Den normale fordeling

Nu undersøger vi den sandsynlighedsfordeling, der er mest central for statistikker: normalfordelingen. Hvis vi er overbeviste om, at vores data er næsten normale, åbner det for mange magtfulde statistiske metoder. Her bruger vi de grafiske værktøjer til R til at vurdere normaliteten af vores data og også lære at generere tilfældige tal fra en normal distribution.

Data

Vi ser på data over legemer. Data indeholder målinger fra 247 men og 260 kvinder, hvoraf de fleste personer er sunde unge mennesker.

```
download.file("http://www.openintro.org/stat/data/bdims.RData", destfile = "bdims.RData")
load("bdims.RData")
```

Vi kigger på de første observationer:

```
head(bdims)
```

Vi har for hver observation 25 variable. Vi vil her fokusere på 3 kolonner: vægt (wgt), højde i cm (hgt) og køn (1 angiver mand, 0 angiver kvinde).
Da mænd og kvinder har en tendens til at have forskellige kropsdimensioner, vil det være nyttigt at oprette to yderligere datasæt: et datasæt kun med mænd og et andet kun med kvinder.

```
mdims <- subset(bdims, sex == 1)
fdims <- subset(bdims, sex == 0)
```

Opgave 1: Lav et histogram over mænds højde henholdsvis kvinders højde. Hvordan vil du beskrive de to fordelinger?

Den normale fordeling

Blev ord som klokkeformet eller normalfordelt anvendt til at beskrive fordelingerne? Det er fristende at sige, at vi står over for en unimodal symmetrisk fordeling.
For at se, hvor nøjagtig denne beskrivelse er, kan vi plotte en normal distributionskurve oven på et histogram for at se, hvor tæt data følger en normal distribution. Denne normale kurve skal have samme gennemsnit og standardafvigelse som data. Vi arbejder med kvinders højder, så lad os

gemme dem som et separat objekt og derefter beregne nogle statistikker, der vil blive henvist til senere.

fhgtmean <- middel (fdims $ hgt)
fhgtsd <- sd (fdims $ hgt)

Derefter laver vi et histogram til brug som baggrund og bruger tæthedsfunktionen til at vise en normal sandsynlighedskurve. Forskellen mellem et "frekvenshistogram" og et "tæthedshistogram" er, at mens vi i et "frekvenshistogram", højderne af søjlerne udgør det samlede antal observationer, så udgør områderne i et tæthedshistogram frekvenserne, således at arealet er lig med 1. Arealet af hver søjle kan beregnes som simpelthen højden gange bredden på stangen. Brugen af et "tæthedshistogram" tillader os at sammenligne histogram med en normal tæthedsfordeling. Eneste forskel mellem "frekvenshistogram" og tæthedshistogram blot y - aksen. Selve formerne er ens. Tjek dette ved brug af nedenstående:

```
hist(fdims$hgt, probability = TRUE)
x <- 140:190
y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)
lines(x = x, y = y, col = "blue")
```

Efter kortlægning af tæthedshistogrammet med den første kommando opretter vi x- og y-koordinaterne for den normale kurve. Vi valgte x-området som 140 til 190 for at spænde over hele rækkevidden af højde. For at oprette y værdierne bruger vi dnorm til at beregne tætheden af hver af disse x-værdier i en fordeling, der er normalfordelt med middelværdien fhgtmean og standardafvigelse fhgtsd. Den endelige kommando tegner en kurve på det eksisterende plot (tæthedshistogrammet) ved at forbinde hvert af de punkter, der er angivet med x og y. Argumentet col angiver blot farven for den linje, der skal tegnes. Hvis vi udeladte det, ville linjen blive trukket i sort.
Øverste del af kurven er afskåret, fordi x- og y-aksernes grænser er indstillet til bedst at passe til histogrammet. For at justere y-aksen kan du tilføje et tredje argument til histogramfunktionen: ylim = c (0, 0,06).

