



Normalfordelingen

Anders Tolver
Institut for Matematiske Fag



Dagens program

- Hvorfor normalfordelingen, og hvad skal den bruges til?
- Egenskaber ved normalfordelingen og beregning af sandsynligheder
- Er data normalfordelt?
- Summer og skalering af normalfordelte variable

Afsnit 4.2 (en stikprøve) og afsnit 4.4 (den centrale grænseværdisætning): først på onsdag

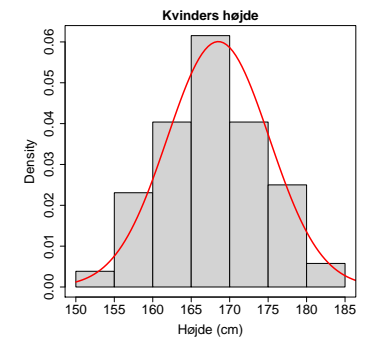


Hvorfor normalfordelingen og hvad skal vi bruge den til?



Kvinders højde fra spørgeskema, 2017

- Selvrapporteret højde (i cm) for 104 kvinder: y_1, \dots, y_{104} .
- Histogram normeret så det samlede areal af rektanglerne er 1.
- $\bar{y} = 168.52$, $s = 6.64$
- Graf for tæthed for normalfordeling

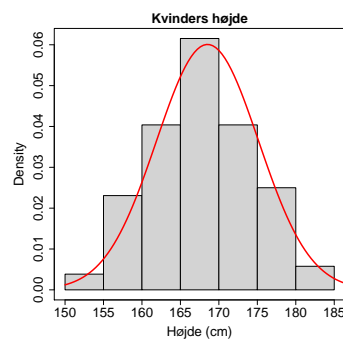


$$f(y) = \frac{1}{\sqrt{2\pi} \cdot 6.64} \exp\left(-\frac{1}{2 \cdot 6.64^2} (y - 168.52)^2\right)$$

Grafen for f er en god approksimation af histogrammet.



Den klokkeformede kurve



- Kurven er **tætheden** (density) for en normalfordeling
- **Kurven ligner histogrammet**. Vi kan bruge normalfordelingen til at beskrive fordelingen af højden



Hvorfor lige normalfordelingen?

Normalfordelingen er vigtig af flere grunde:

- Passer ofte godt til data, evt. efter transformation
- Pæne matematiske egenskaber — giver pæne resultater for estimater, statistiske test mm.
- Centrale grænseværdisætning (CLT): gennemsnit af næsten hvad som helst er approksimativt normalfordelt!

Kaldes også den Gaussiske fordeling.

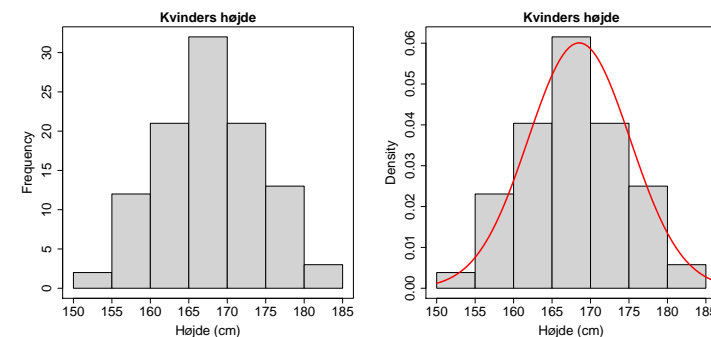
Opkaldt efter Carl Friedrich Gauss, tysk matematiker og fysiker, 1777–1855.



