

2. Numeriske variable – beskrivelse af et datamateriale

2.1 Grafisk præsentation af et numerisk datamateriale – prikdiagram

	køn	kondital	alder	højde	træning
1	pige	27,1	14	151	nej
2	pige	47,4	15	169	ja
3	pige	38,3	16	164,5	ja
4	pige	39,1	16	173	nej
5	pige	27,1	16	169	ja
6	pige	49,2	17	170,5	ja
7	pige	38,3	16	168	ja
8	pige	27,1	16	172,5	ja
9	pige	34,6	16	165	nej
10	pige	36,7	16	162	nej
11	pige	34,0	16	163	ja
12	pige	29,9	15	174,5	nej
13	pige	28,0	16	166	ja
14	dreng	58,7	16	186	ja
15	dreng	37,5	16	185	nej
16	dreng	37,5	16	185	nej
17	dreng	54,2	16	170	ja
18	dreng	47,9	16	175	ja
19	dreng	51,4	16	174	ja
20	dreng	35,5	17	173	nej
21	dreng	47,9	17	177	nej
22	dreng	61,7	16	190	ja
23	dreng	56,0	16	172	ja
24	dreng	35,5	16	176	nej
25	dreng	37,5	16	186	nej

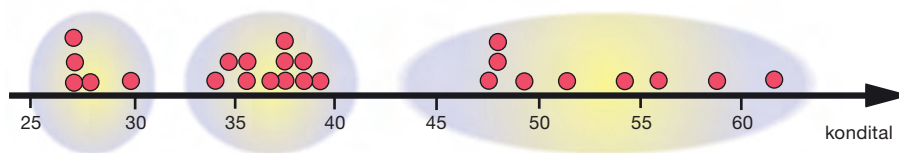
Når vi indsamler data ud fra observationer fx indhenter svar på spørgeskemaer eller lignende, skelner vi mellem *kategoriske* og *numeriske* variable. Vi vil i det følgende beskrive, hvorledes man kan skabe sig overblik over og præsentere et datamateriale ved at gennemgå et større eksempel. I en 1.g-klasse med 25 elever, der har idræt og matematik i et samarbejdsprojekt om sundhed, har man indsamlet en lang række data om elevernes sundhedstilstand. Tabellen, der gengives her, er kun et udsnit.

Vi kan se, at tabellen indeholder den kategoriske variabel køn (pige eller dreng), og den kategoriske variabel træning (går du regelmæssigt til træning? Ja eller nej). Endvidere indeholder tabellen tre numeriske variable, kondital, alder og højde. Her vil vi fokusere på den numeriske variabel kondital.

Det kan være svært at få et indtryk af konditalenes fordeling ud fra tabellen. Derfor foretages først en sortering efter størrelse i stigende rækkefølge fra det mindste kondital til det største. Det resulterer i rækkefølgen nedenfor.

27.1 27.1 28.0 29.9 34.0 34.6 35.5 35.5 36.7 37.5 37.5
37.5 38.3 38.3 39.1 47.4 47.9 47.9 49.2 51.4 54.2 56.0 58.7 61.7

Det mindste kondital (*minimum*) er altså 27.1, og det største kondital (*maksimum*) er 61.7. For at få et visuelt indtryk af fordelingen præsenterer vi dernæst datasættet grafisk. Da der er tale om en numerisk variabel med talværdier, kan vi afsætte observationerne som prikker langs en talakse. Derved fremkommer *prikdiagrammet* for fordelingen af kondital:



Vi ser da, at observationerne falder i tre grupper: En lille gruppe mellem 25 og 30, en større gruppe mellem 34 og 40 og endelig en langstrakt gruppe fra 47 og helt ud til den maksimale værdi på 61.7.

2.2 Sproglig præsentation af niveauet for et datasæt – median og middeltal

Prikdiagrammet rummer al den information, der er i datasættet. Det er mere overskueligt end tabellen, men det er ikke så velegnet, hvis vi ønsker at give en kort sproglig beskrivelse af, hvordan det står til med klassens kondital.

En sådan beskrivelse skal udtrykke noget karakteristisk om, hvilket *niveau* klassens kondital er på. Hvis et enkelt tal skal beskrive niveauet for klassens kondital, så bør denne værdi afbalancere datasættet, dvs. i en eller anden forstand bør der ligge lige så meget på den ene side af denne værdi som på den anden side. Observationerne skal altså ligge symmetrisk omkring den værdi, der i ét tal kan udtrykke klassens kondital.

Der findes to forskellige karakteristiske tal, der hver på sin måde lever op til dette krav.

Definition: Median og middeltal

Ved *medianen* for et datasæt forstår vi den midterste observation. Hvis der er et lige antal observationer udregnes medianen som gennemsnittet af de to midterste observationer.

Ved *middeltallet* eller *gennemsnittet* for et datasæt forstår vi det tal m , som ville give samme samlede sum, hvis alle konditallene blev erstattet af denne værdi m .

Eksempel: Median og middeltal for et datasæt

Ved at opstille konditallene i rækkefølge ser vi, at den midterste observation er 37.5, idet der er 12 observationer under medianen og 12 observationer over medianen.

27.1 27.1 27.1 28.0 29.9 34.0 34.6 35.5 35.5 36.7 37.5 37.5
 37.5
 38.3 38.3 39.1 47.4 47.9 47.9 49.2 51.4 54.2 56.0 58.7 61.7

Konklusion: Medianen = den midterste observation = 37.5.

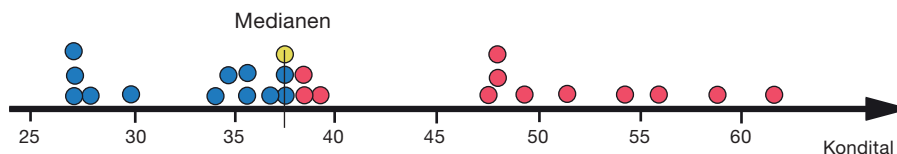
Middeltallet m skal opfylde, at:

$$25 \cdot m = 27.1 + \dots + 61.7$$

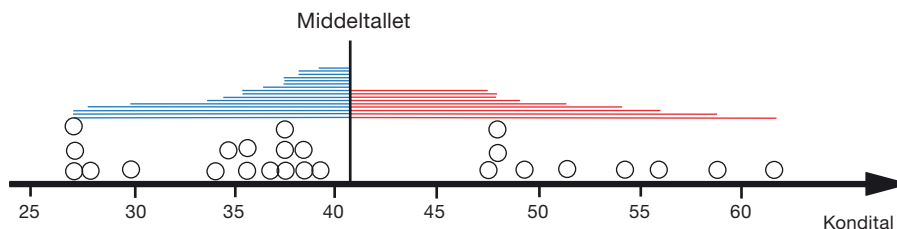
$$m = \frac{27.1 + \dots + 61.7}{25} = 40.724$$

Konklusion: Middeltallet = 40.724.

