Statistik

Statistik er analyse af indsamlet data. Det vil sige at man bearbejder et datamateriale som i matematik næsten altid er tal.

Derved får man et samlet overblik over talmaterialet, og man kan konkludere noget vedr. dette talmateriale.

Rundt omkring i samfundet bliver statistik meget ofte brugt som baggrund for forskellige beslutninger. Derfor er statistik også en vigtig del af matematik i skolen.

Eksempel
Alberte dyrker gymnastik i gymnastikforeningen ”De muntre badutspringere”.

I tabellen her til højre kan man se alderen for de 40 medlemmer i foreningen .

Tabellen er opstillet i regnearket Excel som er godt arbejdsredskab når der skal arbejdes med et datamateriale i statistik.

	A	B	C
1	**Medlemsnr.**	**Alder**	
2	19901	23	
3	19902	22	
4	19903	23	
5	19904	20	
6	19905	22	
7	19906	22	
8	19907	23	
9	19908	22	
10	19909	21	
11	19910	25	
12	19911	25	
13	19912	21	
14	19913	19	
15	19914	21	
16	19915	24	
17	19916	19	
18	19917	22	
19	19918	21	
20	19919	24	
21	19920	25	
22	19921	25	
23	19922	27	
24	19923	21	
25	19924	25	
26	19925	26	
27	19926	20	
28	19927	24	
29	19928	22	
30	19929	20	
31	19930	26	
32	19931	23	
33	19932	22	
34	19933	22	
35	19934	19	
36	19935	24	
37	19936	22	
38	19937	19	
39	19938	27	
40	19939	19	
41	19940	22	
42			

Observationer

Datamaterialet i en statistisk undersøgelse er som nævnt knyttet til en række værdier der som regel er forskellige tal. Disse tal kalder man for *observationer*.

Observationerne i eksemplet ”De muntre badutspringere” er alderen på medlemmerne i gymnastikforeningen, og altså ikke deres medlemsnummer.

Alle observationer udgør tilsammen et *observationssæt*.

Den statistiske værktøjskasse

I en statistisk analyse er der rigtig mange måder at bearbejde observationerne på, og her får vi brug for en statistisk værktøjskasse. Hvis man skal arbejde rigtigt med statistik, skal man både vide hvordan og hvornår man skal bruge de forskellige værktøj, og alt dette lærer man bedst gennem træning.

Herunder gennemgås det vigtigste værktøj som man vil få brug for i grundlæggende statistik, og som er grundlag for statistik på højere niveau.

Typetallet *Typetallet* er det tal som er "typisk" for observationssættet.

- Det vil sige den observation som forekommer flest gange i observationssættet.
- Der kan godt forekomme flere typetal i et observationssæt.

Størsteværdi *Størsteværdien* er den største observation i observationssættet.

- NB. Ikke det største antal gange en observation forekommer

Mindsteværdi *Mindsteværdien* Den mindste observation i observationssættet.

- NB. Ikke det mindste antal gange en observation forekommer

Variationsbredde *Variationsbredden* er forskellen på størsteværdi og mindsteværdi

- Variationsbredden = størsteværdien - mindsteværdien.

Gennemsnittet *Gennemsnittet* er det tal som man får hvis man lægger alle observationer sammen og dividerer dette tal med det samlede antal observationer.

- Gennemsnittet kaldes også for **middeltallet**

Medianen *Medianen* er den observation som står i midten hvis man stiller alle observationer op i rækkefølge med de mindste tal først. Hvis der er et lige antal observationer, så der ikke er et tal i midten, er medianen tallet til venstre for midten.

- Medianen hedder også *2. kvartil* eller *0,50-kvartil*.

Kvartilsæt *Kvartilsættet* består af 1. kvartil, 2. kvartil og 3. kvartil til observationssættet

- *1. kvartil eller 0,25-kvartilen* eller *nedre kvartil* kaldes sådan fordi det er her de første 25% af observationerne ligger indenfor hvis observationerne sættes i rækkefølge med de mindste først.
- *2. kvartil eller 0,50-kvartilen* eller *median* kaldes sådan fordi det er her de første 50 % af observationerne ligger indenfor hvis observationerne sættes i rækkefølge med de mindste først.
- *3. kvartil eller 0,75-kvartilen* eller *øvre kvartil* kaldes sådan fordi det er her de første 75% af observationerne ligger indenfor hvis observationerne sættes i rækkefølge med de mindste først.

