

Ugrupperede observationer

Inden for deskriptiv statistik skelner vi mellem grupperede og grupperede observationer. (det handler kort sagt om, hvorvidt data er samlet i grupper eller ikke).

Eksempel: En undersøgelse om antallet af besøgte lande

25 personer bliver spurgt om, hvor mange lande, de har besøgt. De har svaret følgende:

8, 4, 16, 8, 9, 6, 16, 19, 7, 6, 4, 8, 11, 8, 9, 6, 9, 10, 11, 8, 14, 4, 6, 7, 10

I statistik kaldes hvert af sådanne tal for en *observation*, og alle observationerne udgør tilsammen et *observationssæt*.

Observationssættet skal fremstilles på en måde, så man let kan danne sig et overblik over tallene. Her kommer den deskriptive statistik til hjælp, for den handler netop om hensigtsmæssige måder at præsentere og beskrive mange observationer på.

Vi kunne opstille en tabel som nedenfor. Bemærk at observationerne er opstillet i voksende rækkefølge i første søjle.

Nedenfor følger forklaringen af de forskellige begreber:

observation	hyppighed	kumuleret hyppighed	frekvens	kumuleret frekvens
4	3	3	12%	12%
6	4	7	16%	28%
7	2	9	8%	36%
8	5	14	20%	56%
9	3	17	12%	68%
10	2	19	8%	76%
11	2	21	8%	84%
14	1	22	4%	88%
16	2	24	8%	96%
19	1	25	4%	100%
total	25		100%	

Observation:

Vi kan se, at der er 10 forskellige observationer i vores observationssæt. Bemærk vi kun skriver hver observation en gang i 1. kolonne, selvom den optræder flere gange. Antallet af gange, de optræder kaldes hyppigheder, som forklares nedenfor. Bemærk også at observationerne sorteres efter størrelse.

Hyppighed:

I vores observationssæt kan vi tælle, hvor mange gange en bestemt talværdi optræder. Fx forekommer tallet 4 i alt 3 gange, og vi siger derfor, at observationen 4 har *hyppigheden* 3.

Kumuleret hyppighed:

Det betyder summen af hyppighederne op til den pågældende observation. At kumulere betyder "at opsummere". Fx den kumulerede hyppighed for observationen 9 fås ved at lægge alle

hyppighederne sammen fra den mindste observation til og med observationen 9. Resultatet heraf er 17, dvs. 17 personer har besøgt 9 eller færre lande.

Bemærk at summen af den kumuleret hyppighed passer med antallet af observationer.

Frekvens:

Frekvensen er hyppigheden divideret med det samlede antal observationer angivet i procent (dvs. der er ganget med 100). Altså hvor mange procent udgør en bestemt observation af det hele.

Fx beregnes frekvensen for observationen 4 ved:

$$frekvens = \frac{3}{25} \cdot 100 = 12\%$$

Fordelen ved at udregne frekvensen for en observation er at kunne sammenligne observationssæt af forskellige størrelser. Fx hvis vi havde haft et observationssæt med 30 personer, ville hyppighederne af observationerne herfra jo ikke være sammenlignelige med hyppighederne fra vores observationssæt, hvor der kun er 25 personer.

Kumuleret frekvens:

Den kumuleret frekvens for observationen 9 fås på tilsvarende måde (som kummeleret hyppighed) ved at lægge frekvenserne sammen til og med observationen 9. Resultatet heraf er 68%, dvs. 68% af de 25 personer har besøgt 9 eller færre lande.

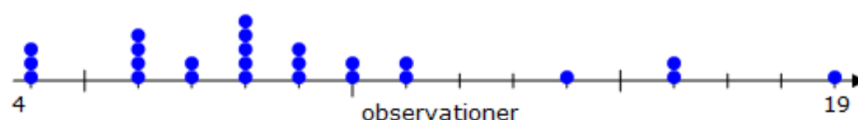
Opgave 1:

a) Hvor stor en del af personerne har besøgt mindst 7 og højst 11 lande?