10 DM



Histogram og relative hyppigheder



I standardiseret histogram er det samlede areal af rektangler lig 1. Så er **relativ hyppighed lig areal af tilhørende rektangler**, fx:

$$\frac{\text{antal højder i interval } [155\text{cm}, 160\text{cm}]}{104} = \frac{12}{104} \approx 0.115 = 11.5\%$$



Tætheder og sandsynligheder

Tilsvarende for tætheden: **Sandsynligheden for at en obs. falder i intervallet fra a til b er lig arealet under kurven**, fx

$$P(155 < Y \leq 160) = \int_{155}^{160} f(y) dy = 0.079 = 7.9\%$$

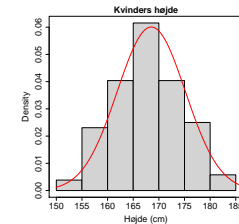
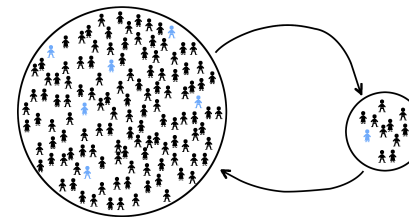
De to sandsynligheder er ikke ens. **Population vs stikprøve.**

- Hvis populationsværdier er fordelt som tætheden beskriver, så vil histogram for stikprøve fra populationen ligne tætheden.
- Normalfordelingstæthed som **model** for histogrammet.

Viser senere hvordan vi faktisk fik beregnet integralet til 7.9%.



Populationer, tæthed vs stikprøve, histogram

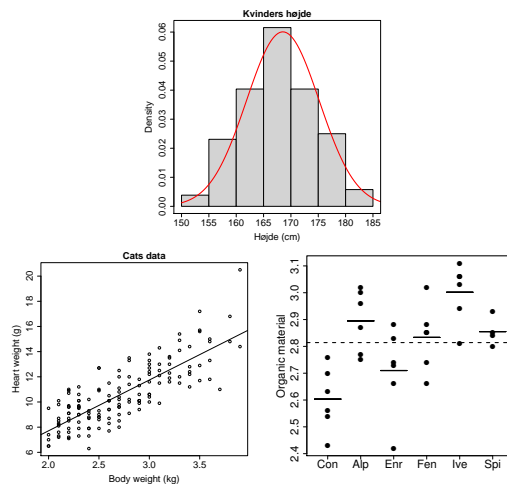


- Population: Normalfordelingstæthed
- Stikprøve: Histogram



Hvad skal vi bruge normalfordelingen til?

Til at beskrive variationen i data når reponsen er kontinuert: En stikprøve, lineær regression, ensidet variansanalyse, ...



Egenskaber ved normalfordelingen og beregning af sandsynligheder



Den generelle normalfordeling

$$f(y) = \frac{1}{\sqrt{2\pi \cdot 6.64^2}} \exp\left(-\frac{1}{2 \cdot 6.64^2}(y - 168.52)^2\right)$$

Udskift tallet 168.52 med μ og tallet 6.64 med σ :

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$$

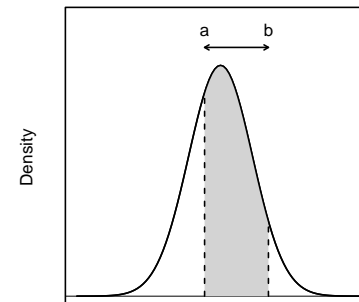
- Siger, at en variabel Y er normalfordelt med middelværdi μ og spredning σ hvis det for alle intervaller $[a, b]$ gælder at

$$P(a < Y \leq b) = \int_a^b f(y) dy.$$

- Vi skriver $Y \sim N(\mu, \sigma^2)$



Tæthed og sandsynligheder



$Y \sim N(\mu, \sigma^2)$ hvis ssh. for at Y lander mellem a og b er lig areal fra a til b under tætheden:

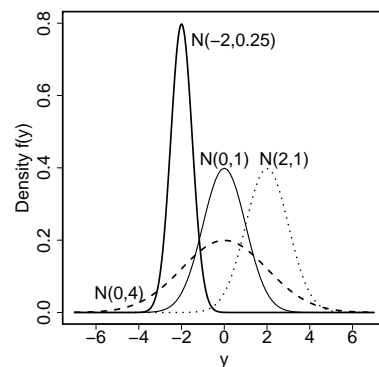
$$P(a < Y \leq b) = \int_a^b f(y) dy$$

- $f(y_1) > f(y_2)$: mere sandsynligt at havne omkring y_1 end y_2 .
- $P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b)$



