

# Statistik



©Erik Vestergaard



## 1. Grupperede observationer

I statistik beskæftiger man sig med *indsamling*, *bearbejdelse* og *fortolkning* af data. Data består af en række såkaldte *observationer*. Disse observationer kan være af *kvalitativ* eller *kvantitativ* art. Hvis det er en afstemning om EU, så er observationerne Ja eller Nej, dvs. kvalitative. Hvis det er data for soldaters vægt, så er der tale om tal, altså data af kvantitativ art. Lad os sige, at vi har indsamlet data for, hvor mange børn, hver deltager på et givet voksenkursus har. Her vil det være hensigtsmæssigt at tælle op, hvor mange personer, som har 0 børn, hvor mange som har 1 barn, hvor mange der har 2 børn etc. Hvor mange, der er af hver observation, udgør de såkaldte *hyppigheder*. Hvis man dividerer hyppighederne med antallet af personer, fås *frekvenserne*. En smart og overskuelig måde at præsentere data på er herefter at lave det velkendte *pindediagram*, hvor man over hver enkel observation på x-aksen har afbildet en linje, med en længde, som er proportional med frekvensen af observationen. Vi vil gå ud fra, at pindediagrammer allerede er kendt fra folkeskolen. Dog vil vi kort vende tilbage til dette emne i afsnittet *ugrupperede observationer* i afsnit 2.

Imidlertid er der mange situationer, hvor det er uhensigtsmæssigt at lave pindediagrammer. Antag for eksempel, at man har gennemført en multiple choice skriftlig køreprøve med 75 spørgsmål, hvor hver enkelt svar enten er rigtigt eller forkert. Man tæller antallet af rigtige op for hver prøvedeltager, og er interesseret i at se, hvordan scorerne fordeles sig på holdet. Her vil det *ikke* være fornuftigt at lave et pindediagram, med en pind for hver score fra 0 til 75. Det vil være uoverskueligt – der er simpelthen for mange observationer! Et andet eksempel er data over vægten af æbler fra en frugtproducent. Der vil være store og små æbler. Måske et, som vejer 165,54g, et andet der vejer 201,13g. Det vil være helt usandsynligt, at to æbler vejer præcist det samme. Derfor vil det være ganske ufornuftigt at lave et pindediagram her. Om det første eksempel kan man sige, at her er mængden af mulige observationer *endelig* (*diskret fordeling*), men mængden er for stor til, at det er fornuftigt at anvende et pindediagram. I det andet eksempel kan et helt interval af tal forekomme som vægt for et givet æble (*kontinuert fordeling*). Derfor vil der i praksis højst være ét æble med en given vægt. Igen upraktisk at bruge et pindediagram.

Løsningen på problemet med præsentationen af data er at *gruppere* observationerne, altså samle observationerne i intervaller og tælle, hvor mange observationer, der er i hvert interval. Der er en række begreber i forbindelse med grupperede observationer, og de lader sig lettest forklare via et eksempel.

### Eksempel 1

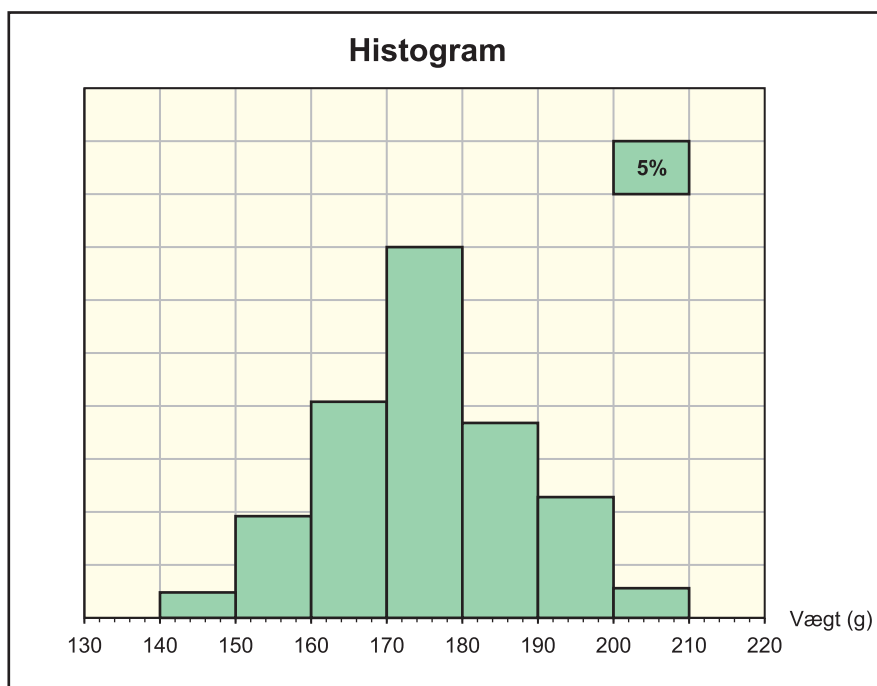
I et gartneri har man på tilfældig vis udvalgt 500 æbler for at undersøge, hvordan vægten af æblerne fordeles sig. Man vælger at inddele i vægt-intervaller af længden 10 gram fra 140 gram til 210 gram. Tabellen på næste side viser intervallhyppighederne. *Frekvenserne* angiver, hvor stor en brøkdel, at æblerne, som befinder sig i det pågældende interval. Frekvensen i intervallet  $]140,150]$  fås således som  $12/500 = 0,024$ . Frekvensen

kan angives som kommatallet 0,024 eller i procent som 2,4%. Dernæst definerer man de såkaldte *kumulerede frekvenser*. Den kumulerede frekvens for intervallet  $]170,180]$  hører til intervallets *højre* endepunkt 180 og angiver, hvor stor en del af æblerne, som har en vægt, som er *mindre end* 180 gram. Det kan udregnes ved at addere alle frekvenserne ”bagud”:  $0,024 + 0,096 + 0,204 + 0,350 = 0,674$ . Da den forrige kumulerede frekvens 0,324 imidlertid repræsenterer summen af alle frekvenserne fra 170 og bagud, så er det nemmere at udregne den nye kumulerede frekvens som summen af den forrige kumulerede frekvens og den nye frekvens:  $0,324 + 0,350 = 0,674$ . Vi skal senere se, hvad man kan bruge de kumulerede frekvenser til.

Figur 1

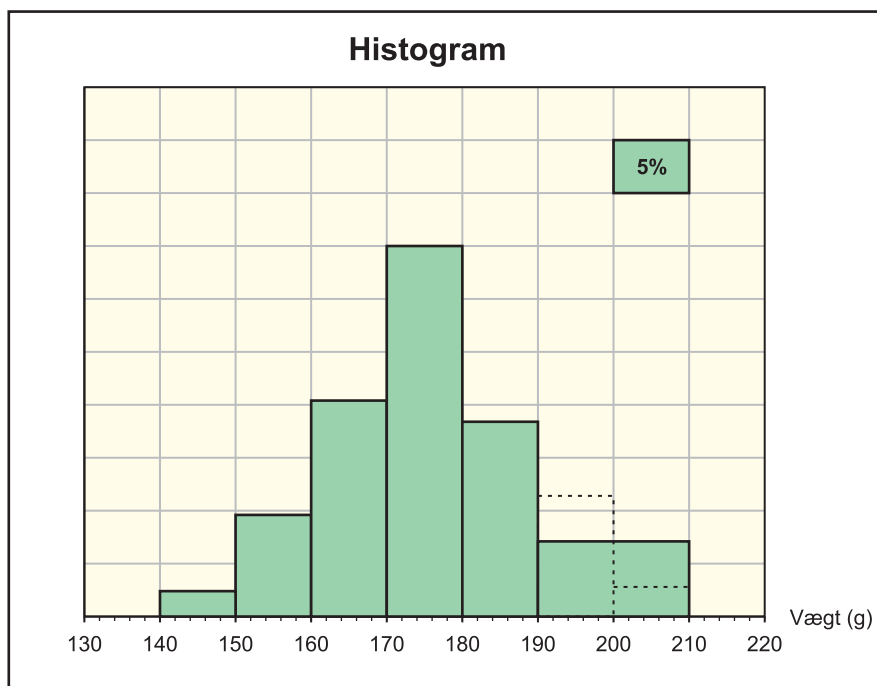
Vægt (gram)	Hyppighed	Frekvens	Kumuleret frekvens
$]140,150]$	12	0,024	0,024
$]150,160]$	48	0,096	0,120
$]160,170]$	102	0,204	0,324
$]170,180]$	175	0,350	0,674
$]180,190]$	92	0,184	0,858
$]190,200]$	57	0,114	0,972
$]200,210]$	14	0,028	1,000
I alt: 500			

Figur 2



*Histogrammet* på figur 2 viser fordelingen af æblernes vægt. Idéen er, at søjlernes *arealer* skal være proportionale med de tilhørende intervalfrekvenser. Man tegner et lille reference-areal, som skal repræsentere en valgt procentsats og tegner herefter hver søjle med en højde, så dens areal bliver korrekt i forhold til reference-arealet. På figur 2 er reference-arealet et tern, som er valgt til at svare til 5%. Der er en grund til, at søjlens areal og ikke dens højde skal være et mål for frekvensen. Man kan for eksempel forestille sig, at man samlede æblerne fra de to sidste intervaller i ét interval  $]190, 210]$ . Intervalfrekvensen vil blive summen  $0,114 + 0,028 = 0,142$ . Hvis højden skulle være et mål for frekvensen, så skulle søjlens højde være lig med summen af højderne af de sidste to søjler. Dette vil være misvisende, da der er tale om et større vægtinterval! Rent visuelt ville histogrammet give indtryk af, at der er temmelig mange æbler med stor vægt. I stedet skal vi regne i arealer. I praksis sker der det, at de to søjlers samlede areal ”spredes” ud over hele intervallet  $]190, 210]$ , hvorved det får en højde svarende til gennemsnittet af de to oprindelige søjlers højde. Dette er illustreret på figur 3.

Figur 3



### Et pædagogisk fif

I praksis har man ofte at gøre med intervaller, som næsten alle har samme bredde  $b$ . I dette tilfælde kan det være en fordel at lave en foreløbig  $y$ -akse med frekvensinddeling, afsætte søjlerne med højder i overensstemmelse med frekvenserne på  $y$ -aksen, lave et passende reference-areal med bredde  $b$ , viske  $y$ -aksen ud og endelig tilrette højderne af de søjler, som ikke har bredden  $b$  ...

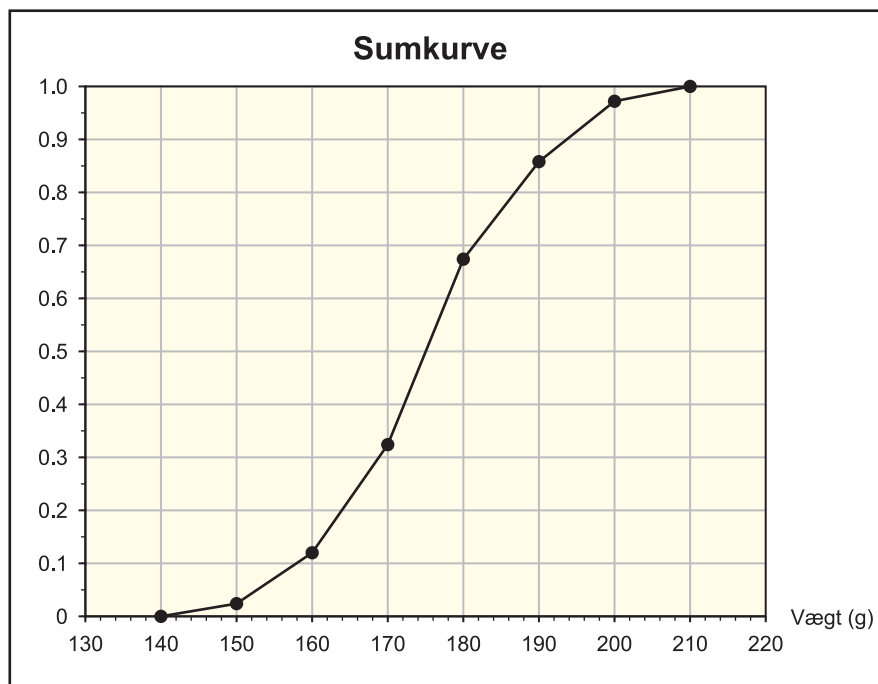
Det næste man kan være interesseret i, er æblernes gennemsnitsvægt. Da vi ikke har præcis viden om hvert æbles nøjagtige vægt, men kun har oplyst antallet af æbler i hvert vægtinterval, så kan vi ikke udregne det korrekte *gennemsnit* eller *middeltal* eller *middeelværdien*, som det også kaldes. Det mest fornuftige, man kan gøre, er at antage, at æblerne er *jævnt fordelte* i hvert interval, hvilket betyder, at gennemsnittet af vægten af æblerne i et givet interval er lig med intervallets midtpunkt. Det forklarer, hvorfor man udregner gennemsnittet  $\mu$  som det *vejede gennemsnit* af intervallernes midtpunkter  $m_i$  med deres respektive frekvenser  $f_i$ :

$$\begin{aligned}
 \mu &= f_1 \cdot m_1 + f_2 \cdot m_2 + \dots + f_n \cdot m_n \\
 &= 0,024 \cdot 145 + 0,096 \cdot 155 + 0,204 \cdot 165 + 0,350 \cdot 175 \\
 &\quad + 0,184 \cdot 185 + 0,114 \cdot 195 + 0,028 \cdot 205 \\
 &= 175,3
 \end{aligned}
 \tag{1}$$

Gennemsnittet kan også beregnes ved at tage det vejede gennemsnit af intervallernes midtpunkter  $m_i$  med deres respektive hyppigheder  $h_i$ , og herefter dividere resultatet med det totale antal observationer (se opgave 2).

