Grupperede fordelinger

For datasæt, der består af målinger, vil typisk være kendetegnet ved at hyppigheden for hver enkelt dataværdi er meget lille. Disse typer af datasæt skal derfor typisk repræsenteres grafisk ved såkaldte *grupperede fordelinger* hvor dataværdierne grupperes i intervaller.

Betragtes f.eks. et datasæt, der er fremkommet ved at måle højden (i cm) af 12 personer, kunne det se således ud:

$$X = \big(176, 172, 170, 184, 186, 175, 181, 177, 168, 179, 192, 171\big)$$

Det sorterede datasæt bliver således:

$$X_s = \big(168, 170, 171, 172, 175, 176, 177, 179, 181, 184, 186, 192\big)$$

Hvis datasættets fordelingen udregnes over samtlige dataværdier i dette talsæt vil metoden ikke have nogen virkning, idet det ses at ingen af dataværdierne er repræsenteret mere end én gang i datasættet. Dette vil betyde at alle værdierne skulle optræde i tabellen og have en hyppighed på 1, hvilket essentielt svarer til at liste datasættet op og der vil derfor ikke være sket nogen statistisk reduktion i informationen.

For at undgå denne situation hvor hyppigheden bliver én for alle dataværdier kan der istedet beregnes en *grupperet fordeling* hvor datasættet grupperes i intervaller af en bestemt længde. For det givne datasæt kunne en passende inddeling for at opnå 5 grupper være at anvende en intervallængde på 5 cm for de første fire grupper og en intervallængde på 10 cm for den sidste. Hyppigheden beregnes nu ved at tælle op hvor mange dataværdier der falder i de enkelte grupper og frekvensen beregnes herefter ved at dividere hyppigheden med antallet af dataværdier.

Datasættets statistiske fordeling kan nu opsummeres i en tabel:

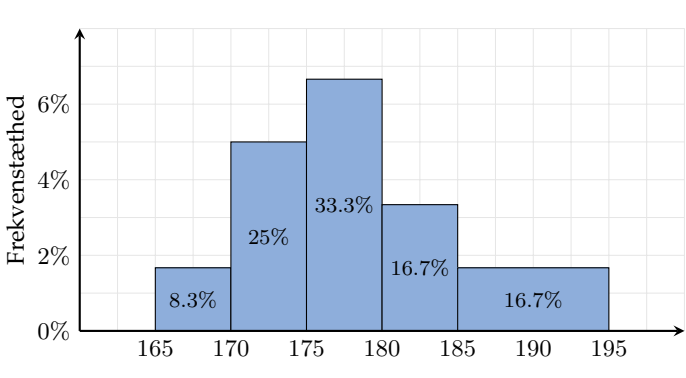
Værdi-interval	Interval-længde	Hyppighed	Frekvens	Frekvens-tæthed
$[165, 170[$	5	1	8.3%	1.67%
$[170, 175[$	5	3	25.0%	5.00%
$[175, 180[$	5	4	33.3%	6.66%
$[180, 185[$	5	2	16.7%	3.34%
$[185, 195[$	10	2	16.7%	1.67%
Sum		$n = 12$	100%	

I tabellen ses det at ud over frekvensen er der også beregnet en *frekvenstæthed* t ved at dividere frekvensen f med intervallængden l.

$$t_1 = \frac{f_1}{l_1} \approx \frac{8.3\%}{5} = 1.67\%$$
$$t_2 = \frac{f_2}{l_2} \approx \frac{25.0\%}{5} = 5.00\%$$
$$\vdots$$

Det bemærkes i tabellen ovenfor, at de to sidste grupper har samme hyppigheder og frekvenser men forskellige frekvenstætheder idet intervallængderne er forskellige.

Tabellen kan nu omsættes til et *histogram* hvor hver interval plottes med en konstant frekvenstætheden som en kasse hvor arealet svarer til frekvensen:



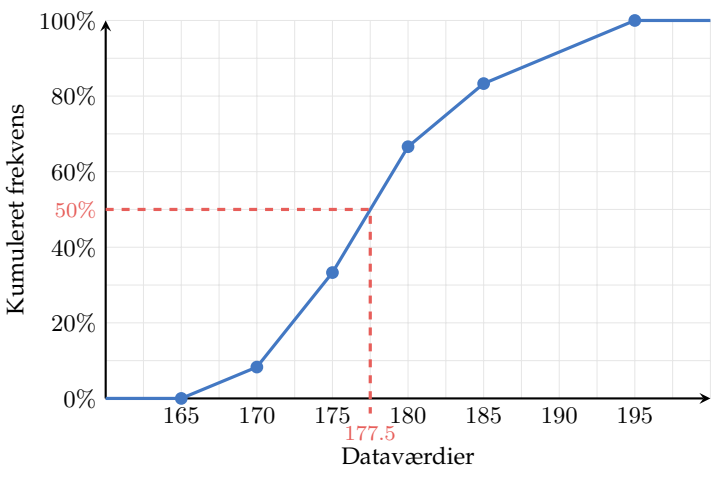
Af histogrammet fremgår det, at størstedelen af datasættet er fordelt i intervallerne $[1.70, 1.75[$ og $[1.75, 1.80[$.

Det bemærkes fra tabellen, at selve informationen om de specifikke dataværdier er forsvundet i grupperingen, og det er derfor ikke muligt at genskabe datasættet ud fra tabellen. Med andre ord antages frekvenstætheden at være konstant over de respektive intervaller, hvilket kan skabe et diagram der ikke er helt retvisende, såfremt der anvendes et for stort intervallængde i grupperingen.

Idet frekvenstætheden i den grupperede fordeling antages at være konstant over de respektive intervaller betyder det effektivt, at den kumulerede frekvens må variere som linjestykker i intervallerne. Det er derfor hensigtsmæssigt for grupperede fordelinger også at afbillede den kumulerede frekvens ved brug af en såkaldt *sumkurve*. For at plotte sumkurven beregnes først den kummulerede frekvens for alle dataværdierne for intervalgrænserne:

Dataværdier	165	170	175	180	185	195
Kum. frekvens	0.0	8.3%	33.3%	66.6%	83.3%	100%

Sumkurven kan nu plottes ved at forbinde punkterne fra tabellen med linjestykker.



Sumkurven kan anvendes til at bestemme hvor stor en del af datasættet, der forventes at ligge under en given værdi. I det ovenstående eksempel kan således aflæses ved den stiplede røde linje, at 50% af datasættet må forventes at ligge under en dataværdi på ca. 177.5 cm.

Geogebra: `Histogram(...)`
Først indtastes data i Geogebras regneark (spreadsheet).
Hvis man har grænseværdierne for histogrammets intervaller indtastet i felterne A1 til A11 og selve datasættet indtastet i felterne B1 til B100 kan man plotte datasættets histogram for frekvenstætheden ved følgende kommando:

`Histogram(A1 : A11, B1 : B100, true, 1/100)`

Her er talværdien $\frac{1}{100}$ den skaleringsfaktoren, der for et datasæt på 100 talværdier omregner hyppigheden til frekvensen.
Hvis man i stedet har en fordelingstabel med grænseværdierne for intervallerne i felterne A1 til A11 og frekvenstætheden i felterne B2 til B11 anvendes kommandoen:

`Histogram(A1 : A11, B2 : B11)`