

Histogram, boksplott, normalfordelingsplott og spredningsplott

TMA4240/TMA4245 Statistikk

Håkon Tjelmeland

Institutt for matematiske fag

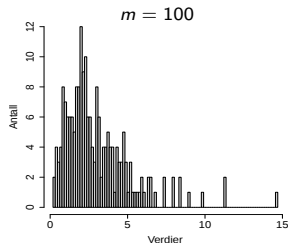
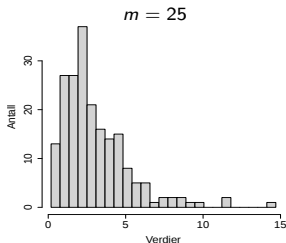
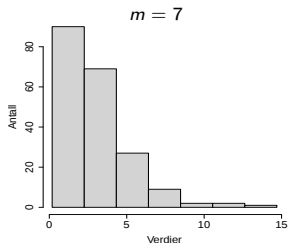
Norges teknisk-naturvitenskapelige universitet

Vanlige plott for å visualisere observerte verdier

- ★ Skal se på fire typer plott som er mye brukt for å visualisere et observert datasett
 - histogram
 - boksplott
 - normalfordelingsplott
 - spredningsplott

Histogram

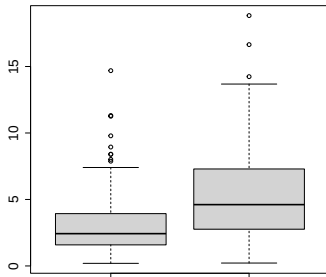
- ★ Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$
- ★ Vi konstruerer et histogram ved
 - velg et intervall $[x_{\min}, x_{\max}]$ som inneholder alle $x_1, x_2, \dots, x_n$
 - velg antall underintervaller m
 - del intervallet $[x_{\min}, x_{\max}]$ inn i m like lange underintervaller
 - tell opp hvor mange av x_i -ene som er i hvert underintervall
 - for hvert underintervall lag et rektangel med høyde lik antall x_i -er i underintervallet
- ★ Eksempel på histogram for $n = 200$ observerte verdier



- ★ Merk:
 - hvis man velger m for stor blir histogrammet «rotete» og dermed lite informativt
 - hvis man velger m for liten blir også histogrammet mindre informativt
 - programvare har gjerne en «default» algoritme for å velge intervallet $[x_{\min}, x_{\max}]$ og verdien til m

Boksplott

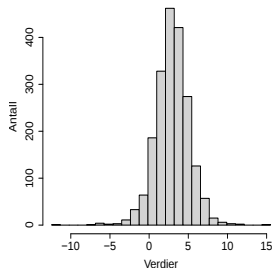
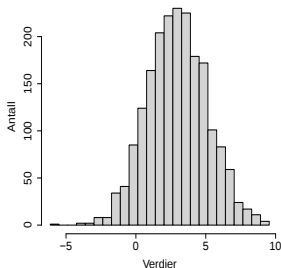
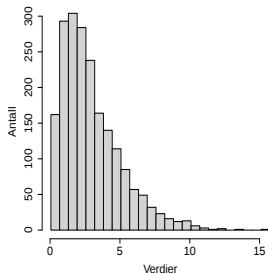
- ★ Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$
- ★ Et boksplott tar utgangspunkt i 0.25-, 0.5- og 0.75-kvantilene.
- ★ Eksempel på boksplott for to datasett:



- ★ Boksen strekker seg fra 0.25-kvantilen til 0.75-kvantilen
- ★ Tversgående linje inne i boksen er medianen (0.5-kvantilen)
- ★ Hver av de to utstikkerne går til de mest ekstreme verdiene som ligger nærmere boksen enn w ganger kvartildifferansen (mest vanlig med $w = 1.5$)
- ★ Observasjoner mer ekstreme enn utstikkerne er separat markert

Normalfordelingsplott

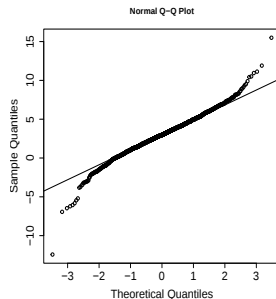
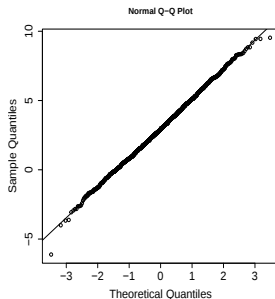
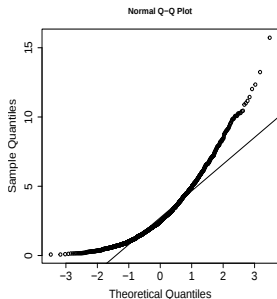
- ★ Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$
- ★ Hvordan vurdere om observasjonene ser ut til å være normalfordelte?
- ★ Ofte vanskelig å vurdere dette ved å se på histogram



- ★ Vi (mennesker) er flinke til å gjenkjenne rette linjer
- ★ Normalfordelingsplott plotter teoretiske normal-kvantiler mot empirisk kvantiler
 - en rett linje indikerer normalfordeling
 - verdier over (under) linja i venstre hale: lettere (tyngre) venstre hale enn normalfordeling
 - verdier over (under) linja i høyre hale: tyngre (lettere) høyre hale enn normalfordeling

Normalfordelingsplott

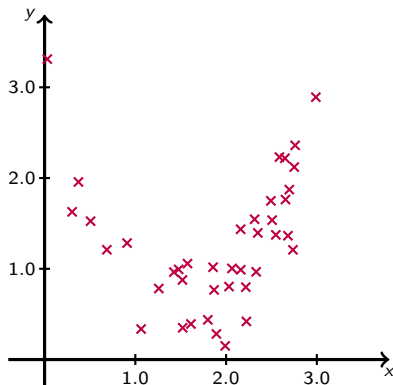
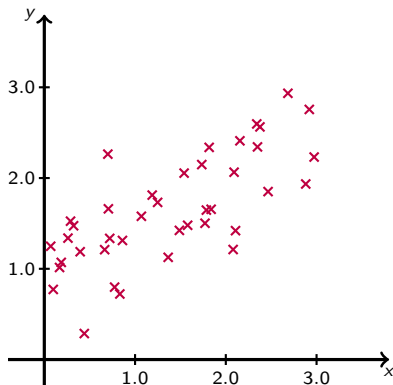
- ★ Anta at vi har observert et utvalg bestående av n observasjoner eller målinger, som vi betegner med $x_1, x_2, \dots, x_n$
- ★ Hvordan vurdere om observasjonene ser ut til å være normalfordelte?
- ★ Ofte vanskelig å vurdere dette ved å se på histogram



- ★ Vi (mennesker) er flinke til å gjenkjenne rette linjer
- ★ Normalfordelingsplott plotter teoretiske normal-kvantiler mot empirisk kvantiler
 - en rett linje indikerer normalfordeling
 - verdier over (under) linja i venstre hale: lettere (tyngre) venstre hale enn normalfordeling
 - verdier over (under) linja i høyre hale: tyngre (lettere) høyre hale enn normalfordeling

Spredningsplott

- ★ Anta at vi har observert n observasjonspar $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$
- ★ I et spredningsplott plotter man x_i -verdiene langs x -aksen og y_i -verdiene langs y -aksen
- ★ Eksempel på spredningsplott for to datasett:



Oppsummering

- ★ Har diskutert fire typer plott for å visualisere observerte verdier fra et utvalg
 - histogram
 - boksplott
 - normalfordelingsplott
 - spredningsplott