Opgave 2: Baseret på dette plot, ser det så ud til, at data følger en næsten normal fordeling?

Evaluering af den normale fordeling

Ved at se på histogrammets form kan det vurderes, om data ser ud til at være næsten normalt fordelt, men det kan være frustrerende at beslutte, hvor tæt histogrammet er til kurven. En alternativ tilgang involverer konstruering af et normalt sandsynlighedsdiagram, også kaldet et normalt Q-Q-plot for "kvantil-kvantil".

```
qqnorm(fdims$hgt)
qqline(fdims$hgt)
```

Et datasæt, der er næsten normalt, vil resultere i et sandsynlighedsdiagram, hvor punkterne tæt følger linjen. Eventuelle afvigelser fra normalitet fører til afvigelser af disse punkter fra linjen. Handlingen for kvindelige højder viser punkter, der har en tendens til at følge linjen, men med nogle forkerte punkter mod halerne. Vi har det samme problem, som vi stødte på med histogrammet ovenfor: hvor tæt er tæt nok?

En nyttig måde at løse dette spørgsmål på er at omformulere det på følgende måde: hvordan ser sandsynlighedsdiagrammer ud for data, som jeg ved, kommer fra en normal distribution? Vi kan besvare dette ved at simulere data fra en normal distribution ved hjælp af rnorm.

```
sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
```

Det første argument angiver, hvor mange værdier, vi vil generere, hvor antallet skal være det sammen aom antallet af højder i fdims-datasættet ved hjælp af længde-funktionen. De to sidste argumenter bestemmer middelværdien og standardafvigelsen for normalfordelingen, hvorfra den simulerede prøve genereres. Vi kan se formen på vores simulerede datasæt, sim_norm, samtidigt med sandsynlighedsdiagrammet.
Lav et sandsynlighedsdiagram sim_norm. Falder alle punkterne på linjen? Hvordan sammenlignes dette plot med sandsynlighedsplottet for de reelle data?

Ud over at vi kan sammenligne det originale plot med plottet, der er genereret ud fra en normal fordeling, så kan vi sammenligne det originale plot med mange forskellige diagrammer ved hjælp af følgende funktion. Det kan være nyttigt at klikke på zoomknappen i plotvinduet.

```
qqnormsim(fdims$hgt)
```

Ser sandsynlighedsplottet for fdims $ hgt ud som de plot, der er udarbejdet ud fra de simulerede data? Synes kvindens højder at være normalfordelt?
Vurder om vægten for kvinder er normalfordelt.

Normalfordelte sandsynligheder

Nu har vi flere værktøjer til at vurdere, om en variabel er normalfordelt eller ej.
Det viser sig, at statistikere ved meget om normalfordelinger. Når vi først har besluttet, at en tilfældig variabel er tilnærmelsesvis normalfordelt, så kan vi besvare alle mulige spørgsmål om den pågældende variabel. Tag for eksempel spørgsmålet om, "Hvad er sandsynligheden for, at en tilfældigt valgt ung voksen kvinde er højere end 182 cm?" Hvis vi antager, at kvindelige højder er normal fordelte (en meget tæt tilnærmelse er også okay), kan vi finde denne sandsynlighed ved at beregne en Z-score og anvende den standardiserede normalfordeling. I R gøres dette i et trin med funktionen pnorm.

```
1 - pnorm(q = 182, mean = fhgtmean, sd = fhgtsd)
```

Bemærk, at funktionen pnorm giver området under normalfordelingskurven under en given værdi, q, med en given middelværdi og standardafvigelse. Da vi er interesseret i sandsynligheden for, at nogen er højere end 182 cm, er vi nødt til at finde 1 - sandsynligheden for den givne værdi.
Vi har nu fundet den teoretiske sandsynlighed, givet vi har en normalfordeling med de givne værdier. Hvis vi vil beregne sandsynligheden empirisk, er vi blot nødt til at bestemme, hvor mange observationer, der falder over 182 og derefter dele dette antal med den samlede prøvestørrelse.

```
sum(fdims$hgt > 182) / length(fdims$hgt)
```

