvor de blandt mange andre ting fik målt deres højde. *Personalestyrelsens* opgørelse *"Udfaldet, Gennemsnitshøjde, BMI (Body Mass Index) på Forsvarets Dag / Sessionen"* [1] fortæller, at *gennemsnitshøjden* af disse personer var 180,82 cm. Som vi vil se senere, så er der god grund til at antage, at højdemålinger er normalfordelte indenfor hvert køn. Hvis *Personalestyrelsen* således også havde oplyst *højdespredningen* (som også er et enkelt tal, ligesom gennemsnittet), så havde disse tre oplysninger

normalfordeling & gennemsnit & spredning

givet en kortfattet og let kommuniker- og brugbar beskrivelse af alle højder[1]. Derimod er de 18237 individuelle højdemålinger en stor datamængde, som kun en computer vil kunne håndtere og bruge.

Formålet med dette notat er at introducere matematiske egenskaber ved normalfordelingen, som blandt andet vil godtgøre hvorfor og hvorledes normalfordelingen kan bruges til at beskrive mange aspekter af verdenen omkring os. Derudover introduceres tilhørende statistiske metoder til at undersøge om et givet konkret datasæt kan beskrives ved en normalfordeling.

1.1 Strukturering af manuskriptet

Før den matematiske formel for normalfordelingen kommer på bordet vil vi i afsnit 2.1 først se på konkrete normalfordelte datasæt fra biologiens verden. Hovedeksemplet er 1986 målinger af kvindelige amerikanske soldaters fodlængde. Dette leder frem til diskussionen og anvendelsen af vores vigtigste værktøj, nemlig *fraktiler* og *fraktildiagrammer*, som detajleres i afsnit 3.

[1] En mere præcis beskrivelse ville dog være at angive andelen af mænd og kvinder blandt de 18237 individer, samt gennemsnitshøjde og højdespredning separat for de to køn. Denne information er dog ikke tilgængelig i [1].

Normalfordelingen kaldes også for den *Gaussiske fordeling* efter den berømte tyske matematiker Carl Friedrich Gauss. Gauss opfandt *"mindste kvadraters metode"* (se [2] for en gennemgang i forbindelse med lineær regression) og beskrev normalfordelingen som den sandsynlighedsfordeling, der naturligt passer sammen med mindste kvadraters metode.

Figur 1: Carl Friedrich Gauss (1777–1855), tysk matematiker, astronom, geodæt og fysiker. Maleri af Christian Albrecht Jensen [3].

Specielt kan vi undersøge om et givet datasæt er normalfordelt ved at sammenligne med fraktilerne for de 1986 fodlængder.

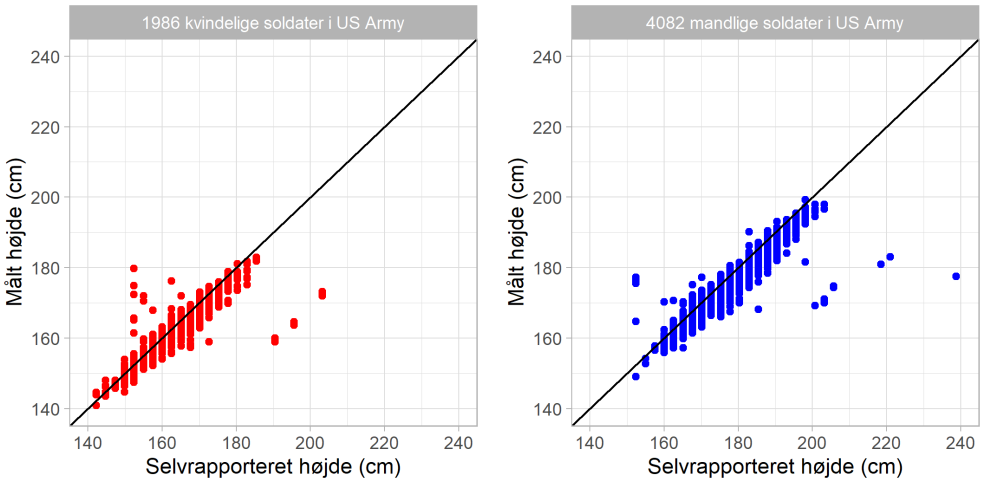
I afsnit 4 kommer den matematiske definition af normalfordelingen endelig på bordet, og i afsnit 4.1 gives et eksempel på en statistisk anvendelse. Den matematiske analyse af egenskaber af og beregninger med normalfordelingen fortsættes i afsnit 5, og i afsnit 6 præsenteres det utroligt vigtige matematiske resultat kaldet *Den Centrale Grænseværdisætning*. Populært sagt fortæller dette matematiske resultat, at vi kan forvente at finde normalfordelte tal i verden omkring os. For at understrege denne pointe vender vi i afsnit 7 tilbage til et konkret datasæt. Nemlig fordelingen af den disponible indkomst i *Danmark* i 2017. Ved brug af fraktildiagrammet er vi i stand til at komme med dybe socio-økonomiske betragtninger baseret på et simpelt datasæt bestående af 13 tal.

2 Opvarmning

For at udvikle vores matematiske og statistiske begrebsapparat vil vi gerne have adgang til et datasæt, som indeholder individuelle målinger. I stedet for at forsøge at få adgang til de 18237 højdemålinger fra *Forsvarets Dag* i første halvår af 2018 har jeg fundet et offentligt tilgængeligt datasæt med antropometriske[2] målinger på amerikanske soldater[3]. Det drejer sig om *"US Army Anthropometric Working Databases (ANSUR II)"* [4] datasættet fra 2012. Dette datasæt indeholder 93 antropometriske mål fra 6068 mennesker, hvoraf 1986 er kvinder og 4082 er mænd. Der er tale om måling af mange størrelser fra længden på ørerne til maveomkredsen, når man sidder ned.

[2]Antropometri betyder opmåling af menneskekroppen.

[3]Selv om vi således bliver i militærets verden, så er målinger af mennesker egentlig at betragte som biologiske data.



Figur 2: Punkterne viser målte mod selvrapporterede højder fra 2012 US Army Anthropometric Working Databases (ANSUR II) opdelt efter køn. De sorte linjer viser referencelinjen $y = x$. For at gøre sammenligningen mellem køn og mellem målte og selvrapporterede højder lettere er x- og y-akserne på begge delfigure ens (fra 140 cm til 240 cm).

Opgave 1. ANSUR II *datasættet indeholder to forskellige angivelser af højden på soldaterne. Dels selvrapporterede højder, hvor man altså har spurgt forsøgsdeltagerne om hvor høje de er. Dels målinger foretaget af militærlægerne. Umiddelbart vil man måske tro, at selvrapporterede og målte højder stemmer godt overens. Men ved at optegne disse to størrelser mod hinanden (se figur 2) kan man se nogle systematiske afvigelser. Diskutér følgende 5 aspekter med udgangspunkt i figur 2:*

1. *Er de mandlige soldater generelt set højere end deres kvindelige kollegaer?*

2. *Er de selvrapporterede højder generelt set større end de målte højder? Gælder dette for begge køn?*

3. *De amerikanske soldater har rapporteret deres højde i enheden* inches, *hvor* 1 inch *svarer til* 2,54 cm. *Dermed svarer* 60 inches *til* 152,4 cm. *Kan denne højde genfindes på figuren? Og hvad med* 80 inches*?*

4. *Hvorfor ligger alle punkterne på nogle lodrette linjer, hvor der altså er synlige spring mellem x-værdierne?*

5. *Er de målte højder mere pålidelige end de selvrapporterede højder? Diskutér også hvad* "mere pålidelig" *egentlig betyder.*

Opgave 1 var en træning i statistisk tankegang, altså det at diskutere omverdenen med udgangspunkt i et datasæt. Som vi senere vil argumentere for så kan de målte højder med god approksimation antages at være normalfordelte (indenfor de to køn hver for sig). De selvrapporterede højder har derimod nogle alt for store værdier til at kunne være normalfordelte.

Vi vil nu arbejde videre med en anden dimension af menneskekroppen, nemlig længden af fødderne[4]. Vi vil ikke opskrive hele datasættet med de 1986 kvindelige og de 4082 mandlige fodlængder. Det vil simpelthen fylde for meget og ikke give noget umiddelbart overblik. Men lad os opskrive de 6 første og de 6 sidste[5] fodlængder (målt i hele antal millimeter) for hvert af kønnene. For kvinderne er dette

$$246,\ 249,\ 265,\ 265,\ 247,\ 270,\ \ldots,\ 253,\ 259,\ 245,\ 249,\ 226,\ 239$$

og for mændene er dette

$$273,\ 263,\ 270,\ 267,\ 305,\ 254,\ \ldots,\ 278,\ 255,\ 263,\ 263,\ 266,\ 295$$

Opgave 2. *For kvinderne er den korteste og længste fodlængde henholdsvis 198 mm og 286 mm, og for mændene er dette 216 mm og 323 mm. Som vi umiddelbart kan se på de konkrete målinger, så er der mindst to kvinder der har samme fodlængde (i hele antal millimeter). Kvinde nummer 2 og kvinde nummer 1984 har nemlig begge en fodlængde på 249 mm. Her kommer tre matematiske spørgsmål som faktisk slet ikke har noget med normalfordelingen at gøre, så spring dem gerne over hvis du ikke synes de er sjove at tænke på.*

1. *Giv et logisk argument for, at der mindst er to kvinder (blandt de 1986 kvinder) der har samme fodlængde (i hele antal millimeter).*

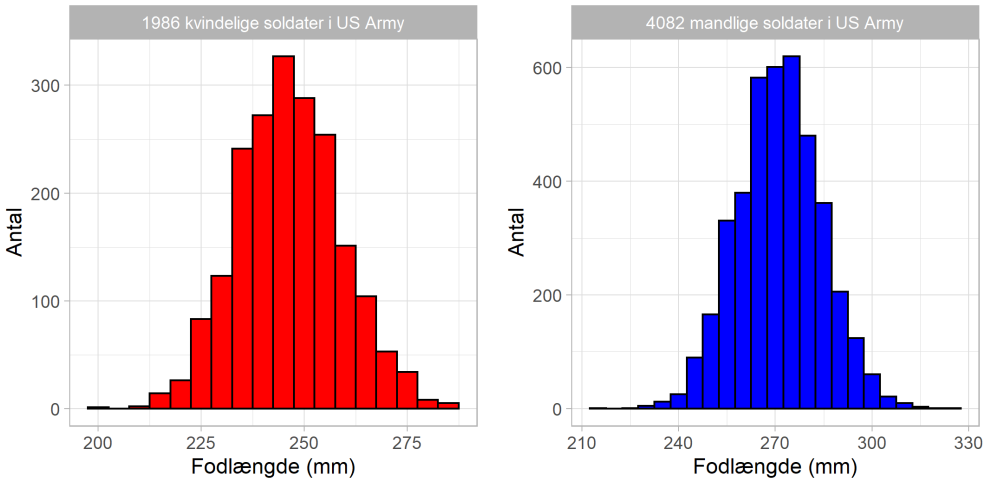
2. *Giv et logisk argument for, at der mindst er to mænd (blandt de 4082 mænd) der har samme fodlængde (i hele antal millimeter).*

3. *Kan man også være matematisk sikker på, at der er en kvinde og en mand der har samme fodlængde (i hele antal millimeter)?*

Hjælp: Svarerne kan findes ved at bruge Dirichlets skuffeprincip [6].

[4]Der er angivet en enkelt fodlængde per person. Jeg formoder, at dette er gennemsnitslængden for de to fødder. Men det kan også være, at man altid måler den samme fod (enten højre eller venste). Nøjagtig hvad der er blevet gjort er ikke beskrevet i datasættet, men det er heldigvis ligegyldigt for vores anvendelse.

[5]De "første" og de "sidste" henviser til observationernes rækkefølge i datasættet, og altså ikke til deres størrelse.



Figur 3: Data fra 2012 US Army Anthropometric Working Databases (ANSUR II). Histogrammerne angiver hyppigheder for bins med bredde 5 mm.

