2. Numeriske variable – beskrivelse af et datamateriale

(I studieretningskapitlerne 11, 12 og 13 og i projekterne til kapitlet er der mulighed for at gennemgå stoffet med brug af andre datasæt)

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

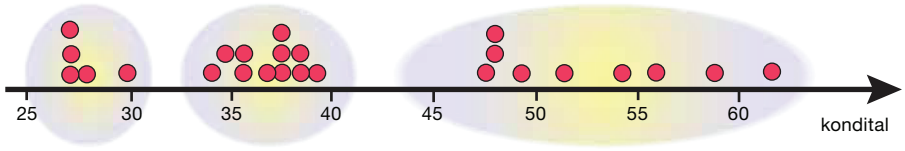
Når vi indsamler data ud fra observationer fx indhenter svar på spørgeskemaer eller lignende, skelner vi mellem *kategoriske* og *numeriske* variable. Vi vil i det følgende beskrive, hvorledes man kan skabe sig overblik over og præsentere et datamateriale ved at gennemgå et større eksempel. I en 1.g-klasse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

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 konditallenes 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 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.

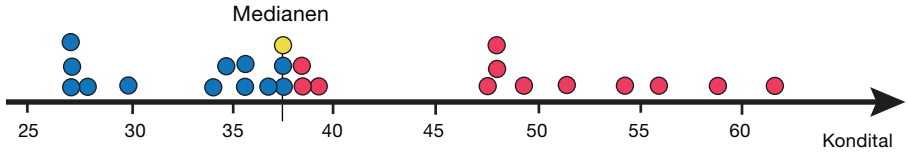
Middeltallet m skal opfylde, at:

$$25 \cdot m = 27{,}1 + \cdots + 61{,}7$$

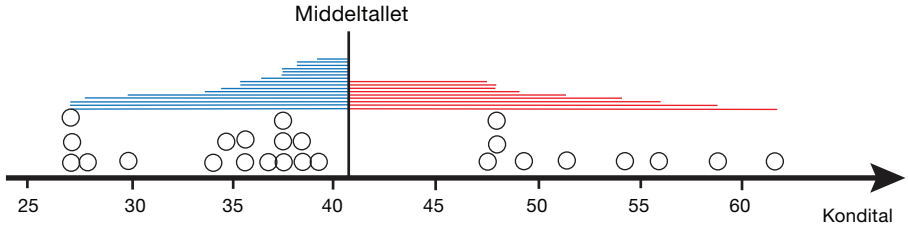
$$m = \frac{27{,}1 + \cdots + 61{,}7}{25} = 40{,}724$$

Konklusion: Middeltallet = 40,724.

Anvender vi prikdiagrammet, får vi følgende grafiske fremstilling:



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



Middeltallet splitter datasættet i to ulige store delsæt 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 højre-skæv 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:

$$25 \cdot m = 27{,}1 + 27{,}1 + \ldots + 58{,}7 + 61{,}7$$
$$m + m + \ldots + m = 27{,}1 + 27{,}1 + \ldots + 58{,}7 + 61{,}7$$

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:

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

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. Du kan via bogens **website** finde en uddybet redegørelse for disse begreber, hentet fra et studieretningskapitel på B-niveau om fagligt samarbejde mellem matematik-samfundsfag.

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 *median*indkomsten. I OECD definerer man fx fattigdom som en indkomst, der er mindre end 50 % af *median*indkomsten. 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.

Øvelse 2.10

a) Konstruer et datasæt med 12 indkomster, som du regner med er repræsentative for borgerne i fx København.

b) Bestem fattigdomsgrænsen for det pågældende datasæt i henhold til OECD's fattigdomsdefinition.

Øvelse 2.11 **(især for B- og A-niveau)**

Vi ser på indkomsterne i en bestemt befolkningsgruppe.

a) Hvad sker der med medianen, hvis alle indkomster bliver dobbelt så store? Tre gange så store?

b) Formuler en regel for dette i ord. Oversæt reglen til en formel, idet der indføres passende variable.

c) Hvad sker der med fattigdomsgrænsen, hvis alle indkomster (og dermed købekraften) bliver dobbelt så store?

d) Hvad sker der med antallet af fattige, hvis alle indkomster bliver fordoblet?