I det følgende skal vi betragte en kurve, den såkaldte *sumkurve*, der angiver, hvor stor en del af observationerne, som er *mindre* end en given værdi. I det aktuelle tilfælde skal man altså kunne aflæse, hvor stor en del af æblerne, som har en vægt mindre end en given vægt. Da der ikke er nogen æbler, som vejer mindre end 140 gram, så har vi straks et datapunkt: (140; 0). Det ses umiddelbart, at man får en række datapunkter på kurven ved at tage sammenhørende værdier af intervallernes *højre* endepunkter og de respektive kumulerede frekvenser: (150; 0,024), (160; 0,120), (170; 0,324) osv. De af sættes i et koordinatsystem og forbindes med rette linjestykker, som vist på figur 4.

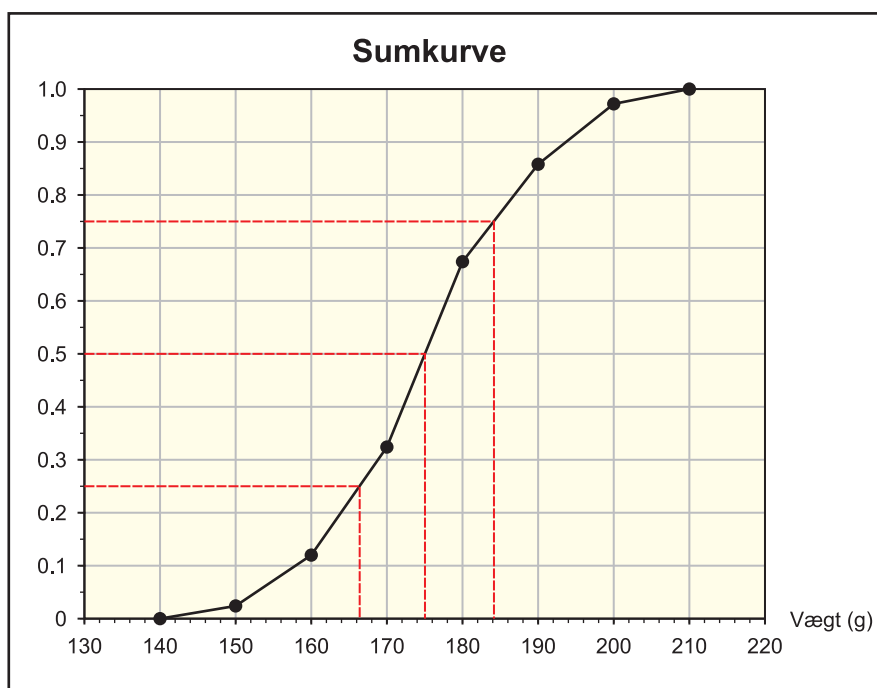
Figur 4



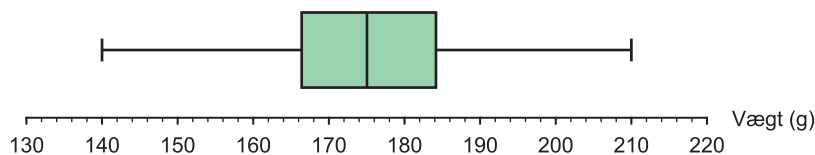
Begrundelsen for at forbinde punkterne med rette linjestykker og ikke tegne en blød kurve igennem dem er igen, at man antager, at observationerne i hvert interval er jævnt fordelte – overvej hvorfor det betyder, at der skal tegnes rette linjestykker? Man indser hurtigt, at en sumkurve altid vil være *voksende*!

På sumkurven kan aflæses tre *kvartiler*: 75%-kvartilen, 50%-kvartilen og 25% kvartilen. De betegnes også henholdsvis *øvre kvartil*, *medianen* og *nedre kvartil*. Tilsammen udgør de datamaterialets *kvartilsæt*. I det aktuelle tilfælde er kvartilerne aflæst til henholdsvis 184,1; 175,0 og 166,4, som det ses på figur 5. Kvartilsættet fortæller os, at der er 75% af æblerne, som har en vægt mindre end eller lig med 184,1 gram, halvdelen har en vægt på mindre end lig med 175,0 gram, og 25% har en vægt, som er mindre end eller lig med 166,4 gram.

Figur 5



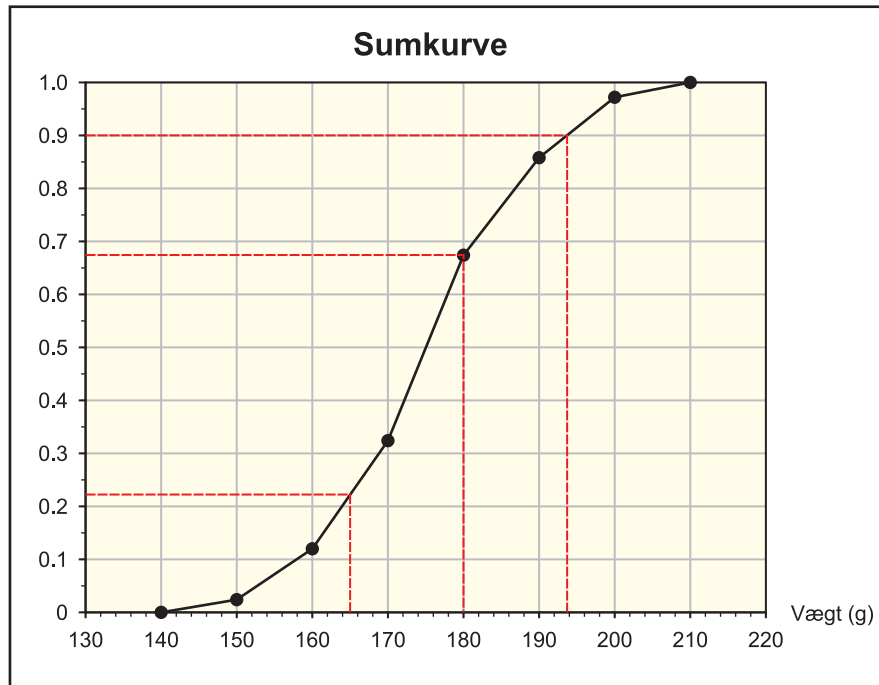
Boxplot



Under sumkurven er afbildet et såkaldt *boxplot* af datamaterialet. Linjens venstre endepunkt angiver den *mindste* observation (140 gram), mens linjens højre endepunkt angiver den *største* observation (210 gram). På linjen er anbragt en *box*, hvis venstre ende angiver nedre kvartil og højre ende viser øvre kvartil. Det lodrette linjestykke i boksen angiver medianen. Box-plottet er en meget visuel måde at præsentere vigtige statistiske *deskriptorer* på, og det fortæller en hel del om datamaterialet. Tilbage til sumkurven. Den kan benyttes til at løse forskellige opgaver. Lad os se på nogle eksempler.

- Hvor mange procent af æblerne har en vægt på *over* 165 gram?
- Hvor mange procent af æblerne har en vægt *mellem* 165 og 180 gram?
- Hvad kan man sige om de 10% tungeste æbler?

Figur 6



*Løsning:* Vi aflæser på sumkurven på figur 6.

- $y$ -værdien svarende til en  $x$ -værdi på 165 aflæses til at være 0,222, så 22,2% af æblerne har en vægt under 165 gram. Dermed har  $100\% - 22,2\% = 77,8\%$  af æblerne en vægt *over* 165 gram.
- Sumkurven viser, at andelen af æbler med en vægt *under* 180 gram er 0,674. Det tilsvarende tal for æbler med en vægt *under* 165 gram er 0,222, så andelen af æbler med en vægt *mellem* 165 og 180 gram er lig med  $0,674 - 0,222 = 0,452 = 45,2\%$ .
- Da en sumkurve altid viser, hvor stor en del af observationerne, som er mindre end en bestemt værdi, så er det bedre at spørge til de 90% af æblerne, som er lettest! En aflæsning på sumkurven giver, at de har en vægt på under 193,7 gram. Altså har de 10% tungeste æbler en vægt på over 193,7 gram.

**Bemærkning 2** (Hvis observationerne er *alder*)

Alder betragtes normalt som en kontinuert variabel, dvs. en person kan for eksempel være 15,32 år gammel. Ofte vil man dog i opgaver se aldersintervaller angivet som for eksempel: 10 – 19, 20 – 29, 30 – 39, osv. Det skal oversættes til følgende intervaller:  $]10, 20]$ ,  $]20, 30]$ ,  $]30, 40]$ , osv. Når en person oplyser at være 29 år gammel, så betyder det nemlig, at personen kan være alt mellem præcist 29 år og et splitsekund før denne fylder 30 år!

**Bemærkning 3** (For viderekommende)

I tilfældet med en kontinuert variabel, hvor observationerne kan antage alle værdier i et helt interval, så er det ikke relevant at stille spørgsmålet om, hvor mange observationer, som er *præcist* lig med en given værdi. Skulle man besvare spørgsmålet, skulle svaret være 0. Tænk for eksempel på eksempel 1 med æblerne. Hvis man spørger om, hvor mange procent af æblerne, som har en vægt på *præcist* 154,239 gram, så er man nødt til at give svaret 0, for det er fuldstændigt usandsynligt, at der skulle være et æble med eksakt denne vægt. Alligevel er det principielt en mulighed! Derfor er det ikke ”tilladeligt” eller relevant at stille et sådant spørgsmål. Dette faktum forklarer også, hvorfor vi ikke går op i om endepunkterne i et interval er med eller ej. Det ændrer ikke på noget! Rent principielt vedtager man dog den konvention, at man grupperer observationerne i intervaller, som er åbne i venstre endepunkt og lukkede i højre endepunkt. Men der er egentligt bare tale om en *konvention*, og den er ikke væsentlig for svarene.

**Bemærkning 4** (Om at gruppere oprindelige data)

Måske har man ikke bare fået udleveret et færdigt grupperet datamateriale, men derimod selv må gruppere rå data. Der er ingen faste regler for, hvordan man gør det. Man må vurdere fra situation til situation. Hvor store skal intervallerne være, skal de have forskellig bredde osv. Med for mange eller for få intervaller, vil histogrammet blive mindre informativt. Hvis man vælger at gruppere diskrete data, så kan man med fordel vælge intervaller, som centrerer observationerne for at undgå, at nogle observationer ligger i interval-endepunkterne. Antag for eksempel, at man har at gøre med følgende række af observationer: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 – hver med givne frekvenser. Så kunne man for eksempel gruppere i følgende intervaller, som centrerer observationerne:  $]0,5; 2,5]$ ,  $]2,5; 4,5]$ , ...,  $]18,5; 20,5]$ . Man kunne også centrere omkring hver enkelt værdi:  $]0,5; 1,5]$ ,  $]1,5; 2,5]$ , ...,  $]19,5; 20,5]$ . I opgave 6 skal vi se, hvordan man grupperer karakterer i Undervisningsministeriet, når man skal undersøge resultatet af årets HF-eksamen.

