

Grupperede observationer

Fra bogen plus c

I nogle situationer skal vi forsøge at danne os et overblik over et stort talmateriale med mange forskellige observationer. Det kan fx være tilfældet, hvis observationerne er afstande eller højder. Vi kan så med fordel samle observationerne i grupper. Vi siger, at observationssættet er *grupperet*.

Nedenstående eksempel gennemgår, hvilke redskaber vi kan bruge til at få overblik over et grupperet observationssæt.

Tabel med intervaller

Et gymnasium har bedt alle sine 400 elever om at svare på, hvor langt de har til skole. Svarene gives på et spørgeskema, hvor eleverne sætter kryds ud for det interval, som indeholder længden af deres skolevej. Målingerne bliver derfor grupperet i intervaller – i dette tilfælde med en intervalbredde på 2 km.

Intervallerne og de tilhørende hyppigheder ses i nedenstående tabel. Frekvensen og den kumulerede frekvens er også udregnet.



iStockphoto.com/Sean Locke



iStockphoto.com/Martin Vegh

afstand (km)	hyppighed	frekvens	kumuleret frekvens
]0;2]	96	24%	24%
]2;4]	152	38%	62%
]4;6]	34	8,5%	70,5%
]6;8]	16	4%	74,5%
]8;10]	68	17%	91,5%
]10;12]	24	6%	97,5%
]12;14]	10	2,5%	100%
total	400	100%	

Tegnene] og [kaldes *intervalparenteser*:

- Hvis parentesernes spidser peger ind mod et tal, er tallet med i intervallet.
- Hvis spidserne peger væk fra tallet, er tallet ikke med i intervallet.

Det betyder fx, at betegnelsen]2;4] angiver intervallet fra, men uden 2 til og med 4. Læg mærke til, at der er uendeligt mange tal i dette interval. Fx ligger tallene 2,01, 3,65 og 4 i intervallet.

Når vi grupperer observationer, vendes intervalparenteserne altid således, at det første tal ikke er med i intervallet, og det sidste tal er med i intervallet (se tabellen). Derfor hører den kumulerede frekvens sammen med et intervals højre endepunkt. I intervallet]2;4] hører den kumulerede frekvens på 62% altså sammen med 4. Den angiver, at der er 62%, der har en skolevej, som er 4 km eller kortere.

Brug tabellen til at finde ud, af hvor stor en procentdel af eleverne, der har højst 6 km til skole.

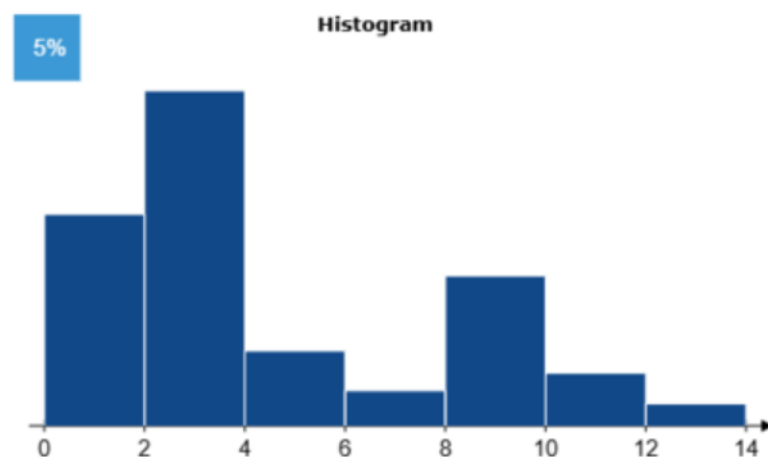
Hvor stor en procentdel af eleverne har over 8 km til skole?

Histogram



Vi kan præsentere observationssættet grafisk ved at bruge frekvenserne og tegne et *histogram*. Et histogram er et søjlediagram, der er specielt på den måde, at det er søjlernes areal, der angiver frekvensen og ikke søjlernes højde. Før vi kan tegne histogrammet, skal vi derfor vælge en arealenhed. Dette gøres typisk ved, at man i histogrammets hjørne viser et lille areal, som man bestemmer skal svare til fx 5% eller 10%. Derefter tegnes søjlerne i de enkelte intervaller, så arealet af dem passer med den frekvens, der hører til intervallet.