Symmetri — centrum — spredning

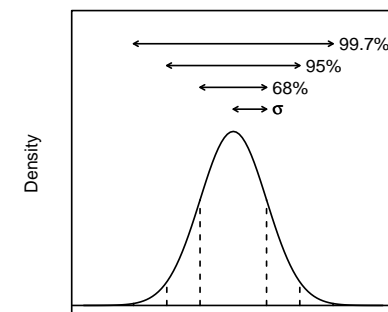
Tæthed for $N(\mu, \sigma^2)$: $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$



Bemærk: Vi skriver $N(\mu, \sigma^2)$ — ikke $N(\mu, \sigma)$. Hvis $Y \sim N(0, 4)$ har Y altså spredning 2.



Sandsynligheder for $\mu \pm k \cdot \sigma$



- 68% mest centrale obs. ligger i intervallet $\mu \pm \sigma$
- 95% mest centrale obs. ligger i intervallet $\mu \pm 2 \cdot \sigma$
- 99.7% mest centrale obs. ligger i intervallet $\mu \pm 3 \cdot \sigma$

Gælder for **alle** normalfordelinger — uanset værdierne af μ og σ .



Beregning af sandsynligheder i normalfordelingen

Som arealer under tæthedsfunktionen, dvs. ved integration, fx.

$$P(155 < Y \leq 160) = \int_{155}^{160} f(y) dy$$

Problem (teoretisk): Man kan ikke finde noget mere eksplicit udtryk end ovenstående.

Hvad så?

- Via omskrivninger til $N(0, 1)$. Sådan står det i bogen.
- Nemmere: Brug funktionen **pnorm** i R med angivelse af mean og sd. Beregner sandsynligheder $P(Y \leq b)$.



Beregning af sandsynligheder i normalfordelingen

Antag at Y er normalfordelt med middelværdi 168.52 og spredning 6.64, altså $Y \sim N(168.52, 6.64^2)$.

Hvad er $P(155 < Y \leq 160)$?

```
> pnorm(160, mean=168.52, sd=6.64)
[1] 0.09972282
> pnorm(155, mean=168.52, sd=6.64)
[1] 0.02086792
> pnorm(160, mean=168.52, sd=6.64) - pnorm(155, mean=168.52, sd=6.64)
[1] 0.0788549
```

Altså:

- $P(Y \leq 160) = 0.0997$ og $P(Y \leq 155) = 0.0209$
- $P(155 < Y \leq 160) = 0.0997 - 0.0209 = 0.0789$



Fraktiler

Hvilken højde opfylder at 90% af kvinder i populationen er lavere end denne højde?

Altså: Antag $Y \sim N(168.52, 6.64^2)$, og find b så

$$P(Y < b) = P(Y \leq b) = 0.90$$

```
> qnorm(0.90, mean=168.52, sd=6.64)
[1] 177.0295
```

Tallet 177.03 kaldes **90% fraktilen** i $N(168.52, 6.64^2)$.



Beregning af sandsynligheder og fraktiler i N

Beregning af sandsynligheder og fraktiler i R

- Givet b , bestem sandsynlighed $P(Y \leq b)$: Brug **pnorm**
- Givet ssh. p , bestem b så $P(Y \leq b) = p$: Brug **qnorm**

I begge tilfælde skal både middelværdi og spredning også angives.



Standardnormalfordelingen

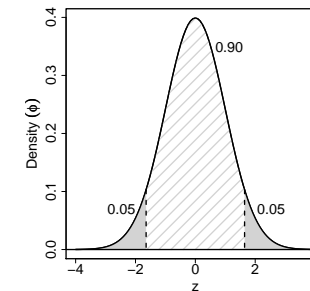
En N -fordelt variabel kan standardiseres:

$$Y \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{Y - \mu}{\sigma} \sim N(0, 1)$$

Vi kalder $N(0, 1)$ for **standardnormalfordelingen**: $\mu = 0$, $\sigma = 1$.