2.3 Sproglig præsentation af spredningen for et datasæt – variationsbredde og kvartilbredde

Der går naturligvis megen information tabt, når man beskriver et helt datasæt med én karakteristisk værdi, som fx median eller middeltal. Eksempelvis kan vidt forskellige datasæt have samme median eller samme middeltal.

Øvelse 2.12

Bestem middeltallet og medianen for følgende to datasæt:

a) {10,20,20,30,70,80,80,90}

b) {45,45,50,50,50,55,55}

Kommenter resultatet.

For at give en mere fyldig og nuanceret sproglig beskrivelse af et datasæt indføres derfor to yderligere begreber. Dels de karakteristiske tal 1. og 3. kvartil. Og dels begreber, der kan beskrive *spredningen* af datasættet.

Definition: 1. og 3. kvartil

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.

Definition: Mål for spredning af et datasæt

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

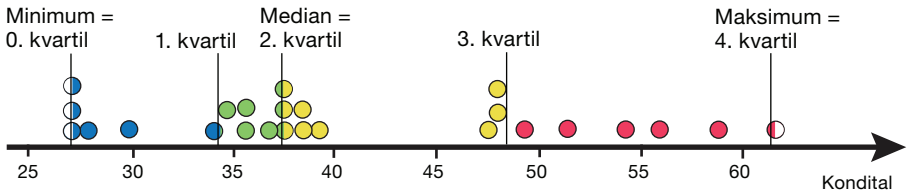
27,1	27,1	27,1	28,0	29,9	34,0	34,6	35,5	35,5	36,7	37,5	37,5
					37,5						
38,3	38,3	39,1	47,4	47,9	47,9	49,2	51,4	54,2	56,0	58,7	61,7

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 delsæt. Hvert delsæt 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.

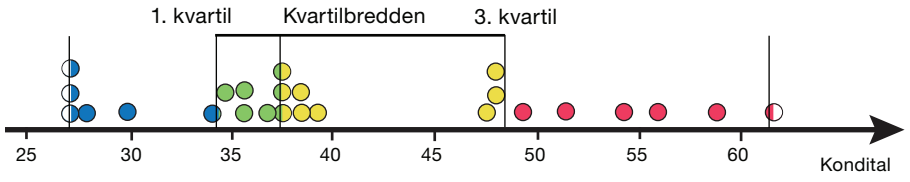
Bemærkning: Den ovenstående opdeling af datasættet i to lige store halvdele beror på en konvention. Vi kunne også have valgt at tælle medianen med i begge halvdele. Forskellige værktøjsprogrammer bruger forskellige konventioner omkring definitionen af første og tredje kvartil.

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:

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

Øvelse 2.14

Hvad er forskellen på definitionen af den første kvartil og definitionen på fattigdomsgrænsen, vi gav i eksemplet side 78 i afsnit 2.2?

Øvelse 2.15

a) Bestem kvartilerne for datasættet {1, 1, 2, 5, 6}.

b) Hvad bliver kvartilbredden?

Øvelse 2.16

a) Konstruer et datasæt med 6 elementer, hvor minimum = 1, nedre kvartil = 2, median = 5, øvre kvartil = 9 og maksimum = 12.

b) Hvilken værdi har middeltallet?

Øvelse 2.17 (især for A-niveau)

Hvor mange elementer må der mindst ligge i kvartilintervallet $[Q_1, Q_3]$, hvis et datasæt indeholder 5 elementer? 6 elementer? 7 elementer? 8 elementer?
Hvilken konklusion vil du drage?

Praxis: Det udvidede kvartilsæt eller 5-punkts-opsummeringen

De karakteristiske tal i en sproglig beskrivelse af datasættet er det *udvidede kvartilsæt*. I eksemplet med kondital er dette:
minimum = 27,1, første kvartil = 34,3, median = 37,5, tredje kvartil = 48,55
maksimum = 61,7
Disse tal kaldes også for *5-punkts-opsummeringen* af datasættet.
Disse statistiske nøgletal er en væsentlig del af den såkaldte enkeltvariabelstatistik, der er indbygget i mange værktøjsprogrammer.

Øvelse 2.18

På bogens **website** ligger en vejledning til, hvordan værktøjsprogrammerne kan udføre *enkeltvariabelstatistik* og dermed udregne det udvidede kvartilsæt og middeltallet på én gang.

a) Indtast listerne fra øvelse 2.12, og benyt et værktøjsprogram til at bestemme det udvidede kvartilsæt samt middeltallet.

b) Benyt et værktøjsprogram til at bestemme det udvidede kvartilsæt samt middeltallet for datasættet med kondital.