Histogrammet for observationssættet fra tabellen ses herunder:



Middelværdi



Ved ugrupperede observationer kunne vi udregne middelværdien eller gennemsnittet af observationerne ved at lægge alle observationerne sammen og dividere med antallet.

For et grupperet observationssæt kan vi ikke umiddelbart bruge samme fremgangsmåde. Vi kender nemlig ikke den enkelte observations præcise værdi, men kun hvilket interval den befinder sig i. Normalt antager man, at observationerne er jævnt fordelt i de enkelte intervaller. Det betyder, at gennemsnittet af observationerne i et bestemt interval er i dette intervals midtpunkt.

Altså fx at

- de 96 elever, der har højst 2 kilometer til skole, i gennemsnit har en skolevej på 1 km
- de 152 elever der har mellem 2 og 4 kilometer til skole i gennemsnit har en skolevej på 3 km
- osv.

Middelværdien kan så udregnes således:

$$\frac{96 \cdot 1 + 152 \cdot 3 + 34 \cdot 5 + 16 \cdot 7 + 68 \cdot 9 + 24 \cdot 11 + 10 \cdot 13}{400} = 4,6$$

Eleverne har således i gennemsnit 4,6 kilometer til skole.

Definition 1: Middelværdi for grupperede observationer



Vi beregner middelværdien for et grupperet observationssæt ved at gøre følgende:

1. Gange hvert intervalmidtpunkt med hyppigheden i intervallet.
2. Lægge resultaterne for de enkelte intervaller sammen.
3. Dividere denne sum med det samlede antal observationer.

Sumkurve og fraktiler



En anden måde at få overblik over observationerne er ved at bruge de kumulerede frekvenser. Dette gøres ved at tegne en såkaldt *sumkurve* i et koordinatsystem.

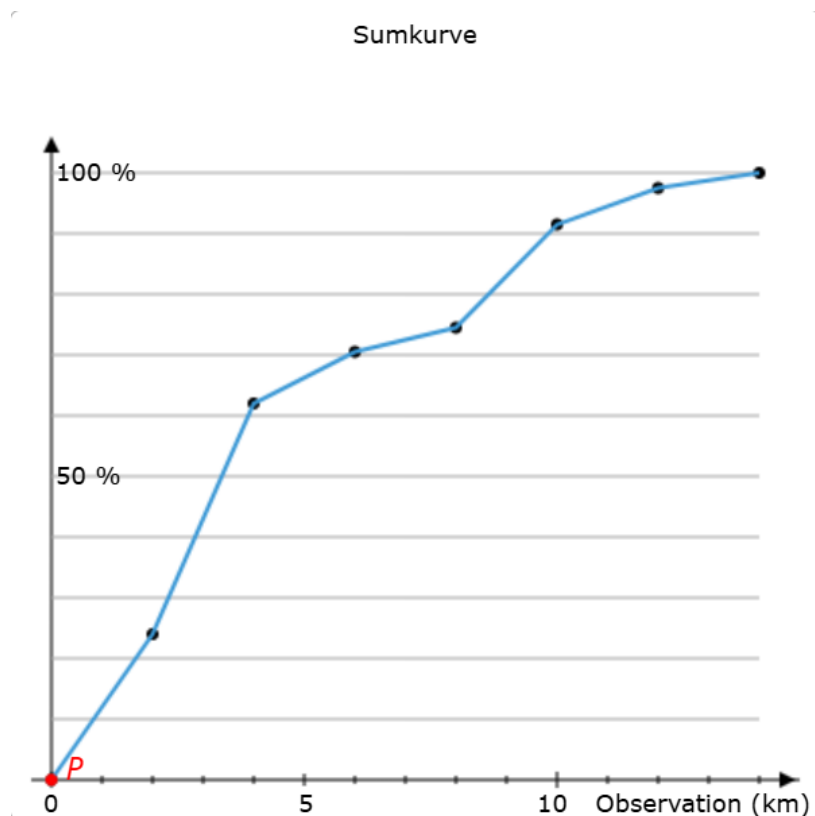
Vi afsætter punkter, der har højre interval-endepunkt som x -værdi og den kumulerede frekvens i intervallet som y -værdi. Selve kurven fremkommer ved, at vi herefter forbinder de afsatte punkter med rette linjestykker.

Med observationerne fra [tabellen ovenfor](#) skal vi derfor afsætte punkterne $(0;0)$, $(2;24)$, $(4;62)$, $(6;70,5)$, $(8;74,5)$, $(10;91,5)$, $(12;97,5)$ og $(14;100)$.