Anvender vi prikdiagrammet, får vi følgende grafiske fremstilling:



Medianen (gul) splitter datasættet i to lige store delsat med 12 observationer (blå) til venstre for medianen og 12 observationer (røde) til højre for medianen.



Middeltallet splitter datasættet i to ulige store delsat med 16 observationer (blå) til venstre for middeltallet og 9 observationer (røde) til højre for middeltallet. Men de 9 observationer ligger længere væk fra middeltallet, så tilsammen vægter de lige så meget som de 16 observationer, der ligger til venstre. Vi siger at fordelingen er højreskæv med en lang hale af observationer til højre.

Det er disse forholdsvis store kondital, der trækker gennemsnittet i vejret i forhold til medianen.

Eksempel: Middeltallet som balancepunkt for afstandene (især for A-niveau)

Den følgende udregning kan gennemføres helt tilsvarende for ethvert datasæt.

Ud fra definitionen er middeltallet m bestemt som det tal, der opfylder:

$$25 \cdot m = 27,1 + 27,1 + \dots + 58,7 + 61,7$$

$$m + m + \dots + m = 27,1 + 27,1 + \dots + 58,7 + 61,7$$

Vi flytter nu alle tal, der er mindre end m , over på venstre side og parrer dem hver for sig sammen med et m . Resten af m 'erne flyttes over på højre side:

$$(m - 27,1) + (m - 27,1) + \dots + (m - 39,1) = (47,4 - m) + (47,9 - m) + \dots + (61,7 - m)$$

Alle parenteser repræsenterer nu afstanden fra en observation hen til middeltallet. Ligningen siger, at summen af afstandene fra de observationer, der ligger til venstre for middeltallet, er lig med summen af afstandene fra de observationer, der ligger til højre for middeltallet.

Da denne udregning kunne gennemføres for ethvert datasæt, er konklusionen generel:
Middeltallet er balancepunktet for afstandene til observationerne.

Øvelse 2.4

- a) Konstruer et datasæt med 4 elementer, hvor minimum er 2, medianen er 4 og maksimum er 7.
Hvad bliver middeltallet for dette datasæt?
- b) Konstruer et datasæt med 4 elementer, hvor minimum er 2, middeltallet er 4 og maksimum er 7.
Hvad bliver medianen for dette datasæt?

Øvelse 2.5

- a) Bestem middeltallet for det følgende datasæt: $\{1, 1, 2, 5, 6\}$.
- b) Bestem afstandene mellem de enkelte data og middeltallet.
- c) Kontroller, at middeltallet er balancepunktet for afstandene.

Øvelse 2.6

Nogle værktøjsprogrammer tillader, at man trækker i et datapunkt i prikdiagrammet. Andre tillader, at man indfører en skyder for en af værdierne i datasættet.

- a) Fremstil et prikdiagram for konditallet, hvor den maksimale observation kan flyttes (dynamisk).
- b) Tilføj medianen og middeltallet til prikdiagrammet.
1. Træk den maksimale observation tættere på resten af observationerne, og beskriv, hvad der sker med medianen, henholdsvis middeltallet.
 2. Træk nu den maksimale observation i modsat retning, og beskriv igen, hvad der sker med medianen, henholdsvis middeltallet.
- Hvori består forskellen på de to situationer?

Praxis: Om brugen af median eller middeltal til at beskrive niveauet for et datasæt

Hvis observationerne er sammenlignelige eller ligger nogenlunde symmetrisk med en stor klump af data i midten, vil vi normalt *foretrække middeltallet*. Det kan eksempelvis være tilfældet med målinger af faldtider, temperaturer eller lignende i et laboratorium.

Hvis observationerne derimod ligger skævt, vil vi normalt *foretrække medianen*. Det kan eksempelvis være tilfældet, når der er tale om en indkomstfordeling med nogle få meget rige personer og mange fattige personer.

Medianen er mindre påvirkelig over for fejlmålinger og atypiske data end middeltallet. En enkelt værdi, der ligger langt udenfor de andre, vil påvirke middeltallet meget, men ikke påvirke medianen. Vi siger derfor, at medianen er *robust*.

Øvelse 2.7

Bestem median og middeltallet for følgende to datasæt:

a) {1,2,3,4,5,6,7,8,9}

b) {1,2,3,4,5,6,7,100,1000}

Kommenter resultatet.

Øvelse 2.8

a) Konstruer et datasæt med 10 tal, hvor det er fornuftigt at bruge middeltallet som mål for niveauet.

b) Konstruer et datasæt med 10 tal, hvor det ikke er fornuftigt at bruge middeltallet som mål for niveauet, og hvor medianen beskriver niveauet bedre.

Øvelse 2.9

En klasse skal i forbindelse med et projekt have målt deres hvilepuls. Der er 6 drenge i klassen. Drengenes hvilepuls fremgår af følgende datasæt: {65,62,73,58,31,69}.

a) Kommenter dette datasæt, og foreslå, hvordan du ville arbejde videre med rapporten.

b) Hvilken indflydelse har dit forslag på medianen? Og på middeltallet?

Eksempel: Fattigdomsbegrebet

Man skelner mellem forskellige slags fattigdomsbegreber. Her vil vi især se på *absolut fattigdom* og *relativ fattigdom*.

Når man vil definere absolut fattigdom, går man ud fra en minimumsstandard for indkomst, under hvilken man ikke kan dække de mest fundamentale eksistensbehov. I FN opererer man fx med grænsen '*en dollar om dagen*' for, hvornår man er fattig i et uland. Når man vil definere relativ fattigdom, ser man i stedet på, hvor indkomsten ligger i forhold til den typiske indkomst i samfundet.