Standardnormalfordelingen



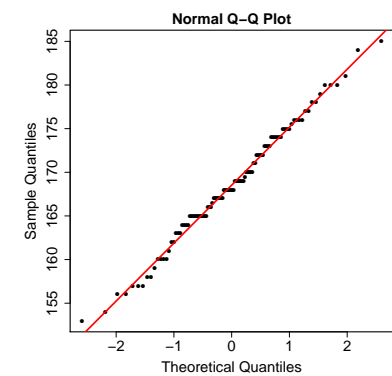
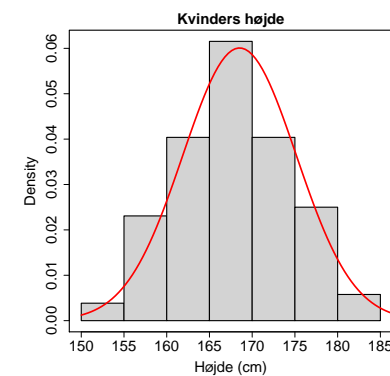
- **95%-fraktilen** er 1.6449: $P(Z \leq 1.6449) = 0.95$... og dermed er $P(-1.6449 < Z < 1.6449) = 0.9$
- **97.5%-fraktilen** er 1.960: $P(Z \leq 1.960) = 0.975$... og dermed er $P(-1.96 < Z < 1.96) = 0.95$



Er data normalfordelt?



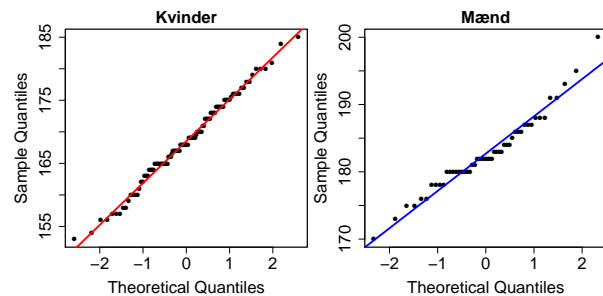
Hvordan checkes om data er normalfordelt?



- Histogram sammen med tæthed for $N(\bar{y}, s^2)$. Kun for n stor.
- **QQ-plot**: Ligger punkterne omkring en ret linie?



QQ-plot



- **Quantile-quantile** (fraktil-fraktil) plot
- x-aksen tilpasset så normalfordelte data ligger på ret linie
- Sammenlign med **ret linie med skæring \bar{y} og hældning s**
- R: QQ-plot med `qqnorm`, linie med `abline`



Vurdering af QQ-plot

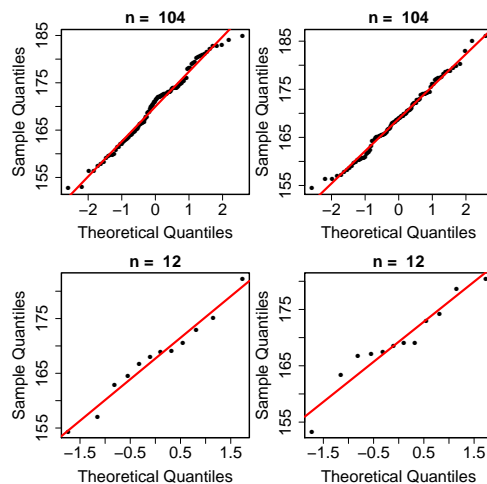
Hvor store skal afvigelserne fra en ret linie være for at man kan konkludere at data **ikke** er normalfordelte?

- Afhænger af antal observationer
- Kan være nyttigt at se på simulerede N -data: Hvordan ser QQ-plots ud når vi **ved** at data er N -fordelte.