Prikdiagram og stolpediagram

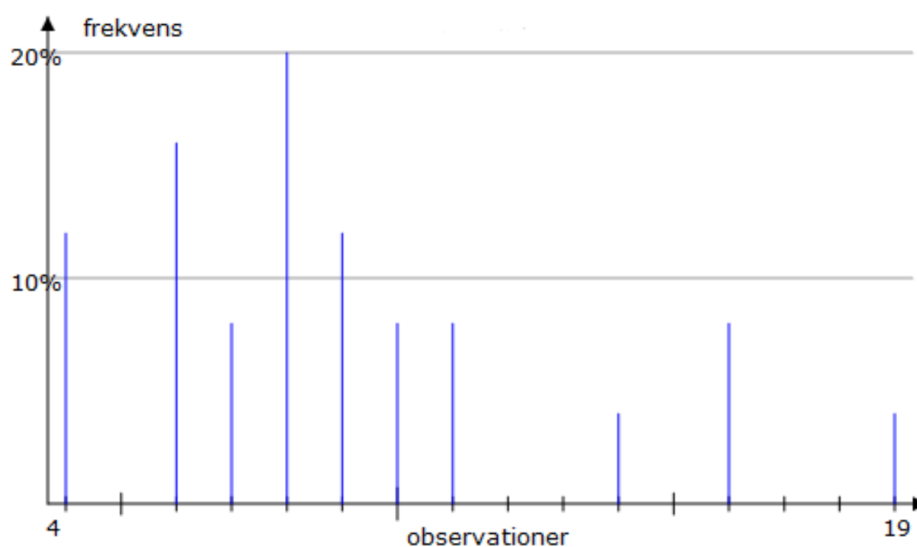
Ofte vil man visualisere observationerne, som giver et overblik over observationssættet enten med et prikdiagram eller et stolpediagram over det ugrupperede observationssæt.

For vores observationssæt ser prikdiagrammet således ud:



Prikdiagrammet illustrerer tallene i de to første søjler i tabellen – altså de forskellige observationer med de tilhørende hyppigheder. Hver prik repræsenterer én af de 25 observationer, og prikkerne stables, når der er flere ens observationer. Prikdiagrammet giver en visuel fornemmelse af, hvordan observationerne er fordelt i intervallet fra 4 til 19 lande. Vi kan se, at observationerne samler sig op i venstre del af intervallet, mens de ligger mere spredt i slutningen.

Alternativt kan vi lave et *stolpediagram*, men her skal der også tegnes en andenakse, hvor hyppigheden eller frekvensen er angivet. I stolpediagrammet nedenfor kan vi aflæse frekvenserne:



Opgave 2

Til et fodboldhold skal der indkøbes fodboldstøvler til 20 spillere. Træneren spørger derfor hver enkelt spiller, hvilken skostørrelse han har. Her er resultatet:

44, 42, 47, 44, 43, 43, 45, 42, 44, 42, 43, 44, 50, 45, 43, 41, 47, 42, 46, 44

- Lav en tabel med observationer, hyppigheder, kumulerede hyppigheder, frekvenser og kumulerede frekvenser.
- Tegn et prikdiagram over observationssættet.
- Tegn et stolpediagram over observationssættet.
- Hvor mange procent af spillerne har en skostørrelse på højst 42?

- e) Hvor mange procent af spillerne har en skostørrelse på mindst 45?
- f) Hvor mange procent af spillerne har en skostørrelse på mindst 43 og højst 46?

Opgave 3

En 2 HF-klasse med 25 elever, der har matematik på B niveau, har været til matematikprøve og har fået følgende karakterer:

00, 00, 02, 02, 02, 02, 4, 4, 4, 4, 7, 7, 7, 7, 7, 7, 10, 10, 10, 10, 10, 10, 10, 10, 12, 12

- a) Overvej, hvorfor denne stikprøve ikke med rimelighed kan bruges til at sige noget om f.eks. alle HF-elever med matematik på B-niveau i Danmark og deres karakterer?
- b) Lav en tabel med observationer, hyppigheder, kumulerede hyppigheder, frekvenser og kumulerede frekvenser.
- c) Tegn et stolpediagram over observationssættet.

Deskriptorer

Tabeller og prikdiagrammer giver et godt overblik over et ugrupperet talmateriale. Alligevel savner man noget. Der er behov for nogle tal, som beskriver observationssættet som helhed, og som kan bruges ved sammenligninger med andre tilsvarende undersøgelser. I den deskriptive statistik kaldes sådanne tal for deskriptorer (= beskrivende tal).