En del datamateriale kan både ansues som værende *diskret* og *kontinuert*. For eksempel lønninger. Diskret, fordi der er en mindste enhed, som der udbetales løn i, fx ører! Da lønninger imidlertid spænder over et meget stort interval i forhold til den mindste

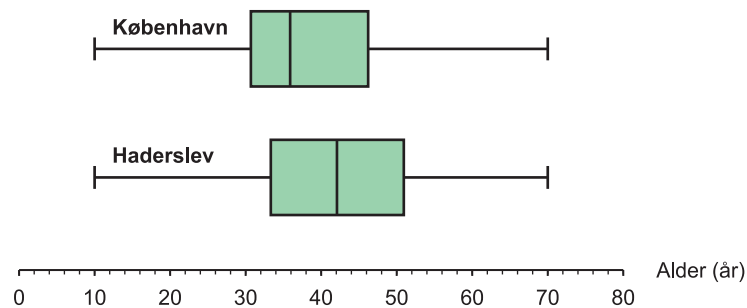
enhed, så er det almindeligvis fornuftigst at betragte lønninger som kontinuerte data. I det følgende skal vi se et eksempel på, hvad boxplot kan fortælle.

### Eksempel 5 (Sammenligning af boxplot)

På figur 7 er afbildet boxplot for aldersfordelingen af mandlige lønmodtagere på højeste niveau i 2005 i henholdsvis København og Haderslev. Medianerne på henholdsvis 35,9 år og 42,1 år viser, at halvdelen af mændene på højeste lønmodtagerniveau er under 35,9 år, mens det tilsvarende tal for Haderslev er 42,1 år. I øvrigt er alle kvartilerne for København mindre end de tilsvarende for Haderslev, hvilket kan tyde på, at forfremmelserne sker noget hurtigere i København, end tilfældet er i Haderslev. Det er dog ikke noget bevis for påstanden, for plottene fortæller ikke noget om, *hvornår* personerne kom op på højeste niveau. Bemærk, at man *ikke* kan sige noget om, hvor *mange*, der er på højeste lønmodtagerniveau. Boxplottene fortæller udelukkende om aldersfordelingen af de mandlige lønmodtagere, som allerede er på højeste niveau. Spændet fra nederste kvartil til øverste kvartil fortæller noget om hvor *spredt* aldersfordelingen er. Forskellen er ikke så stor. I København har den ”midterste halvdel” af mændene på højeste lønmodtagerniveau en alder mellem 30,7 og 46,2 år – et spænd på 15,5 år. I Haderslev er nederste kvartil på 33,3 år og øverste kvartil er 50,9 år – et spænd på 17,6 år. For København ligger medianen tættere på nederste kvartil end øverste kvartil, hvilket betyder, at koncentration af personer er større i den nederste del af den midterste halvdel.

Figur 7

Aldersfordeling af mandlige lønmodtagere på højeste niveau i 2005

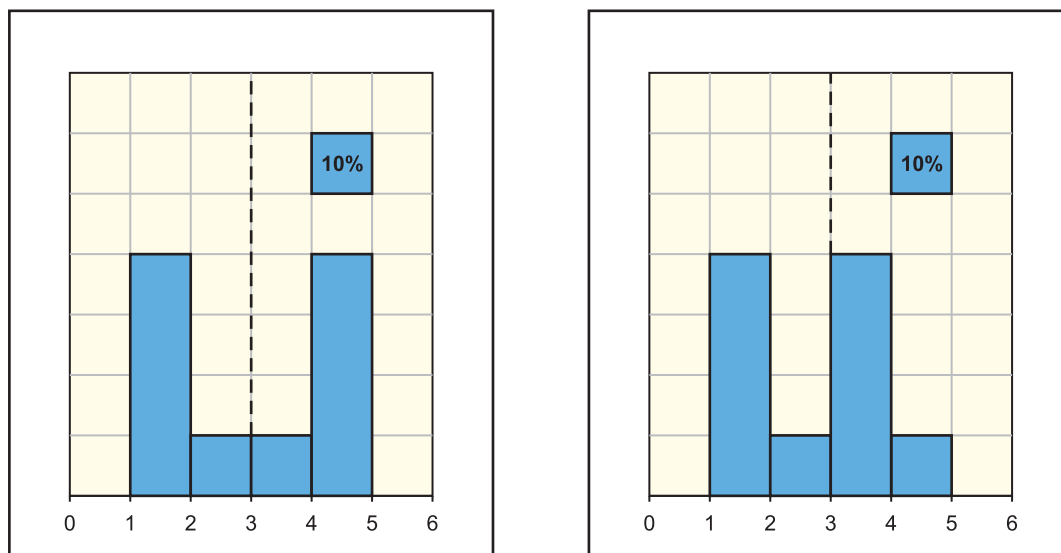


I øvrigt kan det oplyses, at den *gennemsnitlige alder* for mænd i højeste lønmodtagerniveau er 38,8 år i København og 42,3 år i Haderslev. Gennemsnitstal er dog ikke størrelser, som kan aflæses ud af et boxplot! Man ser, at gennemsnittet (middeltallet) ikke altid er lig med medianen. Især for København er de en del forskellige. Eksempel 6 nedenfor giver en forståelse for, hvorfor medianen og gennemsnittet for et grupperet datamateriale ikke behøver være ens.

### Eksempel 6 (Forskellen på gennemsnittet og medianen)

I dette eksempel skal vi forsøge at forstå forskellen mellem gennemsnittet (middeltallet) og medianen. Betragt histogrammet i figur 8. Da histogrammet er symmetrisk omkring den lodrette stiplede akse i 3, så er det klart, at gennemsnittet er lig med 3 – man behøver ikke en gang regne det ud! Medianen kan man selvfølgelig finde ved at lave en sumkurve, men vi kan også bruge definitionen, som siger, at medianen er den værdi, for hvilket 50% af observationerne er under denne værdi. Det ses af histogrammet, at der netop er 50% af observationerne, der er mindre end 3, så medianen er altså også 3. Det gælder i øvrigt altid, at middeltallet og medianen er ens, når histogrammet er symmetrisk om en akse! Antag nu, at vi som vist på figur 9, bytter om på tredje og fjerde søjle: Det ændrer åbenlyst ikke på medianen. Middeltallet er derimod blevet mindre, da nogle observationer er flyttet fra intervallet  $]4, 5]$  ned i intervallet  $]3, 4]$ .

Figur 8 og 9



## 2. Ugrupperede observationer

Som omtalt i indledningen er det mest kendte fra folkeskolen nok ugrupperede observationer, hvor man har en række observationer, hvoraf man kan tegne pindediagrammer m.m. Emnet er ikke særligt interessant, men skal lige omtales, da der kan komme skriftlige eksamensopgaver i det. Det er bedst at se på et eksempel.

### Eksempel 7

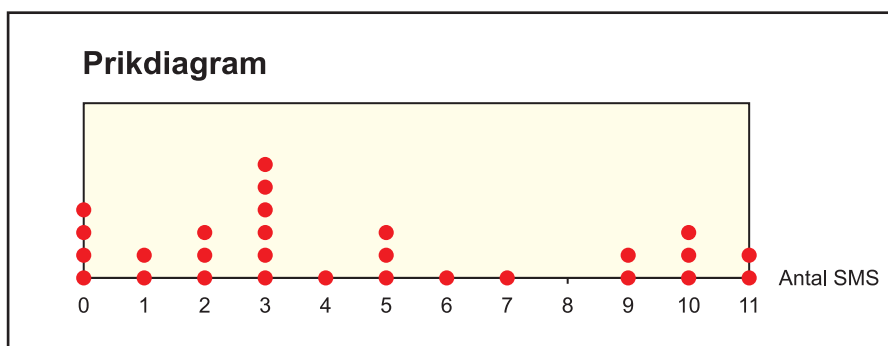
I en gymnasieklasse med 28 elever har man opgjort, hvor mange SMS-beskeder hver elev i klassen har sendt i løbet af en dag. Følgende antal blev opgivet: 3, 5, 2, 0, 10, 11, 10, 1, 2, 5, 5, 1, 3, 10, 7, 0, 0, 9, 3, 3, 3, 9, 6, 2, 0, 3, 4, 11. En sådan række af tal er svære at overskue, så det er hensigtsmæssigt at tælle op, hvor mange, der er af hver observation, dvs. udregne *hyppigheder*. Dernæst kan man beregne *frekvenser* og *kumulerede frekvenser*, som vi har gjort tidligere. Den kumulerede frekvens for en observa-

tion er som tidligere nævnt lig med den samlede frekvens for de observationer, som er mindre end eller lig med den pågældende observation. Vi får:

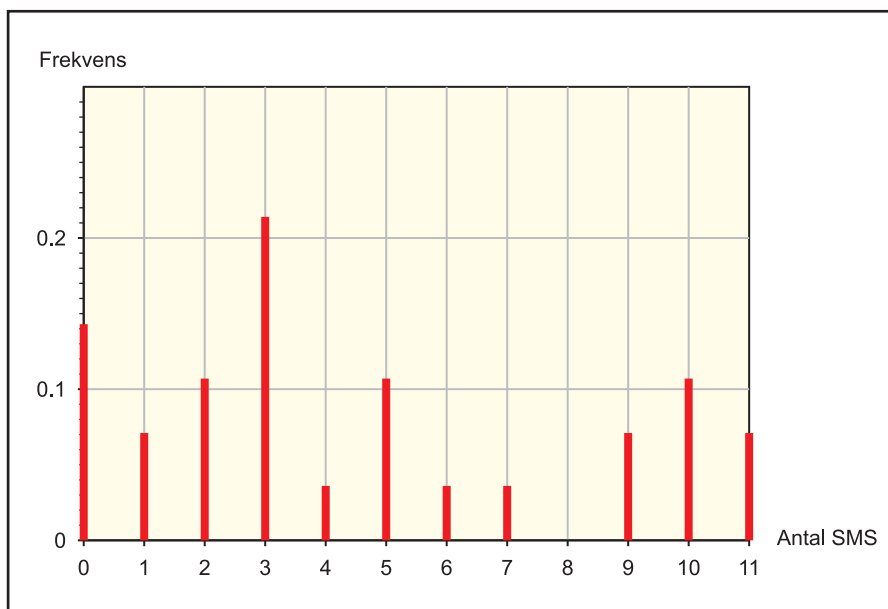
Antal SMS	0	1	2	3	4	5	6	7	8	9	10	11
Hyppighed	4	2	3	6	1	3	1	1	0	2	3	2
Frekvens	0,143	0,071	0,107	0,214	0,036	0,107	0,036	0,036	0,000	0,071	0,107	0,071
Kum. frekvens	0,143	0,214	0,321	0,535	0,571	0,678	0,714	0,750	0,750	0,821	0,928	0,999

Man kan lave et *pindediagram* eller et såkaldt *prikdiagram* over hyppighederne. Sidstnævnte er vist på figur 10. På figur 11 er vist et pindediagram over frekvenserne.

Figur 10



Figur 11



Man kan definere en sumkurve for et ugrupperet observationssæt, nemlig en såkaldt *trappekurve*, men vi vil undlade det her, da det kan føre til forvirring. Man kan nemlig sagtens finde *kvartilsættet* for et ugrupperet observationssæt direkte ved at kigge på de kumulerede frekvenser. Men først skal vi definere, hvad der i det hele taget menes med kvartilsættet for et ugrupperet datasæt:



nen er dermed lig med  $\frac{1}{2} \cdot (4 + 5) = 4,5$ . For at finde øvre og nedre kvartil gør man følgende: Man opdeler følgen af observationer på midten i to lige store delfølger. Hvis der er et ulige antal observationer, så kasserer man dog først den midterste observation, så delfølgerne kan blive lige lange. Da vi har 28 observationer, deler vi følgen op i to lige store følger med 14 observationer i hver. *Nedre kvartil* defineres herefter som medianen i den venstre delfølge, mens *øvre kvartil* defineres som medianen i den højre delfølge. Vi får følgende:

0, 0, 0, 0, 1, 2, 2,3, 3, 3, 3, 3, 3, 4 og 5, 5, 5, 5, 6, 7 7,9, 9, 10, 10, 10, 11, 11.