Den typiske indkomst defineres som *medianindkomsten*. I OECD definerer man fx fattigdom som en indkomst, der er mindre end 50 % af *medianindkomsten*. I et samfund med stor ulighed vil der derfor være mange fattige. Her er der tale om social fattigdom, idet de pågældende – og ikke mindst deres børn – ligger i fare for at blive udstødt socialt, eftersom de ikke kan opretholde et typisk livsmønster.

For at give en mere fyldig og nuanceret sproglig beskrivelse af et datasæt indføres derfor nogle karakteristiske tal, der beskriver *spredningen* af datasættet.

Definition: Mål for spredning

Ved *variationsbredden* for et datasæt forstås tallet: *maksimum – minimum*

Ved datasættets 1. kvartil Q_1 (eller den *nedre kvartil*) forstås medianen for den del af datasættet, der ligger til venstre for medianen.

Ved datasættets 3. kvartil Q_3 (eller den *øvre kvartil*) forstås medianen for den del af datasættet, der ligger til højre for medianen.

Ved *kvartilbredden* for et datasæt forstås tallet: $Q_3 - Q_1$

Medianen kaldes af og til for 2. kvartil og betegnes m eller Q_2 . Talsættet bestående af 1., 2., og 3. kvartil kaldes for *kvartilsættet*.

Eksempel: Kvartilsættet for et datasæt

I vores eksempel med kondital er der 25 elever. Vi skal bestemme *nedre* og *øvre* kvartil og splitter datasættet i to halvdele, nemlig dem, der går forud for medianen, og dem der følger efter medianen:

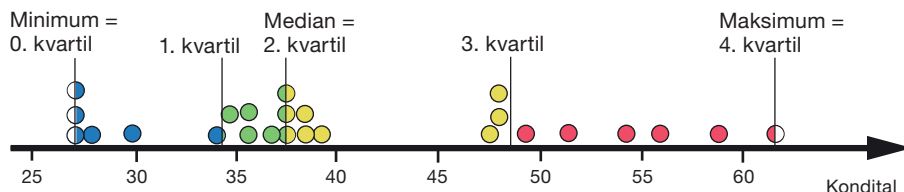
27.1	27.1	27.1	28.0	29.9	34.0	34.6	35.5	35.5	36.7	37.5	37.5
					37.5						
38.3	38.3	39.1	47.4	47.9	47.9	49.2	51.4	54.2	56.0	58.7	61.7

Der er derfor 12 observationer i hver halvdel. Den *nedre kvartil* er medianen for den halvdel af tallene, der er mindre end medianen, og den *øvre kvartil* er medianen for den halvdel af tallene, der er større end medianen. Da vi her har et lige antal observationer, bliver disse to tal udregnet som gennemsnit af de midterste værdier:

27.1	27.1	27.1	28.0	29.9	34.0	34.6	35.5	35.5	36.7	37.5	37.5
					34.3						
					37.5						
38.3	38.3	39.1	47.4	47.9	47.9	49.2	51.4	54.2	56.0	58.7	61.7
					48.55						

Den nedre kvartil er altså 34.3 og den øvre kvartil er 48.55.

Vi inddrager igen prikdiagrammet og afsætter kvartilsættet. Af og til betegnes *maksimum* og *minimum* som *0. kvartil* og *4. kvartil*, og disse tal har vi også markeret:



Datasættet er hermed opdelt i fire delsat. Hvert delsat indeholder mindst $\frac{1}{4}$ af observationerne, dvs. mindst 6 observationer, idet nogle indeholder flere, fordi nogle af observationerne ligger præcist på grænsen mellem to områder og derfor tælles med begge steder. Den nederste fjerdedel af målingerne indeholder således 6 observationer, den næste fjerdedel af målingerne indeholder 7 observationer, den næste fjerdedel indeholder hele 9 observationer, mens den øverste indeholder 6 observationer.

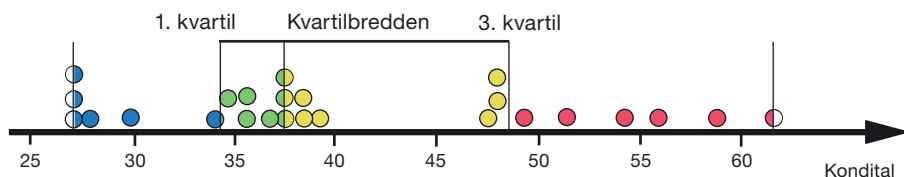
Bemærkning: Den ovenstående opdeling af datasættet i to lige store halvdele beror på en konvention. Vi kunne også have valgt at tælle medianen med i begge halvdele. Forskellige værktøjsprogrammer bruger forskellige konventioner omkring definitionen af første og tredje kvartil.

Øvelse 2.13

Find ud af, hvilken konvention dit værktøjsprogram anvender ved at bestemme kvartilsættet for datasættene:

$\{1,2,3,4,5\}$ $\{1,2,3,4,5,6\}$ $\{1,2,3,4,5,6,7\}$ $\{1,2,3,4,5,6,7,8\}$

Når vi har fastlagt kvartilerne, vil intervallet fra den nedre kvartil til den øvre kvartil indeholde (mindst) halvdelen af datasættets elementer. På prikdiagrammet har vi angivet *kvartilbredden*, der netop er bredden af dette interval:



I eksemplet med kondital fås således:

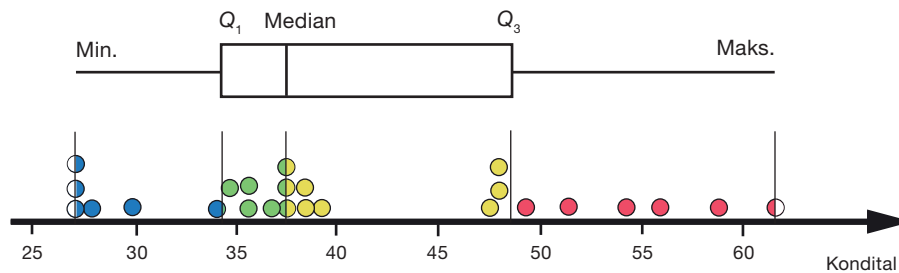
$$\text{Kvartilbredde} = Q_3 - Q_1 = 48,55 - 34,3 = 14,25$$

Praxis: Om brugen af variationsbredde og kvartilbredde

Hvis data er jævnt fordelt, er variationsbredden et rimeligt mål for spredningen af data. Men hvis datamaterialet har tendens til at klumpe sammen i midten, er den robuste kvartilbredde et bedre bud på spredningen af data.