Histogrammer for fodlængderne findes i figur 3. For at lave et histogram skal man først vælge de såkaldte *bins*. Den danske oversættelse af det engelsk ord *"bin"* er *"beholder"*. Man kan altså forestille sig nogle beholdere (matematisk set er dette disjunkte intervaller) under x-aksen, således at datapunkterne (her målinger af fodlængder) hver falder i netop én af beholderne. Ofte vil man enten beslutte sig for antallet eller for bredden af bins. Vi beslutter os for at bruge bins af bredde 5 mm. Fra opgave 2 husker vi, at den korteste og længste fodlængde for kvinderne var henholdsvis 198 mm og 286 mm, hvormed bins bliver de 19 intervaller[6]:

$$
\begin{array}{llll}
]197{,}5\ ;\ 202{,}5] &]202{,}5\ ;\ 207{,}5] &]207{,}5\ ;\ 212{,}5] &]212{,}5\ ;\ 217{,}5] \\
]217{,}5\ ;\ 222{,}5] &]222{,}5\ ;\ 227{,}5] &]227{,}5\ ;\ 232{,}5] &]232{,}5\ ;\ 237{,}5] \\
]237{,}5\ ;\ 242{,}5] &]242{,}5\ ;\ 247{,}5] &]247{,}5\ ;\ 252{,}5] &]252{,}5\ ;\ 257{,}5] \\
]257{,}5\ ;\ 262{,}5] &]262{,}5\ ;\ 267{,}5] &]267{,}5\ ;\ 272{,}5] &]272{,}5\ ;\ 277{,}5] \\
]277{,}5\ ;\ 282{,}5] &]282{,}5\ ;\ 287{,}5] &]287{,}5\ ;\ 292{,}5] &
\end{array}
\tag{1}
$$

Vi tæller nu simpelthen hvor mange af de 1986 kvinders fodlængder der ligger

[6] Af æstetiske hensyn har jeg valgt intervallerne således, at deres midtpunkter er heltal delelige med 5. Videre har jeg inkluderet intervallet (287,5 ; 292,5] selv om der ikke er nogen af kvinderne, der har fodlængder over 287,5 mm. Dette er gjort for at lette diskussionen af symmetri på side 8.

i de forskellige intervaller. Dette viser sig at være

$$\begin{array}{cccc} 1 & 0 & 2 & 14 \\ 26 & 83 & 123 & 241 \\ 272 & 327 & 288 & 254 \\ 151 & 104 & 53 & 34 \\ 8 & 5 & 0 & \end{array} \tag{2}$$

Intervallet]242,5 ; 247,5], som er fremhævet med rødt i panel (1) ovenfor, indeholder kvindernes *median*[7] fodlængde på 246 mm. Vi ser, at *"median binnen"* indeholder flere kvinder end nogen af de andre bins, nemlig 327, som også er fremhævet med rødt ovenfor.

Opgave 3. *Hvad er summen af de 19 antal i panel* (2) *ovenfor? Svar meget gerne uden at lave selve udregningen!*

Alt efter hvorledes optællingen bruges fås forskellige varianter af histogrammet. I figur 3 har vi lavet histogrammer hvor højden af søjlerne (på y-aksen) angiver antal individer i de enkelte bins. For at den visuelle fortolkning af sådanne histogrammer bliver bedst mulig er det vigtigt, at bredden er den samme på alle bins. Her har vi altså valgt at bruge en binbredde på 5 mm = 0,5 cm. De to histogrammer synes at have samme *"form"*, nemlig den karakteriske normalfordelings *klokkeform*. Og som vi senere skal se, så kan fodlængderne indenfor de to køn hver for sig også med god approksimation antages at være normalfordelte.

Ved nærmere eftersyn ses, at de to histogrammer i figur 3 har forskellige y- og x-akser. Vi har nemlig valgt en præsentation, hvor histogrammerne udfylder deres respektive koordinatudsnit. Dette gør det lettere at se, at fordelingerne (visualiseret ved histogrammerne) har samme klokkeagtige form. Idet fordelingerne for de to køn har samme form, og der er målinger fra ca. dobbelt så mange mænd som kvinder, så vil vi umiddelbart forvente, at y-aksen for mændene går ca. dobbelt så langt op i forhold til y-aksen for kvinderne. Dette ser vi også på figur 3, hvor y-aksen går til lidt over 300 for kvinderne og lidt over 600 for mændene.

Som allerede angivet er median fodlængden for kvinderne 246 mm, og for mændene er den 271 mm. En normalfordeling er *symmetrisk* omkring sin *median*. Dette betyder, at hvis vi vælger to bins af samme bredde og som ligger lige langt fra medianen, men henholdsvis til venste og til højre for medianen, så vil de indeholde cirka lige mange datapunkter. Vi kan eftervise dette for

[7]Medianen er en talværdi, som adskiller de 50% laveste værdier fra de 50% højeste værdier. For kvindernes fodlængder er det således en talværdi, der adskiller de 1986/2 = 993 korteste fodlængder fra de 993 længste fodlængder.

kvindernes fodlængder med bins af bredde 5 mm. Tager vi antallene indefra og ud på hver side af antallet 327, som er markeret med rødt i panel (2), fås nemlig

side udfra	afstand i antal bins fra medianbinnen								
medianbinnen	1	2	3	4	5	6	7	8	9
til venstre	272	241	123	83	26	14	2	0	1
til højde	288	254	151	104	53	34	8	5	0

Den opmærksomme læser vil måske indvende, at i dette konkrete eksempel er antallet i binnen til venstre for medianen konsekvent lavere end antallet i binnen til højre for medianen (pånær for den sidste bin). Men dette passer faktisk perfekt med, at

1. Der er færre kvinder jo længere vi kommer væk fra medianen.

2. Medianen 246 ligger ikke midt i intervallet (242,5 ; 247,5], men derimod lidt til højre for intervallets midtpunkt på 245 mm.

Opgave 4. *Hvilke intervaller skal bruges som bins, hvis vi både ønsker en binbredde på 5 mm og at medianen på 246 mm er midtpunkt for et af intervallerne?*

Opgave 5. *Hvis vi bruger de bins, som er svaret på opgave 4, så viser der sig af være 310 kvinder i "medianbinnen". Videre bliver sammenligningen af antal på henholdsvis venstre og højre side som følger:*

side udfra	*afstand i antal bins fra medianbinnen*									
medianbinnen	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
til venstre	*275*	*255*	*147*	*97*	*38*	*13*	*3*	*0*	*0*	*1*
til højde	*307*	*226*	*137*	*95*	*43*	*28*	*9*	*2*	*0*	*0*

Passer dette bedre med at fordelingen af kvindernes fodlængder er symmetrisk omkring medianen? Diskutér også hvorfor vi ikke kan forvente at optælle nøjagtig det samme antal kvinder i tilsvarende bins på venstre og højre side af medianen.

Opgave 6. *Argumentér for, at hvis fordelingen af kvindernes fodlængder er symmetrisk omkring sin median, så er median og gennemsnit ens!*

Opgave 6 postulerer et matematisk udsagn, nemlig at median og gennemsnit er ens for symmetriske fordelinger. Vi har indset, at fordelingen af de 1986 kvinders fodlængder er næsten symmetrisk. Dermed vil vi forvente, at gennemsnitsfodlængden er cirka den samme som median fodlængden. Så

lad os prøve at beregne gennemsnittet for hvert af de to køn. Gennemsnit betegnes ofte med det græske bogstav μ (staves *mu*, men udtales *my*), og for fodlængderne fås

$$\begin{aligned} \mu_{\text{kvinde}} &= \frac{246 + \cdots + 239}{1986} = 246{,}29 \\ \mu_{\text{mand}} &= \frac{273 + \cdots + 295}{4082} = 271{,}18 \end{aligned} \tag{3}$$

Vi ser, at gennemsnitene stort set svarer til medianerne for henholdsvis kvinder (246 mm) og mænd (271 mm).

Lad os stoppe op et kort øjeblik og opsummere hvad vi har fundet ud af indtil nu. Vi har konstateret, at fordelingerne af to forskellige biologiske størrelser (nemlig fodlængderne af henholdsvis kvinder og mænd) har ensartede klokkeagtige former, som er næsten symmetriske omkring deres midtpunkter. Videre giver matematiske overvejelser, at medianen for en *symmetrisk fordeling* kan findes som fordelingens gennemsnit. For at angive *placeringen af en normalfordeling på x-aksen* er det således lige meget om man bruger medianen eller gennemsnittet (de er nemlig ens). Men traditionen er, at man angiver placeringen på x-aksen via gennemsnittet.

For at specificere en normalfordeling mangler der ét aspekt udover gennemsnittet, nemlig bredden af fordelingen. Bredden af en normalfordeling er proportional med den tilhørende *spredning*, som er givet ved kvadratroden af *variansen*. Spedningen betegnes ofte med det græske bogstav σ (staves og udtales *sigma*). Vi bemærker, at både gennemsnitsværdien μ og spredningen σ har samme fysiske enhed som selve målingerne. I datasættet med fodlængder er dette mm (millimeter). Spredningen på fodlængderne for kønnene hver for sig findes til at være[8]

$$\begin{aligned} \sigma_{\text{kvinde}} &= \sqrt{\frac{\big(246 - 246{,}29\big)^2 + \cdots + \big(239 - 246{,}29\big)^2}{1986}} = 12{,}43 \\ \sigma_{\text{mand}} &= \sqrt{\frac{\big(273 - 271{,}18\big)^2 + \cdots + \big(295 - 271{,}18\big)^2}{4082}} = 13{,}10 \end{aligned} \tag{4}$$

Vi ser, at der er lidt større spredning på målingerne for mændene. Mere præcist er spredningen

$$\frac{13,10 - 12,43}{12,43} \cdot 100\% = 5{,}4\%$$

[8]I formelsamlingen [7] er der opgivet to forskellige formler for beregning af spredning. Indtil videre ser vi på datapunkterne for sig selv, og ikke som stikprøver fra to populationer. Derfor bruger vi formlen for spredningen, hvor der *"divideres med antal datapunkter"*.

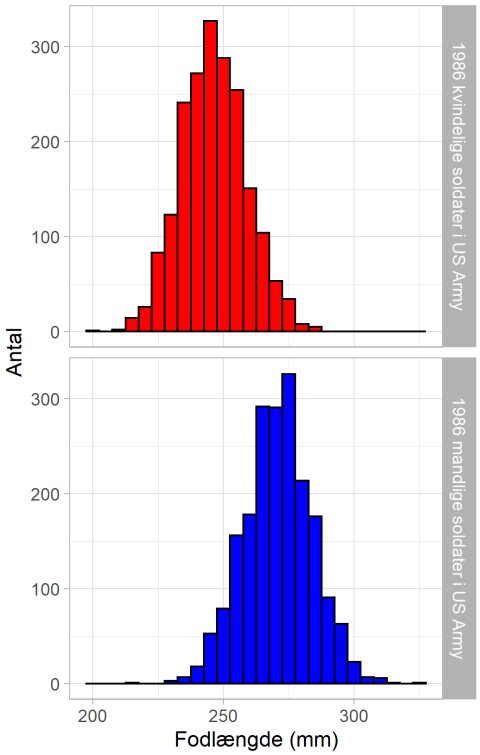
større for mændene i forhold til kvinderne.

Opgave 7. *På figur 3 går x-aksen for kvinderne fra ca.* 200 *til* 285*, mens x-aksen for mændene går fra ca.* 220 *til* 320*. Bredden af x-aksen er altså ca.* 85 *for kvinderne og* 100 *for mændene, dvs.*

$$\frac{100-85}{85}\cdot 100\% = 17{,}6\%$$

*større for mændene i forhold til kvinderne. Forholdet mellem bredderne i figuren (*17,6%*) er altså større end forholdet mellem spredningerne (*5,4%*). Argumentér for, at dette kan forklares med, at der er målinger fra flere mænd end kvinder.*

Som det blev diskuteret i opgave 7, så kan man forvente at se mere ekstreme observationer (usædvanlig korte eller lange fødder) jo flere målinger der er lavet. Så det vil være lettest umiddelbart at sammenligne fodlængderne mellem de to køn, hvis der var lavet målinger på lige mange kvinder og mænd. Det er ikke nemt for os at afstedkomme målinger af flere kvindelige soldater i US Army. Men heldigvis er det tilsvarende let blot at bruge færre af målingerne på de mandlige soldater. Vi udvælger simpelthen 1986 tilfældige målinger blandt de 4082 mænd (så der er lige mange målinger på mænd og kvinder) til brug for den visuelle sammenligning af kønnenes fodlængder. Videre bruges de samme y- og x-akser for histogrammerne for de to køn, og endelig placeres histogrammerne ovenpå hinanden for at fodlængderne (som er på x-aksen) umiddelbart kan sammenlignes. Dette er gjort i figur 4, hvor vi tydeligt kan se, at mændene som gruppe har længere fødder end kvinderne.



Figur 4: Histogrammer tilsvarende figur 3, men med sammen antal målinger og ens y- og x-akser for kvinder og mænd.

Opgave 8. *Sammenlign de to forskellige visualiseringer i figur 3 og figur 4. Hvilken visualisering synes du giver den bedste sammenligning af de to datasæt?*

Opgave 9. *I både figur 3 og figur 4 angiver y-aksen antallet af datapunkter i de forskellige bins. Men ved at bruge y-aksen anderledes fås andre varianter af histogrammet, f.eks. kan man skalere optællingerne således at*

$$\textit{arealet af søjlen} = \textit{andelen af datapunkter i binnen}$$

I et sådant histogram skal højden på søjlen ved "median binnen" i optællingen fra panel (2) *således være*

$$y = \frac{1}{5} \cdot \frac{327}{1986} = 0{,}0329$$

Bemærk, at andelen $\frac{327}{1986}$ *skal divideres med bredden på søjlen, som er 5. Og hvis man f.eks. havde brugt centimeter som enhed på x-aksen, så skulle man dividere med 0,5 (=gange med 2). Lav en tegning med denne variant af histogrammet for kvindernes fodlængde! Overvej også, om denne variant af histogrammet er bedre til at sammenligne datasæt af forskellig størrelse, f.eks. de 1986 kvinder mod de 4082 mænd.*

Opgave 10. *Indtil videre har alle vores histogrammer haft bins med bredde 5, men man kan naturligvis vælge andre binbredder. Lav histogrammer af kvindernes fodlængde med binbredde 1, 2, 5, 10 og 15 (mm), og overvej hvilket af disse 5 histogrammer du synes giver den bedste visualisering af datasættet. Hvad er fordele og ulæmper ved visualiseringerne med de forskellige binbredder? For at lave histogrammerne anbefales det, at man bruger et passende computer program. Bemærk videre, at datasættene med fodlængderne på de 1986 kvinder og de 4082 mænd er vedhæftet dette dokument.*

Indtil videre har vi undersøgt fordelinger ved at beregne tal såsom gennemsnit og spredning, og ved at lave histogrammer. Vi fandt således ud af, at mænd i gennemsnit har længere fødder end kvinder, mens at spredningen af fodlængderne indenfor kønnene er næsten ens (omend spredningen er lidt større for mænd). Ved at se på histogrammerne konkluderede vi desuden, at fordelingen af fodlængden har samme klokkeagtige form for de to køn. Men strengt taget kan det være svært at se, om to histogrammer har nøjagtig samme klokkeform. Et bedre visualiseringsværktøj til at afgøre om to fordelinger har samme form er et såkaldt *fraktildiagram*. I et fraktildiagram optegnes *fraktilerne* for to fordelinger mod hinanden. For at kunne gøre dette skal vi først vide, hvordan fraktiler beskrives og beregnes.