Punktet $(0;0)$ angiver, at der ikke er nogen observationer under 0 km. Punktet $(14;100)$ angiver, at alle observationerne (100%) er på 14 km eller mindre.

Vi antager, at observationerne inde i et interval er jævnt fordelt i intervallet. Derfor forbindes de indtegnede punkter med rette linjestykker.

Sumkurven for observationerne i eksempel 1 ses nedenfor:

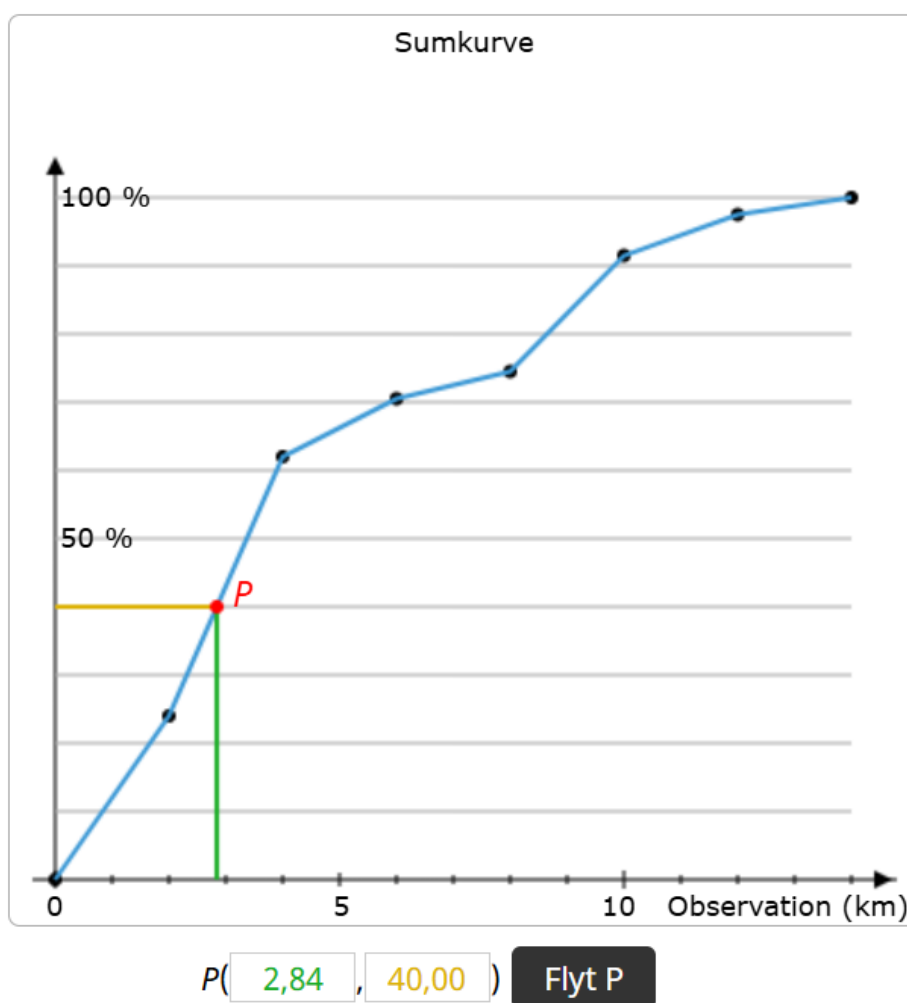


Bemærk, at

- hvis sumkurven er stejl i et interval, så er der mange observationer i dette interval
- hvis sumkurven er flad i et interval, så er der få observationer i dette interval.

Et punkt (x,y) på sumkurven angiver, hvor stor en procentdel, y , af observationerne der ligger under værdien x . Vi kan på sumkurven fx aflæse punktet $(2,84; 40)$, som betyder, at 40% af eleverne har højst 2,84 km til skole.

Tallet 2,84 kaldes her for *40%-fraktilen*. Fraktil betyder brøkdel, og 40%-fraktilen angiver således den grænse, hvorunder 40% af observationerne befinder sig.