De deskriptorer, der er tale om, er:

- observationssættets *størrelse*
- den *mindste observation* (Min)
- den *største observation* (Max)
- variationsbredden (Max-Min).

For vores observationssæt er:

- observationssættets *størrelse* = 25 (*der er 25 observationer*)
- Min = 4 (den mindste observation)
- Max = 19 (den største observation)
- variationsbredden $19 - 4 = 15$

Isoleret set beskriver disse tal ikke observationssættet fyldestgørende. Tallene fortæller fx ikke noget om, hvordan de 25 observationer er fordelt i intervallet fra 4 til 19 lande. Fx er observationerne jævnt fordelt? Eller er der flest i begyndelsen, i midten eller i slutningen af intervallet? Osv.

Middelværdi

Vi skal se på endnu en deskriptor for et observationssæt, nemlig *middelværdien* (også kaldet *gennemsnittet*). Den bestemmes som summen af alle observationerne delt med antallet af observationer:

$$\text{middelværdi} = \bar{x} = \frac{4 + 4 + 4 + 6 + \dots + 19}{25} = \frac{224}{25} = 8,96$$

Eleverne har altså i gennemsnit besøgt 8,96 lande. Bemærk at vi bruger symbolet $\bar{x}$ for middelværdi.

Bemærk, at middelværdien ikke er det samme som medianen (denne vil blive beskrevet i næste afsnit).

Middelværdien kan også beregnes ud fra frekvenserne (der her er skrevet som et decimaltal):

$$\text{middelværdi} = \bar{x} = 0,12 \cdot 4 + 0,16 \cdot 6 + 0,08 \cdot 7 + \dots + 0,04 \cdot 19 = 8,96$$

Læg mærke til, at leddene efter første lighedstegn (fx $0,12 \cdot 4$) er af typen "frekvens gange observation". De forskellige observationer indgår alle på denne måde sammen med deres frekvens. I vores talmateriale er der mange forskellige frekvenser, så nogle observationer har altså større betydning end andre.

Kvartiler

Til at beskrive fordelingen af observationerne har vi tre andre deskriptorer, der opdeler observationssættet i kvarte. Disse kaldes

- nedre kvartil (Q_1) (som udgør de laveste 25% af observationerne)
- medianen (m) (som udgør den midterste værdi af observationerne)
- øvre kvartil (Q_3) (som udgør de højeste 25% af observationerne)

For at bestemme dem stiller vi samtlige observationer op i voksende rækkefølge:

Vores datasæt var:

8, 4, 16, 8, 9, 6, 16, 19, 7, 6, 4, 8, 11, 8, 9, 6, 9, 10, 11, 8, 14, 4, 6, 7, 10

Disse skal nu opskrives i voksende rækkefølge (husk alle observationerne - tæl efter):

4, 4, 4, 6, 6, 6, 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 11, 11, 14, 16, 16, 19

Medianen er det midterste tal i rækken. Da der er 25 observationer, er medianen altså observationen på den 13. plads:

Median = $m = 8$

I observationssæt, hvor der er et lige antal observationer, er medianen gennemsnittet af de to midterste observationer.

Nedre kvartil defineres som det midterste tal (medianen) af observationssættets "nederste halvdel", dvs. af de 12 nederste observationer i vores tilfælde. Nedre kvartil er altså gennemsnittet af observationerne på plads nr. 6 og nr. 7:

$$\text{nedre kvartil} = Q_1 = \frac{6 + 6}{2} = 6$$

Øvre kvartil defineres som medianen af observationssættets "øverste halvdel", der her består af de 12 øverste observationer. Øvre kvartil er altså gennemsnittet af observationerne på plads nr. 19 og nr. 20:

$$\text{øvre kvartil} = Q_3 = \frac{10 + 11}{2} = 10,5$$

Ud fra Q_1 og Q_3 kan vi beregne kvartilbredden:

$$\text{Kvartilbredde} = KB = Q_3 - Q_1 = 10,5 - 6 = 4,5$$

kvartilsæt:

De tre kvartiler kan sammenfattes i kvartilsættet:

$$\text{kvartilsæt} = (Q_1, m, Q_3) = (6, 8, 10,5)$$

Kvartilsættet fortæller her, at ca. 25% af observationerne er 6 eller mindre, at ca. halvdelen af observationerne er 8 eller mindre, og at ca. 75% af observationerne er 10,5 eller mindre.