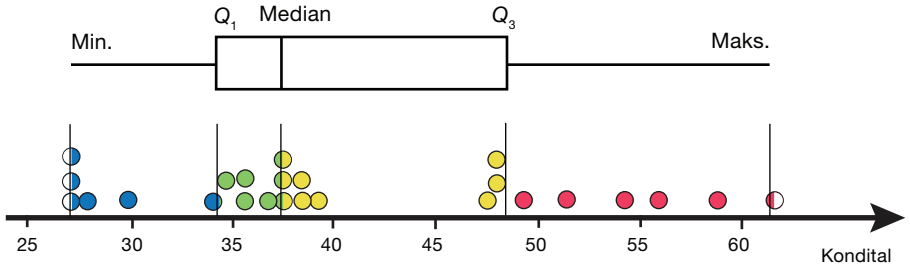
Praxis: Om brugen af variationsbredde og kvartilbredde

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.

Bemærkning: I slutningen af dette afsnit, på side 90-91, præsenteres det mere avancerede spredningsbegreb, som anvendes i sandsynlighedsregning og i den statistik, der bygger på sandsynlighedsregning.

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:

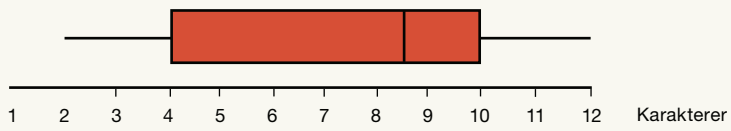


Boksplottet giver i ét blik en visuel information om såvel *niveauet* for konditallet (repræsenteret ved medianen) som konditallets spredning i form af *kvartilbredden*, som jo netop er længden af kvartilboksen (”kassen”). Boksen indeholder altid (mindst) halvdelen af observationerne. På bogens **website** ligger en vejledning til, hvordan værktøjsprogrammerne tegner boksplot.

Det er valgfrit, om boksplottet tegnes lodret eller vandret.

Øvelse 2.19

På figuren ses et boksplot for en karakterfordeling.



a) Aflæs kvartilsættet, og bestem kvartilbredden.

b) 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

a) Hvad er den mindst mulige værdi for gennemsnitskarakteren, dvs. middeltallet?

b) Hvad er den størst mulige værdi for gennemsnitskarakteren, dvs. middeltallet?

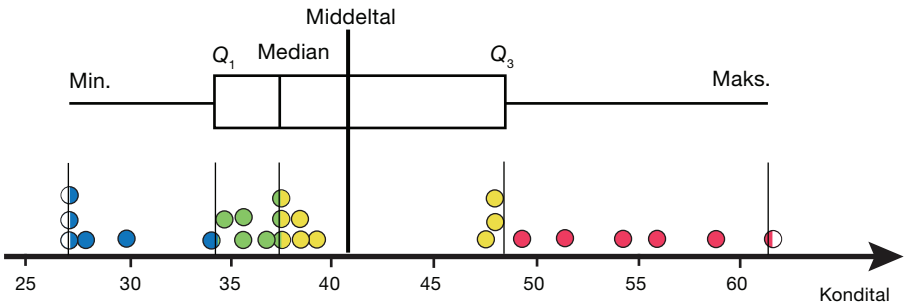
2.5 Sproglig præsentation af formen for en fordeling – symmetri og outliers

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 fordelingens 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

a) 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.

b) Hvad bliver medianen henholdsvis middeltallet for dette datasæt? Formuler i ord en regel for median og middeltal for symmetriske datasæt.

c) 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 forstår 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 2.9, {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?

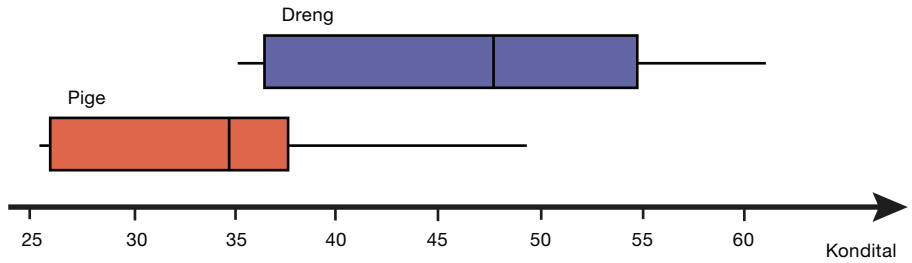
Halley Ozon målestation på Antarktis.