2.4 Grafisk præsentation af et numerisk datamateriale – boksplot

Det udvidede kvartilsæt kan også anvendes i en særlig grafisk fremstilling af datasættet, som vi kalder *boksplot*. Et boksplot tegnes ud fra det udvidede kvartilsæt:

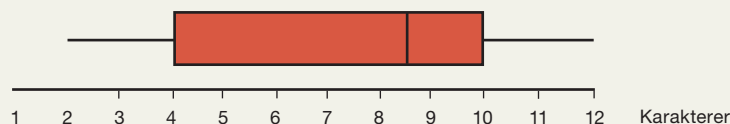


Boksplottet giver i ét blik en visuel information om såvel *niveauet* for konditallet (repræsenteret ved medianen) som konditallets spredning i form af *kvartilbredden*, som jo netop er længden af kvartilboksen ("kassen"). Boksen indeholder altid (mindst) halvdelen af observationerne.

Det er valgfrit, om boksplottet tegnes lodret eller vandret.

Øvelse 2.19

På figuren ses et boksplot for en karakterfordeling.



- Aflæs kvartilsættet, og bestem kvartilbredden.
- Hvordan ville datasættet se ud, hvis vi ydermere får oplyst, at der er 8 karakterer i datasættet?

Øvelse 2.20 (især for A-niveau)

Et karaktersæt for en klasse med 17 elever har det udvidede kvartilsæt:

minimum = 02, nedre kvartil = 4, median = 7, øvre kvartil = 10, maksimum = 12

- Hvad er den mindst mulige værdi for gennemsnitskarakteren, dvs. middeltallet?
- Hvad er den størst mulige værdi for gennemsnitskarakteren, dvs. middeltallet?

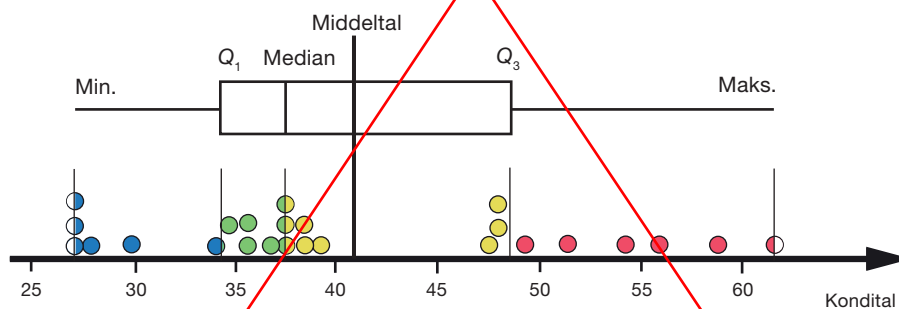
2.5 Sproglig præsentation af formen for en fordeling – symmetri og outliers

Et boksplot kan give et godt fingerpeg om, hvorvidt fordelingen er symmetrisk eller asymmetrisk. I tilfældet med konditallet er det fx tydeligt, at såvel den højre halvdel af kvartilboksen som den højre hale er meget større end den venstre halvdel af kvartilboksen henholdsvis den venstre hale. Fordelingen er altså tydeligt *højreskæv*.

Billedet er imidlertid ikke altid så klart. Fx kan den højre halvdel af kvartilboksen være større end den venstre halvdel, mens den venstre hale måske samtidig er længere end den højre hale. Man får derfor et bedre mål for fordelingsens symmetri eller mangel på samme ved at inddrage middelværdien. Hvis middelværdien er større end medianen, betyder det, at den halvdel af observationerne, der ligger over medianen, vejer tungere – ligger længere væk – end den halvdel af observationerne, der ligger under medianen. Det giver anledning til følgende:

Definition: Højreskæv og venstreskæv

En fordeling med en middelværdi, der ligger tydeligt over medianen, kaldes *højreskæv*, mens en fordeling med en middelværdi, der ligger tydeligt under medianen, kaldes *venstreskæv*. Som et mål for skævheden bruges forskellen mellem middeltallet og medianen.



Øvelse 2.21

- Konstruer et datasæt på 9 tal, der er symmetrisk omkring værdien 5, idet to tal siges at ligge symmetrisk omkring 5, hvis de ligger lige langt fra 5 på hver sin side af 5.
- Hvad bliver medianen henholdsvis middeltallet for dette datasæt? Formuler i ord en regel for median og middeltal for symmetriske datasæt.
- Konstruer et datasæt med 6 elementer, der har fælles median og middeltal, men som ikke er symmetrisk omkring denne fælles værdi.

Styrken ved et boksplot (evt. suppleret med en middelværdi) er, at det giver en god oversigt over fordelingen. Svagheden er, at det skjuler mange detaljer, da vi jo ikke kan se selve observationerne i boksplottet, og derfor ikke kan se, om de fx samler sig i tydelige klumper med tydelige huller imellem. Boksplottet fortæller altså ikke så meget om, hvad der fx foregår inde i kvartilboksen.

Boksplottet giver dog mulighed for at sætte et særligt fokus på fjerntliggende observationer langt ude i halerne. Hvis observationerne ligger tilstrækkelig langt fra den centrale boks, skilles de normalt ud som enkeltobservationer. Sådanne observationer betegner vi med det engelske ord *outliers* (på dansk: *perifere observationer*).

Definition: Outliers

Ved outliers forstås vi observationer, som ligger mere end halvanden kvartilbredde væk fra enten øvre eller nedre kvartil (dvs. fra kvartilboksen).

I eksemplet med kondital er outliers observationer, der ligger mindst $1,5 \cdot 14,25 = 21,375$ væk fra kvartilboksen, dvs. enten 21,375 under første kvartil eller 21,375 over tredje kvartil. Det er der imidlertid ingen af observationerne, der gør. I denne klasse er der derfor ingen outliers (ingen perifere kondital).