Vi kan også lave et udvidet kvartilsæt, hvor min og max er med:

$$\text{udvidet kvartilsæt} = (Min, Q_1, m, Q_3, Max) = (4, 6, 8, 10,5, 19)$$

Outlier

En *outlier* er en observation, der ligger mere end 1,5 gange kvartilbredden fra nedre kvartil eller mere end 1,5 gange kvartilbredden større end øvre kvartil.

$$Q_1 - 1,5 \cdot (Q_3 - Q_1)$$

$$Q_3 + 1,5 \cdot (Q_3 - Q_1)$$

I vores observationssæt har vi:

$$Q_1 - 1,5 \cdot (Q_3 - Q_1) = 6 - 1,5 \cdot 4,5 = 6 - 6,75 = -0,75$$

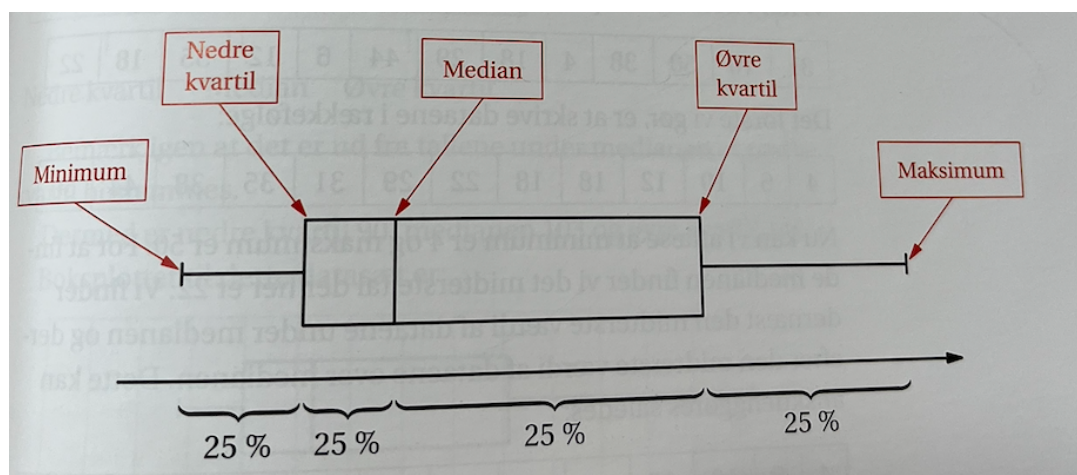
$$Q_3 + 1,5 \cdot (Q_3 - Q_1) = 10,5 + 1,5 \cdot 4,5 = 10,5 + 6,75 = 17,25$$

En observation under $-0,75$ eller over $17,25$ er dermed en outlier.

I vores datasæt har vi observationen 19 som den eneste outlier i observationssættet.

Boksplot

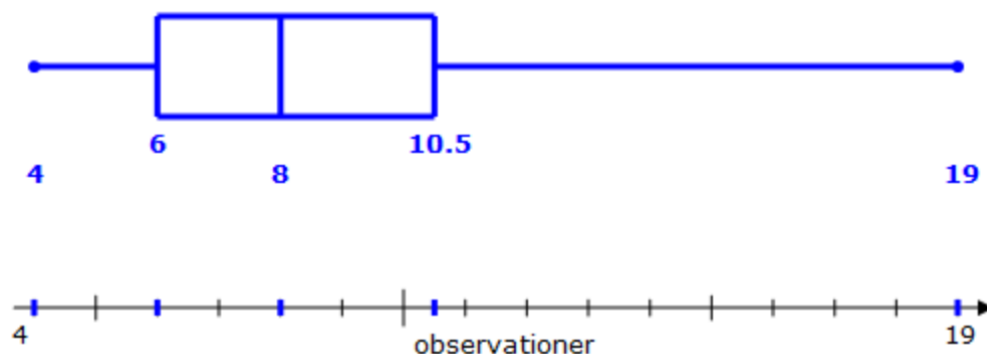
Et boksplot er en grafisk måde at beskrive et datasæt på. Et boksplot indeholder: Min, Max, medianen og nedre kvartil og øvre kvartil.



s. 247 i Matema10k-bogen

Nederst i boksplottet tegnes en tallinje, hvorpå min, max og de tre kvartiler kan aflæses.