QQ-plots fra simulerede data

Her **er** data normalfordelte:



**Summer og skalering af
normalfordelte
variable mm**



Sum af uafhængige normalfordelte variable

Infobox 4.2(a): Hvis Y_1 og Y_2 er **uafhængige**, $Y_1 \sim N(\mu_1, \sigma_1^2)$ og $Y_2 \sim N(\mu_2, \sigma_2^2)$, så er **summen**

$$Y_1 + Y_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

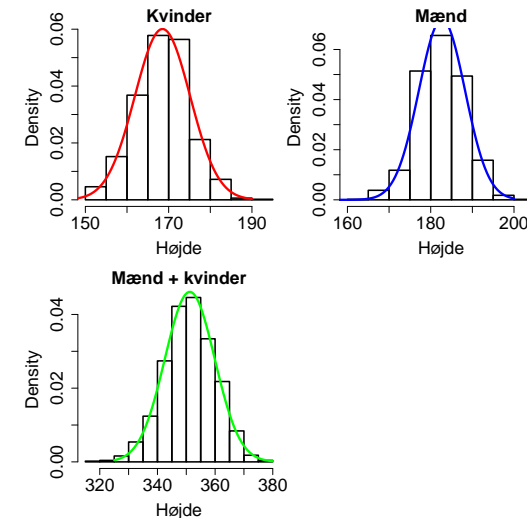
Specielt gælder: $\text{sd}(Y_1 + Y_2) = \sqrt{\sigma_1^2 + \sigma_2^2}$.

Eksempel:

- Antag at kvinders højde er $N(168.52, 6.64^2)$ fordelt
- Antag at mænds højde er $N(182.70, 5.54^2)$ fordelt
- Vælg mand og kvinde tilfældigt fra populationerne. Deres **samlede højde** er $N(351.22, 74.78)$. Spredning 8.65.



Simulationer



Skalering og flytning af normalfordelt variabel

Infobox 4.2(b) Hvis $Y \sim N(\mu, \sigma^2)$ og a og b er kendte tal, så er

$$a + b \cdot Y \sim N(a + b \cdot \mu, b^2 \cdot \sigma^2)$$

Specielt gælder: $\text{sd}(a + b \cdot Y) = |b| \cdot \text{sd}(Y)$

(Idiotisk) Eksempel:

- Antag at kvinders højde **i cm** er $N(168.52, 6.64^2)$ fordelt
- Vælg kvinde tilfældigt fra populationen, og stil hende på en kasse der er 1 m høj. Samlet højde af kvinde og kasse, i meter:

$$Z = \frac{Y}{100} + 1 \sim N(2.6852, 0.0664^2)$$



Gennemsnit af normalfordelte variable

Infobox 4.3 Hvis Y_1, \dots, Y_n er uafhængige og alle $Y_i \sim N(\mu, \sigma^2)$, så er gennemsnittet \bar{Y} også normalfordelt:

$$\bar{Y} = \frac{1}{n}(Y_1 + \dots + Y_n) \sim N(\mu, \sigma^2/n)$$

Specielt gælder: $\text{sd}(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$.

Eksempel:

- Antag at kvinders højde i cm er $N(168.52, 6.64^2)$ fordelt
- Vælg 25 kvinder tilfældigt fra populationen og se på deres gennemsnitshøjde. Gennemsnitshøjden er $N(168.52, 1.33^2)$



Nyttige R-kommandoer

```
pnorm(160, mean=168.52, sd=6.64) ## Beregn sandsynlighed
qnorm(0.90, mean=168.52, sd=6.64) ## Beregn fraktil

qqnorm(hojde)                ## Lav QQ-plot
abline(168.52, 6.64)         ## Indlæg linie

hist(hojde, prob=TRUE)        ## Histogram
f1 <- function(y) dnorm(y, 168.52, 6.64) ## Tætheden smom funktion
plot(f1, 145, 190, add=TRUE)  ## Indtegn tæthed
```

Husk: Beregn gerne sandsynligheder og fraktiler som ovenfor i stedet for at regne tilbage til $N(0, 1)$.



Opsummering — til eget brug

- Hvad vil det sige at Y er normalfordelt?
- Hvor mange procent af en normalfordeling ligger i intervallet "middelværdi ± 2 gange spredning"?
- Hvordan beregner man sandsynligheder i normalford. i R?
- Hvordan checker man om data kommer fra en normalfordeling?
- Hvad er fordelingen af $X + Y$ hvis både X og Y er normalfordelte?
- Hvad er fordelingen af gennemsnittet af ens fordelte normalfordelte variable?