For en procentdel α mellem 0% og 100% er en α-fraktil løseligt beskrevet et tal, der adskiller de α laveste værdier fra de $100\% - \alpha$ højeste værdier. Medianen, som adskiller de 50% laveste værdier fra de 50% højeste værdier, er således 50%-fraktilen. Den præcise matematiske definition er lidt kompliceret:

Et tal x er en α-fraktil netop hvis

$$\boxed{\begin{array}{l}\text{andelen af datapunkter}\\ \text{strengt mindre end } x\end{array}} \leq \alpha \leq \boxed{\begin{array}{l}\text{andelen af datapunkter}\\ \text{mindre end eller lig med } x\end{array}} \tag{5}$$

For et datasæt bestående af N reelle tal er de *i voksende rækkefølge sorterede værdier* således $\frac{n-\frac{1}{2}}{N} \cdot 100\%$-fraktiler, hvor n er et heltal mellem 1 og N.

Den matematiske definition af en α-fraktil er formodentlig temmelig overvældende at læse og forstå for de fleste. Så lad os afprøve definitionen på følgende datasæt bestående af $N = 5$ datapunkter:

$$x_1 = 17 \qquad x_2 = 3 \qquad x_3 = 83 \qquad x_4 = 88 \qquad x_5 = 77$$

Vi skal altså først sortere datapunkterne i voksende rækkefølge. Ovenfor blev det i'te datapunkt kaldt for x_i, og det viser sig at være praktisk at indføre notationen $x_{(i)}$ for det i'te mindste datapunkt. Med denne notation har vi

$$x_{(1)} = 3 \qquad x_{(2)} = 17 \qquad x_{(3)} = 77 \qquad x_{(4)} = 83 \qquad x_{(5)} = 88$$

For de $N = 5$ datapunkter er $x_{(2)}$ således en $\frac{2-\frac{1}{2}}{5} \cdot 100\%$-fraktil, altså 17 er en 30%-fraktil for datasættet ovenfor. Tilsvarende er medianen lig med 77.

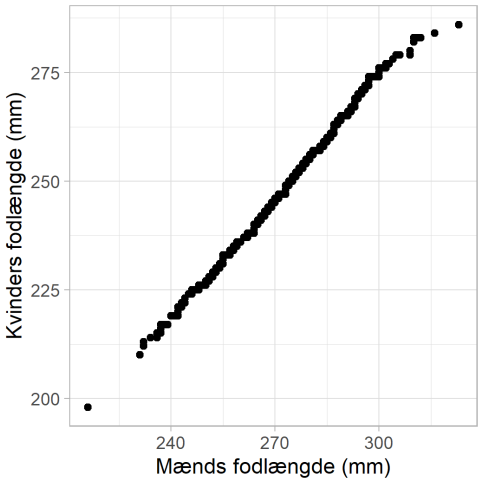
Opgave 11. *Bestem en 25%-fraktil, en 45%-fraktil, en 55%-fraktil og en 75%-fraktil for følgende datasæt bestående af $N = 10$ datapunkter*

57, 44, 52, 63, 1, 25, 51, 12, 81, 83

Antag nu, at vi har to datasæt af samme størrelse, og vi gerne vil vide om de tilhørende fordelinger har samme "form". Dette kunne f.eks. være de $N = 1986$ målinger af kvindelige soldaters fodlængde, og de $N = 1986$ tilfældigt udvalgte målinger af mandlige soldaters fodlængde som blev brugt i figur 4. For at undersøge om fordelingerne er ens laver vi et fraktildiagram ved at gøre følgende

1. Sorter fodlængderne i det første datasæt i voksende rækkefølge. Vi vil kalde disse værdier for $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$.

2. Sorter fodlængderne i det andet datasæt i voksende rækkefølge. Vi vil kalde disse værdier for $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(N)}$.



Figur 5: Fraktildiagram der sammenligner de to datasæt i figur 4.

3. Lav et punktplot af punkterne $(x_{(1)}, y_{(1)})$, $(x_{(2)}, y_{(2)})$, $\ldots$, $(x_{(N)}, y_{(N)})$.

4. Fordelingerne da samme "form" hvis punkterne på punktplottet ligger på en ret linje.

5. Hvis punkterne på punktplottet ikke ligger på en ret linje, så har fordelingerne ikke samme "form".

Bemærk, at udsagnet i punkt 5. ovenfor er ækvivalent med

> Hvis fraktildiagrammet viser systematiske afvigelser fra en ret linje, så vil de to fordelinger ikke have samme form.

Figur 5 viser fraktildiagrammet for fodlængderne. Vi ser, at næsten alle punkterne (på nær nogle enkelte mænd med usædvanlig store fødder) ligger på en næsten perfekt ret linje. Dette betyder, at de i voksende rækkefølge ordnede målinger fra det *andet datasæt* stort set kan fås ved at indsætte de i voksende rækkefølge ordnede målinger fra det *første datasæt* i udtrykket for en *ret linje*. Hvilket er nøjagtigt det samme, som at fordelingerne for de to datasæt har samme form.

Opgave 12. *I denne opgave betragter vi tre datasæt, som alle har 15 datapunkter. Det første datasæt er*

647, 141, 130, 124, 98, 249, 322, 629, 479, 86, 84, 693, 30, 94, 236

Det andet datasæt er

43, 155, 380, 138, 218, 126, 200, 280, 333, 251, 94, 204, 321, 269, 157

Det tredje datasæt er

245, 131, 98, 125, 109, 350, 79, 319, 169, 101, 139, 129, 482, 188, 314

Fordelingerne af to af de tre datasæt kan antages at have samme form, men hvilke af de tre datasæt er det? Vink: Undersøg dette spørgsmål ved hjælp af fraktildiagrammer.

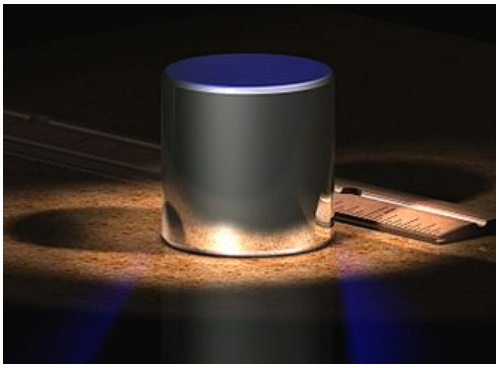
2.1 Opsamling på opvarmningen

Indtil videre har vi set på to konkrete datasæt, beregnet deres gennemsnit og spredning, visualiseret datafordelingerne ved histogrammer og sammenlignet fordelinger via fraktildiagrammer.

Men egentlig var det jo meningen, at vi skulle tale om normalfordelingen. Og bare rolig, der skal nok komme matematiske formler og beregninger på papiret. Men før vi kommer til formlerne vil vi bruge flere kræfter på at opdage normalfordelinger i verden omkring os. Vi har allerede løftet lidt af sløret ved at fortælle, at normalfordelingen har en helt særlig klokkeagtig form og at den derudover er beskrevet ved gennemsnit og spredning. I stedet for at opskrive det matematiske udtryk for denne særlige form, så vil vi i første omgang bruge fodlængderne for de 1986 kvinder i ANSUR II undersøgelsen som reference for formen af normalfordelingen. Denne normalfordelingsreference bruges således:

> Antag vi har et datasæt til vores rådighed og ønsker at undersøge, om dette datasæt er (næsten) normalfordelt. Dette spørgsmål afgøres ved at lave et fraktildiagram mod de 1986 datapunkter fra normalfordelingsreferencen og se efter, om punkterne ligger på en (næsten) ret linje — altså på nær lidt statistisk variation ligesom vi så i figur 5.

Nogle læsere vil måske være oprørte over en sådan "statistisk" tilgang; at afgøre om et datasæt er normalfordelt ved at sammenligne det med fodlængderne på 1986 tilsyneladende arbitrært udvalgte kvindelige amerikanske soldater. Men en sådan tilgang har faktisk været anvendt i fysikkens verden for noget så essentielt som længde og vægt. Indtil 1960 var afstanden *1 meter* defineret som længden på en plantin-iridium stang, der blev opbevaret hos *Det Franske Videnskabsakademi*. Så hvis man skulle være helt sikker på om en given genstand var kortere eller længere end en meter, så måtte man tage til Paris og lave sammenligningen med pågældende metalstang. Indtil den



Figur 6: Et computer grafikbillede [8] af kilogram prototypen, som blev brugt som reference indtil d. 20. maj 2019.

20. maj 2019 var *1 kilogram* tilsvarende defineret som vægten på en platin-iridium cylinder (se figur 6), der også blev opbevaret i Paris. Enhederne for både afstand og vægt er nu blevet erstattet af definitioner baseret på fysiske naturkonstanter. Tilsvarende lover jeg at give en matematisk definition af normalfordelingen senere i dette manuskript.

3 Endnu mere om fraktildiagrammer

For at få metoden med at sammenligne med en referencefordeling til at fungere i praksis, er det nødvendigt generelt

1. At kunne sammenligne to datasæt, som ikke har det samme antal observationer.

2. At kunne sammenligne et datasæt med en sandsynlighedsfordeling.

Hvordan dette kan gøres vil blive beskrevet i de to følgende underafsnit.

3.1 Sammenligning af to datasæt af forskellig størrelse

Lad der være givet et datasæt $\{y_1, \ldots, y_M\}$ bestående af M tal, som skal bruges som normalfordelingsreference. De i rækkefølge ordnede datapunkter benævnes

$$y_{(1)} \leq \cdots \leq y_{(M)}$$

Hvis dette f.eks. er de $M = 1986$ fodlængder for kvinderne i *ANSUR II* undersøgelsen, så gælder helt konkret

$$y_{(1)} = 198, \qquad y_{(2)} = 210, \qquad y_{(3)} = 212, \qquad \ldots, \qquad y_{(1986)} = 286$$

Antag nu, at vi har et nyt datasæt $\{x_1, \ldots, x_N\}$ bestående af N tal, og at vi vil undersøge om dette datasæt er (næsten) normalfordelt. Dette gøres ved at lave et fraktildiagram af de to datasæt mod hinanden. Hvis fraktildiagrammet viser en (næsten) ret linje, så konkluderes at fordelingerne af de to datasæt har (næsten) samme form. Idet vi har besluttet os for at $\{y_1, \ldots, y_M\}$ er normalfordelt, så vil dette betyde, at $\{x_1, \ldots, x_N\}$ er (næsten) normalfordelt. For at lave fraktildiagrammet starter vi med at sortere datapunkterne i det nye datasæt. De i rækkefølge ordnede datapunkter benævnes

$$x_{(1)} \leq \cdots \leq x_{(N)}$$

Hvis de to datasæt har lige mange datapunkter, altså hvis $N = M$, så laves fraktildiagrammet som et punktplot af datapunkterne

$$(x_{(n)}, y_{(n)}) \text{ for } n = 1, \ldots, N.$$

Grunden til at dette giver et fraktildiagram er, at det n'te par $(x_{(n)}, y_{(n)})$ her svarer til $\frac{n-\frac{1}{2}}{N} \cdot 100\%$ fraktilerne i de to datasæt.