Da antallet af observationer i hver af delfølgerne igen er lige, er medianen lig med gennemsnittet af de to midterste værdier. Altså er den nedre kvartil lig med 2,5 og den øvre kvartil lig med 8. Vi får altså ikke helt samme værdier for kvartilsættet, som det vi fik ved brug af definition 9. Alligevel kan begge betragtes som korrekte.

□

### 3. Statistik i samfundet

Statistik er et emne, som vi alle bliver præsenteret for utallige gange i det daglige, mere eller mindre ubevidst. Og denne proces er taget til i takt med, at samfundet er blevet mere komplekst og kontrolleret. Mange diskussioner og beslutninger tager udgangspunkt i statistikker: Er bivirkninger ved et medicinsk præparat markante, så produktet skal forbydes, skal man screene for brystkræft, skal man sætte ekstra ind mod kriminaliteten, skal man lette på skatten for at få flere i arbejde etc. Virksomheder benytter statistiske analyser til at finde ud af, hvordan de bedst benytter deres reklamepenge til at få deres produkter gjort kendte. TV-stationerne laver seeranalyser, for at afgøre hvilke programmer, som skal anbringes hvornår på sendefloden. Aviser bestiller undersøgelser af folks holdninger til utroskab. I sportens verden bruger man statistik til at gøre sporten interessant: Hvem er den mest scorende, den med flest assists etc. Politikerne lader sig påvirke af meningsmålinger i deres valg af lovforslag ... vores moderne samfund er gennemsyret af statistik! Men statistik er også nødvendig i forbindelse med videnskabelige afhandlinger. Er forsøgsresultaterne statistisk signifikante? Kan konklusionerne i afhandlingen stå for en nærmere statistisk analyse? Hvor stor er usikkerheden på målingerne?



#### Danmarks Statistik

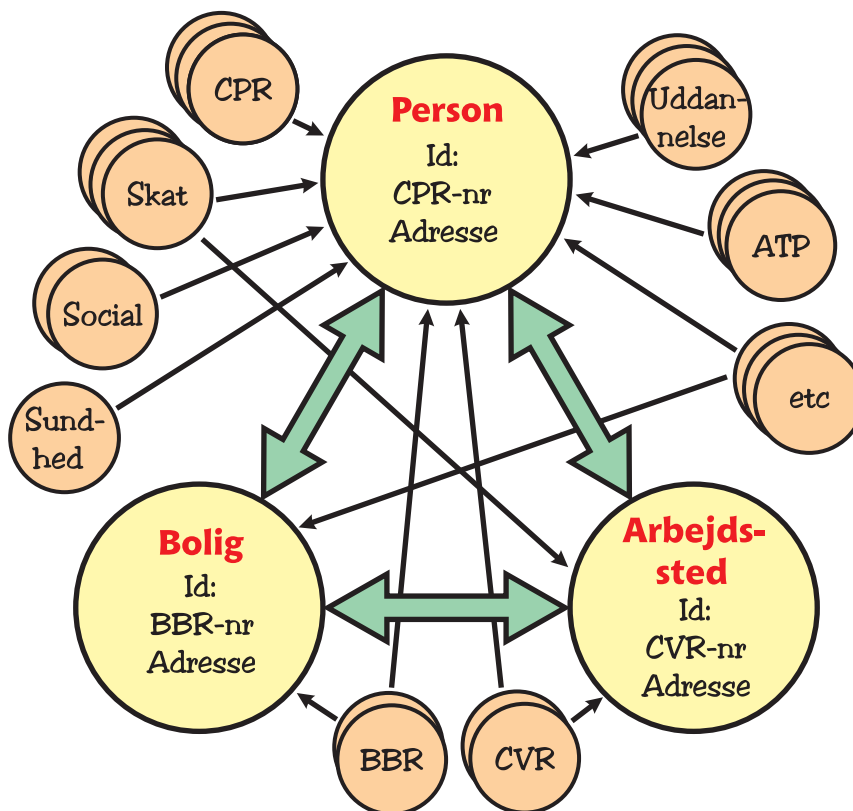
*Danmarks Statistik* (DST) har til huse på Sejrsgade i København. Institutionens historie går tilbage til 1850, hvor dens første opgave var at foretage en folketælling. Siden er institutionens opgaver vokset kraftigt i omfang, så der i skrivende stund (forår 2006) er omkring 570 ansatte. Danmarks Statistik udarbejder upartisk statistik om samfundet som grundlag for demokratiet og samfundsøkonomien. DST bidrager med sin statistik til viden, debat, analyser, forskning, planlægning og beslutninger hos de vigtigste brugere, som er: Befolkningen, statens og kommunernes politikere og administration, er-

hvervslivet og organisationerne, forskerne, pressen og medierne, EU, FN, OECD, IMF og andre internationale organisationer.

En vigtig kendsgerning ved Danmarks Statistik er, at institutionen stort set kun indsamler materialer på hele *populationer*, dvs. den har normalt al data til rådighed, i modsætning til de virksomheder, som anvender stikprøver. Sagen er nemlig, at en række af samfundets instanser har indberetningspligt: Kommunale, amtslige og statslige institutioner har pligt til at indberette om diverse forhold til brug i statistikkerne. Det samme er tilfældet for de fleste private erhvervsvirksomheder. Så alle data står i princippet til rådighed! I visse andre lande, herunder USA, har man af rent principielle grunde valgt ikke at have den meget strenge indberetningspligt, som man har i Danmark. Amerikanske statsborgere er for eksempel ikke registreret med et CPR-nummer eller lignende, som kan identificere dem. En ulempe ved denne øgede "borgerlige frihed" er, at man med års mellemrum må foretage folketællinger. Det samme er tilfældet på en lang række andre områder i det amerikanske samfund, og for at få et overblik over samfundets udvikling, er man derfor ofte nødsaget til at benytte stikprøver i de statistiske undersøgelser. Men sådan er det som sagt ikke i Danmark! Nedenstående figur illustrerer det statistiske informationssystem i Danmark.

Figur 12

## Det statistiske informationssystem



## Stikprøver og meningsmålinger

Nogle institutioner eller virksomheder får deres viden ud fra brug af *stikprøver*. Her kan nævnes *Gallup, Observa, Megafon, Vilstrup, Epinion, etc.* Opgaverne kan være meget forskelligartede: Det kan være en prognose til et folketingsvalg, en brugerundersøgelse for et givet produkt, en analyse af hvilke programmer TV-seerne ser osv. Stikprøver hentyder til, at man forsøger at sige noget om hele populationen ud fra et begrænset udpluk. I mange tilfælde er man simpelthen nødt til at bruge denne form, da man ikke kan spørge hele Danmarks befolkning, hvad den synes om økologiske varer eller hvilke TV-udsendelser, den ser. Det er ikke praktisk muligt, og selv om man forsøgte, ville man ikke kunne komme i kontakt med alle. Det ville også være alt for dyrt og tidskrævende. Derfor benytter man stikprøver. Hvis der er tale om en meningsmåling, så er der for eksempel følgende omkostninger:

1. Nøje overvejelse af hvordan persongruppen sammensættes/udvælges.
2. Man skal være omhyggelig med, *hvordan* man stiller spørgsmålene.
3. Resultaterne af spørgeundersøgelsen skal efterbehandles statistisk.



TNS Gallup har til huse på Masnedøgade i København

Nogle kommentarer til ovenstående punkter:

- 1) Det er meget vigtigt, at man sammensætter gruppen af personer, som man udspørger, så den er *repræsentativ* for hele populationen, hvad enten det er en bestemt del af befolkningen eller hele befolkningen. Det er ingenlunde nemt. Man skal passe meget på, at der ikke kommer *bias*, dvs. en *skævhed*, i stikprøven. Ønsker man for eksempel at finde ud af befolkningens holdning til krigen i Afghanistan, så er det ikke særligt fornuftigt at gå ned på den lokale gågade og spørge folk ud, om soldaterne skal trækkes hjem. Folk på gågaden er nemlig ikke særlig repræsentative for Danmarks befolkning. Der vil formentlig være et underskud af folk i arbejde og et

overskud af husmødre. Der er altså en skævhed i sammensætningen. Man kan formode, at der vil være et overskud af kvinder tilstede og kvinder er gennemsnitligt mere imod krig end mænd er. En anden ting, man også skal passe på er, når udspørgeren selv vælger den, der udspørges. Der kan nemlig være en tendens til at spørge personer på gaden, som ser venlige og imødekomne ud, og der er en mulighed for, at denne gruppe af personer kan have en anden holdning til et spørgsmål, end gennemsnittet af befolkningen. At spørge på Internettet skal man også være påpasselig med, da den ældre del af befolkningen er underrepræsenteret her. Hvis man undersøger de unges præferencer, så gør dette måske ikke så meget.

Den klassiske bommert, som ofte nævnes i forbindelse med udvælgelse af stikprøver er den, der blev begået af *Literary Digest* i deres opinionsundersøgelse for valget i USA i 1936: Franklin D. Roosevelt havde fuldført sine første 4 år som præsident, og genopstillede mod republikaneren Alfred Landon fra Texas. Magasinet *Literary Digest* forudsagde en overvældende sejr til Alfred Landon, med kun 43% af stemmerne til Roosevelt. Undersøgelsen var endda baseret på den største stikprøve nogensinde: 2,4 millioner! Magasinet havde et godt ry: det havde udpeget den rigtige præsident siden 1916. Imidlertid vandt Roosevelt overvældende: med 62% mod 38% og *Literary Digest* gik fallit kort efter. Hvordan kunne magasinet begå sådan en kæmpe fejl – den største nogensinde af et etableret og vigtigt meningsmålingsinstitut? Man havde jo udspurgt en kæmpe gruppe. *George Gallup* var netop ved at grundlægge sit meningsmålingsinstitut og fik sit gennembrud ved at forudsige resultatet af valget med en afvigelse på kun 1 procent, og han havde endda kun udspurgt 50.000 personer. Det var altså ikke stikprøvens størrelse, som var altafgørende, her var det den nye markedsanalyse-teknik med anvendelse af den repræsentative stikprøve, der havde bestået sin prøve! Magasinet fejlede i at sendte spørgsmål ud til 10 millioner mennesker med posten. Navnene fra de 10 millioner mennesker kom fra kilder som telefonbøger og medlemmer af klubber. Denne fremgangsmåde havde en tendens til at frasortere de fattige, hvoriblandt der ikke var mange, som var medlemmer af klubber. Og dengang havde kun 1/4 af befolkningen telefon. Grunden til, at en sådan fejl først skete i 1936 og ikke før var, at i 1936 fulgte de politiske holdninger mere økonomiske linjer ... det havde ikke været tilfældet tidligere, hvor rige og fattige stemte mere ensartet. Så læren af dette eksempel er følgende: *Når en udvælgelsesprocedure er skæv, så hjælper det ikke at tage en større stikprøve. Det vil blot gentage fejltagelsen i større målestok!*

En anden ting, som stikprøver kan risikere at lide under er *non-response bias*, hvorved menes skævhed på grund af for mange personer, som nægter at svare på spørgsmål. Det viser sig nemlig, at gruppen af personer, som ikke svarer, undertiden adskiller sig fra resten på vigtige områder. Faktisk led *Digests* undersøgelse netop heraf, idet kun 2,4 millioner ud af de 10 millioner svarede! Undersøgelser har vist, at lav-indkomst og høj-indkomstgrupperne har en større tendens til ikke at svare, så mellem-indkomstgrupperne er overrepræsenteret. Gode meningsmålingsinstitutter kender dette problem og har metoder til at tage højde for det. Hvis man

ringer til folk, så kan man for eksempel ringe tilbage gentagne gange til de folk der ikke træffes umiddelbart.