	A	B
1	**Medlemsnr.**	**Alder**
2	19913	19
3	19916	19
4	19934	19
5	19937	19
6	19939	19
7	19904	20
8	19926	20
9	19929	20
10	19909	21
11	19912	21
12	19914	21
13	19918	21
14	19923	21
15	19902	22
16	19905	22
17	19906	22
18	19908	22
19	19917	22
20	19928	22
21	19932	22
22	19933	22
23	19936	22
24	19940	22
25	19901	23
26	19903	23
27	19907	23
28	19931	23
29	19915	24
30	19919	24
31	19927	24
32	19935	24
33	19910	25
34	19911	25
35	19920	25
36	19921	25
37	19924	25
38	19925	26
39	19930	26
40	19922	27
41	19938	27

Hyppighed - h(x) *Hyppigheden* angiver hvor ofte (hyppigt) de forskellige observationer forekommer. Det er altså det antal gange at en observation forekommer.
Normalt angiver man hyppigheden med "h(x)"
Hyppighederne til de enkelte observationer sættes ind i en tabel som vist her nedenunder. Igen er det en god hjælp at bruge regneark.

Observationer	h(x)	f(x)	H(x)	F(x)
19	5	12,5	5	12,5
20	3	7,5	8	20
21	5	12,5	13	32,5
22	10	25	23	57,5
23	4	10	27	67,5
24	4	10	31	77,5
25	5	12,5	36	90
26	2	5	38	95
27	2	5	40	100
	40			

Tabellen ovenover er for vores eksempel med statistisk analyse af aldersfordelingen i foreningen "De muntre badutspringere". Af tabellen læses at der 5 medlemmer der er 19 år, 3 er 20 år, 5 er 21 år, osv.

Frekvens - f(x) *Frekvensen* er den procentdel som hyppigheden af hver observationen forekommer med i forhold til det samlede antal observationer. Frekvensen skrives "f(x)"
Frekvensen findes ved at dividere hyppighed med det samlede antal observationer for derefter at gange med 100.

I tabellen over "De muntre badutspringere" findes frekvensen til 19 år ved at dividere 5 med 40 og gange med 100, dvs. f(x) = 5 / 40 * 100 = 12,5 %

Summeret hyppighed - H(x) Den *summerede hyppighed* er hyppighederne lagt sammen med (summeret) de foregående hyppigheder når der begyndes nede fra med den mindste observation. Den summerede hyppighed skrives "H(x)"

I tabellen ovenfor findes H(21) ved at summere hyppighederne for 19, 20 og 21 år.
H(19 =h(19)=5, H(20)=h(19)+h(20)=5+3=8, H(21)=h(19)+h(20)+h(21)=5+3+5=13
H (21) bestemmes lettere på følgende måde: H(21) = H(20) + h(21) = 8+5 = 13

Summeret frekvens - F(x) Den *summerede frekvens* udregnes på samme måde som summeret hyppighed, men her et det bare frekvenserne der skal lægges sammen nede fra den mindste observation og opefter. Den summerede frekvens skrives "F(x)".

I tabellen ovenfor er F(19) = f(19) = 12,5, F(20) = f(19) + f(20) = 12,5 + 7,5 = 20,
F(21) = f(19) + f(20) + f(21) = 12,5 + 7,5 + 12,5 = 32,5.
En lettere måde at beregne F(21): F(21) = F(20) + f(21) = 20 + 12,5 = 32,5

Grupperede observationer

I en del statistisk analyse er det en fordel at dele observationerne ind i grupper. Hvis man fx skulle lave en statistik over en skoleklasse med 25 elever som springer længdespring i en idrætstime, så vil man godt kunne få 25 forskellige resultater hver med en hyppighed på 1. Enkeltobservationer vil i dette tilfælde give et dårligt overblik over datamaterialet. Derfor vil det her være hensigtsmæssigt at datamaterialet bliver inddelt i grupper. F.eks. 0-1 meter, 1-2 meter osv. Disse grupper kalder man i statistik for ***intervaller.***

I eksemplet ovenfor med "De muntre badutspringere" er der lavet statistik på baggrund af enkelt-observationer hvor observationerne altså ikke er inddelt i intervaller. Nedenfor vil vi bruge det samme observationssæt, hvor observationerne nu blot være inddelt i intervaller.