Hvis de to datasæt ikke har lige mange datapunkter, altså hvis $N \neq M$, så er parringen af fraktilerne ikke helt så umiddelbar. Lad os først se på tilfældet med $N < M$. Så har vi flere fraktiler $\frac{m-\frac{1}{2}}{M} \cdot 100\%$ for referencedatasættet $\{y_1, \ldots, y_M\}$ end der skal matches med fraktilerne $\frac{n-\frac{1}{2}}{N} \cdot 100\%$ for det nye datasæt $\{x_1, \ldots, x_N\}$. For et givet $x_{(n)}$ vælger vi at bruge det $y_{(m)}$, hvor de to fraktiler passer bedst muligt sammen. Vi finder dermed m ved at løse ligningen

$$\frac{m-\frac{1}{2}}{M} = \frac{n-\frac{1}{2}}{N},$$

altså

$$m = \frac{1}{2} + M \cdot \frac{n-\frac{1}{2}}{N}.$$

Der er dog et problem med denne løsning, nemlig at m skal være et heltal for at $y_{(m)}$ findes som et datapunkt. Dette løses ved afrunding, hvor vi bruger det mindste heltal der er størrer end eller lig med $M \cdot \frac{n-\frac{1}{2}}{N}$. Den matematiske notation for det mindste heltal større end eller lig med et decimal tal z er $\lceil z \rceil$. På engelsk kaldes dette for *ceiling*, og der gælder f.eks.

$$\lceil 20{,}81 \rceil = 21, \qquad \lceil 101 \rceil = 101, \qquad \lceil \tfrac{22}{7} \rceil = 4, \qquad \lceil 1985{,}1111 \rceil = 1986.$$

Konklusionen på ovenstående matematiske analyse er, at hvis $N < M$, så kan fraktildiagrammet laves som et punktplot af punkterne

$$\left(x_{(n)}, y_{\left(\left\lceil M \cdot \frac{n-\frac{1}{2}}{N} \right\rceil\right)}\right) \text{ for } n = 1, \ldots, N.$$

Hvis $N > M$, så er situationen helt tilsvarende pånær, at der nu er flere datapunkter i det nye datasæt. Dermed laves fraktildiagrammet som et punktplot af punkterne

$$\left(x_{\left(\left\lceil N\cdot\frac{m-\frac{1}{2}}{M}\right\rceil\right)}, y_{(m)}\right) \text{ for } m = 1, \ldots, M.$$

Opgave 13. *Betragt tilfældet hvor $N < M$, og lad n være et heltal mellem 1 og N. Brug definitionen af fraktiler i* (5) *og argumentér for, at*

(1) $x_{(n)}$ er en $\frac{n-\frac{1}{2}}{N} \cdot 100\%$-fraktil for datasættet $\{x_1, \ldots, x_N\}$,

(2) $y_{\left(\left\lceil M\cdot\frac{n-\frac{1}{2}}{N}\right\rceil\right)}$ er en $\frac{n-\frac{1}{2}}{N} \cdot 100\%$-fraktil for datasættet $\{y_1, \ldots, y_M\}$.

Forklar hvorfor dette betyder, at punktplottet af punkterne $\left(x_{(n)}, y_{\left(\left\lceil M\cdot\frac{n-\frac{1}{2}}{N}\right\rceil\right)}\right)$ for $n = 1, \ldots, N$ giver et fraktildiagram, altså en visuel sammenligning af fraktilerne for de to datasæt.

Opgave 14. *Lav fraktildiagrammer mod normalfordelingsreferencen for hver af de tre datasæt i opgave 12. Hvilket af de tre datasæt kan antages at være normalfordelt?*

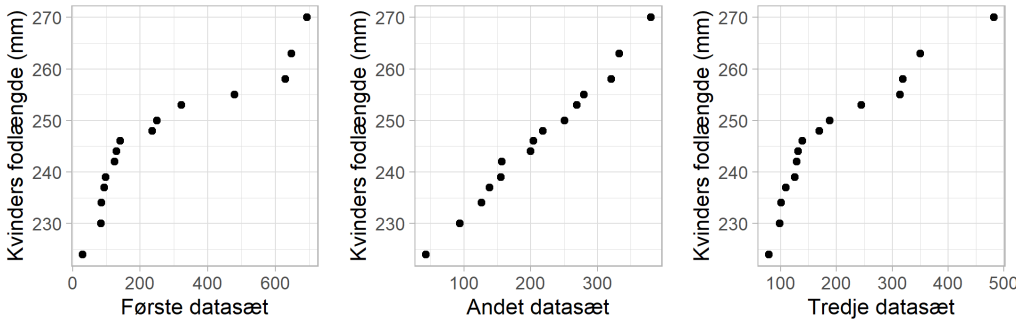
3.2 Sammenligning af et datasæt med en sandsynlighedsfordeling

Metoden givet i afsnit 3.1 kan også udvides til at blive brugt i den situation, hvor $x_{(1)} < x_{(2)} < \cdots < x_{(N)}$ beskriver en voksende række af mulige udfald med tilhørende sandsynligheder $p_1, \ldots, p_N$. Der skal gælde, at $p_n > 0$ og

$$\sum_{n=1}^{N} p_n = p_1 + \cdots + p_N = 1 \tag{6}$$

Som et eksempel lad os se på antal øjne efter et slag med en almindelig terning. Her er der $N = 6$ mulige udfald, nemlig heltallene fra 1 til 6. Vi går ud fra, at terningen er ærlig. Dette betyder, at alle udfald er lige sandsynlige. Da synlighederne skal summere til 1 må disse sandsynligheder altså være $\frac{1}{6}$. Sammenhørende udfald er således givet ved tabellen

n	1	2	3	4	5	6
$x_{(n)}$	1	2	3	4	5	6
p_n	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$



Figur 7: Opgave 14 besvares med tre fraktildiagrammer. Første og tredje datasæt er altså ikke normalfordelte, mens det andet datasæt synes at være normalfordelt.

Opgave 15. *Argumentér for, at fordelingen af summen af øjnene efter et slag med to almindelige terninger har* $N = 11$ *mulige udfald, hvor sammenhørende udfald og sandsynligheder givet ved tabellen*

n	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
$x_{(n)}$	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
p_n	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Opgave 16. *Argumentér for, at en binomialfordeling med antalsparameter* $= 20$ *og sandsynlighedsparameter* $= \frac{1}{2}$ *har* $N = 21$ *mulige udfald, hvor sammenhørende udfald og sandsynligheder er givet ved*

$$x_{(n)} = n - 1, \qquad p_n = \frac{20!}{n!(20-n)!} \cdot \left(\tfrac{1}{2}\right)^{20}$$

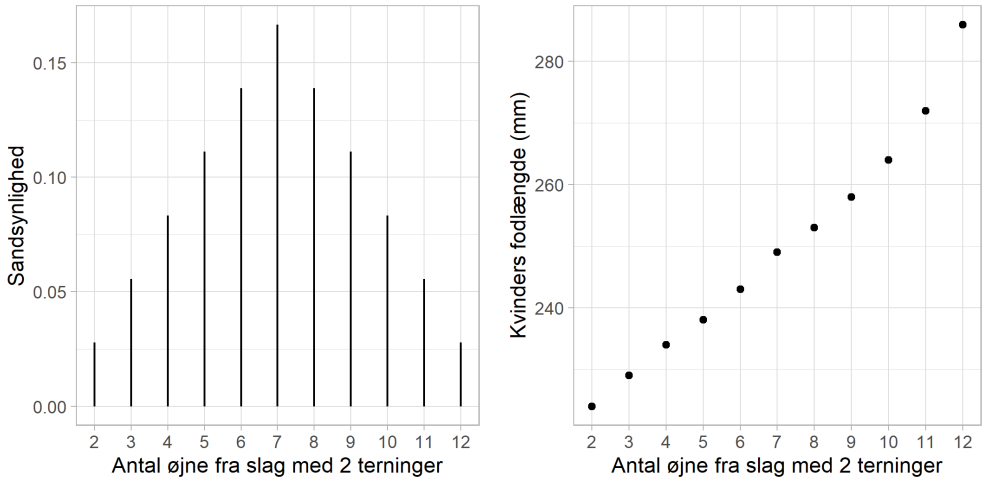
for $n = 1, \ldots, 21$.

I praksis vil vi sammenligne sandsynlighedsfordelinger mod fordelingen af fodlængderne for de $M = 1986$ kvinder i *ANSUR II* undersøgelsen. Og da 1986 er et forholdsvis stort antal, så vil vi antage, at $N \leq M$. I givet fald laves fraktildiagrammet som et punktplot af punkterne

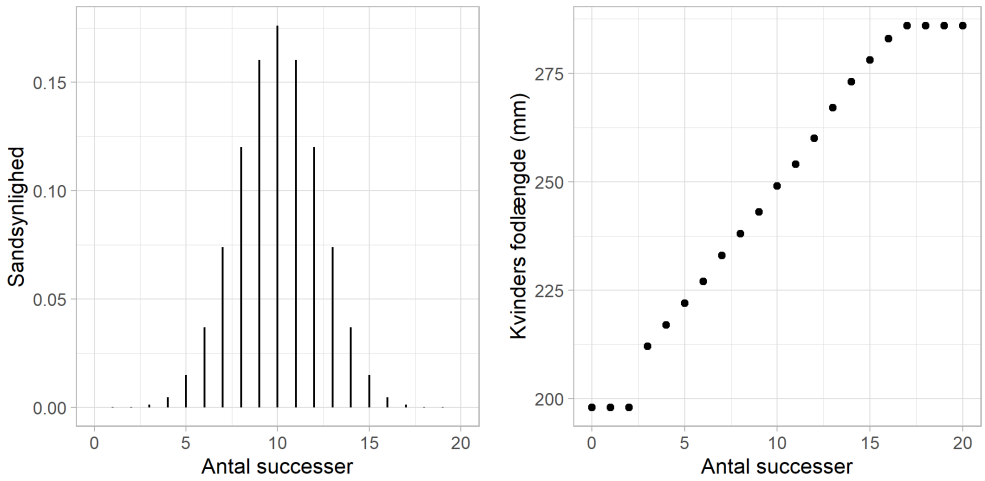
$$\left(x_{(n)}, y_{\left(\left\lceil M \cdot \sum_{i=1}^{n} p_i \right\rceil\right)}\right) \text{ for } n = 1, \ldots, N,$$

hvor $\lceil M \cdot \sum_{i=1}^{n} p_i \rceil$ er det mindste heltal, som er større eller lig med $M \cdot \sum_{i=1}^{n} p_i$.

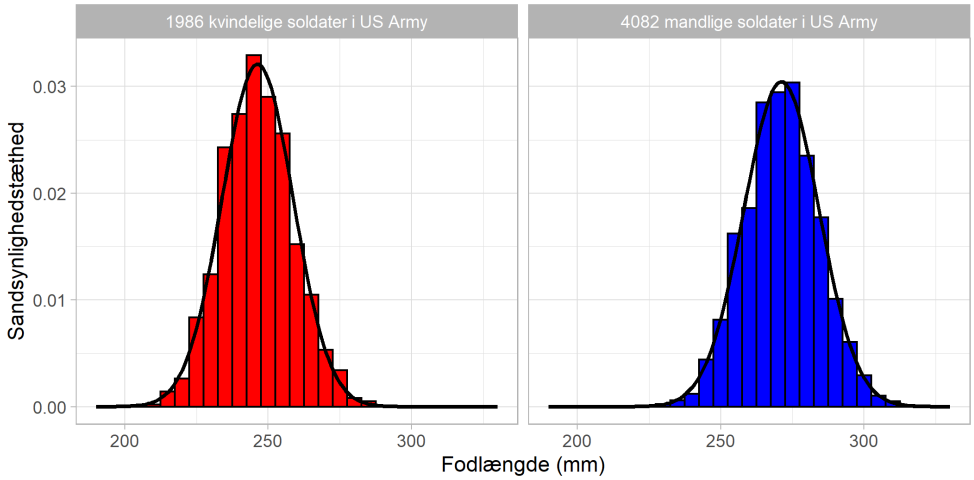
Figur 8 og 9 viser pinde- og fraktildiagrammer for henholdsvis *"summen af antal øjne ved slag med to terninger"* og for *"binomialfordelingen med antalsparameter = 20 og sandsynlighedsparameter = 0,5"* mod normalfordelingsreferencen. Idet fraktildiagrammer viser (næsten) rette linjer konkluderes, at begge disse fordelinger er (næsten) normalfordelinger.



Figur 8: Punktsandsynligheder og normalfordelingsreference-fraktildiagram for summen af antal øjne ved slag med 2 almindelige terninger.



Figur 9: Punktsandsynligheder og normalfordelingsreference-fraktildiagram for binomialfordelingen med antalsparameter = 20 og sandsynlighedsparameter = 0,5.



Figur 10: Sandsynlighedstæthedshistogrammer for fodlængderne på kvindelige og mandlige soldater i *ANSUR II* undersøgelsen.

Opgave 17. *Fraktildiagrammet i figur 9 viser en tydelig afvigelse fra normalfordelingsreferencen når der er få eller mange successer. Diskutér om man kan forvente at de "kun" 1986 målinger af de kvindelige soldater kan matche de mest ekstreme sandsynligheder i binomialfordelingen? F.eks. er sandsynligheden for at få 20 successer givet ved* $(\frac{1}{2})^{20} \approx \frac{1}{1.000.000}$.

4 Normalfordelingen

Opgave 9 introducerede en variant af histogrammet, hvor y-aksen angiver *sandsynlighedstætheder* i stedet for *antal*. En besvarelse af opgave 9 findes i figur 10, der viser sådanne sandsynlighedstæthedshistogrammer for fodlængderne på henholdsvis kvinder og mænd. Det ses, at de to histogrammer har sammenlignelige y-akser (og derfor umiddelbart kan sammenlignes) selv om der er mere end dobbelt så mange mænd som kvinder i *ANSUR II* undersøgelsen.