Men hvilke metoder benyttes da? Besøgsinterviews, telefoninterviews, postomdelte interviews eller Internet-interviews? Svaret er, at det kommer an på formålet og undersøgelsens form. Lange og teksttunge undersøgelser egner sig ikke til oplæsning. Her er det bedre, hvis den spurgte har noget at kigge på. Det kan også være, at respondenterne skal reagere på et logo etc. Besøgsinterviews benyttes også, men ikke så meget som tidligere. De kræver mange resurser. Det skal dog også nævnes, at nogle undersøgelser kræver helt andre former. For eksempel TV-seer undersøgelser, hvor et panel af personer har monteret en måler på deres TV, eller Internetbrugere, som har installeret et særligt program for at kunne registrere deres vaner på Internettet. Internettet er godt til at måle folks reaktion på reklamer, radiospots eller andet audiovisuelt materiale.



En væsentlig årsag til en anden type fejl er, hvis man giver interviewerens lov til selv at vælge, hvem der skal interviewes, eventuelt indenfor en bestemt undergruppe. Det var faktisk årsagen til en anden kendt fejlbedømmelse ved præsidentvalget i 1948 i USA. Et problem ved at overlade for meget til menneskets valg er, at interviewerens vil udspørge dem, der er lettest at få fat i. I 1948 resulterede det i, at man udvalgte for mange republikanere, da de var en smule nemmere at interviewe. Løsningen på dette problem er, at man indfører et element af *tilfældighed* ved at trække lod. Det stiller dog nogle spørgsmål: Har man en liste med alle indbyggere? Hvordan håndterer man rent praktisk, hvis en udvalgt person ikke er hjemme eller bortrejst? etc. For at undgå for mange praktiske problemer kan man vælge at lave *klyngeprøver*, hvormed menes, at man vælger et antal områder ud, eventuelt inddeler i et antal undergrupper, hvori man så udtager personer ved simpel tilfældig lodtrækning. Denne metode er især nyttig ved besøgsinterview, hvor det også gælder om at begrænse transportomkostningerne. Der er mange variationsmuligheder her.

### **TNS Gallup**

TNS Gallup i København blev grundlagt i 1939 i Danmark. Den legendariske reklamemand Haagen Wahl Asmussen havde fulgt Gallups virke i USA, og efter et møde med grundlæggeren fik han retten til at bruge Gallups navn i Danmark. Siden

har Gallup lavet tusindvis af undersøgelser, lige fra prognoser for folketingsvalg til undersøgelser af danskernes TV-vaner. Man kan mod betaling bestille undersøgelser hos Gallup. I den ene bygning på hovedafdelingen på Masnedøgade sidder en række interviewere, typisk studerende, som ringer folk op. Hvad angår meningsmålinger til folketinget, så skelner Gallup mellem *prognoser*, som gennemføres dagen før et folketingsvalg, og *Gallup Politisk Indeks*, som er månedlige opinionsmålinger, og som bliver bragt i Berlingske Tidende. Ved udarbejdelse af prognoser benyttes ikke en ny stikprøve, men et panel af personer, som Gallup før har talt med. Fordelen er, at man her kender til repræsentativiteten af gruppen og nemmere kan korrigere for eventuelle skævheder. Man er dog opmærksom på ikke at forstyrre de samme respondenter igen og igen. De løbende opinionsmålinger foretages derimod via friske stikprøver. Telefonnumrene er baseret på basis af tilfældige tal.

- 2) Man skal være omhyggelig med, at man stiller spørgsmål, som er *klare* og *utvetydige*. Og så skal spørgsmålene ikke være *ledende*. Det duer for eksempel ikke at spørge en person, om han/hun motionerer meget, for hvad er ”meget motion”? Man skal heller ikke anvende fremmedord, som mange ikke kender. Spørgsmål, som lægger op til politisk korrekte svar bør undgås.
- 3) Den tredje omkostning er, at de indsamlede data skal behandles statistisk. Man kan ikke bare uden videre tælle sammen, som hvis man har data for en hel population. I ret stort omfang kan man for eksempel korrigere for skævheder i stikprøven. Og det har for eksempel Gallup stor gavn af. Lad os se på et eksempel.

### Eksempel 10

Lad os gøre det tankeeksperiment, at man udspørger et antal personer, om de vil stemme JA eller NEJ til den nye EU-traktat. Lad os antage, at man i stikprøven fik spurgt 47% kvinder og 53% mænd og at der blandt mændene var en gennemsnitlig Ja-procent på 58%, mens der blandt kvinderne var en gennemsnitlig Ja-procent på 45%. Hvis man kritikløst havde godtaget denne stikprøve som værende repræsentativ, så ville man altså få en total Ja-procent ved at udregne det vejede gennemsnit:

$$0,47 \cdot 0,45 + 0,53 \cdot 0,58 = 0,519 = 51,9\%$$

Imidlertid er der relativt flere mænd end kvinder i stikprøven i forhold til hele den stemmeberettigede del af befolkningen, hvor der er 48,8% mænd og 51,2% kvinder. Vi korrigerer derfor ved at benytte de korrekte vægte frem for stikprøvens:

$$0,512 \cdot 0,45 + 0,488 \cdot 0,58 = 0,513 = 51,3\%$$

I praksis vil man selvfølgelig også skulle korrigere for andre størrelser end køn. Gallup benytter også vejning til at korrigere for, at der er nogle persongrupper, som det er sværere at få fat i end andre pr. telefon. Eksempelvis har Gallup lidt sværere ved at få fat i unge mænd. Forhold, der typisk korrigeres for ved prognoser til folketingsvalg er køn, alder, valgkreds, husstandsstørrelse og partivalg ved forrige folketingsvalg. Det er ikke nødvendigvis alle skævheder, man kan veje sig ud af, så målingerne kan – udover den statistiske usikkerhed – godt være behæftet med

mindre fejl, som påvirker resultatet *systematisk*. Fejl af denne type er i sagens natur ukendte, men meget tyder på, at der er tale om ret små ting.

Endelig skal det siges, at man også forsøger at tilrettelægge (*stratificere*) sammensætningen af stikprøven, *før* undersøgelsen foretages, men det kan kun lade sig gøre, hvis man på forhånd har de relevante oplysninger om respondenterne. Det haves ikke, hvis man ringer til tilfældige telefonnumre. Hvis der derimod er tale om en undersøgelse baseret på et medlemsregister af en slags, gøres det ofte. Også i tilfældet med Internet-undersøgelser, hvor Gallup på forhånd har en masse baggrundsoplysninger om det panel af personer, man har til rådighed.

### Hvor store skal stikprøverne være?

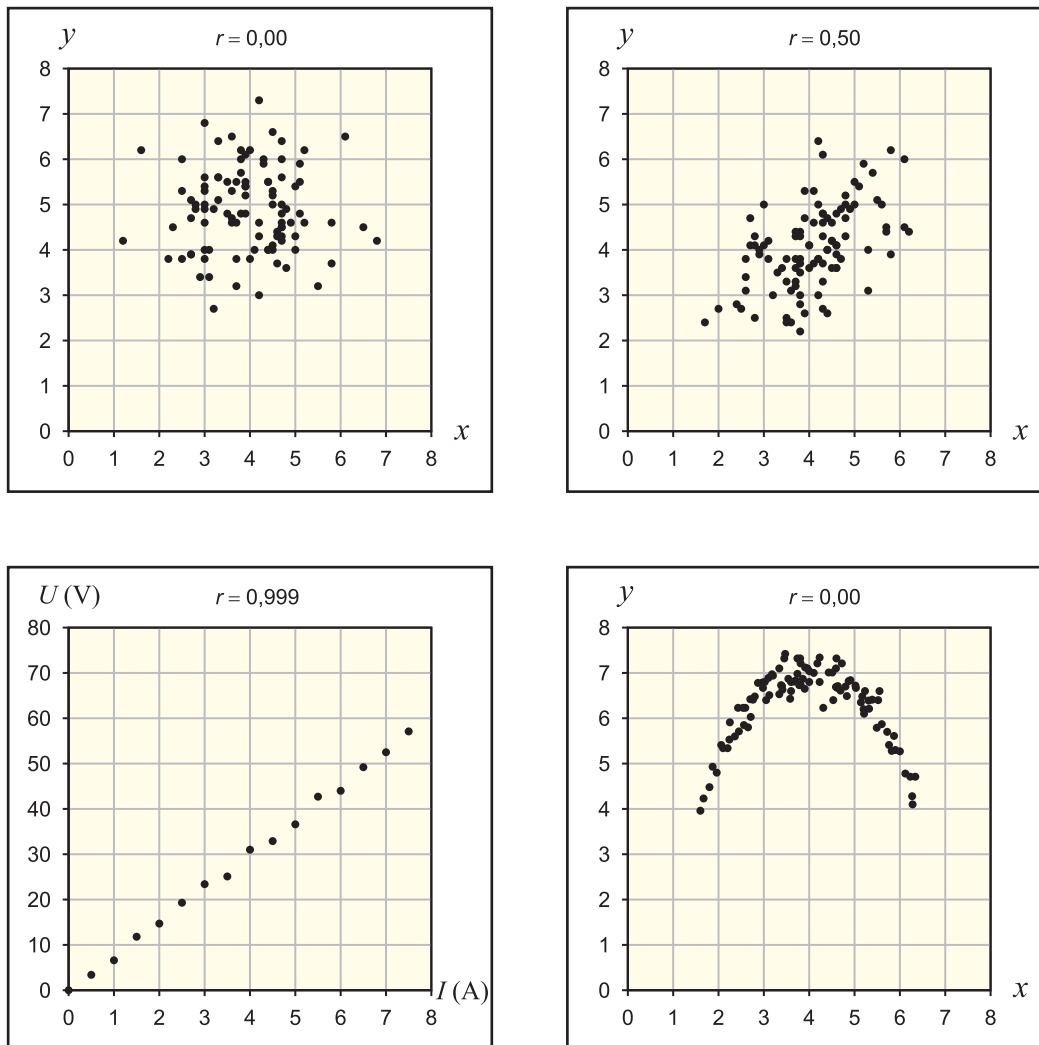
I stikprøveundersøgelser i dag anvender man typisk stikprøver på 1000-1500 personer, altså endnu færre end George Gallup gjorde i 1936! Man skal helst op på mindst nogle hundrede personer, for at den statistiske usikkerhed bliver acceptabel. men omvendt får man ikke meget ekstra ud af at interviewe mere end 1500 personer. *Marginalnyttten* aftager nemlig kraftigt, da usikkerheden falder med kvadratroden af stikprøvens størrelse - som man kan vise.

### Sammenhængen mellem to variable

Man kan være interesseret i at studere sammenhængen mellem to variable i et statistisk materiale. Som et eksempel kan nævnes, at den berømte engelske statistiker *Karl Pearson* (1857–1936) studerede sammenhængen mellem højden af 1078 fædre og deres respektive sønners højde. Data kan afbildes i et koordinatsystem med fædrenes højde på  $x$ -aksen og sønnernes højde på  $y$ -aksen. Det viste sig, at der var en vis tendens til at sønnerne af høje fædre var højere end gennemsnittet. Men sammenhængen var ikke helt tydelig, idet der var stor variation. Vi skal ikke diskutere dette eksempel nærmere, kun se på principperne bag emnet. På figurerne 13-16 på næste side ser vi på fire tænkte tilfælde, hvor to variable stilles op mod hinanden:  $x$  og  $y$ . Med moderne it-hjælpemidler kan man finde på at udføre *lineær regression* på datapunkterne. Programmet *Microsoft Excel* er et eksempel. I den forbindelse er der en størrelse  $r$ , kaldet *korrelationskoefficienten*. Det er en størrelse, der siger noget om, i hvor høj grad, der er tale om en lineær sammenhæng imellem de to variable. Hvis korrelationskoefficienten er lig med  $+1$  eller  $-1$ , så er der tale om en eksakt lineær sammenhæng. Hvis koefficienten derimod er  $0$ , så er sammenhængen længst muligt væk fra at være lineær. Sidstnævnte situation kan ses på figur 13, hvor der ikke ser ud til at være nogen sammenhæng mellem de variable  $x$  og  $y$  overhovedet: Hvis man kender  $x$ , så kan man ikke sige noget fornuftigt om  $y$ . På figur 14 er korrelationskoefficienten  $0,50$  og vi ser, at der er en vis tendens til, at jo større  $x$  er, jo større er  $y$ , men tendensen er ikke tydelig. Figur 15 viser resultaterne fra et fysikforsøg, hvor man har målt sammenhørende værdier mellem spændingen  $U$  i Volt og strømstyrken  $I$  i Ampere igennem en resistor. Vi ser en meget tydelig lineær sammenhæng og korrelationskoefficienten er da også lig med  $0,999$ . Figur 16 viser et tilfæl-

de, hvor korrelationskoefficienten er lig med 0,00, men hvor man alligevel kan sige, at der er tale om en form for sammenhæng, bare ikke lineær. Eksemplet indikerer, at man til enhver tid også bør kigge på det grafiske billede og ikke bare korrelationskoefficienten, når man udtaler sig om en sammenhæng mellem to variable.