Selvom sandsynlighederne ikke er nøjagtigt de samme, så ligger de rimeligt tæt på hinanden. Jo tættere fordelingen er på at være normalfordelt, des mere præcise vil de teoretiske sandsynligheder være.
Skriv to spørgsmål: et spørgsmål om kvinders højder og et spørgsmål vedrørende kvinders vægt. Beregn disse sandsynligheder ved hjælp af både den teoretiske normale fordeling og den empiriske fordeling (i alt fire sandsynligheder). Hvilken variabel, højde eller vægt, var tættest på de normalfordelte sandsynligheder?

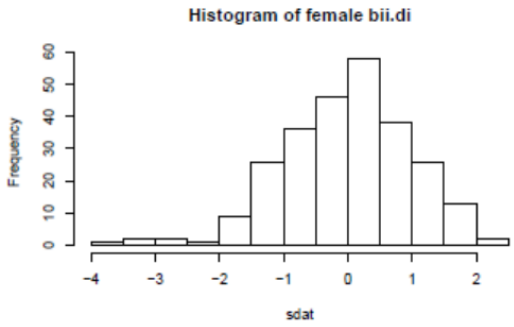
Opgaver:

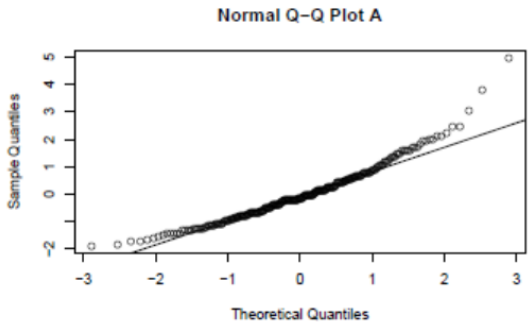
Lad os nu anvende nogle af de andre variable i datasættet for kropsdimensioner. Brug histogrammer, og sammenlign med normalfordelinger. Alle variable er blevet standardiseret (træk først gennemsnittet, divider derefter med standardafvigelsen). Plot nu variablene i R for at kontrollere dette.

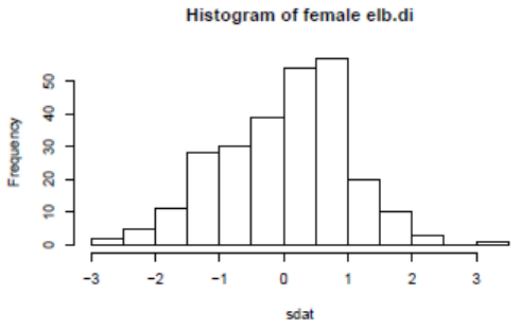
a. Histogrammet for kvindelig bækken diameter (bii.di) hører til normalfordelingsdiagrammet med bogstav _____.
b. Histogrammet for kvindelig albue diameter (elb.di) hører til normalfordelingsdiagrammet med bogstav __.
c. Histogrammet for generel alder (alder) hører til normalfordelingsdiagrammet med bogstav ____.
d. Histogrammet for kvindelig brystdybde (che.de) hører til normalfordelingsdiagrammet med bogstav ____.

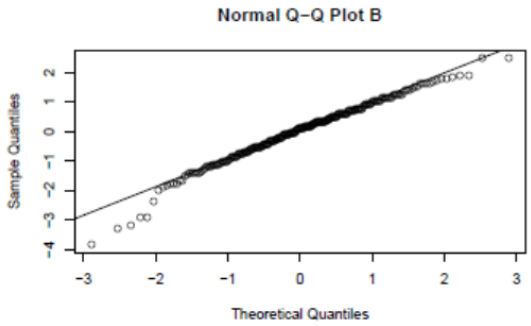
Bemærk, at normalfordelingsdiagrammerne C og D har et let trinvist mønster.
Hvorfor tror du, at dette er tilfældet?