- Ved grupperede observationer vil man ofte ikke kunne finde hverken typetal, størsteværdi, mindsteværdi og variationsbredde, hvis man ikke kender de enkelte observationer. I nogle sammenhænge kan man dog tale om ***typeinterval*** som det interval der har den største hyppighed. Man kan også bestemme et gennemsnit og kvartilerne for observationssættet, og det skal vi se på nedenfor.

- Ved grupperede observationer bruges **firkantede parenteser** omkring intervallerne "[" og "]". Disse parenteser angiver om tallet er med eller ej. Hvis parentesen vender ind mod tallet, er tallet med i intervallet. Vender parentesen væk fra tallet, er tallet ikke med i intervallet, men tallene op til tallet er med.

  Eksempel: I intervallet [2;4[er tallet 2 med og så er tallene op til 4 også med. Det vil sige at tallet 3,9 er med i intervallet, men tallet 4 er ikke med. Sprogligt benævnes intervallet som "fra og med 2 og til og ikke med 4".

Interval	max.	Intervalmidt	h(x)	f(x)	H(x)	F(x)	Intervalmidt*h(x)
	19						
19-21	21	20	8	20	8	20	160
21-23	23	22	14	35	22	55	308
23-25	25	24	9	22,5	31	77,5	216
25-27	27	26	7	17,5	38	95	182
27-29	29	28	2	5	40	100	56
			40				922

Gennemsnit

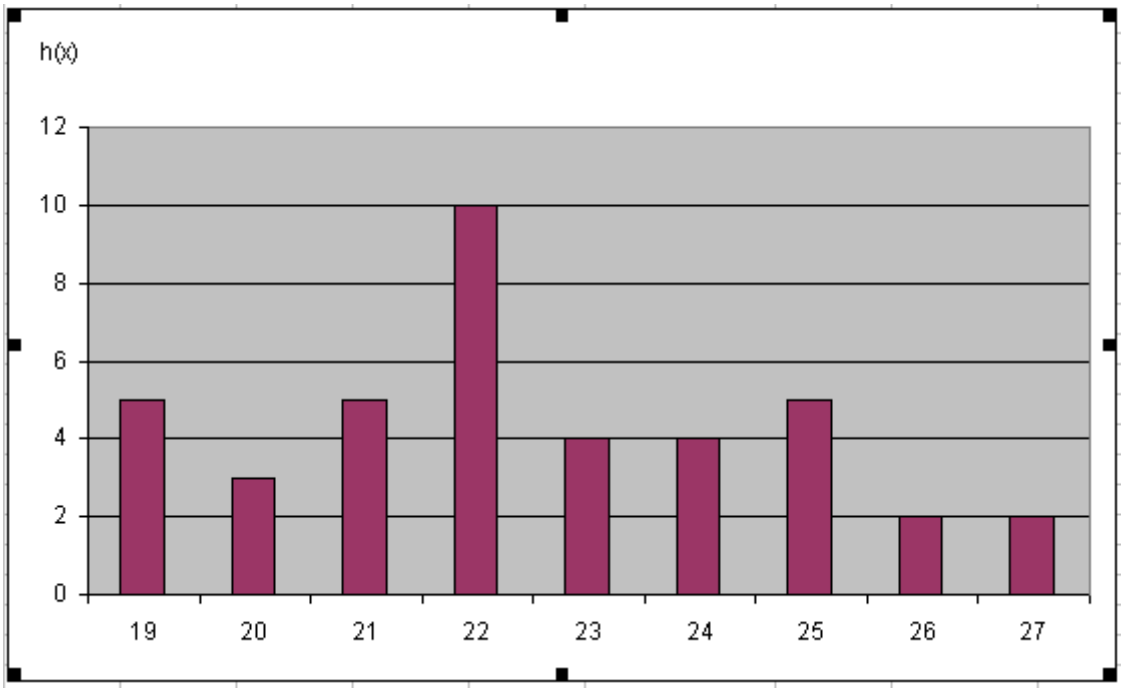
Hvis vi skal finde gennemsnittet af observationer inddelt i intervaller (og hvor vi ikke kan finde tilbage til de oprindelige observationer), skal vi i første omgang finde **intervalmidtpunktet.** Det vil sige at vi finder den midterste værdi i intervallet. Hvis et interval går fra 0 til 10, så er midtpunktet 5. Vi finder intervalmidtpunktet fordi vi ikke ved hvordan observationerne fordeler sig i intervallet. Derfor går vi ud fra at observationerne fordeler sig jævnt omkring midten af intervallet. Det kan vi dog ikke være sikker på at de gør, og derfor er gennemsnittet beregnet på denne måde forbundet med en vis usikkerhed.