Øvelse 2.22

Vend tilbage til datasættet for drengenes hvilepuls i øvelse 2.10, {65,62,73,58,31,69}. Er observationen 31 en outlier?

Øvelse 2.23

Et datasæt har kvartilsættet:

nedre kvartil = 5, median = 8, øvre kvartil = 13.

Hvor lille skal minimumsværdien være, for at den er en outlier? Hvor stor skal den største værdi være, for at den er en outlier?

I nogle undersøgelser vil man vælge at se bort fra outliers og vedtage, at sådanne målinger er udtryk for *måleusikkerhed* eller deciderede fejl i målingerne. Man skal dog altid være varsom med at fjerne målinger.

Det klassiske eksempel på en grov fejltagelse er den manglende opdagelse af ozonhullet, hvor de alarmerende data blev fjernet automatisk af satellitten i 1980'erne, fordi de afveg alt for meget fra det forventede.

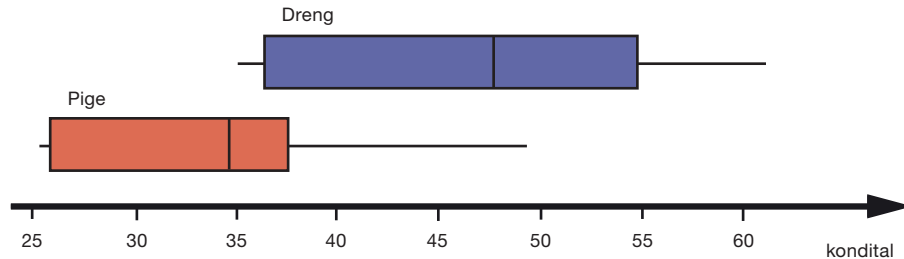


Halley Ozon målestation på Antarktis

2.6 Anvendelse af boksplot til sammenligning af datasæt

Boksplot har deres store styrke når vi ønsker at sammenligne forskellige datasæt.

I eksemplet med kondital indgik også den kategoriske variabel: elevens køn. Vi opdeler nu datasættet efter piger og drenge og tegner, som vist på figuren nedenfor, særskilte boksplot for hvert køn for at sammenligne dem.



Øvelse 2.24

- Tegn selv de to boksplot ved først at bestemme det udvidede kvartilsæt for drenge og for piger.
- Bestem middeltallet for henholdsvis drenge og piger.

Boksplottet giver i ét blik et visuelt indtryk af den markante forskel på konditallene for piger og drenge.

Praxis: Sproglig beskrivelse af grafisk præsentation

Når vi skal give en *sproglig* beskrivelse af denne *grafiske præsentation*, koncentrerer vi os normalt først om *niveauet* i de to datasæt. Har vi selv konstrueret boksplottet, kender vi jo datasættet og kan inddrage *middeltallet*, men er det et boksplot, vi får forelagt, har vi kun mulighed for at anvende *medianen* som mål for niveauet. Dernæst ser vi på spredningen i datasættet ud fra kvartilbredden og halerne, herunder eventuelle outliers. Endelig kan vi se på formen for fordelingen, hvor vi kan inddrage symmetri eller mangel på samme.

Eksempel: Sproglig beskrivelse af en sammenligning af to boksplot

Drengenes *niveau* ligger et godt stykke over pigernes niveau, idet drengenes *median* er 47,9 og pigernes er 34,6.

Faktisk er de to *kvartilbokse* tæt på at være helt adskilte. Drengenes kvartilboks ligger over pigernes median, pigernes kvartilboks ligger under drengenes median, idet drengenes *nedre kvartil* er 37,5, og pigernes *øvre kvartil* er 38,7.

Det betyder, at mens mindst 75 % af drengene har et kondital på 37,5 eller mere, har mindst 75 % af pigerne et kondital på 38,7 eller mindre.

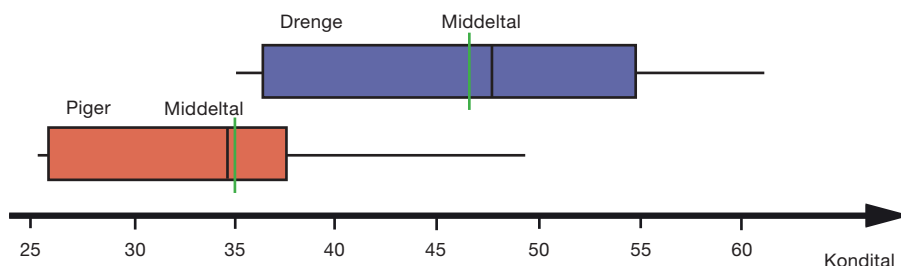
Vi ser endvidere, at drengenes boks er noget længere end pigernes, hvilket betyder, at drengenes kondital ligger mere spredt end pigernes. Som mål for spredningen kan vi bruge *kvartilbredden*, og denne er for drengene: $55,1 - 37,5 = 17,6$, mens kvartilbredden for pigerne er: $38,7 - 27,55 = 11,15$.

Maksimum og *minimum* for drengene er henholdsvis 61,7 og 35,5. Der er ikke tale om *outliers*, da $1,5 \cdot \text{kvartilbredden} = 1,5 \cdot 17,6 = 26,4$. Så outliers ville have kondital *under* $37,5 - 26,4 = 11,1$ eller *over* $55,1 + 26,4 = 81,5$.

For pigerne er *maksimum* og *minimum* henholdsvis 49,2 og 27,1, og der er heller ikke tale om outliers, da vi her har: $1,5 \cdot \text{kvartilbredden} = 1,5 \cdot 11,15 = 16,725$. Så outliers hos pigerne ville være kondital *under* $27,55 - 16,725 = 11,825$ og *over* $38,7 + 16,725 = 55,425$.

Vi kan ikke umiddelbart besvare spørgsmål om skævhed uden at kende middeltallet. Men det gør vi heldigvis i dette tilfælde.

