

Procentiler og kvartilsæt

Procentiler (også kaldet *fraktiler*) angiver en talværdi hvor en give procentdel af datasættets elementer har en dataværdi der er mindre eller lig med procentilværdien.

For at illustrere brugen af procentiler betragtes igen datasættet for højden af 10 personer hvor det sorterede datasæt blev beregnet til:

$$X_s = (1.68, 1.70, 1.72, 1.74, 1.75, 1.76, 1.81, 1.82, 1.84, 1.86)$$

Ud fra datasættet skal det bestemmes hvilken højde en døråbning skal have således at 80% af personerne kan gå gennem døren uden at skulle bukke sig. Idet 80% af 10 er 8 kan man således vælge en værdi mellem det 8. og 9. element i det *sorterede* datasæt (f.eks.. 1.83) som en 80-procentilværdi og herved være sikker på at mindst 80% af personerne kan gå igennem en døråbning af højden 1.83 uden at skulle bukke sig.

Den mest anvendte procentil er 50%-procentilen der også kaldes *medianen* der repræsenterer den midterste dataværdi i datasættet og opdeler således et datasæt i to lige store dele.

Datasættet kan også opdeles i fire lige store dele ved de procentiler, der kaldes *kvartiler*. Idet datasættet opdeles i fire dele består *kvartilsættet* af fem værdier $Q = (Q_0, Q_1, Q_2, Q_3, Q_4)$ der er defineret ved:

- Q_0 kaldes *0. kvartil* og er defineret som den mindste dataværdi som forekommer i datasættet.
- Q_1 kaldes *1. kvartil* og er defineret som 25%-procentilen.
- Q_2 kaldes *2. kvartil* og er defineret som 50%-procentilen som ligeledes er medianen af datasættet.
- Q_3 kaldes *3. kvartil* og er defineret som 75%-procentilen.
- Q_4 kaldes *4. kvartil* og er defineret som den største dataværdi som forekommer i datasættet.

I tillæg til kvartilsættet kan det ligeledes være relevant at nævne datasættets *variationsbredde*, som er differencen mellem datasættets maksimum og minimumværdi (dvs. $Q_4 - Q_0$), samt datasættets *typeværdi*, som er værdien af den hyppigst forekommende dataværdi.

Tages f.eks. udgangspunkt i følgende sorterede datasæt:

$$X_s = (1, 2, 2, 3, 3, 4, 4, 4, 5, 5)$$

0. kvartil og 4. kvartil er hhv. datasættets mindste og største dataværdi kan aflæses direkte af det sorterede datasæt til:

$$Q_0 = \min(X_s) = 1 \quad \text{og} \quad Q_4 = \max(X_s) = 5$$

Medianen kan bestemmes ud fra at datasættet består af 10 dataværdier hvilket betyder at det kan deles i to lige store dele af 5 dataværdier hver.

$$X_s = (\underbrace{1, 2, 2, 3, 3}_5, \underbrace{4, 4, 4, 5, 5}_5)$$

Talværdien, der deler datasættet i to lige store dele, ligger altså mellem de to dataværdier 3 og 4. Hvis dette er tilfældet anvendes gennemsnittet af værdierne:

$$Q_2 = \frac{3 + 4}{2} = 3.5$$

For de to halvdele af datasættet, som begge består af 5 dataværdier, kan man nu bestemme den dataværdi der opdeler halvdelene i to lige store dele:

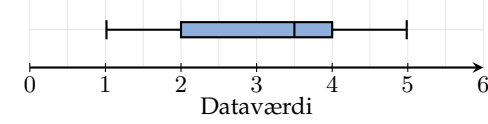
$$(\underbrace{1, 2}_2, \underbrace{2, 3}_2) \quad \text{og} \quad (\underbrace{4, 4}_2, \underbrace{4, 5}_2)$$

Som det kan ses er de to midterværdier, som hhv. udgør 1. og 3. kvartil:

$$Q_1 = 2 \quad \text{og} \quad Q_3 = 4$$

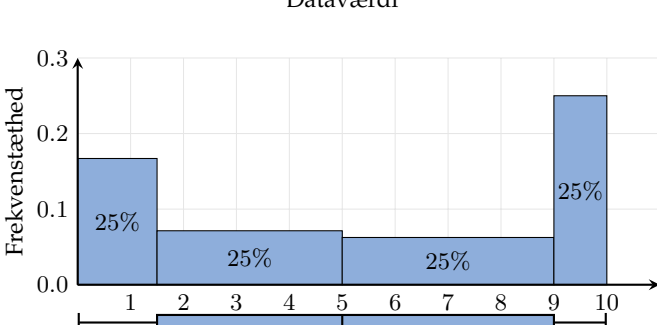
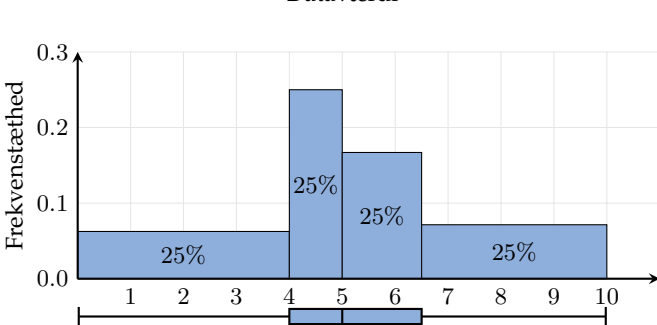
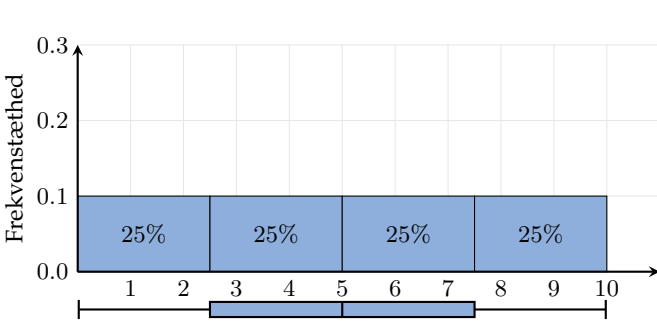
Kvartilsættet er således bestemt til $Q = (1, 2, 3.5, 4, 5)$.