Figur 13-16



I fysik får man ofte resultater, som viser en tydelig sammenhæng. Indenfor samfundsvidenskaberne er sammenhængene derimod mere ”grumsede” og her oplever man ofte korrelationskoefficienter på mellem 0,3 og 0,7.

### Statistisk sammenhæng er ikke det samme som årsagssammenhæng

Det er en nærliggende tanke, at når der konstateres en statistisk sammenhæng mellem to variable eller to faktorer, så er der tale om en *årsagssammenhæng*, hvor den ene variabel eller faktor medfører den anden eller har en indflydelse på den. Her skal man imidlertid passe overordentligt meget på, for der kan nemt være en tredje *skjult variabel* eller *skjult faktor*, som påvirker de to øvrige.

Lad os se lidt abstrakt på situationen. Givet to variable eller to faktorer  $A$  og  $B$ . Med symbolikken  $A \rightarrow B$  vil vi mene at  $A$  er årsag til  $B$ . Med udtrykket ”er årsag til” er vi nødt til at være lidt løse: Vi vil mene at  $A$  har en virkning på  $B$  eller at  $A$  har en betydning for  $B$ . Det vil ikke være hensigtsmæssigt her at benytte det kontante matematiske udtryk *medfører* ( $\Rightarrow$ ). Vi ved at rygning har en betydning for cancer, men det er ikke tilfældet at *alle* rygere får cancer. Hvis  $A$ : Personen ryger og  $B$ : Personen får cancer, så har vi altså med ovenstående definition, at  $A \rightarrow B$ .

Når man observerer en statistisk sammenhæng mellem to variable eller to faktorer  $A$  og  $B$ , så er indholdet af ovenstående altså, at det ikke er sikkert, at  $A \rightarrow B$ . Der er faktisk mindst fem muligheder:

- 1)  $A \rightarrow B$
- 2)  $B \rightarrow A$
- 3) Der er en tredje skjult variabel eller skjult faktor  $C$ , så  $C \rightarrow A$  og  $C \rightarrow B$ .
- 4) Der er en årsagssammenhæng begge veje:  $A \rightleftarrows B$ .
- 5) Den statistisk sammenhæng er en simpel tilfældighed, et sammenfald. For at undersøge det nærmere kan det være hensigtsmæssigt at tage en større stikprøve eller undersøge for systematiske fejl.

En påstand om en årsagssammenhæng bør i øvrigt altid følges op af et plausibelt argument! Vi skal kigge på en række eksempler.

### Eksempel 11

I USA viste talrige epidemiologiske studier, at kvinder, som fik hormonbehandlinger i overgangsalderen (*Hormone Replacement Therapy*), blev ramt af færre hjertesygdomme (*Coronary Heart Disease*) end gennemsnittet. Ledende læger foreslog derfor, at hormonbehandlingerne havde en beskyttende virkning. En nærmere analyse viste imidlertid, at forklaringen var en helt anden. De kvinder, der modtog hormonbehandlinger kom overvejende fra de højere sociale grupper, hvor spisevaner og motionsvaner er bedre. Så der er altså tale om en tredje faktor  $C$ , som spiller ind: *Fordelene ved en høj socioøkonomisk status*.

- $A$ : Kvinden får en hormonbehandling
- $B$ : Kvinden får en hjertesygdom
- $C$ : Kvinden har en høj socioøkonomisk status

Vi konkluderer, at der *ikke* gælder  $A \rightarrow B$ , men derimod  $C \rightarrow A$  og  $C \rightarrow B$ . Så den umiddelbare statistiske sammenhæng mellem  $A$  og  $B$  var altså ikke en årsagssammenhæng. Ordentlige tests kræver typisk anvendelse af *randomiserede kliniske undersøgelser*. Heri indgår *tilfældige udvælgelser*. Problemet med de blotte tal fra ovenstående statistik var at der i stikprøven var en overrepræsentation af kvinder med høj socioøkonomisk status. Det gav anledning til den fejlslutning, at hormonbehandlinger skulle kunne nedsætte forekomsten af hjertesygdomme. Faktisk påviste man senere, at hormonbehandlinger tværtimod har en svag tendens til at *øge* antallet af problemer med hjertet. Vi skal i næste afsnit se nærmere på de metoder, der anvendes i medicin industrien.

### Eksempel 12

Hvis man for eksempel på en skole afbilder børns skostørrelse som funktion af deres læsekundskaber, så er der en tydelig sammenhæng. Men det betyder ikke, at bedre læsekundskaber medfører større skostørrelser! Her er den skjulte variabel *alder*. Jo ældre elever, jo større skostørrelse og jo bedre bliver de til at læse. I dette tilfælde var fejlslutningen nem at spotte; dette er ikke altid tilfældet. Lad os se på et andet eksempel.

### Eksempel 13

I lande, hvor indbyggerne spiser meget fedt, er raterne for visse kræftsygdomme høje. Spørgsmålet er, om fedt har en tendens til at give kræft? Sammenhængen er tydelig, men alligevel er beviserne for en årsagssammenhæng svage, eftersom de lande, som har høje rater af kræft, og de lande, der har lave rater af kræft, adskiller sig på mange andre måder. For eksempel indtager de personer, som får meget fedt i føden, også relativt meget sukker i forhold til de personer, som spiser mindre fedt. Fedt og sukker er relativt dyrt. I rige lande har folk råd til at spise fedt og sukker, frem for fx kornprodukter. Så det er uklart hvad der forårsager kræft i de mere udviklede lande. Det kan også være livsstil, herunder mindre motion.

### Eksempel 14

En artikel i Weekendavisen 9.-15. september 2005 præsenterer i en artikel med titlen ”Skrækhistorier”, adskillige eksempler på fejlslutninger på baggrund af registrerede sammenhænge. En dansk avis skal have fortalt om en undersøgelse, hvor udsagnet var, at børn, der sover for lidt som treårige, bliver overvægtige som voksne. En kendt dansk børnelæge fandt det indlysende, at søvnunderskud sætter sig som overvægt. Forklaringen skulle være den, at børn, der får lov til at være længe oppe, sikkert sidder foran fjernsynet og spiser masser af kager og slik. Men der er jo andre mulige forklaringer: For eksempel at et stort sukkerindtag gør det sværere at sove! Eller måske er forklaringen, at forældre, der bekymrer sig mindre om, hvad deres børn spiser, er de samme forældre, som bekymrer sig mindre om, hvornår børnene går i seng. Det er altså typen af forældre, der er afgørende for både sengetider og overvægt. Der er ikke en årsagssammenhæng imellem de to.

### Medicinske forsøg

At undersøge, om et givet medicinsk præparat har en virkning, er ingenlunde nemt. Dels vil midlet virke forskelligt på forskellige personer, dels vil forsøgspersoners udsagn være farvet af deres viden om, at de har modtaget et lægemiddel. Sidstnævnte betegnes den såkaldte *placebo effekt*. Forsøgspersonerne har en tendens til at synes, at lægemidlet har en positiv virkning, selv i tilfælde hvor midlet er virkningsløst. For at afskærme for dette, tilrettelægges man ofte medicinske forsøg, så der er en *kontrolgruppe*, som får et falsk og virkningsløst middel, fx en kalktablet, som ligner den rigtige pille. Ingen af deltagerne i forsøget ved, om de får en rigtig eller en falsk pille. Samtidigt sørger man for, at forskeren, som forestår undersøgelsen, heller ikke ved, hvilke personer, som har fået den rigtige medicin. Man taler om et *dobbelt blindforsøg*. Så videt muligt bør man også

sikre sig, at gruppen, der modtager rigtig medicin, er udvalgt *tilfældigt* i hele gruppen af forsøgspersoner – for at forhindre en eventuel skævhed.



## Sagt om statistik

Der findes tre slags løgn:

- 1) almindelig løgn,
- 2) forbandet løgn og
- 3) statistik!

## Opgaver

### Opgave 1

Skemaet nedenfor viser vægtfordelingen af pærer fra en stikprøve fra en frugtplantage.

- Udregn frekvenser og kumulerede frekvenser for observationssættet.
- Tegn et histogram for datamaterialet.
- Beregn middeltallet for pærernes vægt.
- Tegn sumkurven.
- Bestem kvartilsættet.
- Hvor mange procent af pærerne har en vægt under 155 gram?
- Hvor mange procent af pærerne har en vægt mellem 155 og 195 gram?
- Hvad kan man sige om de 20% tungeste pærer?

Vægt (gram)	Hyppighed	Frekvens	Kumuleret frekvens
]120,130]	17		
]130,140]	34		
]140,150]	76		
]150,160]	156		
]160,170]	268		
]170,180]	210		
]180,190]	147		
]190,200]	64		
]200,210]	28		
	I alt:		

### Opgave 2

(1) side 6 viser formelen for middelværdien  $\mu$ . Vis, at middelværdien også kan bestemmes ved at tage det vejede gennemsnit af intervallernes midtpunkter  $m_i$  med deres respektive hyppigheder  $h_i$ , og herefter dividere resultatet med det totale antal observationer, kaldet *sum*. Du skal altså argumentere for følgende formel:

$$\mu = \frac{h_1 \cdot m_1 + h_2 \cdot m_2 + \dots + h_n \cdot m_n}{sum}$$

**Opgave 3** (Aldersfordelinger)

Nedenfor en aldersfordeling for den kvindelige del af den danske befolkning, som er født i Danmark, opgjort pr. 1. januar 2006. Tallene er opgjort med udgangspunkt i en tabel fra Danmarks Statistik.

- Udregn frekvenserne.
- Udregn de kumulerede frekvenser. Husk, at de kumulerede frekvenser hører til højre endepunkt af intervallerne, som er  $]0,10]$ ,  $]10,20]$ ,  $]20,30]$ , osv.
- Tegn sumkurven og bestem kvartilsættet.
- Tegn et boxplot for datamaterialet.
- Bestem den gennemsnitlige alder for en kvinde i den danske befolkning.
- Når man skal udregne middeltallet (gennemsnittet), så benyttes midtpunkterne af intervallerne som bekendt, fordi man antager, at fordelingen af observationerne er jævnt fordelte i hvert interval. Er denne antagelse rimelig i for eksempel intervallet fra 80 – 89? Kommenter! Tror du den begåede fejl er stor?

Hvis du har lyst, kan du arbejde videre med opgaven og lave det til et helt lille projekt. Gå eventuelt ind på Danmarks Statistiks hjemmeside på [www.dst.dk](http://www.dst.dk). Gå ind i *statistikbanken*, og vælg *Befolkning og valg > Folketal > BEF5: Folketal pr. 1. januar efter køn, alder og fødeland (1990-2006)*. Her kan du få de tilsvarende data for den mandlige del af befolkningen. Grupper materialet i intervaller på 10 år, ligesom det er gjort i tabellen nedenfor. Sammenlign data for kvinder og mænd. Hvad kan du konkludere? En anden ting, du kan gøre, er at studere befolkningens aldersfordeling i 1990. Hvad er der sket i de 16 år?

Alder (år)	Hypighed	Frekvens	Kumuleret frekvens
0 – 9	313849		
10 – 19	300825		
20 – 29	263223		
30 – 39	338678		
40 – 49	347514		
50 – 59	343854		
60 – 69	284503		
70 – 79	184948		
80 – 89	116599		
90 – 99	24935		
100 – 109	568		
110 – 119	0		
I alt:			

**Opgave 4** (Gennemsnitstemperaturer i Danmark)

Skemaerne nedenfor indeholder data om fordelingen af gennemsnitstemperaturer i Danmark i henholdsvis januar og juli måned i perioden 1874–2005. Der er tale om bearbejdet data fra *Danmarks Metrologiske Institut* (DMI). Udregn frekvenser og kumulerede frekvenser og tegn sumkurver. Bestem kvartilsæt og tegn boxplot. Sammenlign boxplot for januar og juni måned. Kommenter.