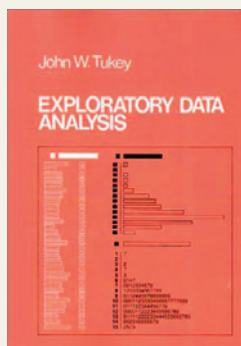
Vi kan konkludere, at drengenes datasæt er en anelse venstreskævt, da middeltallet ligger under medianen, og omvendt for pigerne. Men tallene ligger så tæt, at der ikke er grundlag for at konkludere noget afgørende om skævhed.



Forklaringen på de store *niveauforskelle* er naturligvis, at drenge har en kropsbygning, som favoriserer et højere kondital, så selv om piger og drenge træner lige meget, vil drengens kondital som regel ligge et stykke over pigernes.

Øvelse 2.25

Foretag en tilsvarende opdeling af talmaterialet efter den kategoriske variable træning. Sammenlign de to datasæt, der her fremkommer, ved hjælp af boksplot, og giv en sproglig konklusion som i eksemplet.



Boksplot blev første gang introduceret af den amerikanske statistiker John Tukey i bogen *Exploratory Data Analysis* fra 1977. Tukey gik nye veje med sin eksperimenterende og meget direkte tilgang til håndtering af datamaterialer. Et karakteristisk citat, der viser hans holdning til statistik, er følgende:

"Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught."

Teknikkerne, Tukey indførte, som fx boksplottet, udmærker sig ved at være lettilgængelige, også uden støtte af computer. Men Tukey havde også et klart blik for, at den stadig større maskinkraft og de stadig billigere computere samtidig gav helt nye muligheder for fx simulering. Det er også Tukey, der er ophavsmand til ord som *bit* og *software*.

På [hjemmesiden](#) kan du læse mere om Tukey.

3. Numeriske variable – beskrivelse af store datasæt

3.1 Grafisk præsentation af grupperede datasæt – histogram

Når man henter store datamaterialer fx fra Danmarks Statistik, vil de normalt være *grupperede*, og man har ofte slet ikke adgang til de oprindelige data. Som et eksempel på behandling af et større datamateriale ser vi på følgende tabel fra kriminalitetsstatistikken over aldersfordelingen for de danskere, der i 2001 var ofre for en kriminel handling:

Ofrets alder (interval)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-
Antal (hyppighed)	419	8176	10553	7175	5452	4274	2475	1995	1399

Intervallerne skal forstås således: Er man 19 år, og blot 1 time fra at fylde 20, placeres man i intervallet 10-19. Da intervallet går lige til skæringspunktet 20, vil vi tillade os at sige, at intervallets højre endepunkt er 20.

Andre gange kan inddelingen være angivet således: 0-10, 10-20, 20-30,

Denne inddeling skal forstås på samme måde som forklaret ovenfor.

Ofte vil vi få et bedre overblik over datamaterialet ved at omregne *antal* til *procent*.

Øvelse 2.27

Kopier tabellen ind i et regneark.

a) Udnyt regnearkets muligheder til at finde summen 41918:

Ofrets alder (interval)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-	
Antal (hyppighed)	419	8176	10553	7175	5452	4274	2475	1995	1399	i alt 41918

b) Opret en ny række med procent:

Procent (frekvens)			25,2							100
-----------------------	--	--	------	--	--	--	--	--	--	-----

Procenttallet 25,2 er fremkommet på følgende måde:

$$\frac{10553}{41918} \cdot 100 = 25,2$$

c) Udnyt nu regnearkets muligheder til at udfylde den nederste række, og kontroller at summen af procenterne giver 100.

Praxis: Betegnelser vedrørende grupperede datasæt

Når vi arbejder med grupperede datasæt, er vi i nogle situationer mest interesserede i hvor mange, der er i hvert interval, dvs. *antallet* er i fokus. I andre situationer er vi mest interesserede i den andel, der er i hvert interval, dvs. *procentandelen* er i fokus.

Antallet kaldes i andre lærebøger for hyppigheden eller intervalhyppigheden. På engelsk kaldes antal for *frequency*.

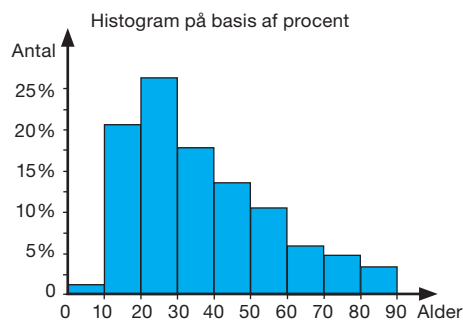
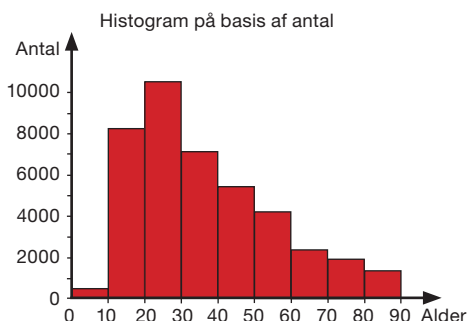
Procentandelen kaldes i andre lærebøger for frekvensen. På engelsk kaldes procentandelen for *relative frequency*.

En anden mulighed er at præsentere datasættet i form af et *histogram*, hvor vi opretter en søjle svarende til *antallet* af observationer i det pågældende interval eller til den *procentandel* af observationerne, der ligger i intervallet. Intervallerne har typisk samme bredde. Histogrammets udseende afhænger meget af, hvordan intervallerne fastlægges.

Praxis: Tegning af et histogram

- Et histogram tegnes i et koordinatsystem, hvor dataværdierne er afsat på første-aksen.
- Histogrammet består af søjler, tegnet over hvert af de intervaller, som datasættet er grupperet i.
- Søjleens bredde bør svare til bredden af det underliggende interval.
- Intervalbredden skal så vidt muligt være den samme i alle intervaller.
- Intervalhøjden repræsenterer antallet af observationer i det pågældende interval eller, hvis man regner i procent, andelen af observationer i intervallet.

Vi tegner nu histogrammet for datamaterialet både ud fra *antal* og ud fra *procent*, hvor førstaksen i begge tilfælde er inddelt efter ofrets alder med delestreger svarende til intervalgrænserne 0, 10, 20, Vi vedtager, at øvre grænse placeres, så intervallet får samme bredde som de øvrige:



Vi ser fx, at ca. en fjerdedel af ofrene tilhører aldersgruppen 20-29 år, som også er den aldersgruppe, der rummer de fleste ofre.