Hvis vi havde kendt alle observationer, ville vi lægge dem sammen og så til sidst dividere med det samlede antal observationer. Faktisk gør vi lidt det samme, når vi har observationerne inddelt i intervaller. Her regner vi blot med at alle observationer i et interval er lig med intervalmidtpunktet.. Eksempel: Hvis intervalmidtpunktet er 5 og hyppigheden af intervallet er 3, så svarer det til at vi har 3 observationer der alle har værdien 5. Dvs. dette interval bidrager med i alt $3 \cdot 5 = 15$ til udregningen af gennemsnittet. Dette tal skrives ind i tabellen i en ny kolonne der benævnes med Intervalmidt $\cdot$ h(x). Alle tal i denne kolonne lægges til sidst sammen, og gennemsnittet fås nu ved at dividere dette resultat med det samlede antal observationer (og ikke antallet af intervaller).

Diagrammer

- **Pindediagram, stolpediagram, søjlediagram**
  Pindediagram, stolpediagram, søjlediagram er blot forskellige navne for samme diagramtype. Kan anvendes som diagram der viser hyppigheder eller frekvenser for enkeltobservationer. Man skal dog lægge mærke til at der kun står ét tal under hver ”pind”, ”stolpe” eller ”søjle”.

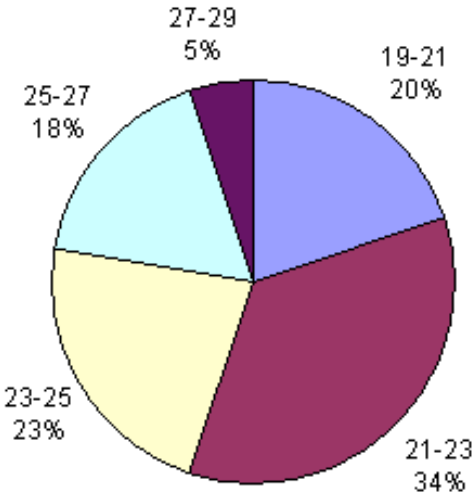
  Nedenfor er vist et stolpediagram for hyppighederne for enkeltobservationer i vores eksempel med ”De muntre badutspringere”.



- **Cirkeldiagram**

  Kan anvendes som diagram til at vise hyppigheder og frekvenser ved såvel enkeltobservationer som grupperede observationer.

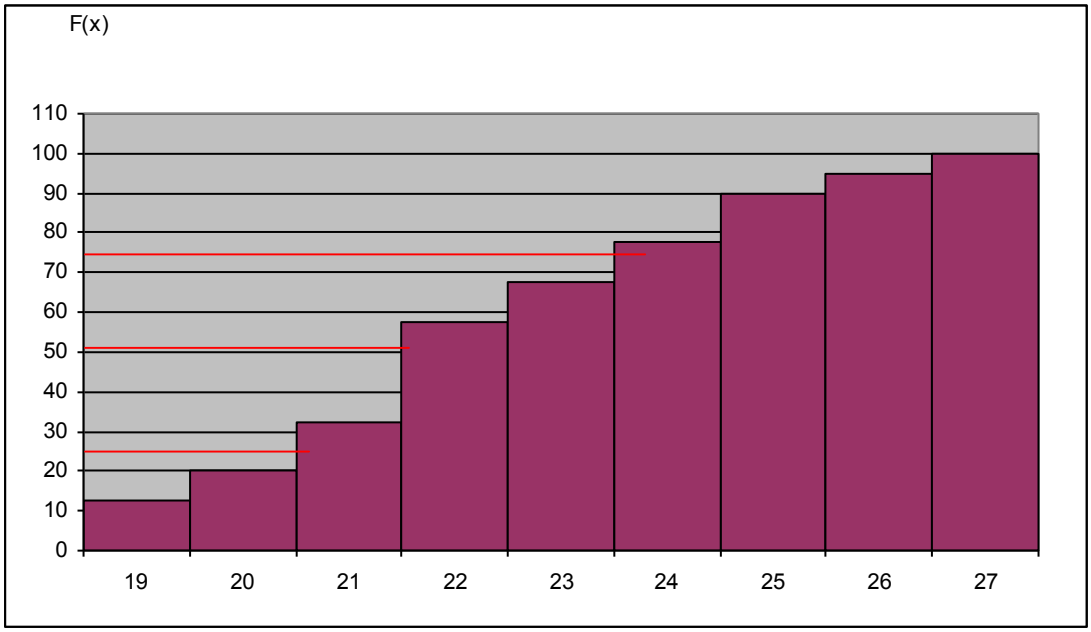
  Her nedenfor er vist et cirkeldiagram for frekvenser ved de grupperede observationer i vores eksempel ”De muntre badutspringere”.