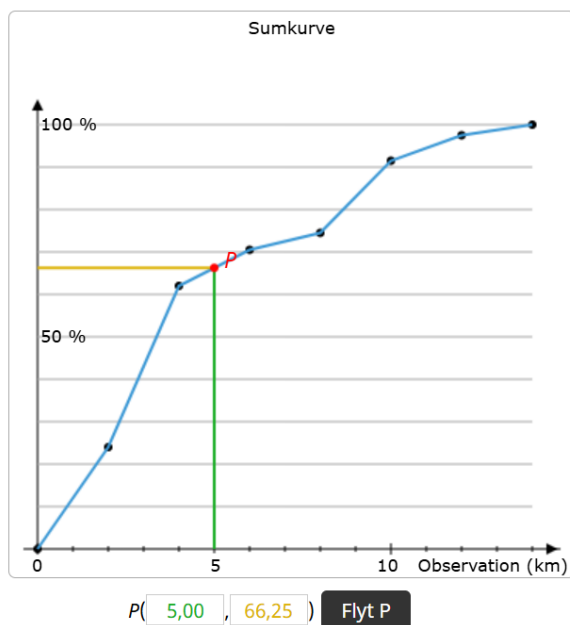
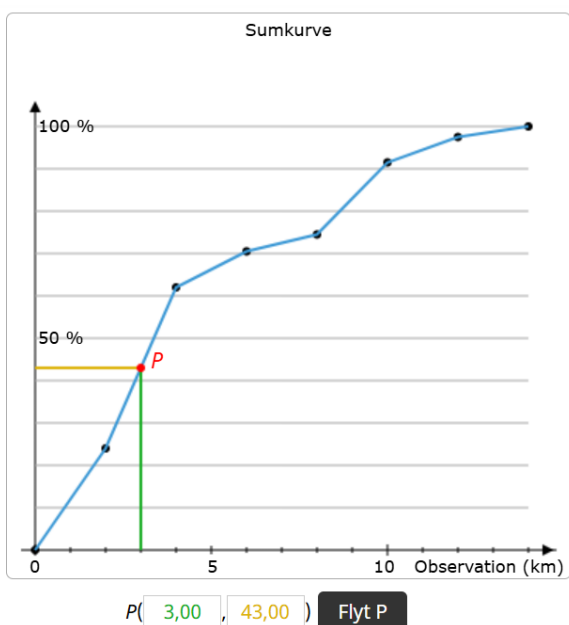
Vi kan også anvende sumkurven til at finde ud af, hvor mange procent af eleverne der har mellem 3 og 5 kilometer til skole.

Vi flytter det røde punkt på kurven, indtil vi er ved 3 på x-aksen, og her aflæser vi y-værdien til 43%.

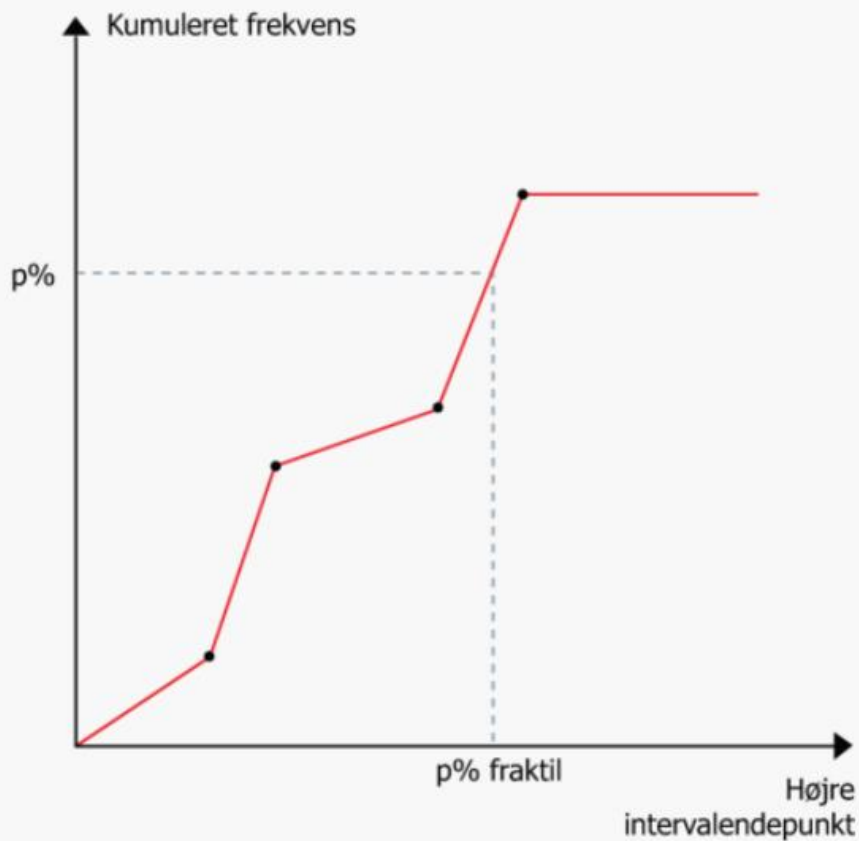
Derpå flytter vi punktet på kurven hen til 5 på x-aksen, og vi aflæser y-værdien til 66,25%.