Det fundne kvartilsæt kan nu afbildes ift. en talakse med et såkaldt *boksdia-*
gram som vist nedenfor:

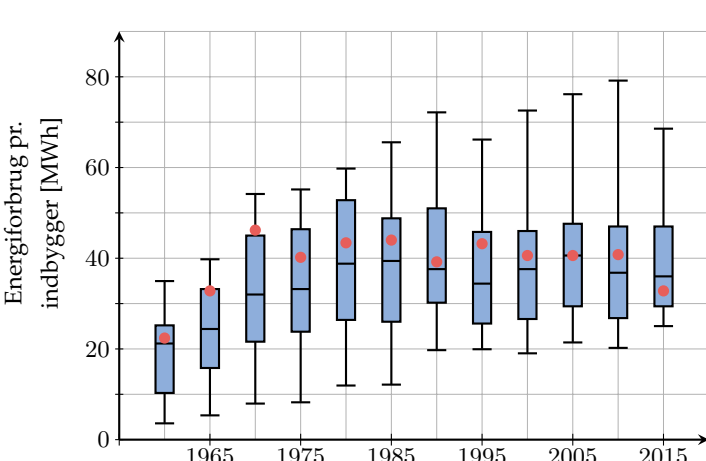


Boksdiaagrammet viser de fem værdier fra kvartilsættet hvor strengen i den midterste boks angiver medianen, boksens ender angiver hhv. 1. og 3. kvartil, og de to pinde ud fra boksen strækker sig til hhv. 0. og 4. kvartil.

For at forstå hvordan man skal aflæse og fortolke et boksdiaagram kan man evt. betragte det som en grupperet fordeling med 4 intervaller, men hvor intervallængderne er tilpasset således at frekvensen af intervallerne er 25%. Dette vil sige at små kvartilintervaller i boksdiaagrammet vil være sammenfaldende med en høj frekvenstæthed i fordelingen. Dette skyldes at hvert kvartilinterval indeholder 25% af dataværdierne som, når fordelt på et lille interval, vil resultere i en høj frekvenstæthed. Af samme årsag ses en tilsvarende sammenhæng mellem store kvartilintervaller i boksdiaagrammet og lave frekvenstætheder i fordelingen. Dette er illustreret i figurene nedenfor der viser tre forskellige datasæt hvor boksdiaagrammet og den tilsvarende grupperede fordeling er sammenlignet.



Idet et boksdiaagram kun har brug for en talakse er det nemt at anvende i større diagrammer hvor man har brug for at sammenligne flere datasæt. Til eksempel er datasættet for 28 EU landendes energiforbrug pr. indbygger vist for hvert 5. år mellem 1960 og 2015 med et boksdiaagram i figuren nedenfor. Herudover er det specifikke energiforbrug for Danmark afmærket med røde punkter. Boksdiaagrammerne, der hver repræsenterer 28 landes energiforbrug, gør det således mere overskueligt at analysere den samlede udvikling af energiforbruget i EU, der ligeledes kan bruges som et sammenligningsgrundlag for hvor Danmarks energiforbrug ligger ift. fordelingen.



Geogebra: Kvartilsæt og BoxPlot(...)

Først indtastes data i Geogebra's regneark (spreadsheet).

Hvis man har selve datasættet indtastet i felterne A1 til A100 kan man beregne talværdierne for kvartilsættet ved følgende kommandoer i Algebra-vinduet:

$$\begin{aligned} Q_0 &= \text{Min}(A1 : A100) \\ Q_1 &= \text{Quartile1}(A1 : A100) \\ Q_2 &= \text{Median}(A1 : A100) \\ Q_3 &= \text{Quartile3}(A1 : A100) \\ Q_4 &= \text{Max}(A1 : A100) \end{aligned}$$

Hvis man har selve datasættet indtastet i felterne A1 til A100 kan man plotte datasættets boksdiaagram ved følgende kommando:

$$\text{BoxPlot}(1, 0.2, A1 : A100)$$

Her er talværdien 1 positionen ift. 2.aksen og talværdien 0.2 er tykkelsen af selve boksdiaagrammet.