Øvelse 2.28

- Konstruer selv de to diagrammer.
- I hvilket interval må medianen befinde sig?



Øvelse 2.29

I tabellen til højre finder du data for herrernes løbetider i Copenhagen Marathon 2009.

- Opstil datamaterialet i en tabel som i øvelse 2.27, og udregn procenttallene.
- Konstruer histogrammerne for dette grupperede datasæt.

I hvilket interval befinder medianen sig?

På *hjemmesiden* findes et projekt som behandler data-materialet fra Copenhagen Marathon 2009.

	Løbetider	Antal
1	2:00 – 2:30	5
2	2:30 – 3:00	215
3	3:00 – 3:30	1399
4	3:30 – 4:00	2648
5	4:00 – 4:30	1575
6	4:30 – 5:00	662
7	5:00 – 5:30	164
8	5:30 – 6:00	52
9	6:00 – 6:30	7
10	6:30 – 7:00	1

3.2 Skøn over middeltal for grupperede datasæt

I afsnit 2 ovenfor beskrev vi *niveauet* for et datasæt ved hjælp af de to karakteristiske værdier, middeltallet og medianen. Det samme ville vi gerne kunne gøre med store datasæt, hvor materialet er præsenteret i form af grupperede observationer.

Da vi ikke kender den eksakte fordeling i eksemplet med kriminalitetsstatistikken, kan vi ikke lave en præcis udregning. Men vi kan komme med et kvalificeret skøn, da vi trods alt ved en hel del om aldersfordelingen. Et sådant skøn bygger på nogle antagelser.

Når vi skal danne os et *skøn over middeltallet* for ofrenes alder, vil vi således antage, at alderen for ofrene i en bestemt aldersgruppe, fx gruppen 20-29, er spredt jævnt ud over hele intervallet, dvs. de ligger symmetrisk omkring midtpunktet i aldersgruppe-intervallet, som her er 25. Der er altså lige så mange over midtpunktet 25, som der er under.

Når vi udregner middeltallet (den gennemsnitlige alder), svarer det samlede bidrag fra denne aldersgruppe derfor til, at alle havde alderen 25. Tilsvarende med de andre aldersgrupper. Bidraget til den samlede sum svarer til, at alle i den første gruppe havde alderen 5, i den næste 15 osv.

For den sidste aldersgruppe har vi et problem, idet der ikke er sat nogen øverste grænse for alderen. Det nemmeste er da at antage, at den spænder over aldersgruppen fra 80 til 90, og at der kun er så få ofre over 90, at det ikke har nogen nævneværdig indflydelse på resultatet.

Vi finder da følgende:

$$\text{den samlede alder for ofrene i 2001} = 5 \cdot 419 + \dots + 85 \cdot 1399 = 1549510 \quad (\text{I})$$

Ovenfor fandt vi, at det samlede antal ofre var 41918.

Skønnet for *middelalderen* er derfor:

$$\frac{\text{Densamlede alder}}{\text{Det samlede antal}} = \frac{1549510}{41918} \approx 37,0 \text{ år} \quad (\text{II})$$

Dette skøn ændres ikke afgørende, selv om midtpunktet for den sidste uafsluttede aldersgruppe fx flyttes til 90. Den gennemsnitlige alder for en person, der er udsat for kriminalitet, er altså 37 år.

Øvelse 2.30 (især for A-niveau)

I antalstabellen erstattes intervallerne med midtpunktet, vi sletter rækken med antal og tilføjer en ny række, hvor procenterne er omregnet til decimaltal. Så har vi følgende tabel over sammenhørende værdier for intervalmidpunkter og procenter:

Interval midtpunkt	5	15	25	35	45	55	65	76	85	i alt
Procent (frekvens)	1,0	19,5	25,2	17,1	13,0	10,2	5,9	4,8	3,3	100
Procent som decimaltal	0,01	0,20	0,25	0,17	0,13	0,10	0,06	0,05	0,03	1

a) Udnyt regnearkets muligheder til at udregne tallet:

$$5 \cdot 0,01 + 15 \cdot 0,20 + 25 \cdot 0,25 + \dots + 85 \cdot 0,03$$

b) Argumenter for, at denne udregning svarer til den samlede udregning af (I) og (II) ovenfor.

c) Argumenter for, at formlen i den følgende definition giver samme resultat som den udregning, vi beskrev ovenfor.

Definition: Middeltallet for grupperede observationer

Middeltallet for et datasæt, der er givet som grupperede observationer inddelt i k intervaller, kan udregnes som et *vægtet gennemsnit* ud fra følgende formel:

$$m = M_1 \cdot f_1 + M_2 \cdot f_2 + \dots + M_k \cdot f_k,$$

hvor M 'erne er midtpunkterne i intervallerne, og f 'erne er procentandelene i de enkelte intervaller, skrevet som decimaltal.

Vægtet gennemsnit kaldes også for *Vejet gennemsnit*.

Øvelse 2.31

Giv et skøn over herrernes gennemsnitsløbetid i Copenhagen Marathon 2009 (data findes i øvelse 2.29).

3.3 Grafisk præsentation af grupperede datasæt – sumkurve

Skal vi danne os et skøn over, hvordan observationerne som helhed fordeler sig i aldersintervallerne, er det for groft at samle observationerne i midtpunkterne for de enkelte aldersgrupper. Her skal vi i stedet udnytte antagelsen om, at observationerne inden for de enkelte aldersgrupper er jævnt fordelt.

Vi udbygger nu tabellen over kriminalitetsstatistikken (fra øvelse 2.30) dels med en række bestående af de højre intervalendepunkter, dels med en række bestående af de såkaldte *kumulerede procenter* (af og til betegnet *kumulerede frekvenser*).