Sandsynlighedstæthedshistogrammer ligner umiddelbart antalshistogrammerne, men y-aksen er skaleret anderledes. I sandsynlighedstæthedshistogrammerne er y-aksen skaleret således at det samlede areal af alle søjlerne er nøjagtig lig med 1, som er det samme som 100%. Lad os f.eks. se på den næsthøjeste søjle i histogrammet for de kvindelige soldater. For denne søjle går x-aksen fra 247,5 til 252,5 (i enheden mm). Søjlen har altså en bredde på 5 (ligesom alle de øvrige søjler), og ender lidt under 0,03 på y-aksen. Fortolkningen af dette er, at lidt under $5 \cdot 0{,}03 = 0{,}15 = 15\%$ af kvinderne har

fodlængder mellem 247,5 og 252,5 (mm).

I figur 10 er der også indtegnet en kurve sammen med hver af histogrammerne. For kvinderne er det grafen for funktionen

$$f_{\text{kvinde}}(x) = \frac{1}{\sqrt{2\pi}\cdot 12{,}43}\cdot \mathrm{e}^{-\frac{1}{2}\cdot\left(\frac{x-246{,}29}{12{,}43}\right)^2},$$

og for mændene er det grafen for funktionen

$$f_{\text{mand}}(x) = \frac{1}{\sqrt{2\pi}\cdot 13{,}10}\cdot \mathrm{e}^{-\frac{1}{2}\cdot\left(\frac{x-271{,}18}{13{,}10}\right)^2}.$$

Begge disse funktioner er eksempler på *tæthedsfunktioner*, som for en normalfordeling med middelværdi μ og spredning σ er givet ved

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\cdot \sigma}\cdot \mathrm{e}^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{7}$$

I funktionsudtrykkene ovenfor har vi således indsat gennemsnit og spredning fundet i (3) og (4), altså

$$\mu_{\text{kvinde}} = 246{,}29 \qquad \sigma_{\text{kvinde}} = 12{,}43 \qquad \mu_{\text{mand}} = 271{,}18 \qquad \sigma_{\text{mand}} = 13{,}10$$

Vi ser, at disse normalfordelingstæthedsfunktioner følger histogrammerne ganske tæt. Dette er ikke en tilfældighed. Som tidligere postuleret, så er fodlængderne rent faktisk normalfordelte med meget god approksimation. Tæthedsfunktionen i ligning (7) er den i slutningen af afsnit 2.1 lovede matematiske definition af normalfordelingen.

Før vi ser nærmere på nogle matematiske egenskaber ved funktionen i (7) vil vi først prøve at bruge den til noget. Hvis vi f.eks. vil vide hvor mange kvinder der har fodlængde mellem 247,5 og 252,5 (mm), så gangede vi tidligere bredden (5 mm) og højden (lidt under 3% per mm) af den tilsvarende søjle i histogrammet. Idet normalfordelingstætheden $f_{\text{kvinde}}(x)$ følger histogrammet tæt, så kunne vi også beregne denne andel ved at integrere $f_{\text{kvinde}}(x)$ over intervallet fra 247,5 til 252,5 (mm), og dermed finde arealet under grafen. Altså

$$\begin{aligned}\int_{247{,}5}^{252{,}5} f_{\text{kvinde}}(x)\,dx &= \int_{247{,}5}^{252{,}5} \frac{1}{\sqrt{2\pi}\cdot 12{,}43}\cdot \mathrm{e}^{-\frac{1}{2}\cdot\left(\frac{x-246{,}29}{12{,}43}\right)^2}\,dx \\ &= 0{,}1525 = 15{,}25\%\end{aligned}$$

Dette er lidt mere end 15%, mens histogrammet viste lidt mindre end 15%. Men hvis vi ser nøje efter på figur 10 så passer dette med, at grafen for $f_{\text{kvinde}}(x)$ ligger lidt over histogrammet på den pågældende del af x-aksen.

235, 263, 258, 246, 247, 241, 243, 255, 243, 279,
250, 228, 264, 238, 266, 228, 253, 247, 248, 267,
259, 245, 236, 271, 260, 261, 244, 241, 232, 237,
231, 242, 238, 244, 256, 255, 271, 236, 235, 257,
244, 240, 254, 252, 247, 252, 234, 237, 260, 252,
244, 242, 236, 237, 228, 272, 247, 237, 230, 259,
257, 250, 237, 253, 265, 258, 267, 236, 248, 249,
259, 252, 231, 236, 235, 262, 232, 243, 243, 243,
247, 260, 259, 223, 249, 269, 259, 274, 238, 274,
237, 251, 253, 246, 248, 245, 243, 249, 241, 248

Tabel 1: 100 tilfældigt udvalgte observationer blandt de 1986 kvindelige fodlængder. De 14 observationer mellem 247,5 og 252,5 (mm) er markeret med rødt.

Integralet ovenfor kan forøvrigt ikke udregnes "i hånden på et stykke papir" ved brug af klassisk integralregning. Udfordringen er, at stamfunktionen hørende til $f_{\text{kvinde}}(x)$ ikke kan skrives på kort form via kendte matematiske funktioner (såsom potens, eksponential og trigonometriske funktioner). Integralet skal derfor beregnes numerisk ved brug af en computer. Man kan så med rette spørge hvorfor man skal erstatte en simpel aflæsning fra et histogram med en besværlig computer beregning af et integral. Det er der heller ikke nogen grund til. Men når måden data er indsamlet på, eller de spørgsmål man stiller til data, bliver mere komplicerede, så viser normalfordelingen sig ofte at være et umådeligt nyttigt og kraftfuldt værktøj. I næste afsnit illustreres dette ved et eksempel.

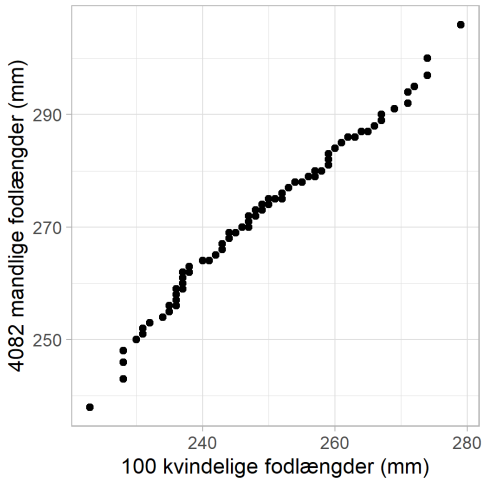
4.1 Et eksempel på brug af normalfordelingen

I dette eksempel ønsker vi at finde ud af, hvor mange af de 1986 kvinder fra *ANSUR II* datasættet der har fodlængder mellem 247,5 og 252,5 (mm). Vi har adgang til hele datasættet og en simpel optælling giver, at dette drejer sig om 288 kvinder. Så langt så godt.

Men hvad nu hvis vi ikke har adgang til hele datasættet? Det kunne f.eks. være, at vi kun har fået opgivet fodlængderne på 100 af de 1986 kvinder. Tabel 1 viser fodlængderne på 100 tilfældigt udvalgte kvinder. Blandt disse 100 kvinder har 14 en fodlængde mellem 247,5 og 252,5 (mm). Ud fra dette vil man gætte på at

$$1986 \cdot \frac{14}{100} = 278{,}04 \approx 278$$

kvinder blandt de 1986 kvinder har fødder af pågældende størrelse. Det er jo ikke så langt fra det *rigtige antal* på 288 kvinder, men spørgsmålet er, om vi



Figur 11: Fraktildiagram over de 100 tal fra tabel 1.

kan komme med et bedre gæt.

Lad os prøve at bruge normalfordelingen i stedet for. Først skridt er at undersøge om de 100 tal fra tabel 1 overhovedet kan antages at være normalfordelte. Vores værktøj til dette er at lave et fraktildiagram. Indtil nu har vi jo brugt de 1986 kvinders fodlængde som *normalfordelingsreference*, men lige nøjagtig i dette eksempel vil dette være at snyde. Vi har jo udtaget de 100 tal fra denne population, så det vil ikke være overraskende hvis de 100 udtagne fodlængder har samme fordeling som den population de kommer fra. For at illustrere fremgangsmåden vil vi i stedet for bruge de 4082 målinger af mænds fodlængder som normalfordelingsreference, og så lave fraktildiagrammet som et punktplot via metoden fra afsnit 3.1. Dette er gjort i figur 11, og idet fraktildiagrammet på nær lidt variation viser en ret linje, så kan vi altså godt bruge normalfordelingen. Derefter beregner vi gennemsnit og spredning[9]

$$\mu_{100\text{ tal}} = \frac{235 + \cdots + 248}{100} = 248{,}23$$

$$\sigma_{100\text{ tal}} = \sqrt{\frac{(235 - 248{,}23)^2 + \cdots + (248 - 248{,}23)^2}{100 - 1}} = 12{,}12$$

[9]Idet vi betragter de 100 målinger som en stikprøve fra populationen af kvinders fodlængde bruges formlen for et estimat for spredningen, altså hvor der divideres med $100 - 1 = 99$ i stedet for med 100. Dette er en spidsfindig teknisk detalje, som vi ikke vil diskutere yderligere!

Normalfordelingsintegralet giver dermed en andel på

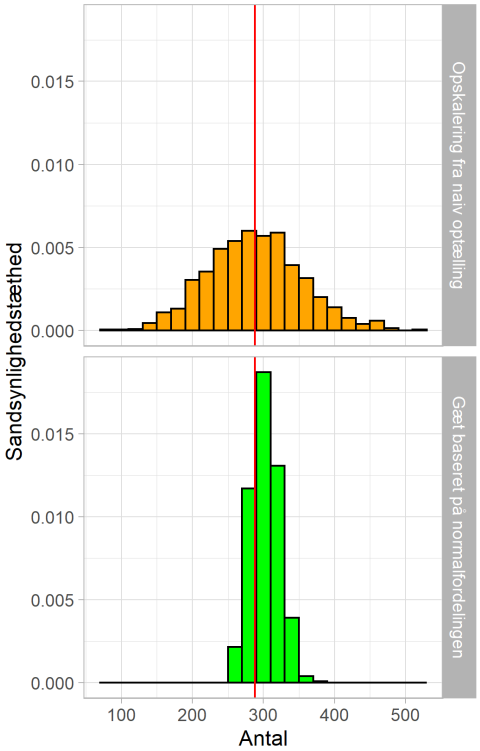
$$\int_{247,5}^{252,5} f_{100\text{ tal}}(x)\,dx = \int_{247,5}^{252,5} \frac{1}{\sqrt{2\pi}\cdot 12{,}12}\cdot \mathrm{e}^{-\frac{1}{2}\cdot\left(\frac{x-248{,}23}{12{,}12}\right)^2}\,dx$$
$$= 0{,}1617 = 16{,}17\%$$

Ud fra dette vil man gætte på, at

$$1986\cdot 0{,}1617 = 321{,}14 \approx 321$$

kvinder blandt de 1986 kvinder har fodlængde mellem 247,5 og 252,5 (mm).

Lad os lave en kort opsummering: Vi har afprøvet to forskellige metoder til at gætte på hvor mange af populationen på 1986 kvinder, der har fodlængder mellem 247,5 og 252,5. Det *rigtige antal* er 288 kvinder. Den første metode, som bestod i en simpel opskalering fra andelen af de 100 tilfældigt udvalgte kvinder til hele populationen på 1986 kvinder, gav et gæt på 278 kvinder. Den anden metode, som først fandt gennemsnit og spredning og derefter brugte normalfordelingen, gav et svar på 321. Selv om det andet gæt er klart længere fra det *rigtige antal* på 288 kvinder end det første gæt, så er den anden metode faktisk bedre! Sagen er den, at den første metode er mere følsom overfor hvilke 100 kvinder, der blev udvalgt. Hvis den sidste måling i tabel 1 f.eks. havde været 247 (mm) i stedet for 248 (mm), så havde første metode gættet på



Figur 12: Opførelse af de to metoder. Den røde linje viser det *rigtige antal* på 288 kvinder.

$$1986\cdot\frac{13}{100} = 258{,}18 \approx 258$$

kvinder, mens gættet fra anden metode kun ville ændre sig ganske lidt (og vil stadigvæk være på 321 kvinder efter afrunding).

For at undersøge hvilken af de to metoder der fungerer bedst har vi lavet et såkaldt *simulationseksperiment*. Eksperimentet består i, at vi 1000 gange

har udtaget et datasæt bestående af 100 tilfældigt udvalgte kvinder blandt de 1986 kvinder i *ANSUR II* undersøgelsen. Og for hver af de 1000 gentagelser har vi brugt de to metoder. Histogrammerne i figur 12 viser en opsummering af de 1000 gæt fra de to metoder. Vi ser, at gættet der baserer sig på normalfordelingen er bedre i den forstand, at det *varierer mindre omkring det rigtige antal på 288 kvinder.*

Opgave 18. *Hvis vi laver normalfordelingsberegningen baseret på middelværdi og spredning fra samtlige 1986 kvinder, og ikke bare 100 tilfældigt udvalgte kvinder, så fås*

$$\begin{aligned} 1986 \cdot \int_{247{,}5}^{252{,}5} f_{kvinde}(x)\,dx &= 1986 \cdot \int_{247{,}5}^{252{,}5} \frac{1}{\sqrt{2\pi}\cdot 12{,}43} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-246{,}29}{12{,}43}\right)^2} dx \\ &= 1986 \cdot 0{,}1525 \approx 303 \end{aligned}$$

kvinder. Diskutér i hvilken forstand dette antal kunne siges at være mere korrekt *end de 288 kvinder fundet ved en simpel optælling i datasættet.*

5 Matematiske egenskaber

Indtil nu har vores overvejelser taget udgangspunkt i et konkret datasæt, nemlig fodlængderne fra *ANSUR II* undersøgelsen. Men i dette afsnit vil vi lægge data og konkrete tal til side, og arbejde abstrakt med de matematiske formler.

