# Kan et datasæt beskrives med normalfordelingen

Vi vil her se på hvordan vi vurderer om et datasæt kan beskrives vha. normalfordelingen. I praksis viser det sig at datasæt ofte er stikprøver af uafhængige stokastiske variable som er tilnærmelsesvis normalfordelte. Det teoretiske argument for dette hedder den centrale grænseværdisætning og ligger desværre uden for vores pensum. For at komme med et bud på fordelingen af de stokastiske variable, er vi først nødt til at estimere deres middelværdi, $μ$, og spredning, $σ$, ud fra datasættet, $x\_{1},x\_{2}…,x\_{n}$:

$$\hat{μ}=\frac{\sum\_{i=1}^{n}x\_{i}}{n}$$

$$\hat{σ}=\sqrt{\frac{\sum\_{i=1}^{n}\left(x\_{i}-\hat{μ}\right)^{2}}{n-1}}$$

Estimatet af middelværdien, $\hat{μ}$, er gennemsnittet af datasættet og estimatet af spredningen, $\hat{σ}$, kaldes *den empiriske spredning* af datasættet.

### Opgave 1 (på klassen)

Vi har datasættet $[1,2,3,4,5]$.

1. Estimér middelværdien og spredningen.

I Maple er kommandoerne heldigvis nemme:



I praksis skal vi bruge et stort datasæt for at kunne beskrive det vha. normalfordelingen og disse kommer ofte i form af et Exceldokument. Så [her er en video](https://eggym.sharepoint.com/%3Av%3A/r/sites/Section_5778/Delte%20dokumenter/General/Maple%2C%20GeoGebra%20og%20Excel/Import%20af%20data%20fra%20Excel%20til%20Maple%20%28video%29.mp4?csf=1&web=1&e=UumBjR) der genopfrisker hvordan man importerer et datasæt til Maple.

### Opgave 2

1. Importér datasættet Normaldata1 og estimér middelværdien og spredningen af den normalfordeling som vi vil beskrive datasættet med.

For at vurdere om normalfordelingen giver en god beskrivelse af datasættet, starter vi med at tegne tæthedsfunktionen sammen med et histogram af datasættet. **Princippet er forklaret** [**i videoen her**](https://eggym.sharepoint.com/%3Av%3A/r/sites/Section_5778/Delte%20dokumenter/General/Normalfordelingen/Kan%20et%20datas%C3%A6t%20beskrives%20med%20normalfordelingen.mp4?csf=1&web=1&e=wbeX1R) **og koden er i Mapledokumentet ’Kan et datasæt beskrives med normalfordelingen’**. 10-tallet i kommandoen *grupperData* nedenfor, er ikke middelværdien, men antallet af intervaller som man typisk sætter til 10.



1. Prøv selv at få figuren ovenfor tegnet i Maple.

Her giver normalfordelingen en okay beskrivelse af datasættet, så nu kan vi bruge den til at beskrive datasættet. Hvis vi yderligere forestiller os at datasættet er stikprøver af uafhængige stokastiske variable med samme fordeling, så har vi at normalfordelingen giver et okay estimat af denne fordeling.

1. Estimér vha. normalfordelingen hvor stor en del af datasættet som ligger inden for intervallet $[4,16]$.
2. Estimér vha. normalfordelingen sandsynligheden for at få en ny stikprøve som er større end eller lig med 10.
3. Estimer vha. normalfordelingen intervallet for de normale udfald.

### Opgave 3

Hvordan skal et datasæt så se ud for at vi kan vurdere at det ikke kan beskrives vha. normalfordelingen? Her skal vi til sidst se et eksempel på det.

1. Importér datasættet Normaldata2 og gør rede for at datasættet ikke kan beskrives vha. en normalfordeling.

### Opgave 4 (valgfri)

For at få en idé om hvor stort et datasæt vi skal have for det kan beskrives med normalfordelingen, skal I selv prøve at simulere et datasæt ud fra normalfordelingen vha. Maplekoden nedenfor.
I eksemplet har jeg valgt en middelværdi på 10, en spredning på 3 og 100 stikprøver.



1. Prøv at simulere nogle datasæt og undersøg hvor godt normalfordelingen fungerer.
2. Undersøg hvor mange stikprøver der skal til før I med en vis sikkerhed vil kunne vurdere om stikprøverne kan stamme fra stokastiske variable som er normalfordelt.