Ofrets alder (interval)	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-	i alt
Højre endepunkt	10	20	30	40	50	60	70	80	90	
Antal (hyppighed)	419	8176	10553	7175	5452	4274	2475	1995	1399	41918
Procent (frekvens)	1,0	19,5	25,2	17,1	13,0	10,2	5,9	4,8	3,3	100
Kumuleret procent	1,0	20,5	45,7	62,8	75,8	86,0	91,9	96,7	100,0	

Kumuleret betyder opsummeret. Eksempelvis er tallet 45,7 fremkommet ved at lægge 25,2 til den opsummerede procent 20,5, der er udregnet for det foregående interval.

Øvelse 2.32

Konstruer en tabel som ovenstående i et regneark. Omregningen fra *antal* til *procent* samt fra *procent* til *kumuleret procent* gennemføres ved at udnytte regnearkets muligheder.

Vi ser da, at den nedre kvartil (25 %) ligger et sted i den tredje aldersgruppe fra 20-29 år, medianen (50 %) ligger et sted i den fjerde aldersgruppe fra 30-39 år, mens den øvre kvartil (75 %) ligger et sted i den femte aldersgruppe fra 40-49 år.

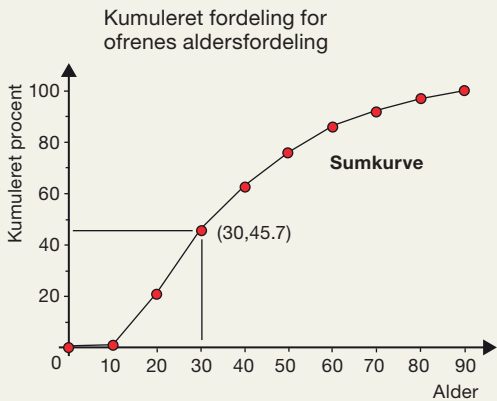
En nærmere beregning af, nøjagtigt hvor vi bør lægge kvartilerne og medianen, er lidt omstændelig, så vi vil i stedet bestemme værdierne *grafisk* ved hjælp af en *sumkurve*.

Definition: Sumkurven for grupperede observationer

Sumkurven for et grupperet datasæt er den grafiske repræsentation af de sammenhørende værdier af *højre intervalendepunkt* og *de kumulerede procenter*. Den grafiske repræsentation fremkommer ved at forbinde punkterne med rette linjestykker.

De kumulerede procenter afsættes (som den afhængige variabel) op ad andenaksen, mens alderen afsættes (som den uafhængige variabel) ud ad førsteaksen. Punkterne forbindes med en graf bestående af rette linjestykker.

Øvelse 2.33



Konstruer i et værktøjsprogram sumkurven som en graf for *den kumulerede procent som funktion af alderen*.

Læg mærke til, at vi har tilføjet begyndelsespunktet (0,0%). Læg også mærke til, at sumkurven er stejlest, dvs. har størst hældning, i aldersgruppen fra 20 til 30 år. Dette svarer til den højeste søjle i histogrammet.

- Argumenter vha. sumkurven for, at 33 % blandt ofrene er yngre end 25 år.
- Bestem vha. sumkurven, hvor stor en andel af ofrene, der er over 45 år.

Øvelse 2.34

Omregn tabellen over løbetider ved herrernes maratonløb Copenhagen Marathon 2009 (data findes i øvelse 2.29) til også at omfatte procenter og kumulerede procenter.

- Konstruer sumkurven for løbetiderne ved at afsætte den kumulerede procent som funktion af løbetiden.

Benyt sumkurven til at besvare følgende spørgsmål:

- Hvor hurtigt løber de hurtigste 10 % af løberne?
- Hvor stor en procentdel af løberne gennemløber maratonløbet på under 3 timer og 15 minutter?
- For hvilket tidsinterval er sumkurven stejlest?
- Hvad fortæller dette om fordelingen af løbetider?

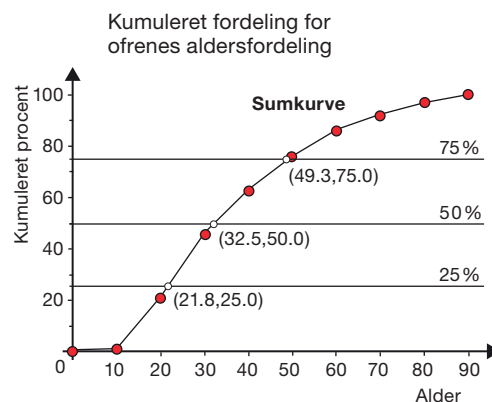
3.4 Skøn over median og kvartiler for grupperede datasæt

På sumkurven kan vi også gå baglæns, dvs. gå fra andenaksen til førsteaksen, dvs. fra kumuleret procent til den uafhængige variabel, der er i spil. Dette udnyttes til at bestemme et kvartilsæt for grupperede datasæt. I vores eksempel med kriminalitetsstatistikken gøres det som vist på figuren.

Vi kan aflæse kvartilerne, dvs. den alder, der svarer til en kumuleret procent på 25 (nedre kvartil), den alder, der svarer til en kumuleret procent på 50 (medianen), og den alder, der svarer til en kumuleret procent på 75 (øvre kvartil).

Betydningen af nedre kvartil er: Mindst 25 % af voldsofrene er under 22 år.

Betydningen af øvre kvartil er: Mindst 25 % af voldsofrene er over 49 år.



Vi aflæser, at nedre kvartil = 21,8, median = 32,5 og øvre kvartil = 49,3

Dermed har vi rådighed over alle fem statistiske nøgletal, der udgør det udvidede kvartilsæt, og kunne nu også skitsere et boksplot for aldersfordelingen. I dette tilfælde har vi imidlertid også histogrammet til rådighed, og vi foretrækker normalt at kommentere et grupperet datasæts fordeling ud fra histogram og middeltal sammen med sumkurve og kvartilsæt.

Vi så af histogrammet, at fordelingen af de grupperede observationer er tydeligt højreskæv. Dette er i overensstemmelse med, at *midde*alderen på 37,0 år er noget større end medianalderen på 32,5 år. I dette tilfælde vil det derfor nok være fornuftigst at anføre medianen som den typiske alder for et offer for kriminalitet.

Øvelse 2.35

- Bestem på tilsvarende måde kvartilsættet for herrernes maratonløb Copenhagen Marathon 2009 (data findes i øvelse 2.29).
- Inddrag middeltallet og kommenter fordelingen.