Først en præcisering af sprogbrugen omkring ordet *"gennemsnit"*. Indtil videre har vi beregnet *gennemsnit* af datapunkter, f.eks. for de 1986 kvinders fodlængder, og vi har sagt at en normalfordeling er specificeret ved sit *gennemsnit* og sin *spredning*. Men den korrekte matematiske terminologi er at bruge ordet *"middelværdi"* for gennemsnittet af en teoretisk sandsynlighedsfordeling. Normalfordelingen er altså specificeret ved sin *middelværdi* og sin *spredning*. Tilsvarende bruges ordet *"gennemsnit"* for middelværdien af et datasæt bestående af konkrete tal. Men nok om disse to ord, det er faktisk slet ikke vigtigt for vores videre overvejelser.

Sandsynlighedstætheden $f_{\mu,\sigma}(x)$ for en normalfordeling med middelværdi μ og spredning σ blev opskrevet i (7). Vi vil se nærmere på de matematiske egenskaber ved denne funktion, som for overskuelighedens skyld gentages her

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

Idet spredningen σ er et positivt tal, kan vi som det første bemærke, at $f_{\mu,\sigma}(x) > 0$ for alle x. Dette betyder, at sandsynligheden for et interval $[a;b]$,

Interval	Normalfordelingssandsynlighed
Middelværdi ± 1 gange spredning	0,682689 ≈ 68%
Middelværdi ± 2 gange spredning	0,954500 ≈ 95%
Middelværdi ± 3 gange spredning	0,997300 ≈ 100%
Middelværdi ± 4 gange spredning	0,999937 ≈ 100%

Tabel 2: Normalfordelingssandsynligheder for forskellige intervaller omkring middelværdien.

hvor $a < b$ er henholdsvis venstre og højre endepunkt, er strengt positiv

$$\int_a^b f_{\mu,\sigma}(x)\,dx > 0,$$

hvilket passer med at sandsynligheder er ikke-negative tal. Tabel 2 oplister normalfordelingssandsynligheder for intervaller af formen

$$[\mu - n \cdot \sigma; \mu + n \cdot \sigma]$$

for $n = 1, 2, 3, 4$, hvor μ er middelværdien og σ er spredningen. Disse sandsynligheder afhænger ikke af de konkrete værdier af μ og σ, hvilket afspejler at normalfordelingerne har samme "form" uanset hvad middelværdien og spredningen er.

Vi bemærk, at sandsynligheden for at være langt væk fra middelværdien (målt i enheder af spredningen) er så forsvindende, at man i praksis godt kan bruge normalfordelingen til at beskrive målinger der ikke ligger på hele den reelle akse. F.eks. brugte vi en normalfordeling med middelværdi $\mu_{\text{kvinde}} = 246{,}29$ og spredning $\sigma_{\text{kvinde}} = 12{,}43$ til at beskrive fodlængden på de kvindelige soldater. Fødder har en positiv længde, og den tilhørende normalfordelingssandsynlighed for at få en negativ fodlængde kan beregnes til

$$\int_{-\infty}^0 f_{\mu_{\text{kvinde}},\sigma_{\text{kvinde}}}(x)\,dx = 1{,}12 \cdot 10^{-87}$$

Selv om dette stadigvæk er et strengt positivt tal, så er det så forsvindende lille, at det ikke gør noget i praksis.

Tabel 2 indikerer også, at der gælder

$$\int_{-\infty}^{\infty} f_{\mu,\sigma}(x)\,dx = 1. \tag{8}$$

I hvert fald er det sådan, at normalfordelingssandsynlighederne for at ligge i intervallerne *middelværdi ± 3 gange spredning* og *middelværdi ± 4 gange*

spredning begge er meget tæt på 1. Dette tyder på at normalfordelingssandsynligheden for at ligge i intervallet *middelværdi* $\pm$ *n gange spredning* nærmer sig 1 når n vokser mod uendelig, hvilket netop er påstanden i (8). Dette er en helt afgørende egenskab, som betyder at funktionen i (7) er en sandsynlighedstæthed, altså en funktion der antager ikke-negative værdier og hvor arealet under grafen er lig med $1 = 100\%$.

Følgende opgave giver et matematisk bevis for den afgørende (8). Beviset består af 4 trin. Det tredje trin bruger dobbelt integraler og polær integration, som ikke er en del af det almindelige gymnasiepensum. Det er ikke afgørende for resten af hovedteksten, at man løser denne opgave. Så man er velkommen til blot at overspringe opgaven.

Opgave 19. *I denne opgave vil vi trin for trin arbejde os igennem et matematisk bevis for* (8).

Trin 1: Argumentér for, at det er tilstrækkeligt at vise (8) *for en normalfordeling med middelværdi* $\mu = 0$ *og spredning* $\sigma = 1$. *En sådan normalfordeling kaldes forøvrigt for en* standard normalfordeling.

Vink: Integration ved substitution $u = \frac{x-\mu}{\sigma}$ *giver*

$$\begin{aligned}\int_{-\infty}^{\infty} f_{\mu,\sigma}(x)\,dx &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\cdot\sigma}\cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}\,dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{u^2}{2}}\,du \\ &= \int_{-\infty}^{\infty} f_{0,1}(u)\,du.\end{aligned}$$

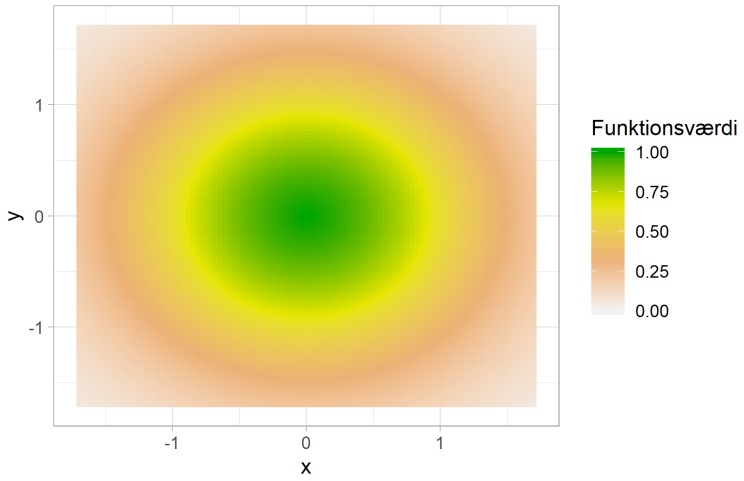
Trin 2: Argumentér for, at det er tilstrækkeligt at vise formlen

$$\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}. \tag{9}$$

Vink: Indsæt $\mu = 0$ *og* $\sigma = 1$ *i* (8).

Trin 3: Umiddelbart er det overraskende, at konstanten π *optræder i* (9). *Fra geometrien genkender vi tallet* 2π *som omkredsen på en cirkel med radius 1. For at slippe af med kvadratroden kvadrerer vi på begge sider af* (9), *som vi skal bevise, hvorefter vi regner videre på venstre siden. Dette giver følgende dobbelt integral*

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)^2 &= \left(\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx\right)\cdot\left(\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}}\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy.\end{aligned}$$



Figur 13: Visualisering af funktionen $\mathrm{e}^{-\frac{x^2+y^2}{2}}$ af de to variable x og y.

Figur 13 viser "grafen" for funktionen $e^{-\frac{x^2+y^2}{2}}$ af de to variable x og y, hvor funktionsværdien er visualiseret ved farvekodning. Et alternativ til en sådan farvegraf er at lave et 3d-plot, men den anvendte visualisering er mere nyttig for os. F.eks. er det lettere at se, at funktionen er konstant på cirkler omkring $(0,0)$. Mere præcist gælder $e^{-\frac{x^2+y^2}{2}} = e^{-\frac{r^2}{2}}$ når (x,y) ligger på omkredsen af cirklen[10] *med radius r.*

Argumentér for, at (x,y)-planen kan skrives som foreningen af omkredsen af cirkler med radius $r \geq 0$, og at dobbelt integralet dermed kan omskrives til

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}\,dx\,dy = \int_0^{\infty} 2\pi r \cdot e^{-\frac{r^2}{2}}\,dr. \tag{10}$$

Vink: Du får brug for, at cirklen med radius r har omkreds $2\pi r$.

Trin 4: Brug omskrivningen (10) til at vise (9), og dermed (8).

Vink: Integration ved substitutionen $u = e^{-\frac{r^2}{2}}$ giver

$$\int_0^{\infty} 2\pi r \cdot e^{-\frac{r^2}{2}}\,dr = \int_0^1 2\pi\,du = 2\pi.$$

[10] Her kom cirklen i spil, så det lader til at vi er på rette vej med vores matematiske bevis!

5.1 Simulering af normalfordelinger

I afsnit 4.1 brugte vi normalfordelingen til at beregne andelen af kvinder med fodlængder i intervallet fra 247,5 mm til 252,5 mm. Dette blev gjort ved at beregne integralet

$$\int_{247,5}^{252,5} f_{\mu,\sigma}(x)\,dx$$

for konkrete værdier af middelværdien μ og spredningen σ. For de 100 tal i tabel 1 havde vi f.eks. gennemsnit $\mu = 248{,}23$ og spredning $\sigma = 12{,}12$, og vi beregnede andelen

$$\int_{247,5}^{252,5} f_{\mu,\sigma}(x)\,dx = 0{,}1617 = 16{,}17\% \quad \text{når } \mu = 248{,}23 \text{ og } \sigma = 12{,}12. \quad (11)$$

Som tidligere nævnt kan normalfordelingsintegraler ikke udregnes "i hånden" ved brug af klassisk integralregning. Integralet skal derimod beregnes ved brug af en computer. Men i stedet for at *"beregne normalfordelingsintegraler"* vil vi i dette afsnit vise hvorledes man kan *"simulere normalfordelte tal"*, og derefter diskutere hvordan dette kan bruges til at lave beregninger med normalfordelingen. At simulere normalfordelte tal betyder, at man f.eks. bruger en computer til at generere nogle tal der matematisk set er normalfordelte.

For at beskrive hvordan dette kan gøres vil vi bruge begrebet *stokastisk variabel*. En stokastisk variabel er en størrelse, der antager en *tilfældig værdi*. Typisk vil man angive *fordelingen* af en given stokastiske variabel X. Hvis vi f.eks. siger, at X er *normalfordelt med middelværdi μ og spredning σ*, så betyder det, at sandsynligheden for at X antager en værdi fra intervallet $]a, b]$ er givet ved integralet

$$P\big(a < X \leq b\big) = \int_a^b f_{\mu,\sigma}(x)\,dx,$$

hvor $f_{\mu,\sigma}(x)$ er givet ved (7). Og hvis vi siger, at den stokastiske variable U er *ligefordelt* på intervallet $]0, 1]$, så betyder det, at sandsynligheden for at U antager en værdi fra intervallet $]a, b]$, hvor $0 \leq a < b \leq 1$, er givet ved

$$P\big(a < U \leq b\big) = b - a.$$

Vi får også brug for begrebet *stokastisk uafhængighed*. To stokastiske variable U og V siges at være stokastiske uafhængige, hvis der gælder

$$P\big(a < U \leq b \text{ og } c < V \leq d\big) = P\big(a < U \leq b\big) \cdot P\big(c < V \leq d\big)$$

for alle $a < b$ og $c < d$.

Figur 13 og resten af opgave 19 viser, at hvis U og V er uafhængige stokastiske variable, der begge er ligefordelte på $]0,1]$, og hvis (X,Y) er givet ved

a) at afstanden R fra $(0,0)$ til (X,Y) opfylder $U = \mathrm{e}^{-\frac{R^2}{2}}$,

b) at vinklen[11] θ fra x-aksen mod urets retning til linjen fra $(0,0)$ til (X,Y) opfylder $\theta = 2\pi V$,

så er X og Y uafhængige stokastiske variable, der begge er normalfordelte med middelværdi 0 og spredning 1.

Opgave 20. *Vis, at (X,Y) kan fås ud fra (U,V) ved formlerne*

$$X = \cos(2\pi V)\cdot\sqrt{-2\cdot\ln(U)}, \qquad Y = \sin(2\pi V)\cdot\sqrt{-2\cdot\ln(U)}.$$

Mange matematikprogrammer og programmeringssprog giver mulighed for at simulerer ligefordelte tal på intervallet $]0,1]$. Dette betyder, at hver gang med "beregner" resultatet, så fås et nyt tilfældigt tal mellem 0 og 1. Og hvis man skal finde to uafhængige ligefordelte tal, så skal man blot "lave beregningen" to gange i træk. Men man kan også simulere approksimative ligefordelinger ved at kaste med en terning.