For vores eksempel ser datasætte ud på følgende måde:



Vi kan derved aflæse følgende ud fra boksplottet:

Min = 4

Max = 19

Nedre kvartil = 6

Median = 8

Øvre kvartil = 10,5

Selve kassen i boksplottet afgrænses af nedre og øvre kvartil - den markerer altså de 50% midterste observationer (fra 25% til 75% af observationerne). Inde i kassen er medianen markeret med en lodret streg.

Fra kassens sider tegnes to linjestykker ud til Min og Max.

Boksplot bruges ofte til at sammenligne datamateriale.

Opgave 4

Betragt følgende 3 observationssæt:

- 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
- 1, 3, 5, 5, 8, 8, 9, 9, 9, 10
- 1, 2, 2, 2, 3, 3, 6, 6, 8, 10

1. Bestem for hvert af observationssættene: Størrelse, Min, Max variationsbredde, kvartilsæt, kvartilbredde og middelværdi.
2. Tegn boksplottene for de tre observationssæt.

Opgave 5

På en arbejdsplads med 15 ansatte har lønforhandlinger betydet, at de ansattes månedsløn er blevet ændret med følgende størrelser (tallene er i tusinde kroner):

-2, -2, -1, -1, -1, -1, -1, -1, 0, 0, 0, 3, 6, 6, 10

Et negativt tal angiver en lønnedgang, og 0 betyder, at lønnen er uændret.

1. Bestem observationsættets middelværdi og median.
2. Hvilket af de to tal, synes du, beskriver observationssættet bedst? Begrund dit svar.
3. Bestem kvartilbredden og find eventuelle outliers.

Opgave 6

Tag udgangspunkt i observationssættet fra opgave 3.

- a) Opstil et kvartilsæt for karakterfordelingen
- b) Tegn et boksplot.

Opgave 7

Vi ser igen på skostørrelserne for 20 fodboldspillere. Observationssættet er her opstillet i numerisk rækkefølge:

41, 42, 42, 42, 42, 43, 43, 43, 43, 44, 44, 44, 44, 44, 45, 45, 46, 47, 47, 50

1. Bestem størrelse, Min, Max, variationsbredde, kvartilsæt, kvartilbredde og middelværdi for observationssættet.
2. Er der nogen outliers?
3. Tegn et boksplot for observationssættet.

Fra jeres formelsamling

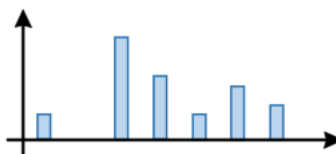
18

STATISTIK

Ugrupperede observationer

Søjlediagram

(59)

Observationen $\min$

(60)

 $\min$: mindste observationObservationen $\max$

(61)

 $\max$: største observationVariationsbredden VB

(62)

 $VB = \max - \min$ Medianen m

(63)

Observationerne sorteres efter størrelse.

Hvis antallet af observationer er ulige, så er m den midterste observation.

Eksempel:

Observationssæt: 1, 4, 5, 9, 14

Medianen er $m = 5$

Hvis antallet af observationer er lige, så er m gennemsnittet af de to midterste observationer

Eksempel:

Observationssæt: 0, 2, 3, 6, 8, 12

Medianen er $m = \frac{3+6}{2} = 4,5$ Nedre kvartil Q_1

(64)

Q_1 er medianen for de observationer, der ligger før medianen m

Øvre kvartil Q_3

(65)

Q_3 er medianen for de observationer, der ligger efter medianen m

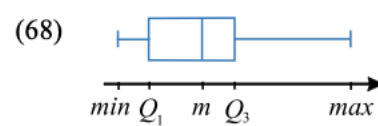
Kvartilbredden KB

$$(66) \quad KB = Q_3 - Q_1$$

Outlier

(67) En outlier er en observation, der er mindre end $Q_1 - 1,5 \cdot KB$ eller større end $Q_3 + 1,5 \cdot KB$

Boksplot



Kvartilsæt

$$(69) \quad (Q_1, m, Q_3)$$

Udvidet kvartilsæt

$$(70) \quad (min, Q_1, m, Q_3, max)$$

Middelværdi (gennemsnittet) $\bar{x}$ for observationssættet $x_1, x_2, \dots, x_n$

$$(71) \quad \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$