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.

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

a) Tegn selv de to boksplot ved først at bestemme det udvidede kvartilsæt for drenge og for piger.

b) 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

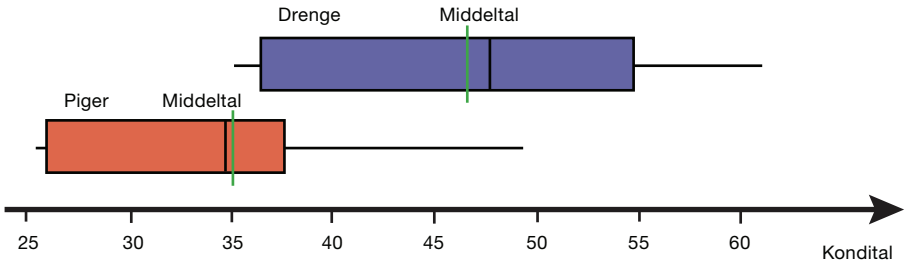
Drengenes *niveau* ligger et godt stykke over pigernes niveau, idet drengenes *median* er 47,9 og pigernes er 34,6.
Faktisk er de to *kvartilbokse* tæt på at være helt adskilte. Drengenes kvartilboks ligger over pigernes median, pigernes kvartilboks ligger under drengenes median, idet drengenes *nedre 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 · kvartilbredden = 1,5 · 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 · kvartilbredden = 1,5 · 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.

Eksempel: Spredningsmål knyttet til middelværdien for et datasæt

De to statistiske deskriptorer *median* og *middeltal* har hver sine styrker og svagheder, når vi ønsker at angive *niveauet* for et datasæt. Det har vi beskrevet i en praxisbox på side 71 (Om brugen af median eller middeltal til at beskrive niveauet for et dataset). Men uanset hvilket af de to tal, vi vælger, så har vi også understreget, at ét tal alene kun kan fange niveauet, og ikke kan fange den karakteristiske fordeling af de pågældende observationer.

I øvelse 2.12 så vi eksempelvis, at de to datasæt:

A = {10, 20, 20, 30, 70, 80, 80, 90} og B = {45, 45, 50, 50, 50, 55, 55}

begge har middeltal og median lig med 50. Men enhver kan jo se, de er meget forskellige.

For at supplere medianen indførte vi deskriptorerne 1. og 3. kvartil, og sammen med dem et karakteristisk mål for hvor spredte observationerne ligger omkring medianen, nemlig *kvartilbredden*. Kvartilsættene for de to datasæt er her:

(20,50,80) med kvartilbredde: 80 – 20 = 60
og (45,50,55) med kvartilbredde: 55 – 45 = 10

Tilsvarende suppleres middelværdien med et *spredningsmål*. Dette er lidt mere kompliceret at bestemme, men heldigvis har værktøjsprogrammerne indbyggede kommandoer, der kan give os disse tal. Det mål, der på dansk kaldes *spredning* er lig med *kvadratroden af det gennemsnitlige afstandskvadrat mellem observationerne og middelværdien*. Lad os først beregne det for *A* og *B*, og dernæst definere det generelt:

spredning (*A*) =

$$\sqrt{\frac{(10-50)^2+(20-50)^2+(20-50)^2+(30-50)^2+(70-50)^2+(80-50)^2+(80-50)^2+(90-50)^2}{8}} = 30{,}82$$

spredning (*B*) =

$$\sqrt{\frac{(45-50)^2+(45-50)^2+(50-50)^2+(50-50)^2+(50-50)^2+(55-50)^2+(55-50)^2}{7}} = 3{,}78$$

Du kan på bogens **website** finde en vejledning i anvendelsen af værktøjsprogrammerne til disse beregninger.

I det datamateriale, vi har arbejdet med i afsnit 2, indgik blandt andet drengenes højder:

Drengehøjder = {186, 185, 185, 170, 175, 174, 173, 177, 190, 172, 176, 186}

Prøv nu selv at bestemme middelværdi og spredning på dette datasæt. Du skal få middelværdi = 179,08 og spredning = 6,53.