- **Trappediagram**

  Ved enkeltobservationer bruger man et trappediagram hvis man skal lave et diagram over den summerede frekvens (eller den summerede hyppighed).

  Et trappediagram kan bl.a. bruges til direkte at aflæse kvartilsættet, dvs. 0,25-kvartilen, O,50-kvartilen og 0,75-kvartilen.

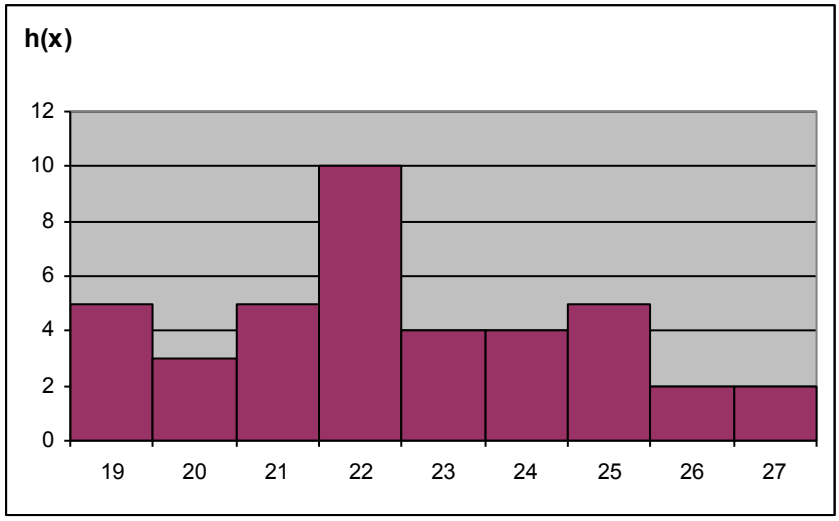


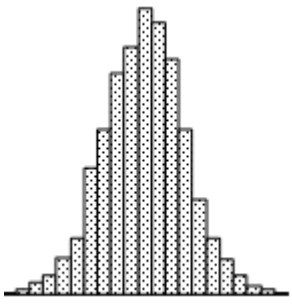
Her ovenover ses et trappediagram over enkeltobservationer for "De muntre badutspringere".
0,25-kvartilen aflæses til 21, dvs. 25 % af medlemmerne er 21 år og derunder.
0,50-kvartilen (medianen) aflæses til 22, dvs. 50 % af medlemmerne er 22 år og derunder.
0,75-kvartilen aflæses til 24, dvs. 75 % af medlemmerne er 24 år og derunder, og 25 % er over 24 år.

- **Histogram**

  Et histogram fremstilles som søjlediagrammer der står helt tæt sammen uden mellemrum.
  Histogrammet anvendes ved såvel enkeltobservationer som grupperede observationer til at vise hvordan hyppighederne fordeler sig, bl.a. for at afgøre om der er tale om en normalfordeling.
  Ved en normalfordeling findes der en tilnærmet symmetriakse i midten af histogrammet.

  Her nedenunder ses et histogram for enkeltobservationer til eksemplet "De muntre badutspringere", og det ses at der ikke er tale om en tydelig normalfordeling.





En tydelig normalfordeling

- **Sumkurve**

Ved afsætning af en sumkurve er det den summerede frekvens F(x) der afsættes op ad y-aksen. Ved enkeltobservationer kunne vi finde medianer og kvartiler ved at kigge på observationssættet eller skemaet. Det er ikke så let ved de grupperede observationer. Her er man nødt til at tegne en sumkurve og aflæse på grafen.

Kvartilsættet aflæses i diagrammet nedenunder til:
0,25 kvartilen: 21,3 år
0,50-kvartilen (medianen): 22,7 år
0,75-kvartilen: 24,8 år

Sprogligt udtrykkes det på følgende måde:

- 25 % af medlemmerne er 21,3 år og derunder.
- Halvdelen af medlemmerne er 22,7 år og derunder, og dvs. at den anden halvdel af medlemmerne er over 22,7 år.
- Endelig er 75 % af medlemmerne 24,8 år og derunder, og dvs. at 25 % er ældre end 24,8 år.