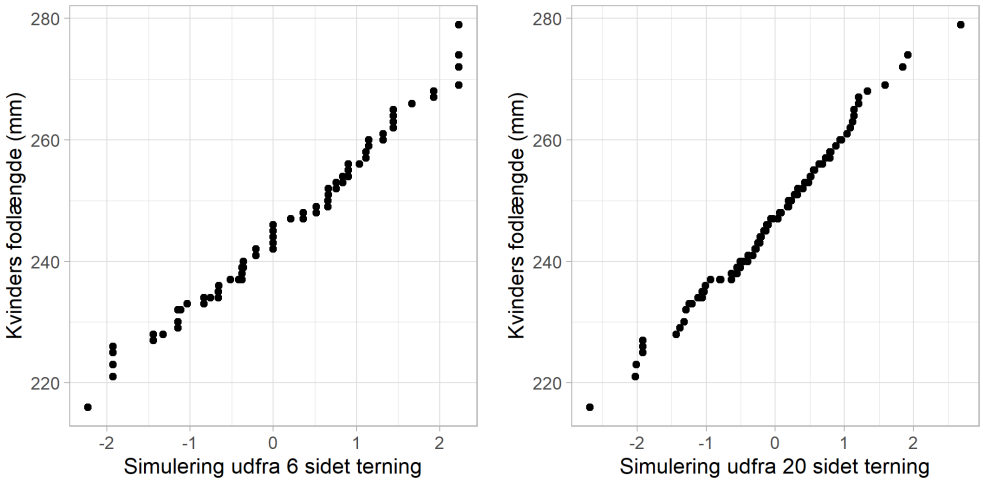
Opgave 21. *Argumentér for, at hvis W er udfaldet af en ærlig terning med K sider, f.eks. $K = 6$ eller $K = 20$, så er $\frac{W-\frac{1}{2}}{K}$ approksimativt ligefordelt på intervallet $]0,1]$. Brug dette til at simulere 100 approksimativt normalfordelte tal. Sammenlign med fordelingen af fodlængderne for de 1986 kvindelige soldater i* ANSUR II *datasættet. Virker metoden?*

Vink: Man skal kaste terningen 100 gange, og samle terningkastene i 50 par af to kast. Dette omregnes til 50 par af U'er og V'er, og derefter til 50 par af X'er og Y'er. Tilsammen giver dette 100 tal, som er approksimativt normalfordelte.

5.2 Beregning af fraktiler

I opgave 21 beskrev vi hvordan man kan simulere approksimativt ligefordelte stokastiske variable U og V ved at kaste med en terning, og derefter bruge dette til at simulere approksimative normalfordelte stokastiske variable X og Y. Sådanne tilfældige og normalfordelte tal kan derefter bruges til at lave sandsynlighedsberegninger, f.eks. at bestemme normalfordelingsfraktiler.

[11] Vinklen måles her i radianer fra 0 til 2π.



Figur 14: Normalfordelingsreference-fraktildiagram for simuleringen af normalfordelinger via terningkast, se opgave 21.

Men hvis man kun ønsker at lave sandsynlighedsberegningerne, så kan man i stedet for beskrive fordelingen af (U, V) ved systematisk at placere punkter jævnt på enhedskvadratet $]0,1]\times]0,1]$. Figur 15 viser 25 punkter jævnt fordelt i et 5×5 grid. Oplistet rækkevis fra nederste venstre hjørne er de 25 punkter

$$(\frac{1}{10}, \frac{1}{10}),\ (\frac{3}{10}, \frac{1}{10}),\ \ldots,\ (\frac{9}{10}, \frac{1}{10}),\ (\frac{1}{10}, \frac{3}{10}),\ (\frac{3}{10}, \frac{3}{10}),\ \ldots,\ (\frac{9}{10}, \frac{3}{10}),$$
$$\ldots,\ (\frac{1}{10}, \frac{9}{10}),\ (\frac{3}{10}, \frac{9}{10}),\ \ldots,\ (\frac{9}{10}, \frac{9}{10})$$

Indsættes disse 25 talpar i formlerne

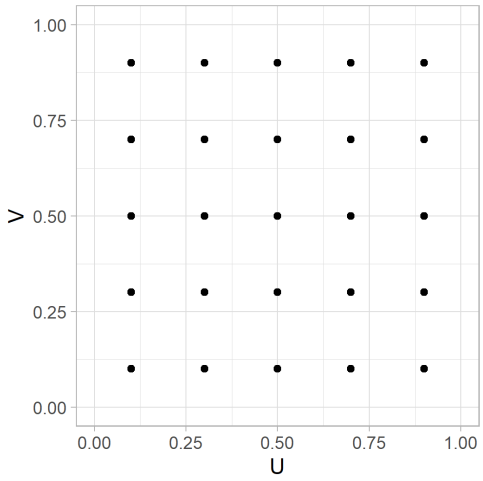
$$X = \cos(2\pi V) \cdot \sqrt{-2 \cdot \ln(U)}, \qquad Y = \sin(2\pi V) \cdot \sqrt{-2 \cdot \ln(U)}$$

fås følgende 50 tal (her afrundet til 3 decimaler)

```
 1,736  1,261 -0,663  2,041 -2,146  0,000 -0,663 -2,041  1,736
-1,261  1,255  0,912 -0,480  1,476 -1,552  0,000 -0,480 -1,476
 1,255 -0,912  0,953  0,692 -0,364  1,120 -1,177  0,000 -0,364
-1,120  0,953 -0,692  0,683  0,496 -0,261  0,803 -0,845  0,000
-0,261 -0,803  0,683 -0,496  0,371  0,270 -0,142  0,437 -0,459
 0,000 -0,142 -0,437  0,371 -0,270
```

Opgave 22. *Efterprøv disse udregninger! Vink: De to første tal er givet ved*

$$\cos\left(\tfrac{2\pi}{10}\right) \cdot \sqrt{-2 \cdot \ln\left(\tfrac{1}{10}\right)} = 1{,}736 \qquad \sin\left(\tfrac{2\pi}{10}\right) \cdot \sqrt{-2 \cdot \ln\left(\tfrac{1}{10}\right)} = 1{,}261$$



Figur 15: Punkterne $(U, V) = (\frac{i-\frac{1}{2}}{K}, \frac{j-\frac{1}{2}}{K})$ for $i = 1, \ldots, K$ og $j = 1, \ldots, K$ med $K = 5$. Dette giver $K^2 = 25$ punkter jævnt fordelt på enhedskvadratet $]0, 1] \times]0, 1]$.

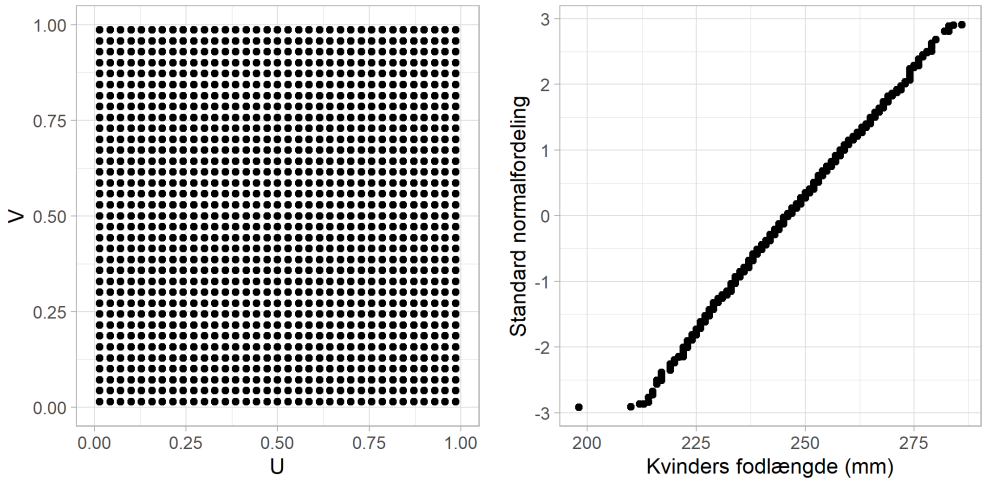
Man kan nu finde approksimationer af normalfordelingsfraktiler ved at ordne de beregnede tal i voksende rækkefølge. Dette giver de 50 tal:

```
-2,146 -2,041 -1,552 -1,476 -1,261 -1,177 -1,120 -0,912 -0,845
-0,803 -0,692 -0,663 -0,663 -0,496 -0,480 -0,480 -0,459 -0,437
-0,364 -0,364 -0,270 -0,261 -0,261 -0,142 -0,142  0,000  0,000
 0,000  0,000  0,000  0,270  0,371  0,371  0,437  0,496  0,683
 0,683  0,692  0,803  0,912  0,953  0,953  1,120  1,255  1,255
 1,261  1,476  1,736  1,736  2,041
```

Opgave 23. *Gør rede for, at de 50 tal ovenfor er approksimationer af* 1%, 3%, ..., 99% *fraktilerne for en normalfordeling med middelværdi 0 og spredning 1. Forklar i særdeleshed, hvorfor der er tale om approksimationer og ikke eksakte beregninger.*

Hvis man vil have en bedre approksimation af ligefordelingen på enhedskvadratet, og dermed af den efterfølgende bestemmelse af normalfordelingsfraktiler, så skal man vælge flere punkter. Venstre panel i figur 16 viser således $35 \cdot 35 = 1225$ punkter jævnt fordelt på enhedskvadratet, hvilket giver $\frac{i-\frac{1}{2}}{2450} \cdot 100\%$ fraktilerne for $i = 1, 2, \ldots, 2450$ for en standard normalfordeling.

Højre panel i figur 16 viser det tilhørende fraktildiagram for sammenligningen med de 1986 målinger af kvinders fodlængder fra *ANSUR II* undersøgelsen. Tidligere har vi brugt kvindernes fodlængder som reference til at afgøre, om andre datasæt er (næsten) normalfordelte. Men nu har vi lavet



Figur 16: Panelet til venstre viser $1225 = 35 \cdot 35$ punkter jævnt fordelt på enhedskvadratet $]0,1]\times]0,1]$. Disse punkter kan transformeres til $2450 = 2 \cdot 35 \cdot 35$ fraktiler for standard normalfordelingen, dvs. en normalfordeling med middelværdi 0 og spredning 1. Højre panel viser sammenligningen af disse fraktiler med fraktilerne for de 1986 kvindelige fodlængder.

en matematisk konstruktion af normalfordelingsfraktiler og kan vende tingene på hovedet! Vi har en *matematisk normalfordelingsreference* (y-aksen på højre panel i figur 16) og kan nu vurdere om *kvindernes fodlængder er normalfordelte!* Dette viser sig at være tilfældet pånær en enkelt kvinde, som har en usædvanlig kort fod (kortere end 200 mm).

Opgave 24. *Lad os også prøve at bestemme integralet fra* (11)*. Altså*

$$\int_{247{,}5}^{252{,}5} f_{\mu,\sigma}(x)\,dx$$

når $\mu = 248{,}23$ *og* $\sigma = 12{,}12$*. Vis, at integration ved substitutionen* $y = \frac{x-248{,}23}{12{,}12}$ *giver*

$$\int_{247{,}5}^{252{,}5} f_{\mu,\sigma}(x)\,dx = \int_{-0{,}06023}^{0{,}35231} f_{0,1}(y)\,dy$$

Beregn de $2450 = 2 \cdot 35 \cdot 35$ *normalfordelingsfraktiler fra figur 16 og optæl, at* 388 *af disse ligger i intervallet fra -0,06023 til 0,35231. Gør rede for, at dette giver*

$$\int_{-0{,}06023}^{0{,}35231} f_{0,1}(y)\,dy \approx \frac{388}{2450} = 15{,}84\%$$

Diskutér om dette giver en god approksimation af den præcise integralværdi på 16,17%.

6 Den Centrale Grænseværdisætning

En af grundene til at normalfordelingen er så utrolig vigtig er, at den dukker op i mange forskellige sammenhæng. Vi har allerede set, at normalfordelingen beskriver højde og fodlængder for mænd og kvinder (afsnit 2.1), "fordelingen af summen af antal øjne ved slag med to terninger" og "fordelingen af succeser" (afsnit 3.2), og til sidst i dette manuskript skal vi se, at normalfordelingen også beskriver fordelingen af logaritmen af danskernes disponible indkomst (afsnit 7). Alle disse eksempler har følgende to ting tilfælles

(a) Normalfordelingen er ikke nødvendigvis en eksakt matematisk beskrivelse. Men derimod er normalfordelingen en approksimativ beskrivelse[12], der f.eks. kan danne grundlag for statistiske analyser af datasæt.

(b) I alle eksemplerne kan målingen forstås som en sum af mange bidrag[13].

"Den Centrale Grænseværdisætning" er en af hovedsætningerne i matematisk sandsynlighedsregning. Ganske kort siger denne sætning, at under passende forudsætninger gælder

$$\text{(b)} \Longrightarrow \text{(a)}$$

Eller skrevet lidt mere udførligt (men stadigvæk løseligt):

> Under passende forudsætninger gælder, at summen af mange stokastiske bidrag er approksimativt normalfordelt. Approksimationen bliver bedre og bedre jo flere led der er i summen.