Udfør en grafisk illustration af dette, fx med brug af et pindediagram, hvori du afsætter middelværdien samt de to værdier: *middelværdi – spredning* og *middelværdi + spredning*. Kommenter, hvad du ser.

Eksempel: Standardafvigelsen knyttet til middelværdien for en stikprøve

Når vi arbejder med observationer i form af datasæt, så skelner vi ofte imellem, om datasættet udgør "hele populationen", eller om datasættet er en stikprøve, der måske kan bruges til at sige noget om en større population. I eksemplet ovenfor udgør drengene hele populationen. Men hvis vi i et naturvidenskabeligt forsøg ønsker at bestemme en værdi for tyngdeaccelerationen, eller temperaturen under et bestemt kemisk forsøg, så kan vi opfatte vores datasæt som stikprøver, hvor vi i virkeligheden er interesseret i den "sande værdi". Der findes jo en bestemt værdi for tyngdeaccelerationen, uanset hvad vi måler. Vores målinger er en tilnærmelse til den "sande værdi", og har vi mange målinger, er det rimeligt at sige, at middelværdien af alle målingerne er et fornuftigt estimat af den "sande værdi".
I et bestemt eksperiment har et hold opdelt i 8 grupper forsøgt at bestemme tyngdeaccelerationen, *g*, ved at måle på pendulsvingninger, idet *g* indgår i formlen for sammenhængen mellem svingningstid og pendullængden. Resultaterne var følgende:

G = {9,73, 9,84, 9,79, 9,77, 9,83, 9,80, 9,91, 9,86}

Regn selv med! Vi bestemmer middelværdien:

middel(G) = 9,816

Dette er nu vores estimat for den sande værdi af tyngdeaccelerationen. Beregningen suppleres som før med at bestemme et mål for, hvor spredte vores målinger ligger. Men med stikprøver er der nu en enkelt detalje, der skal justeres i beregningen. Vi har lagt en binding på dataværdierne, idet vi selv har beregnet den middelværdi, der indgår i beregningen af spredningen. Derfor skal vi for stikprøver ikke dividere med antallet, her 8, men med 1 mindre end antallet, dvs. 7 i dette tilfælde. Dette spredningsmål har fået sit eget navn: *Standardafvigelsen* (engelsk: *Standard Deviation*):

Standardafvigelse (G) =

$$\sqrt{\frac{(9{,}73-9{,}816)^2+(9{,}84-9{,}816)^2+(9{,}79-9{,}816)^2+(9{,}77-9{,}816)^2+(9{,}83-9{,}816)^2+(9{,}80-9{,}816)^2+(9{,}91-9{,}816)^2+(9{,}86-9{,}816)^2}{7}}$$

Værktøjsprogrammerne kan også bestemme dette tal: Du skal få 0,056.

Vi sammenfatter i følgende:

Praxis: Spredning og standardafvigelse

Har vi givet et datasæt, $\{y_1, y_2, ..., y_n\}$, der kan betragtes som hele den population, vi arbejder med, så suppleres middeltallet, $\bar{y}$ med *spredningen*, s, i beskrivelsen af datasættet. *Spredningen*, der beregnes af værktøjsprogrammet, har formlen:

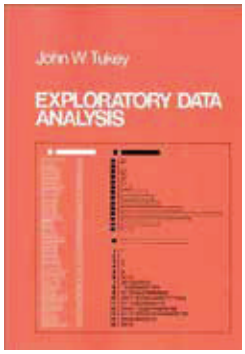
$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2}{n}}$$

Har vi givet et datasæt, der kan betragtes som en stikprøve af en større population, og hvor vi er interesseret i at bestemme den større populations "sande middelværdi" μ, så opfatter vi $\bar{y}$ som et estimat for μ, og middeltallet suppleres med *standardafvigelsen* σ i beskrivelsen af datasættet. *Standardafvigelsen*, der beregnes af værktøjsprogrammet, har formlen:

$$\sigma = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2}{n-1}}$$

Vi anvender således følgende symboler: Datasættets middeltal: $\bar{y}$. Den sande middelværdi: μ. Den empiriske spredning: s. Standardafvigelsen: σ.

Eksempel: Tukeys eksperimenterende analyse af data



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*.

Du kan på bogens **website** læse mere om Tukey.