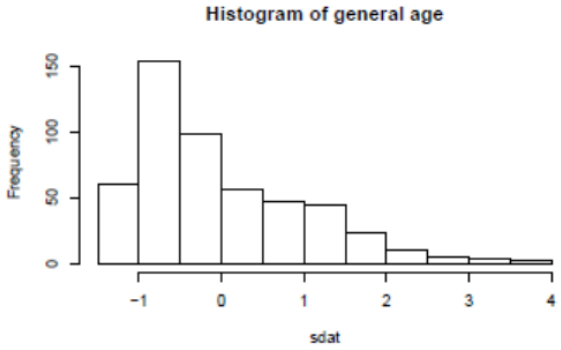
Som det kan ses, så kan sandsynlighedsdiagrammer bruges både til at vurdere normalitet og til at vurdere skævheder. Lav et tæthedsfordelingsdiagram for kvindelig knædiameter (kne.di). Baseret på, er denne variabel venstreskæv, symmetrisk eller højre skæv? Brug et histogram til at bekræfte dine resultater.

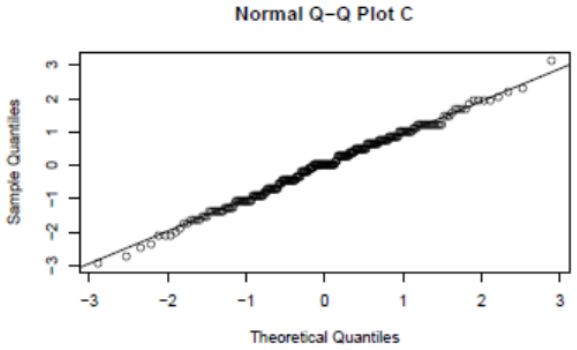

Histogram of female bii.di
Frequency
sdat

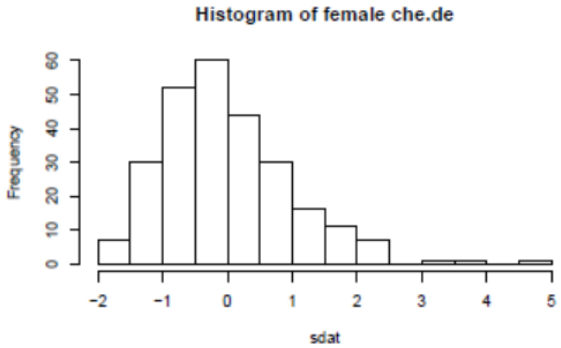

Normal Q–Q Plot A
Sample Quantiles
Theoretical Quantiles

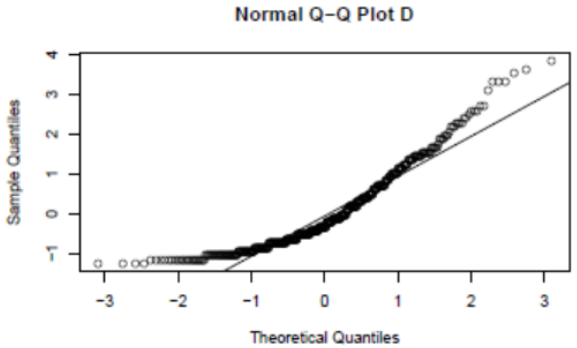

Histogram of female elb.di
Frequency
sdat


Normal Q–Q Plot B
Sample Quantiles
Theoretical Quantiles


Histogram of general age
Frequency
sdat


Normal Q–Q Plot C
Sample Quantiles
Theoretical Quantiles


Histogram of female che.de
Frequency
sdat


Normal Q–Q Plot D
Sample Quantiles
Theoretical Quantiles