Den første version af den centrale grænseværdisætning er *De Moivre-Laplace's sætning* publiceret i 1738, som giver approksimationen

$$\frac{n!}{k!(n-k)!}p^k(1-p)^k \approx \frac{1}{\sqrt{2\pi np(1-p)}}\exp\left(-\frac{(k-np)^2}{2np(1-p)}\right)$$

når $0 < p < 1$ og $k \approx np$.

Opgave 25. *Gør rede for, at* De Moivre-Laplace's sætning *siger, at binomialfordelingssandsynligheder kan approksimeres ved normalfordelingstætheder.*

Efterfølgende er den centrale grænseværdisætning blevet udvidet og forbedret mange gange. Sådanne forbedringer handler dels om at præcisere betydningen af *"under passende forudsætninger"*, som vi gerne vil have til at

[12] Vi har gentagne gange beskrevet noget som "(næsten) normalfordelt".

[13] På side 35 gives der 4 eksempler på hvad dette betyder.

være så lidt restriktive som muligt. For jo svagere forudsætningerne er, jo oftere vil normalfordelingen kunne bruges til beskrivelse af omverdenen. Dels om at kvantificere betydningen af *"approksimativt normalfordelt"*, således at vi kender begrænsningerne ved normalfordelingsapproksimationen. Det vil dog føre for vidt at komme nærmere ind på de matematiske detaljer her. I stedet for vil vi argumentere for egenskab (b) for de forskellige eksempler:

Højde: En persons højde er bestemt både af arv og miljø. Angående den genetiske arv, så er der mange forskellige gener der påvirker højde. Tilsvarende er højden påvirket af mange faktorer under opvæksten, både som foster, barn og ung. Alle disse forskellige bidrag er med til at bestemme højden som voksen.

Fodlængde: Samme argumentation som for højde.

Summen af antal øjne ved slag med to terninger: Ja, vi har allerede sagt at det er summen af to bidrag.

Binomialfordelingen: Binomialfordelingen med antalparameter = 20 beskriver antal *successer* ud af 20 forsøg, hvor sandsynligheden for succes er givet ved sandsynlighedsparameteren. Hvis hvert forsøg indkodes som *1 = succes* og *0 = fiasko*, så beskriver binomialfordelingen *"summen af forsøgene"*. Hvert af forsøgene giver dermed et bidrag til binomialfordelingen.

Disponibel indkomst: En persons disponible indkomst afhænger blandt andet af personens baggrund og karriereforløb. Elementerne i disse ting giver hver for sig et bidrag til indkomsten. Men til forskel for biologiske størrelser såsom højde og fodlængde, så er mange økonomiske fænomener multiplikative i deres natur[14]. F.eks. er en lønstigning på 2% ikke det samme beløb i kroner og ører for direktøren som for sekretæren. Men idet $\log(x \cdot y) = \log(x) + \log(y)$, så transformerer logaritmen *"multiplikative fænomener"* til *"additive fænomener"*. Som udgangspunkt vil man som statistiker derfor forvente, at det er logaritmenen af den disponible indkomst, der er normalfordelt.

[14]Mange størrelser fra fysik er også multiplikative i deres natur. Således skal man i langt de fleste fysikformler gange størrelser sammen for at beregne output.

Disponibel indkomst	Antal personer
Under 100.000 kr	816.953
100.000 – 199.999 kr	1.562.916
200.000 – 299.999 kr	1.332.199
300.000 – 399.999 kr	624.411
400.000 – 499.999 kr	220.366
500.000 – 749.999 kr	131.034
750.000 – 999.999 kr	27.867
1.000.000 – 1.999.999 kr	21.032
2.000.000 – 2.999.999 kr	3.954
3.000.000 – 3.999.999 kr	1.504
4.000.000 – 4.999.999 kr	763
5.000.000 – 9.999.999 kr	1.118
10.000.000 kr og derover	484
Personer ialt	4.744.601

Tabel 3: Disponibel indkomst for personer over 14 år som har boet i Danmark hele året i 2017. Forbrugerprisindeks (2016-priser). Data downloadet fra *Statistikbanken* [10].

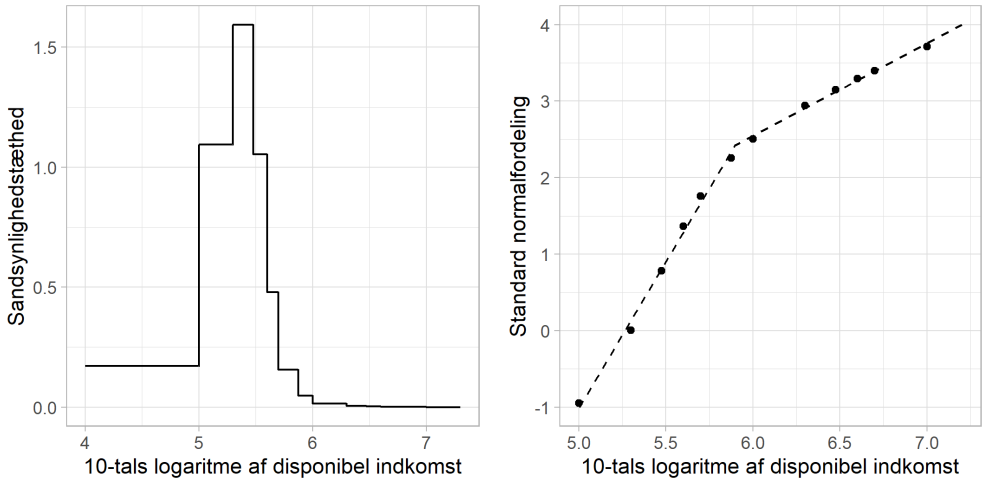
7 Disponibel indkomst

Danmarks Statistik definerer *disponibel indkomst* som (citat):

> For at beregne disponibel indkomst finder man summen af erhvervsindkomst, offentlige overførsler, private pensioner, formueindkomst samt anden personlig indkomst. Alle disse indkomster er før skat. Dernæst trækkes skatter, renteudgifter og underholdsbidrag fra og lejeværdi af egen bolig lægges til.

Danmarks Statistik har en portal på internettet, *Statistikbanken*, hvorfra man kan downloade en bred vifte af forskellige statistisk information omkring Danmark og danskerne. F.eks. kan man downloade grupperet information omkring danskernes disponible indkomst [10]. Da dette manuskript blev skrevet omhandlede de yngste data indkomståret 2017. Disse data findes i tabel 3. Vi har ingen umiddelbar forventning om, at den disponible indkomst skulle være normalfordelt. Men som argumenteret for i afsnit 6, så er der grund til at tro, at logaritmen af indkomsten kunne være normalfordelt. Så lad os undersøge dette.

Venstre panel i figur 17 viser et bud på sandsynlighedstætheden for indkomstfordelingen på en logaritmisk skala. Jeg har valgt at bruge 10-tals logaritmen, hvor f.eks. $\log(100.000) = 5$ og $\log(1.000.000) = 6$. Men der skal



Figur 17: Sandsynlighedstæthed og normalfordelingsfraktildiagram for logaritmen af danskernes disponible indkomst i året 2017. Den stiplede linje i højre panel viser tilpasningen af punkterne med en ret linje der knækker i et enkelt punkt.

træffes lidt flere valg før vi kan lave figuren. For at give specifik mening til første og sidste datarække i tabel 3 *vælger* jeg således, at den *laveste* og den *højeste* disponible indkomst er henholdsvis 10.000 og 20.000.000. Videre vælger jeg at lave konstante sandsynlighedstætheder mellem indkomst grænserne fra tabel 3.

Opgave 26. *Lav de beregninger der skal til for at bestem sandsynlighedstæthederne i venstre panel i figur 17. Hvordan kan det være, at man ikke kan bruge 0 kr som den nedre grænse for den disponible indkomst i denne visualisering?*

Vink: For de 1.332.199 personer med en disponibel indkomst mellem 200.000 og 300.000 kr gælder

$$\frac{1}{\log(300.000) - \log(200.000)} \cdot \frac{1.332.199}{4.744.601} = 1{,}59$$

Venstre panel i figur 17 ser jo ikke super klokkeformet ud. Men sådanne histogramlignende visualiseringer kan også være svære at se på. Desuden afhænger denne visualisering også af de intervalgrupperinger som *Danmarks Statistik* har besluttet for os, og i særdeleshed også den nedre indkomstgrænse på 10.000 kr, som vi selv har valg. For at afgøre om den logaritmiske disponible indkomst er normalfordelt, er det således meget bedre af bruge et fraktildiagram mod fraktilerne for standard normalfordelingen. Dette findes i højre panel i figur 17.

Opgave 27. *I afsnit 5.2 har vi udviklet en approksimativ metode til at finde fraktilerne for en standard normalfordeling*[15]*. Således viser figur 16 beregningen af* $2450 = 2 \cdot 35 \cdot 35$ *fraktiler. Men andelen af personer i den øverste indkomstgruppe er*

$$\frac{484}{4.744.601} \approx \frac{1}{9803}$$

Argumentér for, at vi som et minimum har brug for approksimative beregninger af 9803 *normalfordelingsfraktiler for at vi kan lave et pålideligt fraktildiagram for indkomstfordelingen.*

I praksis viser det sig, at den approksimative metode fra afsnit 5.2 med $N = 101$ og dermed $20402 = 2 \cdot 101 \cdot 101$ fraktiler ikke giver tilstrækkelig præcise beregninger af de ekstreme fraktiler til at give et præcist billede af indkomstfordelingen. Heldigvis giver mange software programmer muligheden for at lave eksakte beregninger af normalfordelingsfraktilerne. Så det er også hvad vi har brugt for at lave figur 17.

Efter disse lidt spidsfindige matematiske overvejelser vil vi vende os mod den spændende socio-økonomiske fortolkning af fraktildiagrammet i figur 17. Umiddelbart ser vi, at logaritmen af den disponible indkomst *ikke* er normalfordelt. For punkterne i fraktildiagrammet ligger ikke på en ret linje. Derimod ligger de pænt på *to rette linjer*, som skærer hinanden i fraktilen $x = 5{,}90$ for log-indkomstfordelingen og $y = 2{,}42$ for standard normalfordelingen. Fortolkningen af dette er, at samfundet er knækket over i to indkomstgrupper der hver for sig har en *"naturlig dynamik"* (ifølge overvejelserne fra afsnit 6). Den nederst gruppe består af de

$$\int_{-\infty}^{2{,}42} f_{0,1}(u)\, du = 0{,}9922 = 99{,}22\%$$

af befolkningen med en disponibel indkomst under

$$10^{5{,}90} = 794.328$$

kroner. Den øverste gruppe består de resterende 0,78% af befolkningen med de højeste disponible indkomster.

Opgave 28. *Højre panel i figur 17 viser, at gruppen med de højeste disponible indkomster har en større spredning (på log-skalaen). Hvad er fortolkningen af dette?*

[15] En standard normalfordeling er en normalfordeling med middelværdi 0 og spredning 1.

Litteratur og ressourcer

[1] *"Statistiske Oplysninger: Udfaldet, Gennemsnitshøjden og BMI (Body Mass Index) på Forsvarets Dag / Sessionen"*, Forsvarministeriet, Personalestyrelsen, Maj 2019. `http://viewer.zmags.com/publication/5caf2b62#/5caf2b62/8`

[2] Bo Markussen & Anders Rønn-Nielsen, *"Lineær regression til A-niveau"*, 2018. Supplerende undervisningsmateriale til gymnasiet. Materialet findes også i en version til B-niveau. Kan downloades fra `https://emu.dk/stx/matematik/om-lineaer-regression-og-statistik-i-gymnasiet`

[3] Potræt af Carl Friedrich Gauss (1777–1855), malet af Christian Albrecht Jensen (1792–1870), Gauß-Gesellschaft Göttingen e.V. (Foto: A. Wittmann). Downloadet fra `https://commons.wikimedia.org/w/index.php?curid=57629`

[4] *"ANSUR II Working Databases"*, Natick Soldier Research Development and Engineering Center, US Army Research, 2012. Datasæt kan f.eks. downloades fra `https://data.world/datamil`

[5] Carl Friedrich Gauss, *"Theoria motus corporum coelestium in sectionibus conicis solem ambientium"*, Hamburgi: sumtibus Frid. Perthes et I.H. Besser, 1809.

[6] *"Dirichlets skuffeprincip"*, se `https://da.wikipedia.org/wiki/Dirichlets_skuffeprincip`.

[7] Jesper Bang-Jensen, Bodil Bruun og Jørgen Dejgaard, *"Matematisk formelsamling, stx, A-niveau"*, Styrelsen for Undervisning og Kvalitet, Undervisningsministeriet, Maj 2018. Se *"Spredning for observationssættet* $x_1, \ldots, x_n$*"* i `https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf19/mar/190325-mat-a-stx-formelsamling-feb-2019.pdf?la=da`

[8] User:Greg L, computer grafikbillede af kilogram prototypen, downloadet fra `https://commons.wikimedia.org/wiki/File:CGKilogram.jpg`

[9] Abraham de Moivre, *"The Doctrine of Chances: or, a method for calculating the probabilities of events in play"*, second edition, Woodfall, 1738.

[10] *"Disponibel indkomst for personer over 14 år efter region, enhed, køn og indkomstinterval"*, Statistik Banken, Danmarks Statistik. Grupperet indkomstdata downloadet fra `www.statistikbanken.dk/INDKFPP3`