Vi ved nu, at der er 43% af eleverne, der har højst 3 km til skole, og 66,25% der har højst 5 km til skole.

Derfor må der være $66,25\% - 43\% = 23,25\%$ af eleverne, der har mellem 3 og 5 kilometer til skole.



Definition 2: Sumkurve og fraktil



En sumkurve for et grupperet observationssæt er en graf, som tegnes ved at afsætte punkter af typen

(højre intervalendepunkt , kumuleret frekvens)

i et koordinatsystem og forbinde de afsatte punkter med rette linjestykker.

Sumkurven angiver, hvor stor en procentdel af observationerne der er mindre end en given værdi.

En p%-fraktil angiver den grænse, hvorunder p% af observationerne ligger.

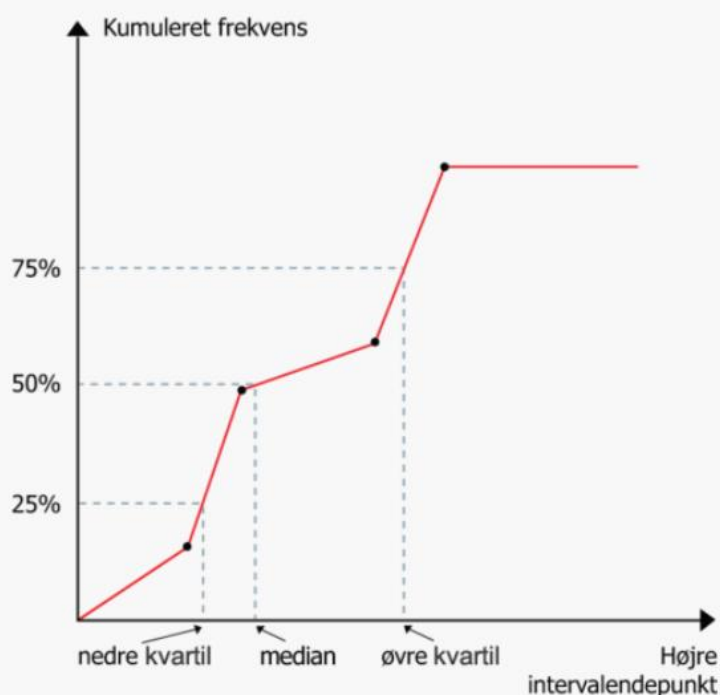
Kvartilsæt

⋮

I afsnittet om ugrupperede observationer har vi allerede arbejdet med 3 specielle fraktiler, som vi kaldte nedre kvartil (Q_1), median (m) og øvre kvartil (Q_3) og tilsammen kvartilsættet. Vi vil også bruge kvartilsættet som deskriptorer i forbindelse med grupperede observationer, idet vi bemærker, at nedre kvartil er 25%-fraktilen, medianen er 50%-fraktilen og øvre kvartil er 75%-fraktilen.

I eksemplet med skolevejen kan vi på [sumkurven ovenfor](#) aflæse kvartilsættet til (2,05 ; 3,37 ; 8,06), hvilket fortæller os, at 25% af eleverne har en skolevej på højst 2,05 km, halvdelen af eleverne har en skolevej på højst 3,37 km, og 75% af eleverne har en skolevej på højst 8,06 km.

Definition 3: Kvartilsæt for grupperede observationer



Kvartilsættet for et grupperet observationssæt udgøres af:

- *nedre kvartil* (Q_1), som er 25%-fraktilen
- *medianen* (m), som er 50%-fraktilen
- *øvre kvartil* (Q_3), som er 75%-fraktilen.

Kvartilbredden er forskellen mellem øvre kvartil og nedre kvartil ($Q_3 - Q_1$).