Gennemsnitstemperatur i januar (°C)	Hypighed	Frekvens	Kumuleret frekvens
]–8,0;–7,0]	2		
]–7,0;–6,0]	3		
]–6,0;–5,0]	2		
]–5,0;–4,0]	5		
]–4,0;–3,0]	7		
]–3,0;–2,0]	13		
]–2,0;–1,0]	22		
]–1,0;0,0]	31		
]0,0;1,0]	13		
]1,0;2,0]	18		
]2,0;3,0]	12		
]3,0;4,0]	4		
I alt:			

Gennemsnitstemperatur i juli (°C)	Hypighed	Frekvens	Kumuleret frekvens
]13,5;14,0]	3		
]14,0;14,5]	7		
]14,5;15,0]	17		
]15,0;15,5]	13		
]15,5;16,0]	15		
]16,0;16,5]	18		
]16,5;17,0]	24		
]17,0;17,5]	19		
]17,5;18,0]	7		
]18,0;18,5]	5		
]18,5;19,0]	3		
]19,0;19,5]	1		
I alt:			

**Opgave 5** (Nedbør i Danmark)

Skemaerne nedenfor indeholder data om fordelingen af nedbør i Danmark i henholdsvis januar og juli måned i perioden 1874–2005. Der er tale om bearbejdet data fra *Danmarks Metrologiske Institut* (DMI). Udregn frekvenser og kumulerede frekvenser og tegn sumkurver. Bestem kvartilsæt og tegn boxplot. Sammenlign og kommenter boxplotene for januar og juni måned.

Nedbør i januar (mm)	Hypighed	Frekvens	Kumuleret frekvens
0 – 10	4		
10 – 20	6		
20 – 30	18		
30 – 40	13		
40 – 50	24		
50 – 60	16		
60 – 70	17		
70 – 80	14		
80 – 90	11		
90 – 100	3		
100 – 110	4		
110 – 120	2		
	I alt:		

Nedbør i juli (mm)	Hypighed	Frekvens	Kumuleret frekvens
0 – 10	0		
10 – 20	4		
20 – 30	3		
30 – 40	14		
40 – 50	19		
50 – 60	14		
60 – 70	22		
70 – 80	16		
80 – 90	8		
90 – 100	16		
100 – 110	10		
110 – 120	3		
120 – 130	2		
130 – 140	0		
140 – 150	1		
	I alt:		

**Opgave 6** (Karakterfordeling i HF fællesfag 2004)

Karaktererne for HF fællesfag matematik i 2004 fordelte sig som følger:

Karakter	00	03	5	6	7	8	9	10	11	13
Frekvens (%)	4,1	15,8	18,5	8,7	10,0	10,0	10,6	12,6	9,0	0,7

For at studere resultatet nærmere behandler Undervisningsministeriet karaktererne som grupperet data, med skillepunkter midt mellem de aktuelle karakterer:

Vægt (gram)	Frekvens	Kumuleret frekvens
]0;1,5]	0,041	
]1,5;4]	0,158	
]4;5,5]	0,185	
]5,5;6,5]	0,087	
]6,5;7,5]	0,100	
]7,5;8,5]	0,100	
]8,5;9,5]	0,106	
]9,5;10,5]	0,126	
]10,5;12]	0,090	
]12;13]	0,007	

- Udregn de kumulerede frekvenser.
- Bestem kvartilsættet.
- Bestem karaktergennemsnittet.

**Opgave 7** (Statistik fra Sundhedsstyrelsen)

Denne opgave er mere løs. Gå ind på Sundhedsstyrelsens hjemmeside [www.sst.dk](http://www.sst.dk), specielt under statistiksiderne (direkte link [www.sundhedsdata.sst.dk](http://www.sundhedsdata.sst.dk)). Her kan du finde meget interessant statistik. Vær dog opmærksom på, at for at du kan lave sumkurver med videre, så skal der være data grupperet på kvantitative intervaller, fx alder. Det duer ikke at gruppere på år – her er det mere *Indekstal*, som er det relevante redskab. Data er organiseret lidt på samme måde som hos Danmarks Statistik, nemlig som en *Matrix*. Her skal du markere de størrelser, du ønsker information om. I nogle felter, fx under aldersintervaller, får du brug for at markere flere eller alle intervallerne. Det kan gøres ved at klikke på den første og klikke på den sidste, mens man holder Shift-tasten nede.

**Opgave 8** (Spiritus- og promillekørsel)

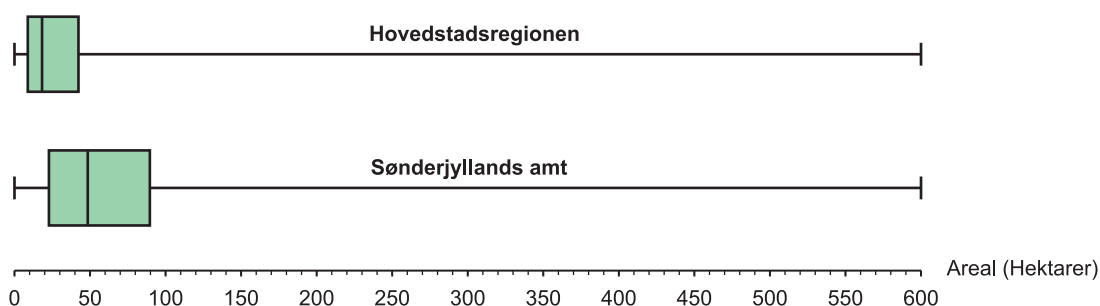
Danmarks Statistik har nedenstående data om strafferetlige afgørelser vedrørende spiritus- og promillekørsel.

Alder (år)	1994		2004	
	Mænd	Kvinder	Mænd	Kvinder
15 – 19	353	10	578	25
20 – 24	993	35	1210	55
25 – 29	1067	71	982	46
30 – 39	2143	250	2201	207
40 – 49	2034	221	2565	269
50 –	1486	137	2570	238

- Lav en sumkurve for hver af de fire datasæt. Du kan eventuelt vælge at lade øvre grænse på alderen være 80 år.
- Bestem kvartilsættet for hver af de fire fordelinger.
- Beregn middeltallet for hver af de fire fordelinger.
- Kommenter tallene.
- Som du forhåbentligt har fundet frem til i c), så er gennemsnitalderen for spiritus- og promilledømte vokset i perioden fra 1994 til 2004. Hvorfor skal man passe på med dermed at slutte, at der er en tendens til at folk er ældre, når de bliver taget for denne forseelse? *Hjælp*: Tænk på befolkningens aldersfordeling.

**Opgave 9**

Nedenfor ser du to boxplot for arealfordelingen af landbrugsbedrifterne i henholdsvis Hovedstadsregionen og i Sønderjyllands Amt i 2004, udregnet på baggrund af data fra Danmarks Statistik. Sammenlign de to boxplot og sig, hvad de fortæller dig om landbrugene i de to områder.

**Størrelsen af landbrugsbedrifter i 2004 efter areal**

**Opgave 10** (Udgifter til konfirmation)

Nedenstående data stammer fra en undersøgelse foretaget af Gallup for *Berlingske Tidende* til en artikel bragt den 25. februar 2006. Der blev stillet mange spørgsmål, hvoraf spørgsmål 5 og 9 er gengivet nedenfor. 1507 personer svarede på spørgsmål 5, mens 1359 besvarede spørgsmål 9.

**Spørgsmål**

Q5: Hvad er efter din mening et rimeligt beløb for forældrene at bruge på en konfirmation, når man ikke tæller gaven med?

Q9: Hvor mange penge regner du med at bruge på selve gaven?

Data kunne ikke vejes (korrigeres), da køn/alder/geografisk fordeling ikke var kendt i denne lidt specielle population: Husstande med børn i 14-16 års alderen. Det antages dog ikke at betyde det helt store her. Data er som følger, fordelt på beløb:

Beløb (kr)	Q5
0 – 10.000	40 %
10.001 – 20.000	45 %
20.001 – 30.000	9 %
30.001 – 40.000	2 %
40.001 – 50.000	1 %
50.001 –	–
Ved ikke	3 %

Beløb (kr)	Q9
0 – 200	1 %
201 – 400	4 %
401 – 600	5 %
601 – 800	2 %
801 – 1.000	8 %
1.001 – 2.000	21 %
2.001 – 4.000	25 %
4.001 – 6.000	18 %
6.001 – 8.000	5 %
8.001 – 10.000	4 %
10.001 – 15.000	2 %
15.001 – 20.000	1 %
20.001 – 25.000	–
25.001 –	–
Ved ikke	4 %

- Lav en sumkurve for hvert af de to datasæt. Bemærk, at der er henholdsvis 3 og 4 procent med svaret ”ved ikke”. Du kan da eventuelt vælge, at din population kun er de personer, der har en mening om beløbet. Det betyder, at du skal korrigere alle procenterne ved at dividere med henholdsvis 0,97 og 0,96 (Overvej!). Det vil dog kun give en lille korrektion her, men de kumulerede frekvenser vil til gengæld gå til 100%, op til afrunding!
- Aflæs kvartilsættet og kommenter tallene.
- Hvad er middeltallene for de to datasæt?
- Du kan eventuelt overveje, hvordan du ville formulere en avisartikel med udgangspunkt i ovenstående statistik.

**Opgave 11** (Er du socialist?)

Nedenstående data stammer fra en undersøgelse foretaget af Gallup for *Ugebrevet A4* i foråret 2006. Der blev stillet mange spørgsmål, hvoraf resultaterne af de to første er gengivet nedenfor. I alt blev 1001 personer udspurgt. Denne opgave er formuleret mere løst, og du skal selv finde ud af fornuftige måder at behandle data på. Forsøg desuden at uddrage konklusioner på baggrund af materialet.

**Spørgsmål 1**

Hvilket af følgende fire udsagn er du mest enig i?

U1: Jeg opfatter mig selv som venstreorienteret.

U2: Jeg opfatter mig selv som højreorienteret.

U3: Jeg opfatter mig hverken som venstre- eller højreorienteret.

U4: Ved ikke.

Efter vejning fik man følgende statistik fordelt på alder:

Alder (år)	U1	U2	U3	U4
18 – 35	29 %	18 %	52 %	1 %
36 – 49	19 %	17 %	61 %	3 %
50 – 65	20 %	12 %	67 %	2 %
66 –	15 %	13 %	65 %	6 %

**Spørgsmål 2**

Hvilket af følgende fire udsagn er du mest enig i?

U5: Jeg opfatter mig selv som socialist.

U6: Jeg opfatter mig selv som borgerlig.

U7: Jeg opfatter hverken mig selv som socialist eller borgerlig.

U8: Ved ikke.

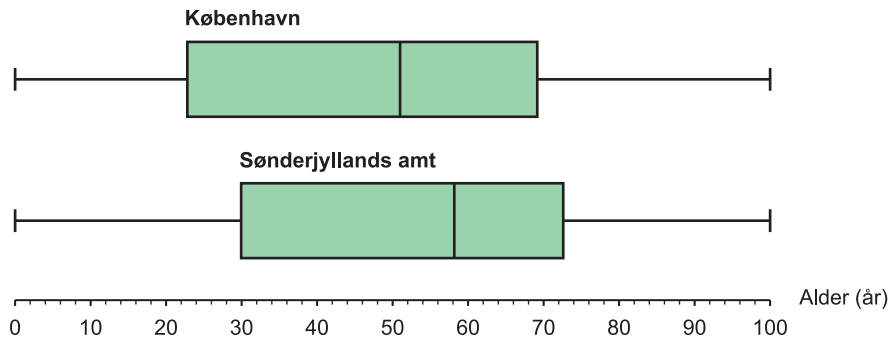
Efter vejning fik man følgende statistik fordelt på alder:

Alder (år)	U5	U6	U7	U8
18 – 35	21 %	36 %	42 %	1 %
36 – 49	14 %	38 %	48 %	1 %
50 – 65	21 %	42 %	36 %	1 %
66 –	16 %	52 %	28 %	4 %

## Opgave 12

Nedenfor ser du to boxplot for aldersfordelingen for mænd indlagt på sygehus i 2004 i henholdsvis København og Sønderjyllands Amt., udregnet på baggrund af data fra Danmarks Statistik. Sammenlign de to boxplot og forklar, hvad de fortæller dig. Fortolk resultaterne.

### Indlæggelser af mænd i 2004 efter alder



## Opgave 13 (Et dårligt slankeråd)

Ekstrabladet havde den 30/7 2005 en artikel med titlen ”Søde kostråd i agurketiden”, forfattet af Arne Astrup. Artiklen beskriver, hvorledes en gruppe læger havde lanceret nyheden ”Slank med sukker”. Lægerne påstod, at man kan slanke sig ved at komme sukker i kaffen! De havde nemlig fundet, at der i midtfirserne var langt færre overvægtige blandt mandlige københavnere, der brugte sukker i kaffen – end blandt dem, der ikke sødede kaffen med sukker. Men hvordan kan man så slanke sig ved at hælde flere kalorier i kaffen? Jo, lægerne tænker, at sukkeret måske stimulerer et særligt mæthedshormon, og konkluderer, at man skal gå og småspise sukker i løbet af dagen for at slanke sig. Arne Astrup kommenterer spørgsmålet om sukker i kaffen slanker eller feder: ”Vi ved, at sukker i sodavand mætter fantastisk dårligt, og derfor feder mere, end hvis de samme sukkerkalorier blev indtaget fra fast føde. Det skulle da være meget mystisk, hvis sukker i sodavand feder, og sukker i kaffe slanker: Medmindre at sukker og koffein i samspil har en speciel virkning. Men den videnskabelige artikel giver ikke belæg for at konkludere, at sukker slanker ....”. Kan du afsløre lægernes fejlargumentation? Hvad viser undersøgelsen af mændene i midtfirserne snarere?

## Opgave 14 (Øjenoperationen gik galt)

I en artikel i BT, tirsdag den 9. august 2005 var der en artikel, hvor en patient havde oplevet, at nogle øjenoperationer var gået galt. Patienten udtaler: ”Lægen fortalte, at der var tre procent risiko, for at operationen kunne gå galt, så man skulle omopereres. Men jeg endte med at få fem operationer, så den statistik holder overhovedet ikke”. Ifølge artiklen mener patienten, Finn Hansen, ikke, at øjenlægen fra en privatklinik i Charlotten-

lund, har fortalt ham hele sandheden om risikoen, da han i år 2000 besluttede sig for at få foretaget en laseroperation for nærsynethed.

- a) Antag, at sandsynligheden for, at hver operation går galt er 3% og at resultatet af de enkelte operationer er *uafhængige* af hinanden. Hvor stor er så sandsynligheden for, at operationen går galt fem gange, som i tilfældet med Finn Hansen.
- b) Er det rimeligt, at antage, at hændelserne er uafhængige, som antaget i spørgsmål a). Kom eventuelt med nogle argumenter for, at det ikke behøver være tilfældet. Hvis hændelserne er afhængige, vil det så øge eller mindske sandsynligheden for det skete, tror du?

### Opgave 15 (Fængsel eller samfundstjeneste?)

Frank Jensen fra Socialdemokratiet argumenterede på et tidspunkt i TV-avisen for at man skulle benytte sig mere af samfundstjeneste frem for fængsel til dømte. Han fremførte, at statistikker viste, at der var større tilbagefald til kriminalitet blandt folk, der kom i fængsel end for folk, som blev henvist til samfundstjeneste.

- a) Hvad er det for en påstand af typen *årsag*  $\Rightarrow$  *virkning*, som Frank Jensen indirekte formulerer?
- b) Hvilket problem kan der være ved argumentet? *Hjælp*: Er gruppen af personer, der får samfundstjeneste, sammenlignelig med gruppen af personer, der må blive i fængsel?
- c) Hvis man virkelig skulle afgøre om samfundstjeneste har en gavnlig virkning på de indsattes tilbagefald, hvilket ”forsøg” kunne man så opstille? Det vil man dog nok ikke gøre i praksis.

### Opgave 16 (Tobaksrygning gør dummere)

En artikel i Weekendavisen 9.-15. september 2005 præsenterer i en artikel med titlen ”Skrækhistorier” adskillige eksempler på fejlslutninger på baggrund af registrerede sammenhænge. Blandt andet anføres det i artiklen, at flere undersøgelser på det seneste har argumenteret for påstanden, at tobaksrygning gør folk dummere. En amerikansk undersøgelse af børn i alderen 6-16 år konkluderer for eksempel, at også passiv rygning gør dig dummere. Børn fra ryger-hjem var således dårligere til at læse og regne, og de klarede sig dårligere i IQ-tests end andre børn. I undersøgelsen rensede man for visse forskelle, såsom forældres uddannelse og sociale status. Denne undersøgelse og to andre blev refereret ukritisk i forskellige danske dagblade, og konklusionerne blev markedsført af blandt andet Kræftens bekæmpelse. Konklusionen om, at rygning medfører, at man bliver dummere, er imidlertid yderst tvivlsom. Gruppen af personer, som ryger og gruppen af personer, som ikke ryger er nemlig meget forskellig. Der er dog korrigeret for nogle sociale faktorer, men langt fra alle. Prøv, om du kan komme på nogle andre faktorer, der adskiller folk i de to grupper – faktorer, som kan have afgørende betydning for hvor godt børnene klarer sig i skolen.

**Opgave 18** (Spørgeundersøgelse på gymnasium)

Overvej, hvordan du ville arrangere en spørgeundersøgelse på dit gymnasium, for eksempel for elevers holdning til alkohol. Tænk på de praktiske aspekter ved gennemførelsen af undersøgelsen, samt om dens pålidelighed. Ville du lave en stikprøve, eller spørge alle? Hvis du benytter stikprøve, hvor mange ville du så spørge og hvordan? Hvilke spørgsmål ville du stille? Ville du korrigere de indsamlede data bagefter? Hvis ja, så hvordan? etc.....

**Opgave 19** (Ugrupperet data)

Til en gymnasiefest med 29 elever har man spurgt, hvor mange øl, som hver elev har drukket i løbet af en aften. Resultaterne fremgår af tabellen nedenfor.

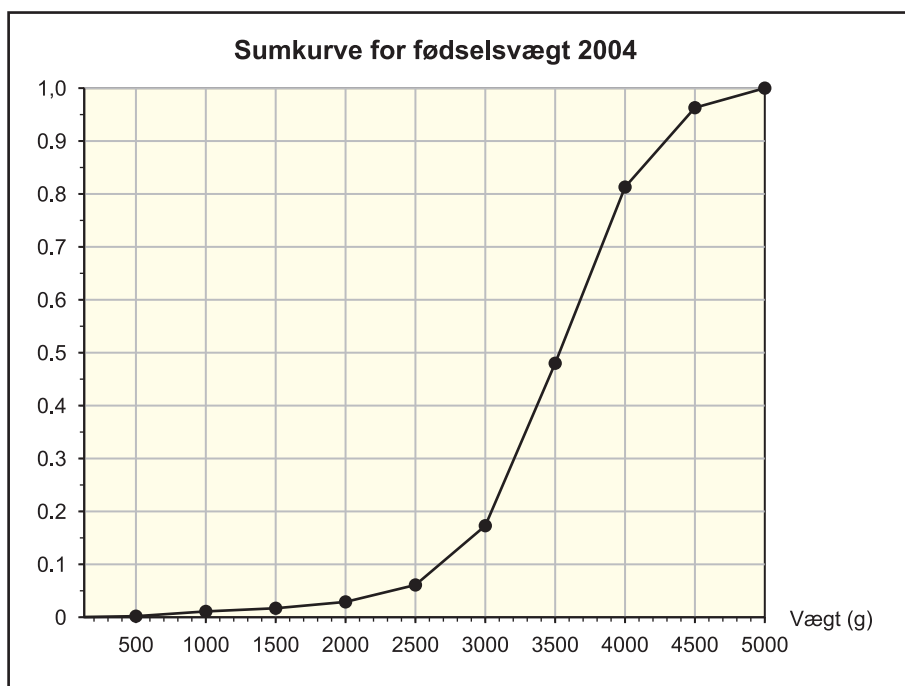
<b>Antal øl</b>	0	1	2	3	4	5	6	7	8	9	10
<b>Hyppeghed</b>	2	5	6	4	1	3	0	4	1	2	1
<b>Frekvens</b>											
<b>Kum. frekvens</b>											

- Udregn frekvenserne og de kumulerede frekvenser.
- Tegn et pindediagram for frekvenserne.
- Bestem kvartilsættet.

**Opgave 20** (Sumkurve til histogram)

Nedenfor er sumkurven og de kumulerede frekvenser for fødselsvægten for børn født i 2004, taget fra Danmarks Statistik. Bestem frekvenserne og tegn et histogram.

<b>Fødselsvægt (g)</b>	<b>Kumuleret frekvens</b>	<b>Frekvens</b>
0 – 500	0,002	
500 – 1000	0,011	
1000 – 1500	0,017	
1500 – 2000	0,029	
2000 – 2500	0,061	
2500 – 3000	0,173	
3000 – 3500	0,480	
3500 – 4000	0,813	
4000 – 4500	0,963	
4500 – 5000	1,000	

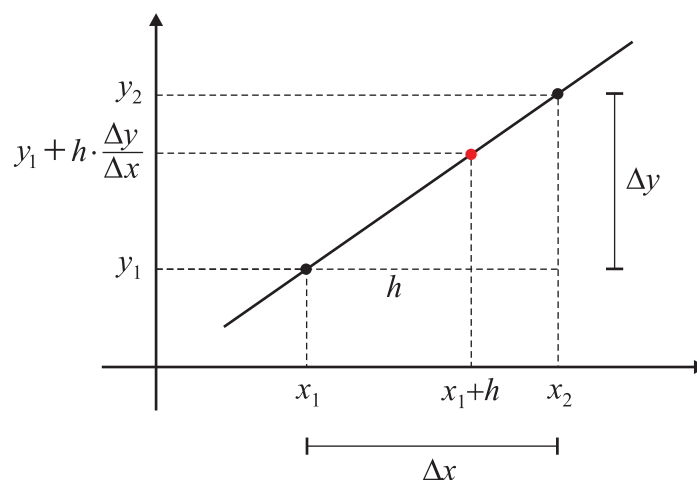


### Opgave 21 (Ugrupperet data)

Følgende kondital blev målt i en gruppe med 24 mænd: 51, 35, 62, 44, 37, 35, 42, 41, 41, 55, 39, 59, 61, 62, 44, 44, 36, 53, 47, 53, 46, 57, 61, 39. Bestem kvartilsættet og lav et boksplot. *Hjælp:* Opskriv de forskellige observationer og deres hyppigheder. Udregn frekvenser og kumulerede frekvenser, etc.

### Opgave 22 (Lineær interpolation)

Lad os sige, at vi kender to punkter  $(x_1, y_1)$  og  $(x_2, y_2)$  på en sumkurve og ønsker at beregne en mellemliggende værdi på sumkurven. Kig på nedenstående figur og prøv at forstå, hvordan det kan gøres.



**Opgave 23** (Stikprøver)

Kurt vil gerne undersøge danskernes holdning til sangprogrammet *X factor* på DR1. Han lægger følgende spørgsmål ud på facebook:

- a) Er X factor er et godt program?
- b) Er sangerne i programmet dygtige sangere?

Kurt får 71 svar og tæller svarene sammen, hvorefter han konkluderer at 68% af danskerne mener at X Factor er et godt program samt at 45% af danskerne mener at sangerne er dygtige. Er det en brugbar undersøgelse? Påpeg flere uheldige ting ved undersøgelsen.

**Litteratur**

David Freedman, Robert Pisani og Roger Purves. *Statistics*. Third Edition. W. W. Norton & Company 1998 (opr. 1978).

Richard J. Larsen, Morris L. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Fourth Edition. Pearson Prentice Hall, 2006 (opr. 1981).

**Websites**

[www.dst.dk](http://www.dst.dk) (Danmarks Statistik)

[www.politi.dk](http://www.politi.dk) (Kig under punktet *Statistik*)

[www.sst.dk/](http://www.sst.dk/) (Sundhedsstyrelsen. Her er en del statistik til rådighed)

[www.gallup.dk/](http://www.gallup.dk/)